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

Anonymous Author(s)
Affiliation
Address
email

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.

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* [1], see Figure 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 [2]) 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 [3], hindering rapid scientific advancement [4].

The growing deluge of large-scale image repositories in biology [5, 6, 7] 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) [8], 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 [9, 10, 11]. We ask the question:

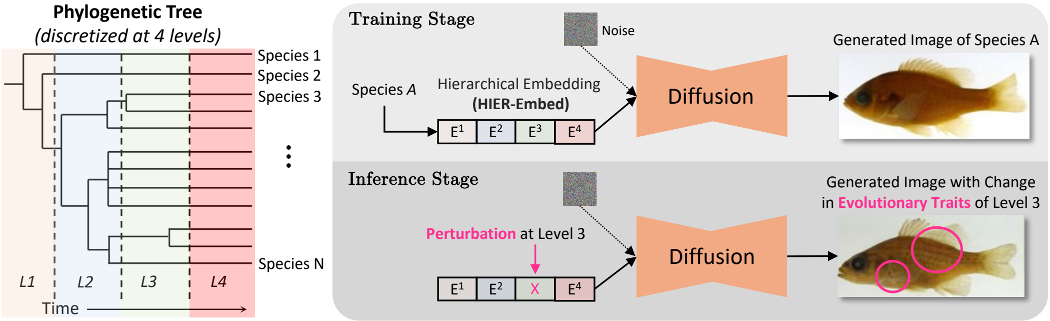


Figure 1: Overview of Phylo-Diffusion framework. Every species in the tree of life (phylogenetic tree) is encoded to a HIERarchical Embedding (HIER-Embed) comprising of four vectors (one for each phylogenetic level), which is used to condition a latent diffusion model to generate synthetic images of 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.

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 Figure 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 [12], 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 [2]. 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) [13, 14, 15] 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 [16]. 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 [17, 18]. 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 involves identifying and interpreting 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 [19, 20, 21]. 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 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 classification tasks, it is non-trivial to integrate them into generative modeling frameworks to produce synthetic images with controlled perturbations in the embedding space, similar to gene knockout and gene editing/swapping experiments

Phylogeny-guided Neural Networks (Phylo-NN): A recent work closely aligned with our goal of discovering evolutionary traits directly from images is Phylo-NN [22]. Phylo-NN uses an encoder-decoder architecture to represent images of organisms as structured sequences of feature vectors termed "Imageomes", that 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. 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. Additionally, Phylo-Diffusion allows for perturbations in 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.

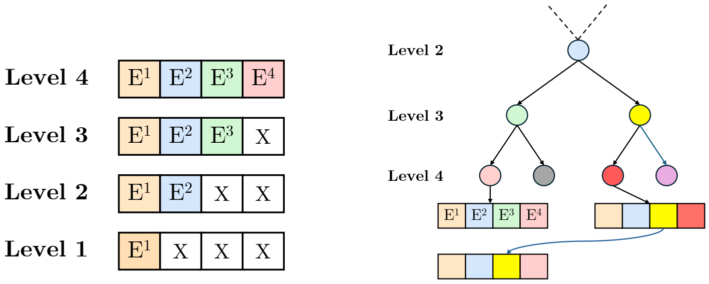
Background on Latent Diffusion Models (LDMs): Diffusion Models [23] learn a target distribution $p(x)$ by incrementally transforming a noisy sample x generated from a Gaussian distribution $\mathcal{N}(0, I)$ into one that is more likely to be generated from $p(x)$ over a series of timesteps T. While early frameworks of diffusion models like DDPM [23], DDIM [24] and ADM [25] suffered from high computational costs and long training/inference times, Latent Diffusion Models (LDMs) [8] 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 B 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:



(a) Trait Masking

(b) Trait Swapping

Figure 2: Schematics of the two proposed experiments for discovering evolutionary traits using Phylo-Diffusion.

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

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 $\texttt{Concat[.]}$ 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.

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 sub-tree 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 as 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-of-subtree species, as empirically demonstrated later in the Results Section.

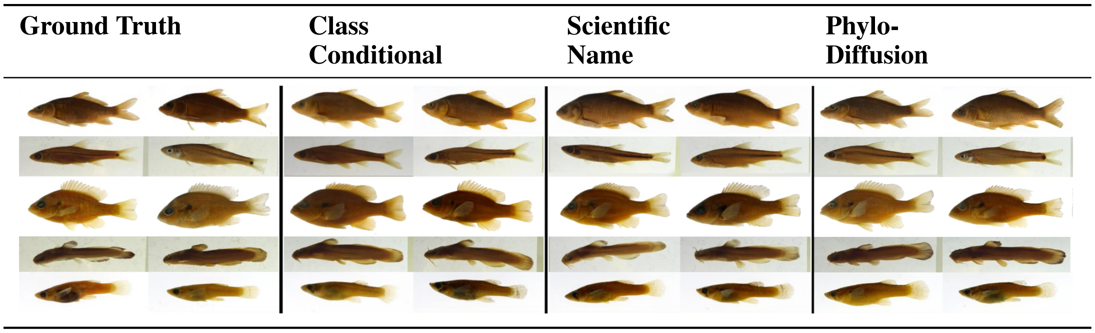


Figure 3: Comparing the quality of synthetic images generated by different conditioning mechanisms in LDMs. Every row corresponds to a different species and we show two samples per species for every conditioning mechanism. The order of species from top to bottom is *Cyprinus carpio*, *Notropis hudsonius*, *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 Figure 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) Project [26], comprising a total of 5434 images spanning 38 fish species. We obtained the phylogenetic tree of fish species from *opentree* [27] python package (see Appendix B 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 [28] of bird species in Appendix H.

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 [29] 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 the combined hierarchical embedding of $d = 512$ dimensions. Phylo-Diffusion uses this d-dimensional embedding to condition LDMs through cross-attention in denoising the U-Net backbone and train LDMs without classifier-free guidance. We used VQGAN [30] 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 Appendix A.

Table 1: Quantitive comparison of generated images sampled using DDIM [24] (100 samples/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

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 [31]. 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 Figure 3, with additional images provided in Appendix F. We also show the robustness of Phylo-Diffusion's results with varying embedding dimensions and phylogenetic levels in Appendix G.

5.2 Classification Accuracy

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

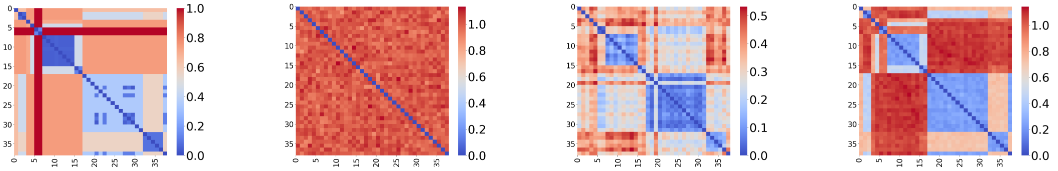
We used a separate model for species classification, specifically a ResNet-18 model [32] trained using the same train/val split as Phylo-Diffusion. The primary objective behind building this classifier is to verify if images generated by Phylo-Diffusion contain sufficient discriminatory information to 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 to unseen species in Appendix G.

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 Figure 4. Ideally, we expect distances in 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, distance matrix exhibits notable similarities to the phylogenetic distances, thanks to the hierarchical nature of information contained in scientific names (i.e., *genus* & *species*). 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.

5.4 Trait Masking Results

To obtain classification probabilities or logits associated with generated images, we employ the classifier detailed in Section 5.2. For masked embeddings of subtrees at level 3, defined as $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{z_{noise}}]$, we analyze the 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, Figure 5 presents probability distributions for species within the *Lepomis* subtree when the full set



(a) GT Phylogeny (b) Class Conditional (c) Scientific Name (d) HIER-Embed

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

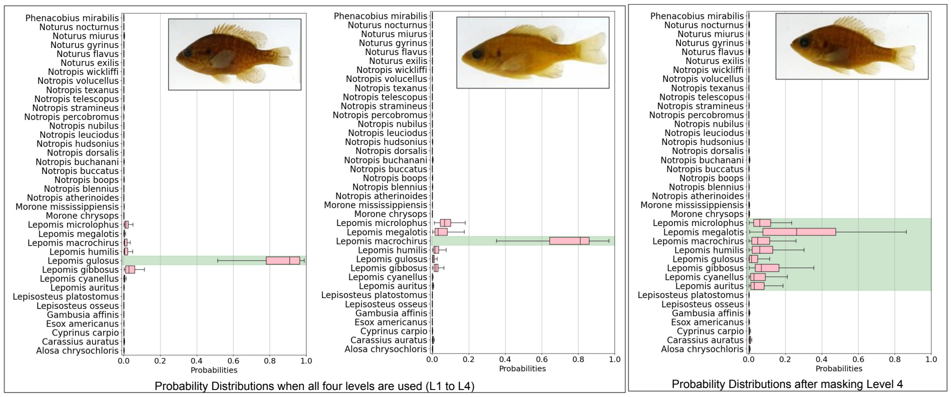


Figure 5: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Lepomis gulosus* and *Lepomis macrochirus* (shown in green) that are part of the same sub-tree till level 3. Right: class probability distributions of images generated by masking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

of hierarchical encodings $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{E^4}]$ are provided. It demonstrates that the probabilities are significantly higher for targeted class, as intended for image generation. After masking, we observe that generated images are very similar and capture common features of the *Lepomis* genus. For all our calculations and plots, we generate 100 images for each subtree and node. Appendix C contains additional histograms that detail logit distributions across all different subtrees at each level, offering comprehensive insights into hoe the model learns the hierarchical structures across different levels.

