Hierarchical Conditioning of Diffusion Models Using Tree-of-Life for Studying Species Evolution

Mridul Khurana[1(✉)], Arka Daw[2], M. Maruf[1], Josef C. Uyeda[1], Wasila Dahdul[3], Caleb Charpentier[1], Yasin Bakış[4], Henry L. Bart Jr.[4], Paula M. Mabee[5], Hilmar Lapp[6], James P. Balhoff[7], Wei-Lun Chao[8], Charles Stewart[9], Tanya Berger-Wolf[8], and Anuj Karpatne[1]

[1] Virginia Tech, Blacksburg, VA, USA
{mridul,karpatne}@vt.edu
[2] Oak Ridge National Laboratory, Oak Ridge, TN, USA
[3] University of California, Irvine, CA, USA
[4] Tulane University, New Orleans, LA, USA
[5] Battelle, Columbus, OH, USA
[6] Duke University, Durham, NC, USA
[7] University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[8] The Ohio State University, Columbus, OH, USA
[9] Rensselaer Polytechnic Institute, Troy, NY, USA

Abstract. A central problem in biology is to understand how organisms evolve and adapt to their environment by acquiring variations in the observable characteristics or traits of species across the tree of life. With the growing availability of large-scale image repositories in biology and recent advances in generative modeling, there is an opportunity to accelerate the discovery of evolutionary traits automatically from images. Toward this goal, we introduce Phylo-Diffusion, a novel framework for conditioning diffusion models with phylogenetic knowledge represented in the form of HIERarchical Embeddings (HIER-Embeds). We also propose two new experiments for perturbing the embedding space of Phylo-Diffusion: trait masking and trait swapping, inspired by counterpart experiments of gene knockout and gene editing/swapping. Our work represents a novel methodological advance in generative modeling to structure the embedding space of diffusion models using tree-based knowledge. Our work also opens a new chapter of research in evolutionary biology by using generative models to visualize evolutionary changes directly from images. We empirically demonstrate the usefulness of Phylo-Diffusion in capturing meaningful trait variations for fishes and birds, revealing novel insights about the biological mechanisms of their evolution. (Model and code can be found at imageomics.github.io/phylo-diffusion)

Keywords: Evolution · Diffusion Models · Hierarchical Conditioning

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73024-5_9.

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2025
A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15147, pp. 137–153, 2025.
https://doi.org/10.1007/978-3-031-73024-5_9

1 Introduction

Given the astonishing diversity of life forms on the planet, an important end goal in biology is to understand how organisms evolve and adapt to their environment by acquiring variations in their observable characteristics or *traits* (*e.g.*, beak color, stripe pattern, and fin curvature) over millions of years in the process of evolution. Our knowledge of species evolution is commonly represented in a graphical form as the "tree of life" (also referred to as the *phylogenetic tree* [14], see Fig. 1), illustrating the evolutionary history of species (leaf nodes) and their common ancestors (internal nodes). Discovering traits that are heritable across the tree of life, termed *evolutionary traits*, is important for a variety of biological tasks such as tracing the evolutionary timing of trait variations common to a group of species and analyzing their genetic underpinnings through gene-knockout or gene-editing/swapping (*e.g.*, CRISPR [22]) experiments. However, quantifying trait variations across large groups of species is labor-intensive and time-consuming, as it relies on expert visual attention and subjective definitions [28], hindering rapid scientific advancement [19].

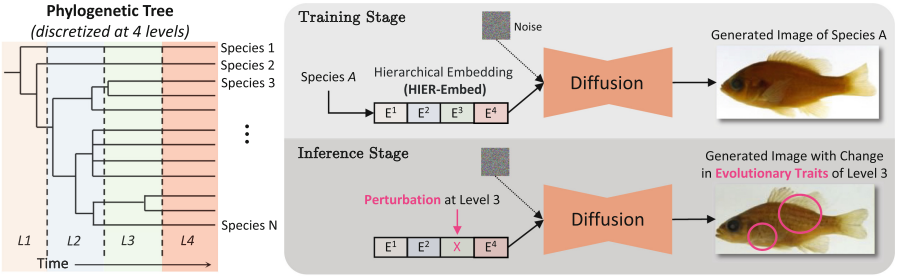


Fig. 1. Overview of Phylo-Diffusion framework. Every species in the tree of life (phylogenetic tree) is encoded to a HIERarchical Embedding (HIER-Embed) comprising of four vectors (one for each phylogenetic level), which is used to condition a latent diffusion model to generate synthetic images of the species. By structuring the embedding space with phylogenetic knowledge, Phylo-Diffusion enables visualization of changes in the evolutionary traits of a species (circled pink) upon perturbing its embedding. (Color figure online)

The growing deluge of large-scale image repositories in biology [9,30,31] presents a unique opportunity for machine learning (ML) methods to accelerate the discovery of evolutionary traits automatically from images. In particular, with recent developments in generative modeling such as latent diffusion models (LDMs) [26], we are witnessing rapid improvements in our ability to control the generation of high-quality images based on input conditioning of text or image prompts. This is facilitating breakthroughs in a variety of commercial use-cases of computer vision where we can analyze how changes in the input prompts affect variations in the generated images [7,27,34]. We ask the question: *can*

we leverage LDMs to control the generation of biological images of organisms conditioned on the position of a species in the tree of life? In other words, can we encode the structure of evolutionary relationships among species and their ancestors as input conditions in LDMs? This can help us analyze trait variations in generated images across different branches in the phylogenetic tree, revealing novel insights into the biological mechanisms of species evolution.

