Discovering Novel Biological Traits From Images Using Phylogeny-Guided Neural Networks

Mohannad Elhamod[1], Mridul Khurana[1], Harish Babu Manogaran[1], Josef Uyeda[1], Meghan Balk[2], Wasila Dahdul[3], Yasin Bakis[4], Henry Bart[4], Paula Mabee[2], Hilmar Lapp[5], James Balhoff[6], Caleb Charpentier[1], David Carlyn[7], Wei-Lun Chao[7], Charles Stewart[8], Daniel Rubenstein[9], Tanya Berger-Wolf[7], and Anuj Karpatne[1]

elhamod@vt.edu,mridul@vt.edu,harishbabu@vt.edu,juyeda@vt.edu,balk@battelleecology.org,wdahdul@uci.edu, ybakis@tulane.edu,hbartjr@tulane.edu,mabee@battelleecology.org,Hilmar.Lapp@duke.edu,balhoff@renci.org, calebc22@vt.edu,carlyn.1@buckeyemail.osu.edu,chao.209@osu.edu,stewart@rpi.edu,dir@princeton.edu, berger-wolf.1@osu.edu,karpatne@vt.edu

[1]Virginia Tech, [2]Battelle, [3]University of California Irvine, [4]Tulane University, [5]Duke University, [6]University of North Carolina Chapel Hill, [7]Ohio State University, [8]Rensselaer Polytechnic Institute, [9]Princeton University

ABSTRACT

Discovering evolutionary traits that are heritable across species on the tree of life (also referred to as the phylogeny) is of great interest to biologists to understand how organisms vary and evolve. However, the measurement of traits is often a subjective and labor-intensive process, making *trait discovery* a highly label-scarce problem. We present a novel approach for discovering evolutionary traits directly from images without relying on trait labels. Our proposed model-agnostic approach, *Phylo-NN*, encodes the image of an organism into a sequence of quantized feature vectors –or codes– , where different segments of the sequence capture evolutionary signals at varying levels in the phylogeny tree. We demonstrate the effectiveness of our approach in producing biologically meaningful results in a number of downstream tasks including species image generation and species-to-species image translation, using fish species as a target example.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; **Computer vision**; **Image representations**; • **Applied computing** → **Imaging**.

KEYWORDS

computer vision, neural networks, phylogeny, morphology, knowledge-guided machine learning

ACM Reference Format:
Mohannad Elhamod[1], Mridul Khurana[1], Harish Babu Manogaran[1], Josef Uyeda[1], Meghan Balk[2], Wasila Dahdul[3], Yasin Bakis[4], Henry Bart[4], Paula Mabee[2], Hilmar Lapp[5], James Balhoff[6], Caleb Charpentier[1], David Carlyn[7], Wei-Lun Chao[7], Charles Stewart[8], Daniel Rubenstein[9], Tanya Berger-Wolf[7], and Anuj Karpatne[1]. 2023. Discovering Novel Biological Traits From Images Using Phylogeny-Guided Neural Networks. In *Proceedings of (KDD '23).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

One of the grand challenges in biology is to find features of organisms– or *traits*–that define groups of organisms, their genetic and developmental underpinnings, and their interactions with environmental selection pressures [16]. Traits can be physiological, morphological, and/or behavioral (e.g., beak color, stripe pattern, and fin curvature) and are integrated products of genes and the environment. The analysis of traits is critical for predicting the effects of environmental change or genetic manipulation, and to understand the process of evolution. For example, discovering traits that are heritable across individuals, or across species on the tree of life (also referred to as the *phylogeny*), can identify features useful for individual recognition or species classification, respectively, and is a starting point for linking traits to underlying genetic factors. Traits with such genetic or phylogenetic signal, termed evolutionary traits, are of great interest to biologists, as the history of genetic ancestry captured by such traits can guide our understanding of how organisms vary and evolve. This understanding enables tasks such as estimating the morphological features of ancestors, how they have responded to environmental changes, or even predicting the potential future course of trait changes [7, 28]. However, the measurement of traits is not straightforward and often relies on subjective and labor-intensive human expertise and definitions [40]. Hence, *trait discovery* has remained a highly label-scarce problem, hindering rapid scientific advancement [27].