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 ($S_k \in \mathcal{S}_i^l$), we compute $P_{ref}(S_k)$ by averaging over $|\mathcal{S}_i^l| - 1$ probability values since we exclude the case when S_k is used to generate images. And when S_k is outside of the sub-tree ($S_k \notin \mathcal{S}_i^l$), 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 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 observe that species within the subtree exhibit a more pronounced increase in probabilities compared to species

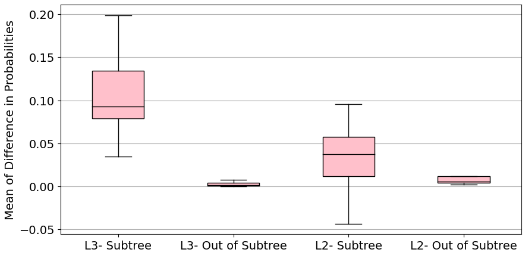


Figure 6: Box plot for mean of difference in probabilities for species within subtree and out of subtree for level-3 & level-2.

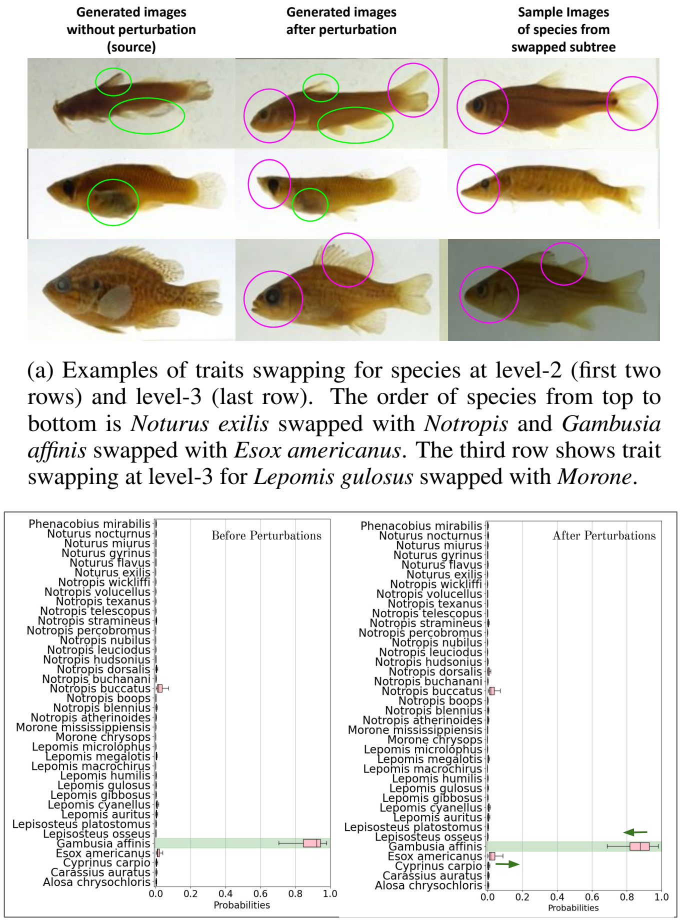
outside the subtree, aligning with our expectations. Notably, this trend is consistently observed across both levels 2 and 3. More details of class-wise probability distribution shifts for each level are provided in Appendix C. The outcomes of these experiments affirm that Phylo-Diffusion effectively 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.

5.5 Trait Swapping Results

Figure 7 shows examples of trait swapping results enabled by Phylo-Diffusion. For the first example (first row of Figure 7a), we swap the level-2 embedding of source species *Noturus exilis* with level-2 embedding of its sibling group *Notropis/ Carassius*. The goal here is to discover traits of *Noturus exilis* inherited at level-2 that differentiate 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, the *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 (Figure 7a, row 2), we swap level-2 information of *Gambusia affinis* (left) with *Esox americanus* (right). The generated images 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 the classifier detailed in Section 5.2. 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 Figure 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 highlighted in purple begins to split. 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 [33, 34]. Additional visualizations for trait swapping are provided in Appendix D.

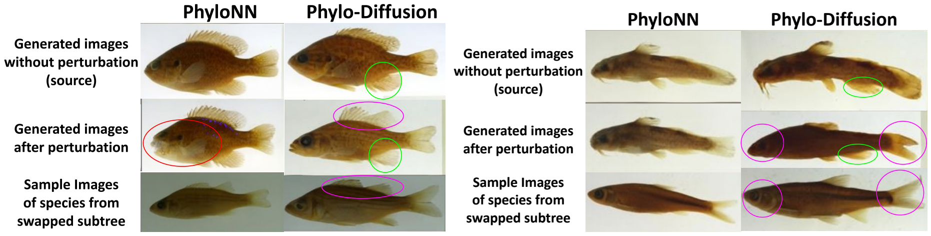
Comparisions with Phylo-NN: Figure 8 compares trait swapping results of Phylo-Diffusion and Phylo-NN for the same set of example species. Figure 8a shows trait swapping at level-3 for the source species of *Lepomis gulosus* (top) and target sub-tree of the *Morone* genus (bottom). In Phylo-NN, images 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, Figure 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 Figure 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 Appendix E.



(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*.

Figure 7: Examples of trait swapping results.


PhyloNN
Phylo-Diffusion
Generated images without perturbation (source)
Generated images after perturbation
Sample Images of species from swapped subtree
PhyloNN
Phylo-Diffusion
Generated images without perturbation (source)
Generated images after perturbation
Sample Images of species from swapped subtree

(a) Swapping L3 traits: *Lepomis gulosus* with *Morone*

(b) Swapping L2 traits: *Notorus mirurus* with *Notropis*

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

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, as the current method is limited to working on discretized trees with a fixed number of levels. Future work can focus on discovering evolutionary traits at every internal node of the phylogenetic tree at varying levels. 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.

References

[1] Paschalia Kapli, Ziheng Yang, and Maximilian J. Telford. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444, May 2020. ISSN 1471-0064. doi: 10.1038/s41576-020-0233-0. URL `http://dx.doi.org/10.1038/s41576-020-0233-0`.

[2] A. A. Nemudryi, K. R. Valetdinova, S. P. Medvedev, and S. M. Zakian. Talen and crispr/cas genome editing systems: Tools of discovery. In *Acta naturae, 6(3)*, page 19–40, 2014.

[3] Tiago R Simões, Michael W Caldwell, Alessandro Palci, and Randall L Nydam. Giant taxon-character matrices: quality of character constructions remains critical regardless of size. *Cladistics*, 33(2):198–219, 2017.

[4] Moritz D Lürig, Seth Donoughe, Erik I Svensson, Arthur Porto, and Masahito Tsuboi. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. *Frontiers in Ecology and Evolution*, 9:642774, 2021.

[5] Grant Van Horn and Oisin Mac Aodha. iNat Challenge 2021 - FGVC8, 2021. URL `https://kaggle.com/competitions/inaturalist-2021`.

[6] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, 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*, pages 19412–19424, 2024.

[7] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, Iuliia Zarubiieva, Joakim Bruslund Haurum, Scott Lowe, Jaclyn McKeown, Chris Ho, Joschka McLeod, Yi-Yun Wei, et al. A step towards worldwide biodiversity assessment: The bioscan-1m insect dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023.

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[12] Anthony JF Griffiths. *An introduction to genetic analysis*. Macmillan, 2005.

[13] Anuj Karpatne, Xiaowei Jia, and Vipin Kumar. Knowledge-guided machine learning: Current trends and future prospects. *arXiv preprint arXiv:2403.15989*, 2024.

[14] Anuj Karpatne, Ramakrishnan Kannan, and Vipin Kumar. *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data*. CRC Press, 2022.

[15] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29 (10):2318–2331, 2017.

[16] Miriam Leah Zelditch, William L Fink, and Donald L Swiderski. Morphometrics, homology, and phylogenetics: quantified characters as synapomorphies. *Systematic Biology*, 44(2):179–189, 1995.

[17] Richard Edmunds, Baofeng Su, James Balhoff, Brian Eames, Wasila Dahdul, Hilmar Lapp, John Lundberg, Todd Vision, Rex Dunham, Paula Mabee, and Monte Westerfield. Phenoscape: Identifying candidate genes for evolutionary phenotypes. *Molecular biology and evolution*, 33, 10 2015. doi: 10.1093/molbev/msv223.

[18] Prashanti Manda, James Balhoff, Hilmar Lapp, Paula Mabee, and Todd Vision. Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. *Genesis (New York, N.Y. : 2000)*, 53, 07 2015. doi: 10.1002/dvg.22878.

[19] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[20] Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2744–2753, June 2023.

[21] Dipanjyoti Paul, Arpita Chowdhury, Xinqi Xiong, Feng-Ju Chang, David Carlyn, Samuel Stevens, Kaiya Provost, Anuj Karpatne, Bryan Carstens, Daniel Rubenstein, et al. A simple interpretable transformer for fine-grained image classification and analysis. *arXiv preprint arXiv:2311.04157*, 2023.