Toward this goal, we introduce **Phylo-Diffusion**, a novel framework for discovering evolutionary traits of species from images by conditioning diffusion models with phylogenetic knowledge (see Fig. 1). One of the core innovations of Phylo-Diffusion is a novel HIERarchical Embedding (**HIER-Embed**) strategy that encodes evolutionary information of every species as a sequence of four vectors, one for each discretized level of ancestry in the tree of life (covering different evolutionary periods). We also propose two novel experiments for analyzing evolutionary traits by perturbing the embedding space of Phylo-Diffusion and observing changes in the features of generated images, akin to biological experiments involving genetic perturbations. First, we introduce **Trait Masking**, where one or more levels of information in HIER-Embed are masked out with noise to study the disappearance of traits inherited by species at those levels. This is inspired by *gene knockout* experiments [10], wherein one or more genes are deactivated or "knocked out" to investigate the gene's function, particularly its impact on the traits of the organism. Second, we introduce **Trait Swapping**, where a certain level of HIER-Embed in a reference species is swapped with the embedding of a sibling node at the same level, similar in spirit to *gene editing/swapping* experiments made possible by the CRISPR technology [22]. The goal of trait swapping is to visualize trait differences at every branching point in the tree of life that results in the diversification of species during evolution.

Here are the main contributions of this paper. Our work represents a novel methodological advance in the emerging field of knowledge-guided machine learning (KGML) [15–17] to structure the embedding space of generative models using tree-based knowledge. Our work also opens a new chapter of research in evolutionary biology by using generative models to visualize evolutionary changes directly from images, which can serve a variety of biological use-cases. For example, Phylo-Diffusion can help biologists automate the discovery of *synapomorphies*, which are distinctive traits that emerge on specific evolutionary branches and are crucial for systematics and classification [33]. Our proposed experiments of trait masking and swapping can also be viewed as novel image-based counterparts to genetic experiments, which traditionally take years. Our work thus enables biologists to rapidly analyze the impacts of genetic perturbations on particular branches of the phylogenetic tree–a grand challenge in developmental biology [5,20]. We empirically demonstrate the usefulness of Phylo-Diffusion in capturing meaningful trait changes upon perturbing its embedding for fishes and birds, generating novel hypotheses of their evolution.

2 Related Works and Background

Interpretable ML: Discovering evolutionary traits from images requires the identification and interpretation of fine-grained features in images that define and differentiate species. Several methodologies have recently been developed in the field of interpretable ML for localizing image regions that contain discriminatory information of classes, including ProtoPNet [3], PIP-Net [21] and INTR [23]. Despite their effectiveness and applicability across a wide range of applications, these methods are not directly suited for our target application of discovering evolutionary traits for two primary reasons. First, they are not designed to incorporate structured biological knowledge (*e.g.*, knowledge of the tree of life) in the learning of interpretable features, and thus are unable to provide biologically meaningful explanations of feature differences across groups of species in the phylogenetic tree, which is key to discovering evolutionary traits. Second, since most methods in interpretable ML are designed for the task of classification, it is non-trivial to integrate them in generative modeling frameworks to produce synthetic images with controlled perturbations in the embedding space similar to gene knockout and gene editing/swapping experiments, in contrast to our proposed experiments in Phylo-Diffusion.

Phylogeny-guided Neural Networks (Phylo-NN): A recent work closely aligned with our goal of discovering evolutionary traits directly from images is Phylo-NN [6]. Phylo-NN uses an encoder-decoder architecture to represent images of organisms as structured sequences of feature vectors termed "Imageomes", where different segments of Imageomes capture evolutionary information from varying levels of ancestry in the phylogenetic tree. While Phylo-NN shares several similarities with our proposed framework, Phylo-Diffusion, in terms of motivations and problem formulations, there are also prominent differences. The primary goal of Phylo-NN is specimen-level image reconstruction, whereas Phylo-Diffusion considers a different goal of controlling image generation at the species-level. As a result, Phylo-NN learns a unique Imageome sequence for every organism, enabling us to study the variability in individuals from the same species and the analysis of similarity in Imageome segments learned at shared ancestry levels. On the other hand, Phylo-Diffusion learns a unique embedding for every species and ancestor node in the tree of life, which serves as input conditions to generate distributions of synthetic images. Phylo-Diffusion thus uses hard constraints to ensure that all species with a common ancestor learn the exact same embeddings at their shared ancestry levels, making it easy to analyze trait commonalities and variations across groups of species, in contrast to Phylo-NN. Additionally, Phylo-Diffusion allows for perturbations in the embedding space of generative models in biologically meaningful ways inspired by gene knockout and gene editing/swapping experiments, going beyond the capabilities of Phylo-NN. We consider Phylo-NN as a baseline in our experiments to compare its performance with Phylo-Diffusion.