With the recent availability of large image repositories containing millions of images of biological specimens [42, 45, 46], there is a great opportunity for the field of data mining and machine learning (ML) to contribute to the problem of trait discovery [27]. Advances in deep learning methods for mainstream ML problems have enabled us to extract useful information from images and to map them to structured feature spaces where they can be manipulated in a number of ways, e.g., for image generation and image-to-image translation. In this work, we ask the question: *can deep learning*

KDD '23,
,
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

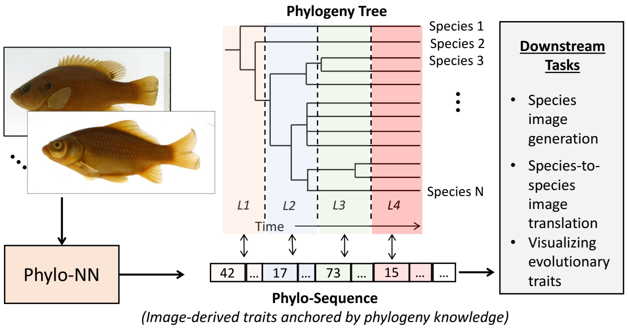

Phylogeny Tree
Species 1
Species 2
Species 3
⋮
Species N
L1
L2
L3
L4
Time
Phylo-NN
42 ... 17 ... 73 ... 15 ... ...
Phylo-Sequence
(Image-derived traits anchored by phylogeny knowledge)
Downstream Tasks
• Species image generation
• Species-to-species image translation
• Visualizing evolutionary traits

Figure 1: Our proposed *Phylo-NN* model converts images to discrete Phylo-sequences where different segments of the sequence (shown in distinct colors) capture evolutionary information at different levels of phylogenetic tree (L1 to L4).

models discover biological traits that capture evolutionary signals automatically from images in an unsupervised manner without relying on trait labels?

Despite the biological relevance of answering this question, discovering evolutionary traits directly from unlabeled images is challenging for two main reasons. First, since traits are determined both by the genes and the environment, not all image features extracted by a deep learning model for reconstructing images or differentiating between species will exhibit evolutionary signals. Hence, it is important to disentangle the image features of an organism that preserve evolutionary information from remaining features influenced by unrelated factors [7]. Second, information about evolutionary signals is not available as a set of known attributes (or trait labels) but rather in the form of structured knowledge of how species are related to each other on the phylogeny (see Figure 1). Without access to trait labels, current methods for feature disentanglement in deep learning [5, 26] are unfit for discovering evolutionary traits. Furthermore, current standards in deep learning for image synthesis [13, 38] or interpretable machine learning (ML) [4, 30] are unable to leverage structured forms of biological knowledge (e.g., phylogenetic trees and pedigrees) in the training and extraction of image features, making them unsuitable for analyzing and manipulating learned features in biologically meaningful ways.

We propose a novel approach for unsupervised trait discovery from images termed *phylogeny-guided neural networks* (*Phylo-NN*), which encodes the image of an organism into a sequence of quantized feature vectors or "codes", where different segments of the sequence capture evolutionary signals at varying levels in the phylogeny tree (see Figure 1). Analogous to gene sequences, our image-derived sequences (termed *Phylo-sequences*) enable us to discover evolutionary traits that one species shares with other species and with ancestor nodes within its lineage, represented by branching points in the phylogenetic tree. Further, by manipulating the space of Phylo-sequences, we can perform a number of downstream tasks such as species image generation, species-to-species image translation, and visualization of evolutionary traits. We demonstrate the effectiveness of *Phylo-NN* in solving these tasks in a biologically meaningful way compared to baseline methods, using fish species as a target example.

Our work, for the first time, provides a bridge between the "language of evolution" represented as phylogenetic trees and the "language of images" extracted from images as Phylo-sequences. This work is part of a larger-scale effort to establish a new field of research in "Imageomics" [31], where images are used as the source of information to accelerate biological understanding of phenotypic traits. Our work also provides a novel methodological advance in the emerging field of knowledge-guided machine learning (KGML) [18–20] by using structured biological knowledge in the learning of latent representations for image generation and translation.

2 BACKGROUND AND RELATED WORK

What is a Phylogeny? The phylogeny of a set of species is a tree that characterizes the evolutionary distances among these species and their common ancestors represented as nodes of the tree. In this tree, the length of every edge is a value that represents the evolutionary distance between two nodes (measured in time intervals representing thousands or millions of years), which is estimated from living species, and time-calibrated ages using dated fossil ancestors. While rates of change along different edges may vary substantially, on average we expect that longer edges will accumulate higher levels of genetic or phenotypic trait change than shorter edges, and species that are recently diverged will be more similar both in their genetics and phenotypic traits. In our work, we consider discretized versions of phylogeny trees with $n_l = 4$ levels, such that every species class (leaf node in the tree) has exactly $n_l - 1$ ancestors. More on our phylogeny preprocessing pipeline in Appendix B.