[22] Mohannad Elhamod, Mridul Khurana, Harish Babu Manogaran, Josef C Uyeda, Meghan A Balk, Wasila Dahdul, Yasin Bakis, Henry L Bart Jr, Paula M Mabee, Hilmar Lapp, 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*, pages 3966–3978, 2023.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[25] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[26] Great lakes invasives network project, https://greatlakesinvasives.org/portal/index.php.

[27] Jonathan A. Rees and Karen Cranston. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*, 5:e12581, 2017. ISSN 1314-2836. doi: 10.3897/BDJ.5.e12581. URL `https://doi.org/10.3897/BDJ.5.e12581`.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Cub-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[31] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Luke Harmon, Jonathan Losos, Jonathan Davies, Rosemary Gillespie, John Gittleman, W. Jennings, Kenneth Kozak, Mark McPeek, Franck Moreno-Roark, Andy Purvis, Robert Ricklefs, Dolph Schluter, James Schulte II, Ole Seehausen, Brian Sidlauskas, Omar Torres-Carvajal, Jason Weir, and Arne Mooers. Early bursts of body size and shape evolution are rare in comparative data. *Evolution; international journal of organic evolution*, 64:2385–96, 04 2010. doi: 10.1111/j.1558-5646.2010.01025.x.

[34] Matthew Pennell, Richard FitzJohn, William Cornwell, and Luke Harmon. Model adequacy and the macroevolution of angiosperm functional traits. *American Naturalist*, 186:E33–E50, 07 2015. doi: 10.1086/682022.

Supplementary Materials

Here is a summary of additional details and experiments included in the Appendices.

1. Appendix A: Hyperparameter Settings and Training Details
2. Appendix B: Details about Phylogenetic Tree
3. Appendix C: Additional Details about Trait Masking Experiments
4. Appendix D: Additional Examples for Trait Swapping Experiments
5. Appendix E: Additional Comparisons with PhyloNN
6. Appendix F: Additional Samples of Generated Images
7. Appendix G: Ablation Results
8. Appendix H: CUB Dataset Results

A Hyperparameter Settings and Training Details

Table 3 lists all the hyperparameters for the models trained. We used cross-attention as the conditioning mechanism for all the models and all the models were trained from scratch. At the inference stage, we used DDIM [24]sampling with 200 steps. For computing metrics like FID, IS, , we use ADM's [25] TensorFlow evaluation script.

Table 3: Hyperparameter settings of the baselines and Phylo-Diffusion.

Model	Class Conditional	Scientific Name	Phylo-Diffusion
z-shape	64 × 64 × 3	64 × 64 × 3	64 × 64 × 3
Diffusion Steps	1000	1000	1000
Noise Schedule	linear	linear	linear
Model Size	469M	902M	469M
Channels	224	224	224
Depth	2	2	2
Channel Multiplier	1,2,3,4	1,2,3,4	1,2,3,4
Attention resolutions	32, 16, 8	32, 16, 8	32, 16, 8
Number of Heads	32	32	32
Dropout	-	-	-
Batch Size	8	8	8
Iterations	400k	400k	400k
Learning Rate	4e-5	4e-5	4e-5
Scale	1	1	1
Embedding Dimension	1 x 512	77 x 768	1 x 512
Transformers Depth	1	1	1

All diffusion models require about 7 days to train on a single A100 GPU for both bird and fish datasets. Inference throughput is 0.9 samples/sec using DDIM with 200 steps computed over generating 100 images per class. We do not have any additional overheads in training and inference time compared to LDMs.

B Details about Phylogenetic Tree

Figure 9 shows the phylogeny tree for all the species in the fish dataset along with the information of the four discrete levels used in our study (marked by different colored circles). Table 4 and 5 list out all the groupings (subtrees) made after discretizing the tree into four levels where the fourth level is the species itself.

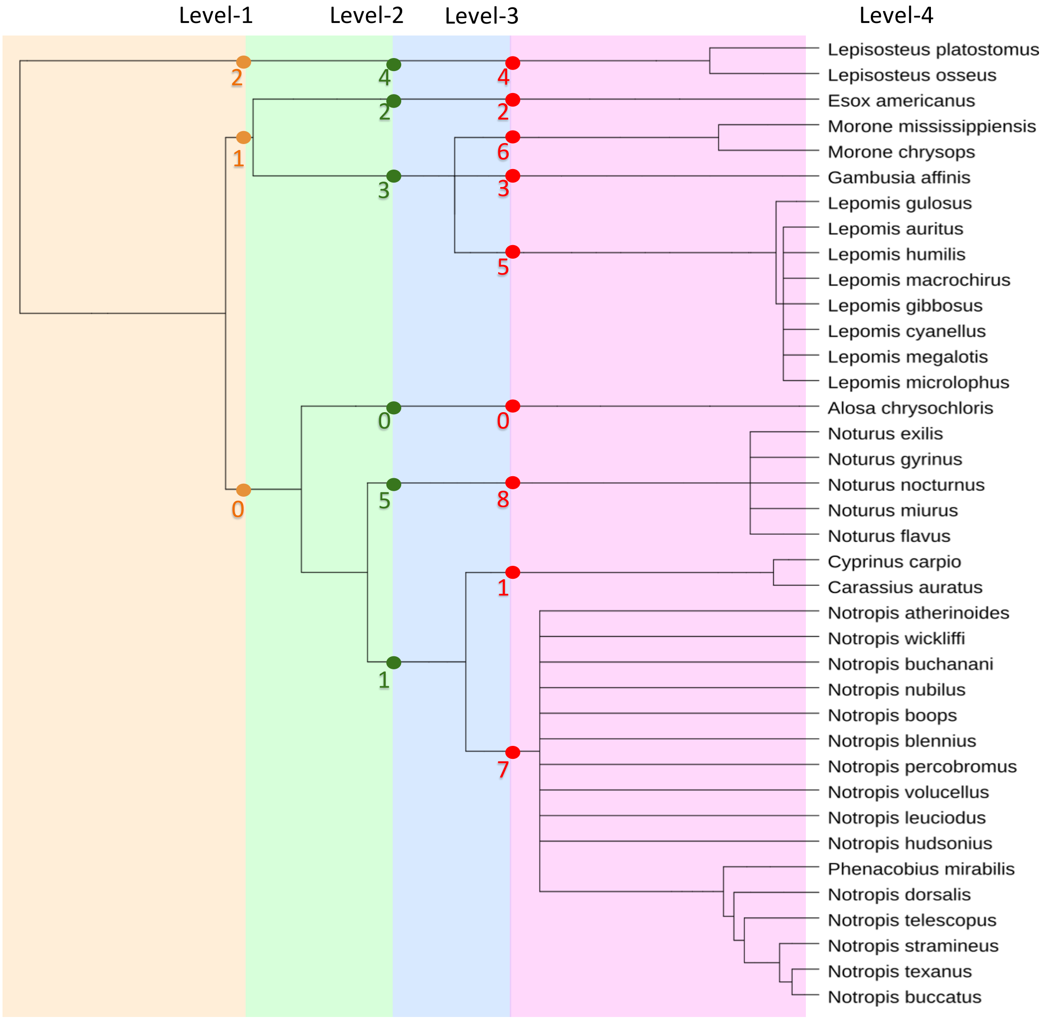


Figure 9: Phylogeny tree for fishes for all 38 species. Filled circles show nodes of the subtrees defined at each of the four levels after discretization.

C Additional Details about Trait Masking Experiments

C.0.1 Additional Visualizations of Changes in Probability Distributions after Masking:

Figure 10, 11, 12, 13 and 14 show additional examples of changes in probability distributions when level-4 information is replaced with *noise*. In each figure, the first two plots display probability distributions (or logits) of images generated using embeddings from all four levels, *i.e.* $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{E^3}, \mathbf{E^4}]$, of two representative species sharing a common ancestry up to level-3 (highlighted in green). We show that the logits are higher for the targeted species as expected. The third plot logits after masking level-4 embeddings, leading to a dispersion of probabilities across all descendant species within the subtree up to level-3 (highlighted in green). The only exception is Figure 13, where there is some skewness in the logits of descendant species, which is likely due to the data imbalance across classes at higher levels of the tree and also due to biases in the classifier (classifier test accuracy is 85% as reported in sec:classifier). In this case, the classifier sometimes misclassifies *Notropis boops* as *Notropis blennius* in the first plot and *Notropis dorsails* as *Notropis buccatus* in the second plot. Consequently, the third plot for the *Notropis* subtree shows a higher probability for *Notropis blennius*. Similarly, Figure 15, 16 and 17 provide examples of trait masking where both Level 3 and 4 are replaced with *noise*, *i.e.* $[\mathbf{E^1}, \mathbf{E^2}, \mathbf{z_{noise}}, \mathbf{z_{noise}}]$. We observe a similar trend in the dispersion of probabilities across all descendant species within the same subtree at level-2. In all trait masking visualizations,