Background on Latent Diffusion Models (LDMs): One of the state-of-the-art approaches in generative modeling is the framework of Diffusion Models

[13], which learns a target distribution $p(x)$ by incrementally transforming a noisy sample x generated from a Gaussian distribution $\mathcal{N}(0, I)$ into one that is more likely to be generated from $p(x)$ over a series of timesteps T. While early frameworks of diffusion models (*e.g.*, DDPM [13], DDIM [29] and ADM [4]) suffered from high computational costs and long training/inference times, Latent Diffusion Models (LDMs) [26] are able to address these concerns to a large extent by operating in a compressed latent space, significantly accelerating their ability to generate high-resolution images. The basic idea of LDMs is to train a separate auto-encoder to map an input image x into its latent representation $z_0 = \mathcal{E}(x)$ using encoder $\mathcal{E}$, which when fed to decoder $\mathcal{D}$ produces a reconstruction of the original image, $\tilde{x} = \mathcal{D}(\tilde{z}_0)$. LDMs employ diffusion models in the compressed latent space z by modeling the conditional probability of the reverse diffusion process as $\tilde{z}_{t-1} \sim p_\theta(\tilde{z}_{t-1}|\tilde{z}_t, y, t)$, where y is the input condition. This is implemented using a conditional denoising U-Net backbone $\epsilon_\theta(z_t, y, t)$ with learnable parameters θ. LDMs also pre-process y using a domain-specific encoder $\mathbf{E} = \tau_\phi(y)$ trained alongside the U-Net backbone ϵ_θ to project y into the intermediate layers of ϵ_θ using cross-attention mechanisms. The learnable parameters of LDMs are trained by minimizing the following loss function:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{z_t, y, t, \epsilon \sim \mathcal{N}(0,I)} \left[\|\hat{\epsilon}_\theta(z_t, \tau_\phi(y), t) - \epsilon\|^2 \right] \tag{1}$$

3 Proposed Framework of Phylo-Diffusion

3.1 Hierarchical Embedding (HIER-Embed)

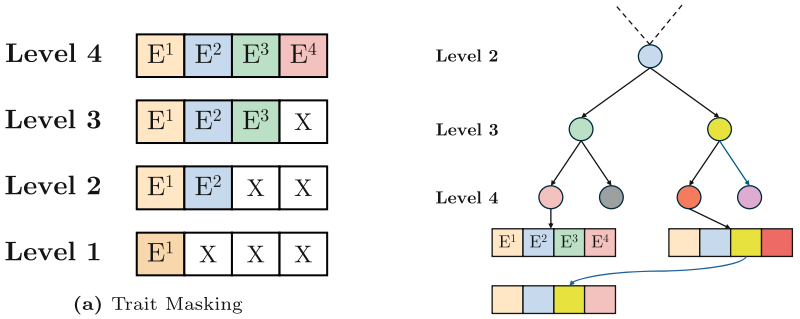
Phylo-Diffusion uses a novel hierarchical embedding (HIER-Embed) strategy to structure the embedding E of every species node using phylogenetic knowledge. As a first step, we consider a discretized version of the phylogenetic tree involving four ancestral levels, level-1 to level-4, where every level corresponds to a different range of time in the process of evolution. (See Appendix for a detailed characterization of the four ancestry levels for fish species used in this study.) Given a set of n species, $\mathcal{S} = \{S_1, S_2, S_3, ..., S_n\}$, let us represent the position of species $S_i \in \mathcal{S}$ in the phylogenetic tree at the four ancestry levels as $\{S_i^1, S_i^2, S_i^3, S_i^4\}$, where S_i^l represents the ancestor node of S_i at level-l. Hence, if two species S_i and S_j share common ancestors till level-k, then $S_i^l = S_j^l$ for $l = 1$ to k. We define the level-l embedding of species S_i as:

$$\mathbf{E}_i^l = \texttt{Embed}(S_i^l) \in \mathbb{R}^{d'}, \tag{2}$$

where $\texttt{Embed}(.)$ is a learnable embedding layer that provides a simple way to store and look-up the trained embeddings of every node. The combined hierarchical embedding (HIER-Embed) of species S_i is obtained by concatenating its embeddings across all four levels as follows:

$$\mathbf{E}_i = \tau(S_i) = \texttt{Concat}\,[\mathbf{E}_i^1, \mathbf{E}_i^2, \mathbf{E}_i^3, \mathbf{E}_i^4] \in \mathbb{R}^d, \tag{3}$$

where `Concat[.]` denotes the concatenation operation and $y = S_i$ is the input condition used in LDMs. Note that different segments of $\mathbf{E}_i$ capture information about the traits of S_i acquired at different time periods of evolution. In particular, we expect the embedding vectors learned at earlier ancestry levels of $\mathbf{E}_i$ to capture evolutionary traits of S_i common to a broader group of species. On the other hand, embeddings learned at later ancestry levels are expected to be more specific to S_i. In the following, we present two novel experiments for studying evolutionary traits by perturbing the embedding space learned by HIER-Embed.



(a) Trait Masking (b) Trait Swapping

Fig. 2. Schematic diagrams of the two proposed experiments for discovering evolutionary traits using Phylo-Diffusion.

3.2 Proposed Experiment of Trait Masking

The goal of this experiment is to verify if HIER-Embed is indeed able to capture hierarchical information in its level-embeddings such that masking information at lower levels of the embedding only erases traits acquired at later stages of evolution while retaining trait variations learned at earlier levels. In other words, we want to verify that the embeddings learned by HIER-Embed at level-l capture information common to all descendant species that are part of the same subtree at level-l. Figure 2a represents a schematic diagram of the process followed for trait masking. We start with the combined embedding containing information at all four levels, $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{E^4}]$. To examine what is learned at the last level of this embedding, we mask it out by substituting it with Gaussian noise defined as $\mathbf{z_{noise}} \sim \mathcal{N}(0, I) \in \mathbb{R}^{d'}$. This results in the perturbed embedding $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{z_{noise}}]$, effectively eliminating the species-level (or $\mathbf{E^4}$) information. This masking should prompt the model to generate images that reflect only the information learned up to the third level while obscuring species-level details. We can extend this experiment by incrementally introducing *noise* at later levels, e.g., at both levels 3 & 4, and so on.