Deep Learning for Image Generation: There exists a large body of work in the field of deep learning for image generation, including methods based on Variational Autoencoders (VAEs) [24], Generative Adversarial Networks (GANs) [14, 21–23, 38], and Transformer networks [10, 15]. While conventional VAEs embed images in continuous feature spaces and sample from a Gaussian distribution, a recent variant of the VAE termed Vector-Quantized VAE (VQVAE) [32] uses discrete feature spaces quantized using a learned codebook of feature vectors (or codes) and employs a PixelCNN [43] model for sampling in the discrete feature space. The rationale for converting images to discrete representations is to allow for easier and more complex manipulations (e.g., in language modeling) than continuous features. This work was extended in [13] to produce VQGAN, which is different from VQVAE in two aspects. First, it adds a discriminator to its framework to improve the quality of the generated images. Second, it uses a Transformer model, namely the GPT architecture [36], to generate images from the quantized latent space instead of a PixelCNN. VQGAN is a state-of-the-art method that generates images of better quality efficiently at higher resolutions than other counterparts such as StyleGAN [21] and Vision Transformers [10, 15]. Our work derives inspiration from VQGAN to embed images in discrete feature spaces but with the grounding of structured biological knowledge available as phylogeny trees.

Interpretable ML: In order to move beyond black-box applications of AI and to make deep learning decisions easy to understand by humans, there is a growing trend to focus on the interpretability and explainability of deep learning features [11]. Some of the earliest works in this direction include the use of saliency scores [41] and

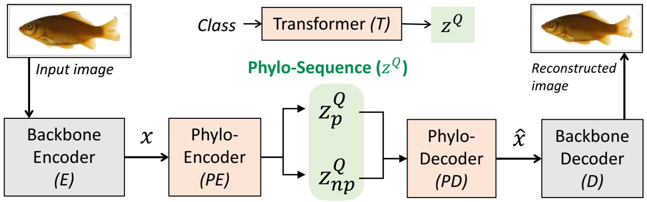


Figure 2: Overview of proposed *Phylo-NN* model architecture.

Class Activation Maps (CAMs) [39] that reveal sensitive regions of an image that affect classification decisions. However, these methods are known to be noisy and often imprecise [1]. Recent work includes the ProtoPNet model [4], which learns a set of template image patches (or prototypes) for each class during training, and then uses those templates to both predict and explain a test image. These methods suffer from two drawbacks. First, they do not allow for structured biological knowledge to guide the learning of interpretable features and hence are not designed to produce results that are *biologically meaningful*. Second, they are mostly developed for classification problems and cannot be directly applied to image generation and translation problems.

Disentangling ML Features: Another line of research that is relevant to our work is on disentangling the features of a deep learning model to align them with target "concepts." This includes the approach of "Concept whitening" [5], where the latent space of a classification model is whitened (i.e., normalized and decorrelated) such that the features along every axis of the latent space corresponds to a separate class. Another approach in this area is that of Latent Space Factorization[26], where the latent space of an autoencoder is linearly transformed using matrix subspace projections to partition it into features aligned with concept attributes and those that capture non-attribute information.

Knowledge-Guided ML: KGML is an emerging area of research that aims to integrate scientific knowledge in the design and learning of ML models to produce generalizable and scientifically valid solutions [20]. Some examples of previous research in KGML include modifying the architecture of deep learning models to capture known forms of symmetries and invariances [2, 48], and adding loss functions that constrain the model outputs to be scientifically consistent even on unlabeled data [8, 37]. In biology, KGML methods have been developed for species classification that leverage knowledge of taxonomic grouping of species [9, 12]. We build upon these methods to develop *Phylo-NN* that incorporates phylogeny knowledge in the learning of discrete feature spaces for a variety of downstream tasks including image generation and translation.