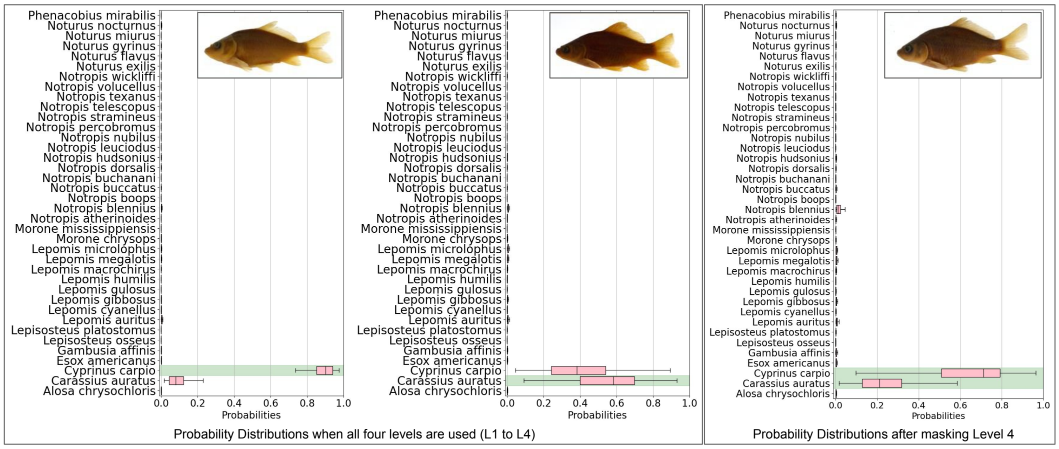
Table 4: Phylogenetic groupings of fish species included in this study at different ancestry levels.

Level	Node at level	Species groupings
3	Node 0	*Alosa chrysochloris*
	Node 1	*Carassius auratus*, *Cyprinus carpi*
	Node 2	*Esox americanus*
	Node 3	*Gambusia affinis*
	Node 4	*Lepisosteus osseus*, *Lepisosteus platostomus*
	Node 5	*Lepomis auritus*, *Lepomis cyanellus*, *Lepomis gibbosus*, *Lepomis gulosus*, *Lepomis humilis*, *Lepomis macrochirus*, *Lepomis megalotis*, *Lepomis microlophus*
	Node 6	*Morone chrysops*, *Morone mississippiensis*
	Node 7	*Notropis atherinoides*, *Notropis blennius*, *Notropis boops*, *Notropis buccatus*, *Notropis buchanani*, *Notropis dorsalis*, *Notropis hudsonius*, *Notropis leuciodus*, *Notropis nubilus*, *Notropis percobromus*, *Notropis stramineus*, *Notropis telescopus*, *Notropis texanus*, *Notropis volucellus*, *Notropis wickliffi*, *Phenacobius mirabilis*
	Node 8	*Noturus exilis*, *Noturus flavus*, *Noturus gyrinus*, *Noturus miurus*, *Noturus nocturnus*
2	Node 0	*Alosa chrysochloris*
	Node 1	*Carassius auratus*, *Cyprinus carpio*, *Notropis atherinoides*, *Notropis blennius*, *Notropis boops*, *Notropis buccatus*, *Notropis buchanani*, *Notropis dorsalis*, *Notropis hudsonius*, *Notropis leuciodus*, *Notropis nubilus*, *Notropis percobromus*, *Notropis stramineus*, *Notropis telescopus*, *Notropis texanus*, *Notropis volucellus*, *Notropis wickliffi*, *Phenacobius mirabilis*
	Node 2	*Esox americanus*
	Node 3	*Gambusia affinis*, *Lepomis auritus*, *Lepomis cyanellus*, *Lepomis gibbosus*, *Lepomis gulosus*, *Lepomis humilis*, *Lepomis macrochirus*, *Lepomis megalotis*, *Lepomis microlophus*, *Morone chrysops*, *Morone mississippiensis*
	Node 4	*Lepisosteus osseus*, *Lepisosteus platostomus*
	Node 5	*Noturus exilis*, *Noturus flavus*, *Noturus gyrinus*, *Noturus miurus*, *Noturus nocturnus*

we consistently observe that logits of generated images of species within the subtree (highlighted in green) are higher than for species outside the subtree. This demonstrates Phylo-Diffusion's ability to effectively capture hierarchical information at various levels of the phylogenetic tree.

Table 5: Phylogenetic groupings of species included in this study at different ancestry levels (continued from Table 4)

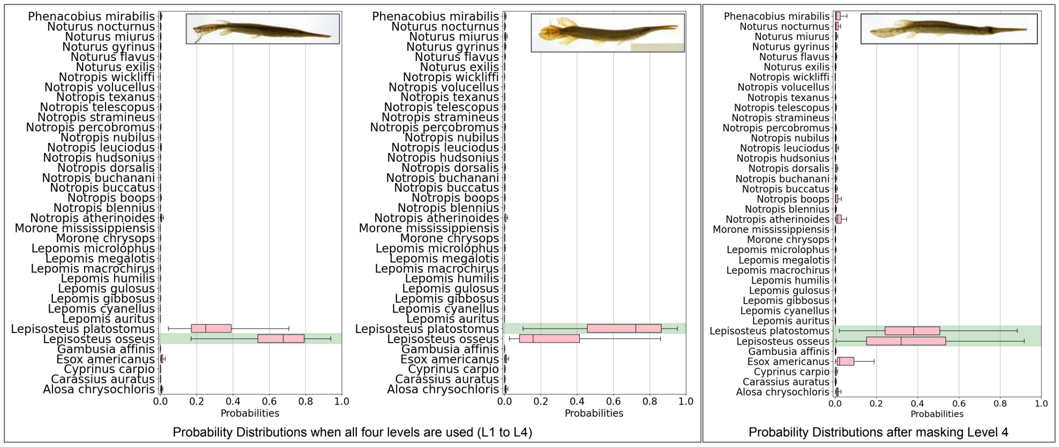
Level	Node at level	Species groupings
1	Node 0	*Alosa chrysochloris*, *Carassius auratus*, *Cyprinus carpio*, *Notropis atherinoides*, *Notropis blennius*, *Notropis boops*, *Notropis buccatus*, *Notropis buchanani*, *Notropis dorsalis*, *Notropis hudsonius*, *Notropis leuciodus*, *Notropis nubilus*, *Notropis percobromus*, *Notropis stramineus*, *Notropis telescopus*, *Notropis texanus*, *Notropis volucellus*, *Notropis wickliffi*, *Noturus exilis*, *Noturus flavus*, *Noturus gyrinus*, *Noturus miurus*, *Noturus nocturnus*, *Phenacobius mirabilis*
	Node 1	*Esox americanus*, *Gambusia affinis*, *Lepomis auritus*, *Lepomis cyanellus*, *Lepomis gibbosus*, *Lepomis gulosus*, *Lepomis humilis*, *Lepomis macrochirus*, *Lepomis megalotis*, *Lepomis microlophus*, *Morone chrysops*, *Morone mississippiensis*
	Node 2	*Lepisosteus osseus*, *Lepisosteus platostomus*



Probability Distributions when all four levels are used (L1 to L4)

Probability Distributions after masking Level 4

Figure 10: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Cyprinus carpio* and *Carassius auratus* (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)



Probability Distributions when all four levels are used (L1 to L4)

Probability Distributions after masking Level 4

Figure 11: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Lepisosteus osseus* and *Lepisosteus platostomus* (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)

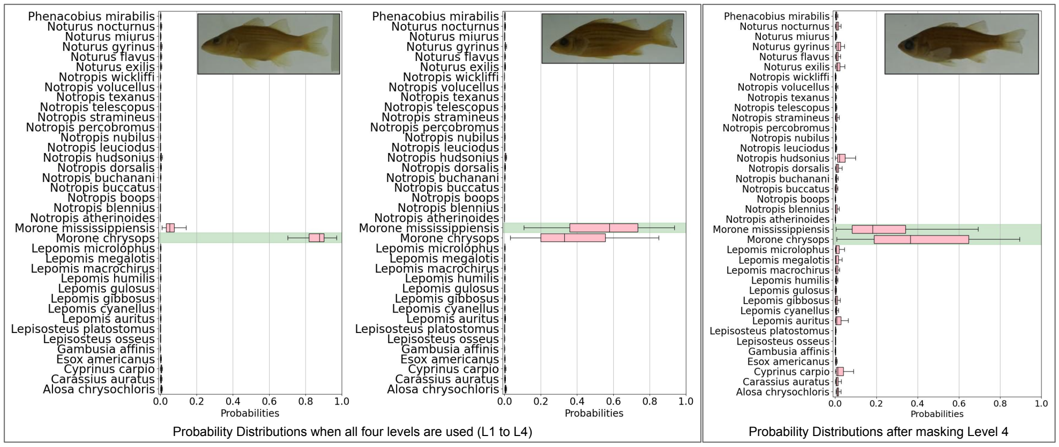


Figure 12: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Morone chrysops* and *Morone mississippiensis* (shown in green) that are part of the same sub-tree till level 3. Right: class probability distributions of images generated by masking level 4 (descendant species that have common ancestry till level 3 are highlighted in green)