Expected Changes in Probability Distributions: Note that when all four level embeddings are used, *i.e.* $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{E^4}]$, the generated images are

expected to be classified to a unique species S_i. In terms of probability distributions, the probability of predicting species S_i should be distinctly higher than the probability of predicting any other species. However, when we mask out certain level embeddings (*i.e.*, mask out information at level 4), we are intentionally removing information necessary to distinguish species S_i from its siblings species that are part of the same sub-tree (e.g., those that share a common ancestor at level 3). For this reason, we expect the generated images to show higher probabilities of being classified to any of the descendant species of the sub-tree, compared to the other species that are outside of the sub-tree. To quantify this behavior, we can measure the change in probability distributions for species within and outside the sub-tree after masking out an internal node. Since we expect probabilities to increase only for species within the sub-tree, the mean increase in probabilities for within-subtree species should be higher than that for out-tree species, as empirically demonstrated later in the Results Section.

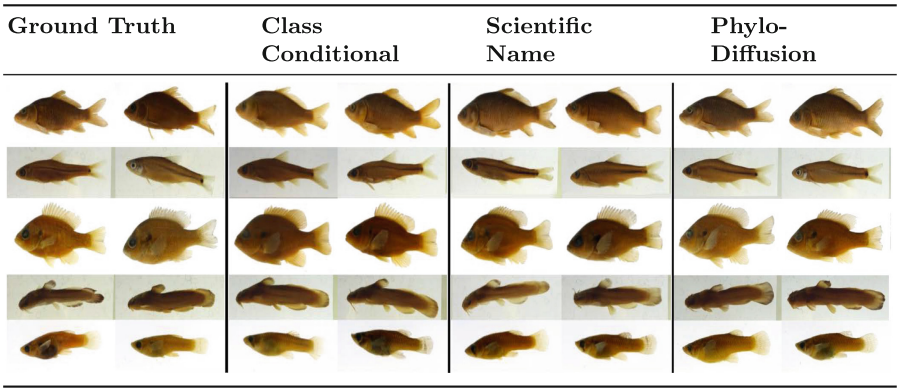


Fig. 3. Comparing the quality of synthetic images generated by different conditioning mechanisms in LDMs. Every row corresponds to a different species and we show two samples per species for every conditioning mechanism. The order of species from top to bottom is *Cyprinus carpio*, *Notropis hudsonius*, *Lepomis auritus*, *Noturus exilis*, and *Gambusia affinis*.

3.3 Proposed Experiment of Trait Swapping

In trait swapping, we substitute the level-l embedding of a source species with the level-l embedding of a sibling subtree at an equivalent level. Figure 2b shows a schematic representation of the trait swapping experiment where the level-3 embedding (green) of a source species is replaced with its sibling level-3 embedding (yellow). Images generated for this perturbed embedding are expected to retain all of the traits of the source species except the swapped embedding, which

should borrow traits from the sub-tree rooted at the sibling node. Visualizing trait differences in the generated images before and after trait swapping can help us understand the evolutionary traits that branched at a certain level (e.g., those leading to the diversification of green and yellow sub-trees at level-3 in the example phylogeny of Fig. 2b). In terms of the probability distribution, similar to trait masking, we expect to see a drop in the probabilities of the source species (*pink*), and simultaneously, we expect an increase in probabilities for all the descendent species in subtree at node *yellow*, *i.e. red* and *purple*.

4 Evaluation Setup

Datasets: We use a collection of fish images as our primary dataset for evaluation. This dataset was procured from the Great Lakes Invasives Network (GLIN) [1] Project, comprising a total of 5434 images spanning 38 fish species. We obtained the phylogenetic tree of fish species from *opentree* [2] python package (see Appendix for details on the phylogenetic tree). The raw museum images were pre-processed and resized to 256×256 pixels and the dataset was partitioned into training and validation sets, following a 75-25 split. We provide additional results on the CUB-200-2011 dataset [32] of bird species in the Appendix.

Table 1. Quantitive comparison of generated images sampled using DDIM [29] (100 samples per class).

Model Type	Method	FID ↓	IS ↑	Prec. ↑	Recall ↑
GAN	Phylo-NN	28.08	2.35	0.625	0.084
Diffusion	Class Conditional	11.46	2.47	0.679	0.359
Diffusion	Scientific Name	11.76	2.43	0.683	0.332
Diffusion	Phylo-Diffusion (ours)	11.38	2.53	0.654	0.367

Baselines of Conditioning Mechanisms: (1) *Class Conditional:* One of the simplest ways of encoding information about a species class is to map class labels $y \in [1, N_c]$ to a fixed d-dimensional embedding vector $e \in \mathbb{R}^d$ using a trainable embedding layer. Note that the resulting embeddings are not designed to contain any hierarchical information in contrast to HIER-Embed. (2) *Scientific Name Encoding:* The scientific name of a species contains valuable biological information typically comprising of a combination of the *genus* name and *species* name. Since species that share their *genus* name are likely to contain common phylogenetic traits, we use them as a baseline for conditioning LDMs for discovering evolutionary traits. Specifically, we employ a pre-trained frozen CLIP model [25] to encode the scientific names of species into fixed d-dimensional embeddings.