3 PROPOSED APPROACH: *PHYLO-NN*

Figure 2 provides an overview of our proposed *Phylo-NN* model. Our method can operate on the latent space of any backbone encoder model E that takes in images as input and produces continuous feature maps $\mathbf{x}$ as output. There are three computing blocks in *Phylo-NN* as shown in Figure 2. The first block, Phylo-Encoder (*PE*), takes continuous feature maps $\mathbf{x}$ as input and generates quantized feature sequences as output. These sequences comprise of two *disentangled* parts: z_p^Q, which captures evolutionary signals at varying levels in the phylogeny (p) tree, and z_np^Q, which captures

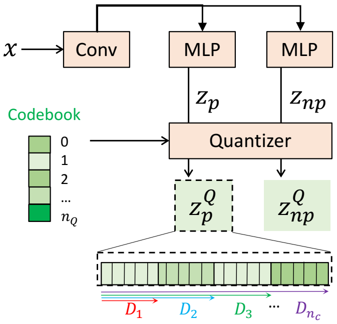


Figure 3: Detailed view of the Phylo-Encoder block in *Phylo-NN*.

non-phylogeny (np) information that is still important for image reconstruction but is not influenced by evolutionary signals. The second block, Phylo-Decoder (*PD*), maps the sequences back to the space of feature maps $\hat{\mathbf{x}}$, such that $\hat{\mathbf{x}}$ is a good reconstruction of $\mathbf{x}$. We then feed $\hat{\mathbf{x}}$ into a backbone decoder model D that reconstructs the original image. Note that in the training of *PE* and *PD* models, both the backbone models E and D are kept frozen, thus requiring low training time. *Phylo-NN* can then be plugged into the latent space of any powerful encoder-decoder framework to perform trait discovery. The third block of *Phylo-NN* is a transformer model T that takes in the species class variable as input, and generates a distribution of plausible Phylo-sequences corresponding to the class as output. These sequences can be fed to the *PD* model to generate a distribution of synthetic images. In the following, we provide details of each of the three blocks of *Phylo-NN*.

3.1 Phylo-Encoder Block

Figure 3 shows the sequence of operations that we perform inside the *PE* block. We first apply a convolutional layer on $\mathbf{x}$ to produce feature maps of size $(H \times W \times C)$. We split these C feature maps into two sets. The first C_p maps are fed into an MLP layer to learn a global set of feature vectors $\mathbf{z}_\mathrm{p}$ capturing phylogeny information. The size of $\mathbf{z}_\mathrm{p}$ is kept equal to $(n_\mathrm{l}\ n_\mathrm{p} \times d)$, where n_l is the number of phylogeny levels, n_p is the number of feature vectors we intend to learn at every phylogeny level, and d is the dimensionality of feature vectors. Similarly, the remaining $C - C_\mathrm{p}$ maps are fed into an MLP layer to produce a set of feature vectors $\mathbf{z}_\mathrm{np}$ capturing non-phylogeny information of size $(n_\mathrm{np} \times d)$

Vector Quantization: Both $\mathbf{z}_\mathrm{p}$ and $\mathbf{z}_\mathrm{np}$ are converted to *quantized* sequences of feature vectors, $\mathbf{z}_\mathrm{p}^Q$ and $\mathbf{z}_\mathrm{np}^Q$, respectively, using the approach developed in VQVAE [32]. The basic idea of this quantization approach is to learn a set (or codebook) of n_q distinct feature vectors (or codes), such that every feature vector in $\mathbf{z}_\mathrm{p}$ and $\mathbf{z}_\mathrm{np}$ is replaced by its nearest counterpart in the codebook. This is achieved by minimizing the *quantization loss*, $L_\mathrm{q} = |\mathbf{z} - \mathbf{z}^Q|$. The advantage of working with quantized vectors is that every feature vector in $\mathbf{z}_\mathrm{p}^Q$ and $\mathbf{z}_\mathrm{np}^Q$ can be referenced just by its location (or index) in the codebook. This allows for faster feature manipulations in the space of discrete code positions than continuous feature vectors.

Using phylogeny knowledge in $\mathbf{z}_\mathbf{p}^Q$: To ensure that the quantized feature sequence $\mathbf{z}_\mathrm{p}^Q$ contains phylogeny information, we design a novel *phylogeny loss* for training $\mathbf{z}_\mathrm{p}^Q$, described in the following.