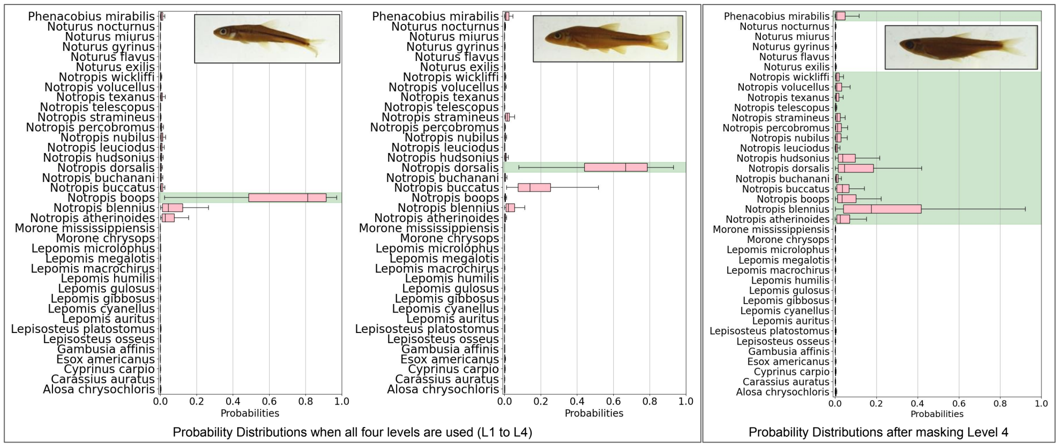


Figure 13: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Notropis boops* and *Notropis dorsalis* (shown in green) that are part of the same sub-tree till level 3. Right: class probability distributions of images generated by masking level 4 (descendant species that have common ancestry till level 3 are highlighted in green)

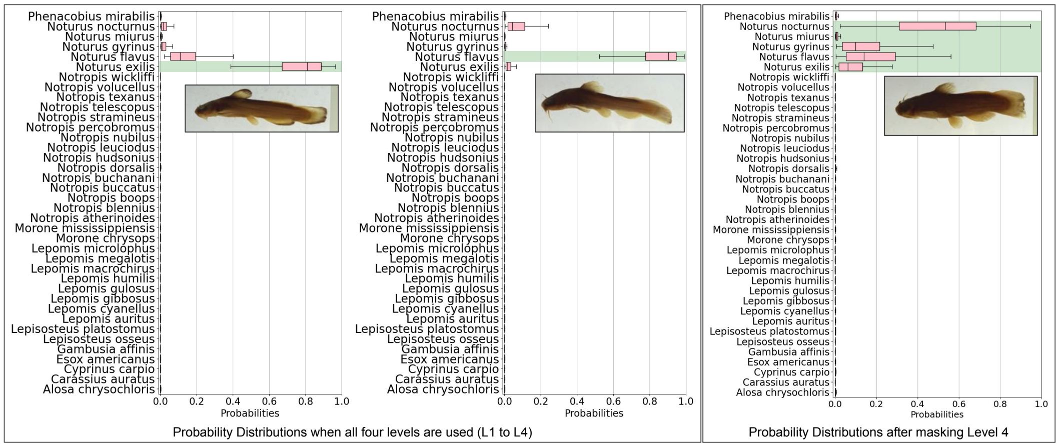


Figure 14: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Noturus exilis* and *Noturus falvus* (shown in green) that are part of the same sub-tree till level 3. Right: class probability distributions of images generated by masking level 4 (descendant species that have common ancestry till level 3 are highlighted in green)

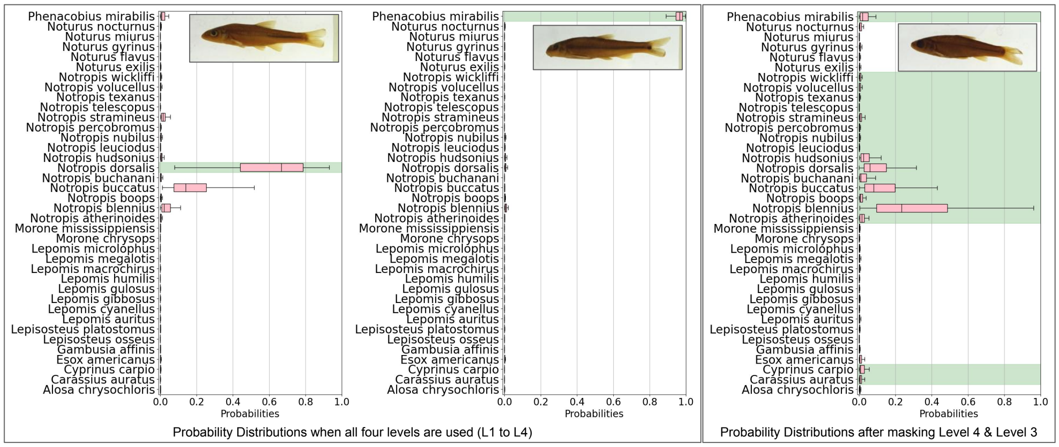


Figure 15: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Notropis dorsalis* and *Phenacobius mirabilis* (shown in green) that are part of the same sub-tree till level 2. Right: class probability distributions of images generated by masking level 3 and level 4 (descendant species that have common ancestry till level 2 are highlighted in green)

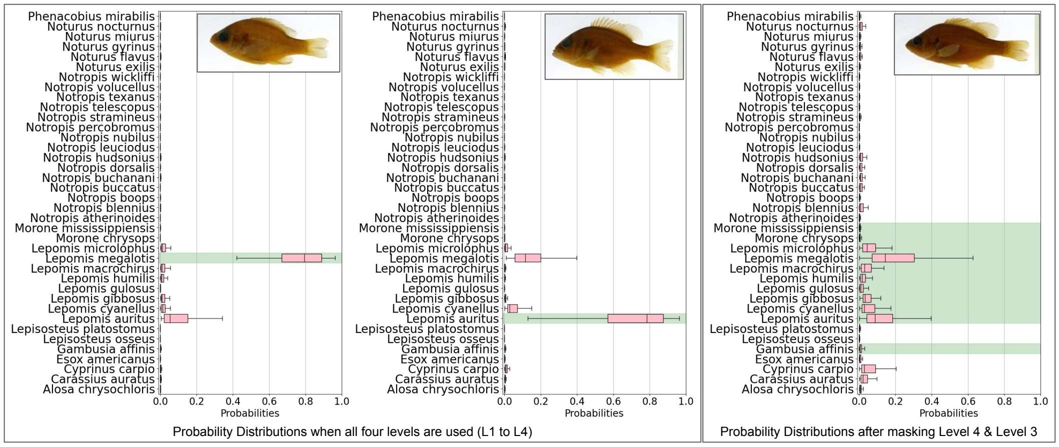


Figure 16: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Lepomis megalotis* and *Lepomis auritus* (shown in green) that are part of the same sub-tree till level 2. Right: class probability distributions of images generated by masking level 3 and level 4 (descendant species that have common ancestry till level 2 are highlighted in green)

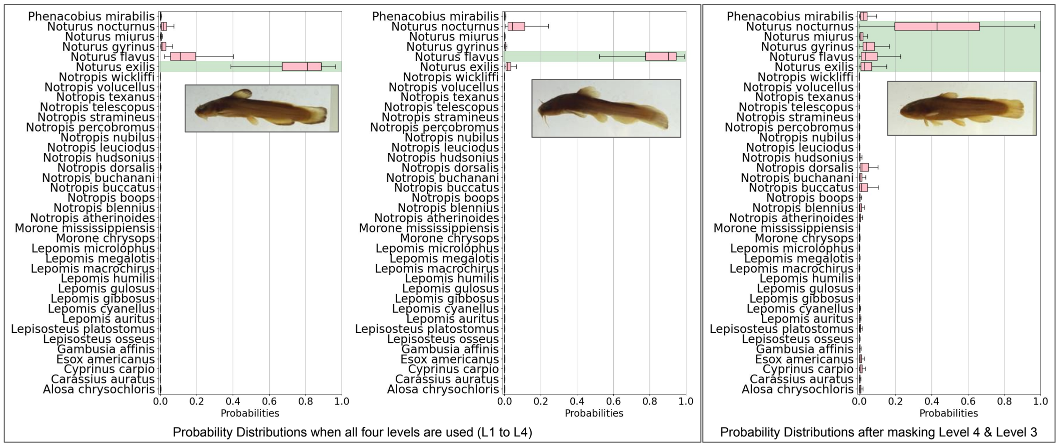


Figure 17: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Noturus exilis* and *Noturus flavus* (shown in green) that are part of the same sub-tree till level 2. Right: class probability distributions of images generated by masking level 3 and level 4 (descendant species that have common ancestry till level 2 are highlighted in green)

C.0.2 Additional Quantitative Results of Trait Masking:

Tables 6 and 7 show the change in probabilities for different nodes at levels 3 and 2, respectively. We can see that indeed P^{sub}_{diff} is larger than P^{out}_{diff} for all internal nodes at levels 2 and 3 (except node 5 at level 2), indicating that Phylo-Diffusion is capturing the necessary hierarchical information required for the dispersion of probabilities after masking. Figure 6 in the main paper shows the box plots of Tables 6 and 7. Table 8 summarizes this information by showing the average P^{sub}_{diff} and P^{out}_{diff} for all nodes at a given level. It is important to note that for this experiment, we focus on nodes that have more than one species in the defined subtree.