Training details: We used $d' = 128$ as the embedding dimension for each level of HIER-Embed, which when concatenated across the four levels produces

Table 2. Classification F1-Score on the 100 samples generated per class. The base classifier has an accuracy of *85%* on the test set.

Method	F1-Score (in %) ↑
Phylo-NN	47.37
Class Conditional	81.99
Scientific Name	70.16
Phylo-Diffusion (ours)	82.21

the combined hierarchical embedding of $d = 512$ dimensions. Phylo-Diffusion uses this d-dimensional embedding to condition LDMs through cross-attention in denoising U-Net backbone and train LDMs without classifier-free guidance. We used VQGAN [8] as the backbone encoder-decoder to achieve the latent representations desired for LDMs with a downsampling factor of 4. All the models with different encoders are trained for 400k iterations, employing the best model checkpoint if convergence occurs early. Additional hyperparameters, such as learning rate, batch size, and U-Net architecture, are detailed in the Appendix.

5 Results

5.1 Quality of Generated Images

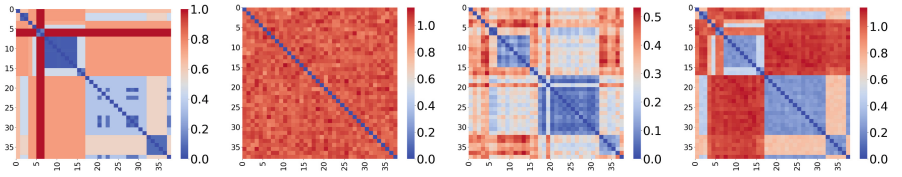
Table 1 compares the quality of generated images of baselines using the metrics of Fréchet Inception Distance (FID) score, Inception Score (IS), and Precision, Recall calculated in the feature space as proposed in [18]. Our results show that Phylo-Diffusion is at par with state-of-the-art generative models, achieving an FID of 11.38 compared to LDM's 11.46. We show a sample of generated images in Fig. 3, with additional images provided in the Appendix. We also show the robustness of Phylo-Diffusion's results with varying numbers of phylogenetic levels and embedding dimensions in the Appendix.

5.2 Classification Accuracy

We used a separate model for species classification, specifically a ResNet-18 model [12] trained using the same training/validation split as Phylo-Diffusion. The primary objective behind building this classifier is to verify if images generated by Phylo-Diffusion contain sufficient discriminatory information to be classified as their correct species classes. Table 2 compares the classification F1-scores over 100 samples generated by baseline conditioning schemes. We can see that the synthetic images generated by Phylo-Diffusion achieve the highest F1 score (82.21%), which is quite close to the F1 score of the base classifier on the original test images (85%). We present additional results showing the generalizability of Phylo-Diffusion in classifying generated images to unseen species in the Appendix.

5.3 Matching Embedding Distances with Phylogenetic Distances

We investigate the quality of embeddings produced by baseline methods by comparing distances in the embedding space with the ground-truth (GT) phylogenetic distances computed from the tree of life, as illustrated in Fig. 4. Ideally, we expect distances in the embedding space of species pairs to be reflective of their phylogenetic distances. For Class Conditional, we can see that the distance matrix does not show any alignment with the GT phylogenetic distance matrix. In the case of Scientific Name Encoding, the distance matrix exhibits notable similarities to the phylogenetic distances, thanks to the hierarchical nature of information contained in scientific names (i.e., *genus*-name & *species*-name). However, one limitation of this encoding is its inability to capture inter-genus similarities or differences. In contrast, HIER-Embed shows a distance matrix that closely aligns with the GT phylogenetic distance matrix, validating its ability to preserve evolutionary distances among species in its embedding space.



(a) GT Phylogenetic Distance **(b)** Class Conditional Encoding **(c)** Scientific Name Encoding **(d)** Hierarchical Embedding

Fig. 4. Comparing Cosine distances in the embedding space of species for varying conditioning mechanisms.

5.4 Trait Masking Results

To obtain classification probabilities or logits associated with generated images, we employ the classifier detailed in Sect. 5.2. For the masked embeddings of subtrees at level 3, defined as $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{z_{noise}}]$, we analyze logits of generated images and compare them with those generated without masking. Figure 5 demonstrates that for a specific subtree, in this case *Lepomis*, logits for species within the subtree are higher compared to those for species outside it. This outcome aligns with the expectation that Phylo-Diffusion, when provided with information up to Level 3, can capture overarching characteristics of all species within the given subtree. Additionally, Fig. 5 presents probability distributions for species within the *Lepomis* subtree when the full set of hierarchical encodings $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{E^4}]$ is provided. It demonstrates that the probabilities are significantly higher for the targeted class, as intended for image generation. After masking, we observe that the generated images are very similar and capture common features of the *Lepomis* genus. For all our calculations and plots, we