Table 6: Average change in probability distributions for every node at Level 3.

Node	Subtree	Out-of-Subtree
Node 1	0.1988	0.0018
Node 4	0.0952	0.0051
Node 5	0.0753	0.0007
Node 6	0.0903	0.0076
Node 7	0.0346	0.0006
Node 8	0.1472	0.0003

Table 7: Average change in probability distributions for every node at Level 2.

Node	Subtree	Out-of-Subtree
Node 1	0.0299	0.0023
Node 3	0.0449	0.0062
Node 4	0.0952	0.0051
Node 5	-0.0434	0.0292

Table 8: Average change in probability distributions across all nodes at a certain level.

Levels	Subtree	Out-of-Subtree
Level 3	0.1070	0.0027
Level 2	0.0316	0.0107

D Additional Examples for Trait Swapping Experiments

D.0.1 Additional Visualizations of Trait Swapping Experiments:

Figure 18 illustrates trait swapping for the source species *Noturus exilis* (left), where the information at Level-2 is swapped with that of a sibling subtree at Node *B* (right). The image in the center is generated using the trait swapped embedding. This visualization of the perturbed species helps us study the trait changes that would have branched out at level-2 between Node *A* and Node *B*.In the generated image (center), we observe the absence of barbels(whiskers), and the caudal fin (tail) is getting forked (or split) highlighted in pink, which are traits adopted from species in the subtree at *B* (*Notropis*). Whereas other fins like the dorsal, pelvic, and anal fin still resemble the source species *Noturus exilis* highlighted in green. The same is also reflected in the change of probability distribution after perturbations; the probability distribution of source species *Noturus exilis* decreases and the probability of it being a *Notropis* increases slightly.

Similarly in Figure 19, for the studied species *Lepomis gulosus* (left), the information at Level 3 is swapped with subtree at Node B (*Morone*). The perturbed species generated (center) captures traits from both lineages. The spotted pattern in the body and fins is retained from *Lepomis* (Node *A*) but these spots now start to follow the horizontal stripes pattern observed in *Morone* (species at Node *B*). Additionally, the dorsal fin highlighted in pink starts to split into two, with the left half retaining the spiny structure from *Lepomis* highlighted in green. This observation suggests that the species at Node *A* and *B* possess distinct traits at this level-3 branching node. The same is reflected in the shift in probability distributions towards species in Node *B* after trait swapping.

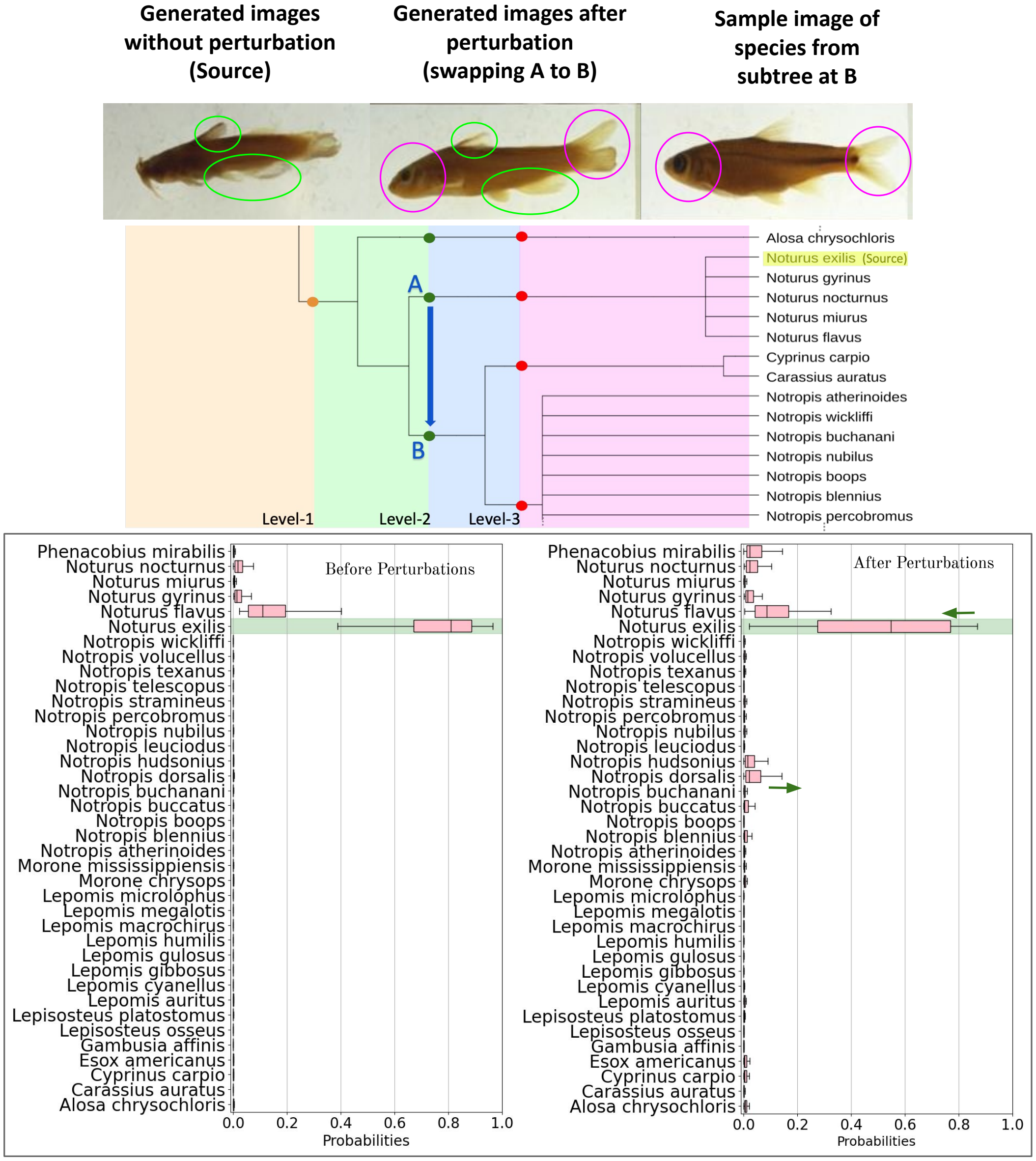


Figure 18: Visualization of changes in traits after swapping information at Level 2 (Node *A*) for *Noturus exilis* (left) with its sibling subtree at Node *B*(right) to generate perturbed species (center). Traits shared with the source species are outlined in green, whereas those shared with the sibling subtree at Node B are outlined in pink.

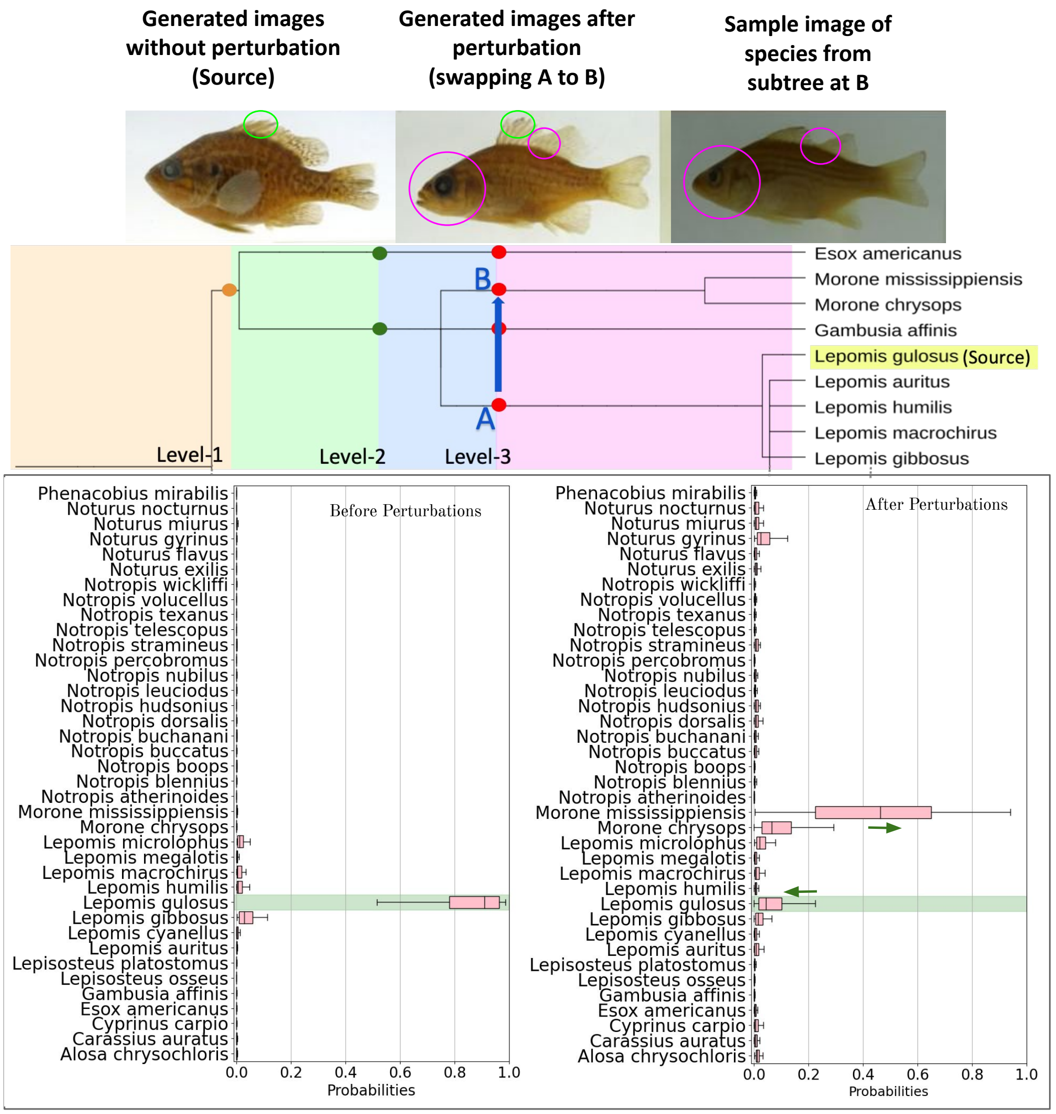


Figure 19: Visualization of changes in traits after swapping information at Level 2 (Node A) for *Lepomis gulosus* (left) with its sibling subtree at Node B(right) to generate perturbed species (center). Traits shared with the source species are outlined in green, whereas those shared with the sibling subtree at Node B are outlined in pink.

E Additional Comparisons with PhyloNN

Figure 20 compares the trait swapping experiment for Phylo-Diffusion with the PhyloNN baseline, where level-2 information of *Gambusia affinis* is replaced with that of *Esox americanus*. In the highlighted pink circle, the face of the image generated after perturbations (center) becomes more pointed, and the body shape flattens to resemble *Esox americanus*. This perturbed image also retains traits like the caudal (tail) fin and the black-spotted pattern towards the bottom (highlighted in green) from the source species, *Gambusia affinis*. The differences observed with Phylo-Diffusion are notable, whereas the PhyloNN generates a perturbed image nearly identical to the original, showing no significant changes.

Similarly, Figure 21 shows a comparison after replacing level-2 information of *Notropis husonius* with that of *Noturus*. For Phylo-Diffusion, the caudal (tail) fin is vibily joining highlighted in pink, resembling the caudal fins of *Noturus*. This change is analogous to Figure 18, where level-2 information of *Noturus* was replaced with *Notropis* (vice-versa), resulting in the caudal (tail) fin getting forked or split. Hence, this helps us understand that at Level-2, the two species diverged to develop different caudal fins. However, for PhyloNN, the generated image after trait-swapping is blurry, and most of the traits still closely resemble close the source species, which is unlikely given that the level-2 embeddings have been replaced.

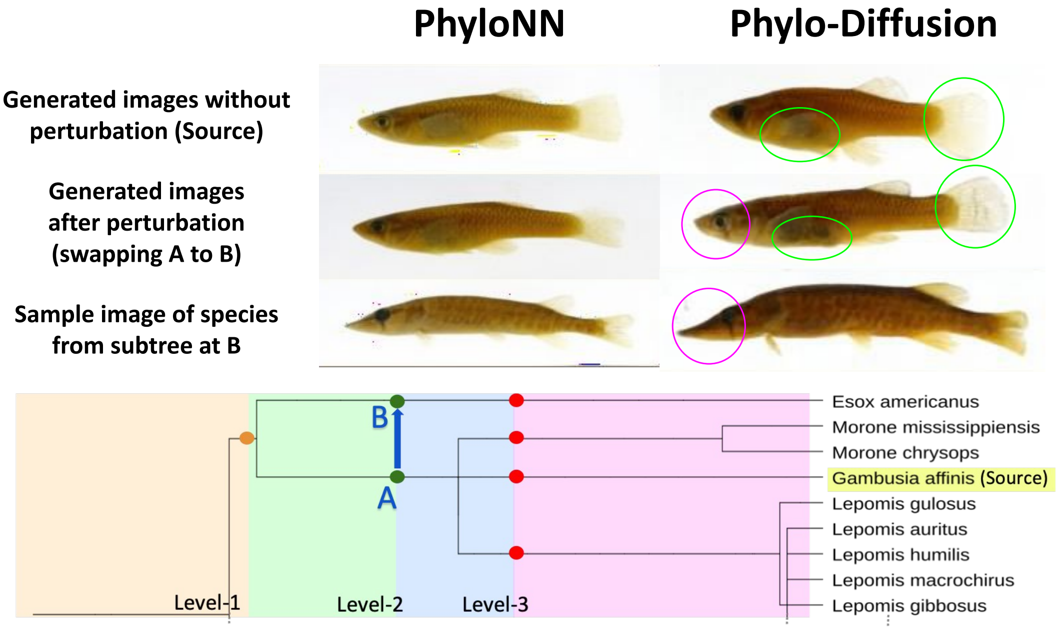


Figure 20: Comparison of PhyloNN with Phylo-Diffusion (ours) for trait swapping where the Level-2 information (Node *A*) of *Gambusia affinis* is swapped with its sibling subtree at Node *B* to generate perturbed species (center). Traits shared with the source species are outlined in green, whereas those shared with the sibling subtree at Node B are outlined in pink.

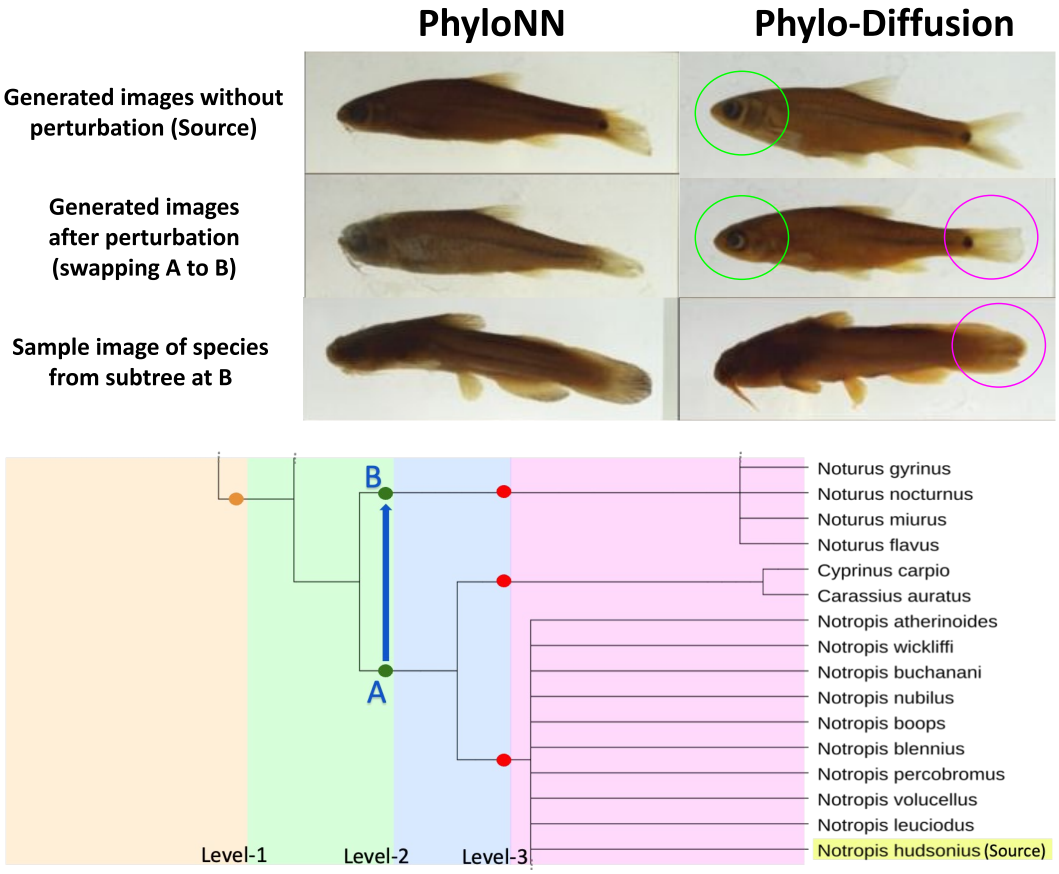


Figure 21: Comparison of PhyloNN with Phylo-Diffusion (ours) for trait swapping where the Level-2 information (Node *A*) of *Notropis husonius* is swapped with its sibling subtree at Node *B* to generate perturbed species (center). Traits shared with the source species are outlined in green, whereas those shared with the sibling subtree at Node B are outlined in pink.

F Additional Samples of Generated Images

Figure 22 shows additional examples of generated images for different species using Phylo-Diffusion. Each row of the figure depicts the generated images for the same species while the different rows represent distinct species. Notably, we observe inter-class variations among species belonging to the same class, such as differences in fish orientation and size.