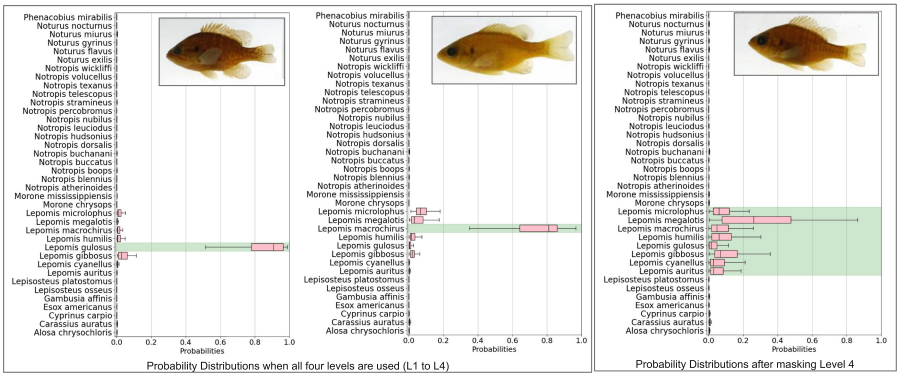


Fig. 5. Left: class probability distributions of images generated by using embeddings at all four levels for two species *Lepomis gulosus* and *Lepomis macrochirus* (shown in green) that are part of the same sub-tree till level 3. Right: class probability distributions of images generated by masking level 4 (descendant species that have common ancestry till level 3 are highlighted in green)

generate 100 images for each subtree and node. The appendix contains additional histograms that detail logit distributions across all different subtrees at each level, offering comprehensive insights into how the model discriminates and learns the hierarchical structure across different levels and nodes.

Quantitative Evaluation of Probability Distrubtions: To quantitatively evaluate the ability of Phylo-Diffusion to capture hierarchical information and show desired changes in probability distributions after masking, we compute the following metrics. Let us denote the set of all species in the data as $\mathcal{S}$ and for a given sub-tree at an internal node i of level l, let us denote the subset of descendant species as $\mathcal{S}_i^l = \{S_1, S_2, \ldots, S_n\}$. We first compute the reference probabilities P_{ref} of every species before masking (i.e., by using all four level embeddings). Let us denote the probability of predicting a generated image using all four embeddings of a descendant species $S_j \in \mathcal{S}_i^l$ into species class S_k as $P_{S_j}(S_k)$. The reference probability of a species S_k can then be given as:

$$P_{ref}(S_k) = \begin{cases} \frac{1}{|\mathcal{S}_i^l|-1} \sum_{S_j \in \mathcal{S}_i^l \setminus S_k} P_{S_j}(S_k), & \text{if } S_k \in \mathcal{S}_i^l, \\ \frac{1}{|\mathcal{S}_i^l|} \sum_{S_j \in \mathcal{S}_i^l} P_{S_j}(S_k), & \text{if } S_k \notin \mathcal{S}_i^l. \end{cases} \tag{4}$$

Note that when S_k is part of the sub-tree, i.e., $S_k \in \mathcal{S}_i^l$, $P_{ref}(S_k)$ is computed by averaging over $|\mathcal{S}_i^l| - 1$ probability values since we exclude the case when S_k is used to generate the images. On the other hand, when S_k is outside of the sub-tree, i.e., $S_k \notin \mathcal{S}_i^l$, we average over all $|\mathcal{S}_i^l|$ probability values. Given these reference probabilities values before masking, we can compute the change in probability of predicting species S_k after masking as $P_{diff}(S_k) = P_{mask}(S_k) - P_{ref}(S_k)$, where $P_{mask}(S_k)$ is the probability of predicting a generated image after masking to S_k. We expect P_{diff} to be larger for descendant species $S_k \in \mathcal{S}_i^l$

compared to species that are outside of the sub-tree because of the dispersion of probabilities in a sub-tree as a consequence of masking. We thus compute the average P_{diff} for species that belong to subtree $\mathcal{S}_i^l$ as $P_{diff}^{sub}(i,l)$ and species that are outside the subtree $\mathcal{S}_i^l$ as $P_{diff}^{out}(i,l)$.

Figure 6 shows the box plot of P_{diff}^{sub} and P_{diff}^{out} across internal nodes at levels 2 and 3. We can see that species within the subtree exhibit a more pronounced increase in probabilities compared to those outside the subtree, aligning with our expectations. Notably, this trend is consistently observed across both Levels 2 & 3. More details of the class-wise probability distribution shifts for each level are provided in the appendix. The outcomes of these experiments affirm that Phylo-Diffusion effectively identifies unique features at Levels 2 and 3, and captures shared features or traits of any chosen subtree at the internal nodes of the phylogeny.

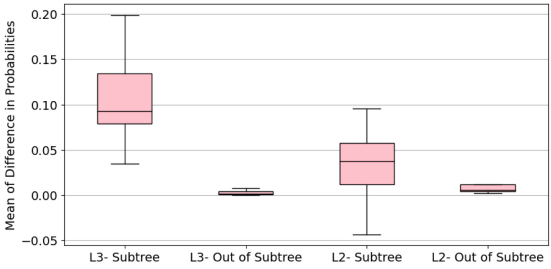
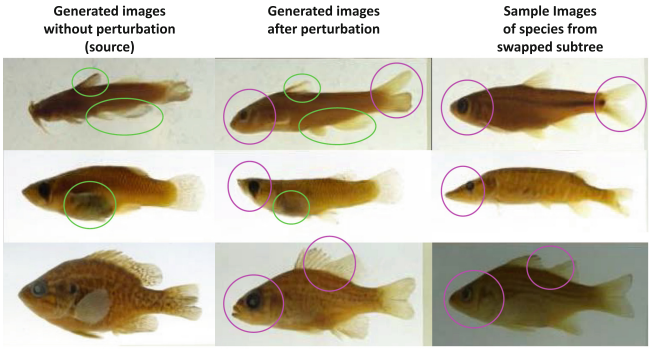


Fig. 6. Box plot for the mean of difference in probabilities for species within the subtree and out of subtree for level-3 and level-2.