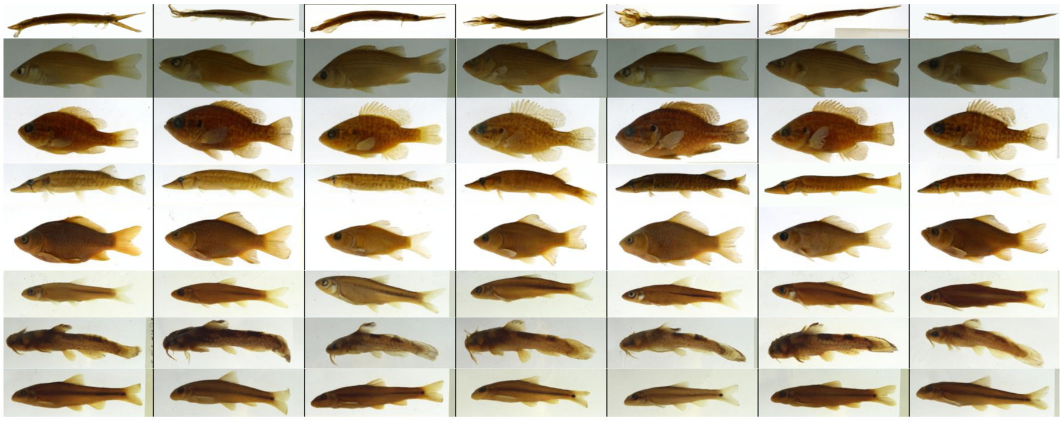


Figure 22: Comparison of images of different species generated by HEIR-Embed where the generated images for a given row depict variations for the same species while the different rows represent distinct species. The order of species from top to bottom is *Lepisosteus osseus, Morone chrysops, Lepomis gulosus, Esox americanus, Carassius auratus, Notropis blennius, Noturus exilis, Phenacobius mirabilis*

G Ablation Results

G.1 Generalization to Unseen Species: Leave-three-out

As an additional ablation experiment, we conduct a leave-three-out experiment by excluding three species from different subtrees during training to test the model's ability to generalize to new species and situate them in the phylogeny. This experiment involves training the model excluding three species, *Notropis blennius, Noturus gyrinus, and Lepomis humilis* that belong to different subtrees as seen in Table 4. The generated images from the three subtrees after trait masking closely resemble the actual images of the 3 species, with an F1 score of 95.6 on a classifier trained to discriminate the 3 species. This experiment underscores the robustness and accuracy of Phylo-Diffusion in embedding and generating phylogenetically consistent images.

G.2 Effect of Varying the Number of Levels in Phylo-Diffusion

To demonstrate the robustness of Phylo-Diffusion, we perform ablation experiments with varying numbers of levels in the discretization of the phylogeny tree. We show that the choice of the number of levels depends on the depth of the phylogenetic tree and the internal nodes to be studied. We train models with $\{2, 4, 6, 8\}$ levels on the phylogeny shown in Figure 9. Table 9 demonstrates that the model is robust to the choice of the number of levels.

Table 9: Quantitative results for Phylo-Diffusion with varying number of levels in the discretized phylogeny tree.

# levels	FID ↓	IS ↑	Prec. ↑	Recall ↑
2	11.84	2.45	0.67	0.36
4	11.38	2.53	0.65	0.37
6	11.41	2.49	0.66	0.37
8	11.77	2.50	0.67	0.37

G.3 Effect of Varying Embedding Dimensions in Phylo-Diffusion

We further evaluate the effect of varying the number of embedding dimensions used in HIER-Embed on the performance of Phylo-Diffusion. In this experiment, we trained Phylo-Diffusion by varying HIER-Embed's dimension in the following range of values: $\{16, 32, 64, 128, 256, 512, 1024\}$. Table 10 shows that Phylo-Diffusion is quite robust to the choice of embedding dimension with minimal drop in performance as we reduce the embedding dimension even to small values.

Table 10: Quantitative results for Phylo-Diffusion with varying embedding dimensions of hierarchical embeddings.

Embedding Dim.	FID ↓	IS ↑	Prec. ↑	Recall ↑
16	11.23	2.45	0.66	0.36
32	11.25	2.45	0.66	0.38
64	11.56	2.47	0.66	0.37
128	11.31	2.45	0.67	0.37
256	11.53	2.42	0.67	0.35
512	11.38	2.53	0.65	0.37
768	11.69	2.49	0.65	0.37
1024	11.51	2.48	0.67	0.36

H CUB Dataset Results

To show the applicability of our approach on other datasets with larger and deeper phylogenies, we perform additional experiments on 190 bird species from the CUB-200-2011 dataset (see Table 11). We selected the set of bird species based on whether we are able to obtain their phylogenetic knowledge from Bird Tree, which are pre-processed similar to the fishes. We removed the background of these images using segmentation masks to focus only on the body of the birds.

Table 11: Quantitative results on the new birds dataset (30 samples/class). The classifier has a base accuracy of 76% on the test set.

Method	FID ↓	IS ↑	Prec. ↑	Recall ↑	F1 ↑
Class Conditional	6.8	3.2	0.70	0.49	0.68
Scientefic Name	8.5	3.1	0.65	0.48	0.18
Phylo-Diffusion (ours)	6.7	3.1	0.72	0.49	0.64

H.0.1 Trait Masking:

Similar to fishes, Figure 23 shows the changes in probability distributions when Level 3 & 4 information is replaced with *noise*. The first two plots show the logits of images generated for *Black-footed albatross* and *Sooty albatross* using embeddings from all the four levels. The third plot shows the dispersion of logits across the three descendant species that are part of the sub-tree defined till level 2, i.e., masking level 3 & 4. We see similar results for CUB as well where the probability of classifying the generated images into any of the descendant species that share a subtree (highlighted in green) is generally greater than the species outside the subtree.

H.0.2 Trait Swapping:

Figure 24 shows an example of the trait swapping experiment on the birds dataset, similar to the experiments for fishes in the main paper. We see that the image generated from the perturbed embedding (center) picks up the trait of black coloration around the eye (purple circle) that is shared by the target sub-tree (right) while traits like pointed beak (green circle) are retained from the source species (left).

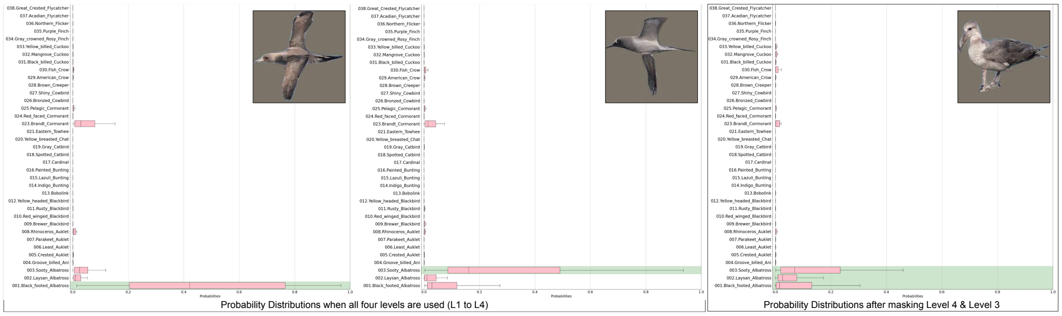
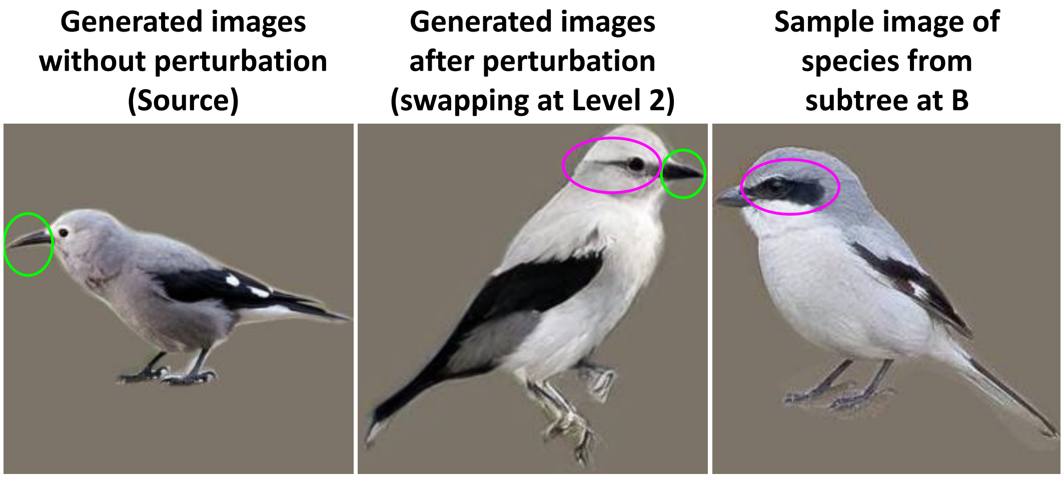


Figure 23: Left: class probability distributions of images generated by using embeddings at all four levels for two species *Black-footed albatross* and *Sooty albatross* (shown in green) that are part of the same sub-tree till level 2. Right: class probability distributions of images generated by masking level 3 and level 4 (descendant species that have common ancestry till level 2 are highlighted in green).



Generated images without perturbation (Source)

Generated images after perturbation (swapping at Level 2)

Sample image of species from subtree at B

Figure 24: Visualization of changes in traits after swapping information at Level 2 for *Clark nutcracker* (left) with its species from its sibling subtree (right) to generate perturbed species (center). Traits shared with the source species are outlined in green, whereas those shared with the sibling subtree at Node B are outlined in pink.