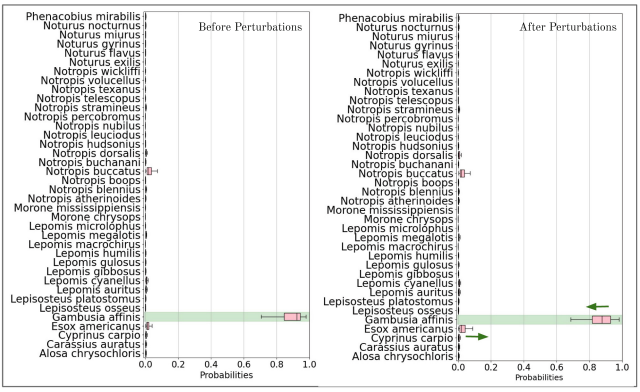
5.5 Trait Swapping Results

Figure 7 shows examples of trait swapping results enabled by Phylo-Diffusion. For the first example (first row of Fig. 7a), we swap the level-2 encoding of source species *Noturus exilis* with level-2 encoding of its sibling group *Notropis/ Carassius*. The goal here is to discover traits of *Noturus exilis* inherited at level 2 that differentiates it from other groups of species that branched out at this point of time in evolution. We can see that the generated images of the perturbed embedding (center) exhibit the absence of barbels (whiskers) highlighted in purple, while the caudal (or tail) fin is beginning to fork (or split), a trait adopted from *Notropis* (right). In contrast, other fins such as the dorsal, pelvic, and anal fins highlighted in green remain similar to those of the source species, *Noturus exilis* (left). This suggests that at level-2, *Notropis* and *Noturus* species diverged by developing differences in two distinct traits, barbels and forked caudal fins while keeping other traits intact.

For the second example (Fig. 7a, row 2), we swap level-2 information of *Gambusia affinis* (left) with that of *Esox americanus* (right). The generated images



(a) Examples of traits swapping for species at level-2 (first two rows) and level-3 (last row). The order of species from top to bottom is *Noturus exilis* swapped with *Notropis* and *Gambusia affinis* swapped with *Esox americanus*. The third row shows trait swapping at level-3 for *Lepomis gulosus* swapped with *Morone*.

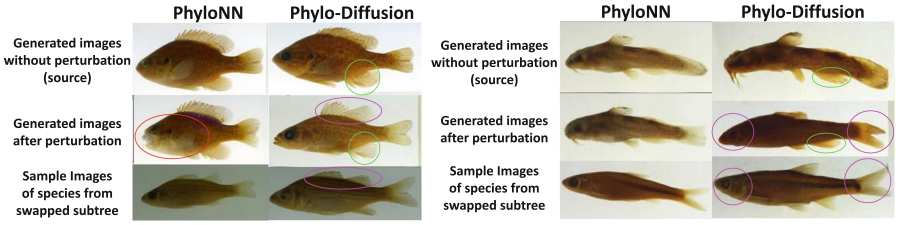


(b) For Row 2 of Figure 7a, we show that the probability distribution of *Gambusia affinis* decreases after the swapping traits at level-2, with an increase in the probability distribution of *Esox americanus*.

Fig. 7. Examples of trait swapping results.

of the perturbed embedding (center) exhibit a more pointed head highlighted in purple, and a slimmer body shape resembling *Esox americanus*. Notably, the perturbed species retains discoloration at the bottom from the source species highlighted in green. Figure 7b presents probability distributions (or logits) of *Gambusia affinis* before and after trait swapping using a separately trained classifier. We observe a slight decrease in logits for *Gambusia affinis* and an increase in logits for *Esox americanus*, consistent with our expectations. In the third row of Fig. 7a, we swap level-3 information of *Lepomis gulosus* (left) with that of *Morone* genus (right). The resulting images from the perturbed embedding (center) capture the horizontal line pattern characteristic of *Morone* genus, and the dorsal fin (purple circle) begins to split. All of these examples suggest novel scientific hypotheses about differences in evolutionary traits acquired by species

at different ancestry levels, which can be validated by biologists in subsequent studies. Note that our experiments are most effective at levels near the species nodes, specifically at levels 2 & 3, since phylogenetic signal is known to diminish as we move toward the root of the tree [11,24]. Additional visualizations of trait swapping results are provided in the Appendix.



(a) Swapping Level 3 traits for *Lepomis gulosus* with *Morone*

(b) Swapping Level 2 traits for *Notorus mirurus* with *Notropis*

Fig. 8. Comparing Phylo-NN with Phylo-Diffusion for examples of trait swapping.

Comparisions with Phylo-NN: Figure 8 compares trait swapping results of Phylo-Diffusion and Phylo-NN for the same set of example species. Figure 8a shows trait swapping at level-3 for the source species of *Lepomis gulosus* (top) and target sub-tree of the *Morone* genus (bottom). In Phylo-NN, images generated by perturbing the Imageome sequences appear blurry (red circle), while Phylo-Diffusion effectively captures the splitting of dorsal fin (purple circle) and the horizontal stripe pattern of the *Morone* genus, while maintaining the fin structure of *Lepomis gulosus* (green circle). Similarly, Fig. 8b compares trait swapping for *Noturus miurus* (top) with the target sub-tree of *Notropis* genus (bottom) at level-2. For Phylo-NN, the perturbed images are almost identical to the source species. However, Phylo-Diffusion shows visible trait differences such as the absence of barbels and the caudal (or tail) fin beginning to fork or split (purple circle), which are traits picked from the target sub-tree of *Notropis* genus. Note that we had considered the same target sub-tree in Fig. 7a row 1 and observed similar trait differences in the generated images after perturbation, further validating the ability of Phylo-Diffusion to discover consistent evolutionary traits. We provide additional results comparing Phylo-Diffusion and Phylo-NN trait swapping results in the Appendix.

6 Conclusions and Future Work

In this work, we introduced Phylo-Diffusion, a novel framework for discovering evolutionary traits from images by structuring the embedding space of diffusion models using tree-based knowledge. In the future, our approach can be extended to work on other applications involving image data linked with phylogenies or pedigrees. Our work also has limitations that need to be addressed in future

research. For example, while our current work is limited to discretized trees with a fixed number of levels, future works can focus on discovering evolutionary traits at every internal node of the phylogenetic tree with varying levels without performing any discretization. Future works can also attempt to capture convergent changes in evolution, i.e., changes that occur repeatedly in different branches of the tree, and perform ancestral state reconstruction with uncertainty estimates.

Acknowledgments. This research is supported by National Science Foundation (NSF) awards for the HDR Imageomics Institute (OAC-2118240). We are thankful for the support of computational resources provided by the Advanced Research Computing (ARC) Center at Virginia Tech.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).

References

1. Great lakes invasives network project.https://greatlakesinvasives.org/portal/index.php
2. Rees, J.A., Cranston, K.: Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodivers. Data J. **5**, e12581 (2017). https://doi.org/10.3897/BDJ.5.e12581
3. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
4. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems, vol. 34, pp. 8780–8794 (2021)
5. Edmunds, R., et al.: Phenoscape: identifying candidate genes for evolutionary phenotypes. Mol. Biol. Evol. **33**, 13–24 (2015). https://doi.org/10.1093/molbev/msv223
6. Elhamod, M., et al.: Discovering novel biological traits from images using phylogeny-guided neural networks. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3966–3978 (2023)
7. Esser, P., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)
9. Gharaee, Z., et al.: A step towards worldwide biodiversity assessment: the BIOSCAN-1M insect dataset. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
10. Griffiths, A.J.: An Introduction to Genetic Analysis. Macmillan (2005)

11. Harmon, L., et al.: Early bursts of body size and shape evolution are rare in comparative data. Evolution **64**, 2385–96 (2010). https://doi.org/10.1111/j.1558-5646.2010.01025.x
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
14. Kapli, P., Yang, Z., Telford, M.J.: Phylogenetic tree building in the genomic age. Nat. Rev. Genetics **21**(7), 428–444 (2020). https://doi.org/10.1038/s41576-020-0233-0
15. Karpatne, A., et al.: Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. **29**(10), 2318–2331 (2017)
16. Karpatne, A., Jia, X., Kumar, V.: Knowledge-guided machine learning: current trends and future prospects. arXiv preprint arXiv:2403.15989 (2024)
17. Karpatne, A., Kannan, R., Kumar, V.: Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data. CRC Press (2022)
18. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
19. Lürig, M.D., Donoughe, S., Svensson, E.I., Porto, A., Tsuboi, M.: Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. Front. Ecol. Evol. **9**, 642774 (2021)
20. Manda, P., Balhoff, J., Lapp, H., Mabee, P., Vision, T.: Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. Genesis **53**, 561–571 (2015). https://doi.org/10.1002/dvg.22878
21. Nauta, M., Schlötterer, J., van Keulen, M., Seifert, C.: Pip-net: patch-based intuitive prototypes for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2744–2753 (2023)
22. Nemudryi, A.A., Valetdinova, K.R., Medvedev, S.P., Zakian, S.M.: TALEN and CRISPR/Cas genome editing systems: tools of discovery. Acta Naturae **6**(3), 19–40 (2014)
23. Paul, D., et al.: A simple interpretable transformer for fine-grained image classification and analysis. arXiv preprint arXiv:2311.04157 (2023)
24. Pennell, M., FitzJohn, R., Cornwell, W., Harmon, L.: Model adequacy and the macroevolution of angiosperm functional traits. Am. Nat. **186**, E33–E50 (2015). https://doi.org/10.1086/682022
25. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
27. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510 (2023)
28. Simões, T.R., Caldwell, M.W., Palci, A., Nydam, R.L.: Giant taxon-character matrices: quality of character constructions remains critical regardless of size. Cladistics **33**(2), 198–219 (2017)

29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
30. Stevens, S., et al.: Bioclip: a vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19412–19424 (2024)
31. Van Horn, G., Mac Aodha, O.: iNat Challenge 2021 - FGVC8 (2021). https://kaggle.com/competitions/inaturalist-2021
32. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Cub-200-2011. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
33. Zelditch, M.L., Fink, W.L., Swiderski, D.L.: Morphometrics, homology, and phylogenetics: quantified characters as synapomorphies. Syst. Biol. **44**(2), 179–189 (1995)
34. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