Note that $\mathbf{z}_\mathrm{p}^Q$ contains n_l sub-sequences of length n_p, where every sub-sequence corresponds to a different level of the phylogeny tree. While the first sub-sequence S_1 should ideally capture information contained in $\mathbf{x}$ that is necessary for identifying ancestor nodes at level 1 of the phylogeny tree, S_2 should contain additional information that when combined with S_1 is sufficient to identify the correct ancestor node of $\mathbf{x}$ at level 2. In general, we define the concept of a Phylo-descriptor $D_i = \{S_1, S_2, \ldots, S_i\}$ of $\mathbf{x}$ that contains the necessary information for identifying nodes at level i (see Figure 3). We feed D_i to an MLP layer that predicts the class probabilities of nodes at level i, which are then matched with the correct node class of $\mathbf{x}$ at level i, $c_i(\mathbf{x})$, by minimizing the following phylogeny loss, L_p:

$$L_\mathrm{p} = \sum_{i=0}^{n_\mathrm{l}} \beta_i \mathrm{CE}(\mathrm{MLP}_i(D_i(\mathbf{x})), c_i(\mathbf{x})), \tag{1}$$

where CE is the cross-entropy loss and β_i is the weighting hyperparameter for level i.

Disentangling $\mathbf{z}_\mathbf{p}^Q$ and $\mathbf{z}_\mathbf{np}^Q$ While minimizing L_p guides the learning of $\mathbf{z}_\mathrm{p}^Q$ to contain phylogeny information, we still need a way to ensure that $\mathbf{z}_\mathrm{np}^Q$ focuses on complementary features and does not contain phylogeny information. To achieve this, we first apply an orthogonal convolution loss L_o (originally proposed in [47]) to the convolutional layer of Phylo-Encoder, to constrain the C convolutional kernels to be orthogonal to each other. To further ensure that $\mathbf{z}_\mathrm{np}^Q$ has no phylogeny information, we also employ an adversarial training procedure to incrementally remove phylogeny information from $\mathbf{z}_\mathrm{np}^Q$. In particular, we apply an MLP layer MLP_adv on $\mathbf{z}_\mathrm{np}^Q$, and then train the parameters of MLP_adv to minimize the following *adversarial loss*:

$$L_\mathrm{adv} = \sum_{i=0}^{n_\mathrm{l}} \beta_i \mathrm{CE}(\mathrm{MLP}_i(MLP_\mathrm{adv}(\mathbf{z}_\mathrm{np}^Q(\mathbf{x}))), c_i(\mathbf{x})), \tag{2}$$

This is aimed at training MLP_adv on the parameters of MLP_adv to detect any phylogeny information contained in $\mathbf{z}_\mathrm{np}^Q$. Simultaneously, we train to maximize L_adv on the rest of *Phylo-NN*'s parameters, making $\mathbf{z}_\mathrm{np}^Q$ irrelevant for the task of identifying nodes in the phylogeny tree.

4 PHYLO-DECODER BLOCK

The goal of the PD block is to convert $\mathbf{z}^Q = \{\mathbf{z}_\mathrm{p}^Q, \mathbf{z}_\mathrm{np}^Q\}$ back to the space of original feature maps, $\mathbf{x}$. The sequence of operations in *PD* are almost a mirror image of those used in *PE*. We first pass $\mathbf{z}_\mathrm{p}^Q$ and $\mathbf{z}_\mathrm{np}^Q$ through two MLPs, and then concatenate the outputs of these MLPs to create feature maps of size $(H \times W \times C)$. These feature maps are then fed into a convolutional layer to produce $\hat{\mathbf{x}}$. Minimizing the reconstruction loss, $L_\mathrm{rec} = |\hat{\mathbf{x}} - \mathbf{x}|$, ensures that $\hat{\mathbf{x}}$ is a good approximation of $\mathbf{x}$. Finally, *PE* and *PD* are jointly trained using a weighted summation of all the losses mentioned above

4.1 Transformer Block

Once *PE* and *PD* have been trained, we can extract Phylo-sequences $\mathbf{z}^Q$ for every image in the training set, and convert any Phylo-sequence to an image. The goal of the Transformer block is to learn the patterns of codes in the extracted Phylo-sequences of different classes (e.g., species class or ancestor node class), and use these patterns to generate synthetic Phylo-Sequences for every class. To achieve this task, we follow the approach used by VQGAN [13] and train a GPT transformer model [36] T_i to generate plausible sequences of $\mathbf{z}^Q$ for every node class at level i. The generated Phylo-sequences can then be converted into synthesized specimen images using *PD* and *D*.

5 EVALUATION SETUP

5.1 Data

In this work, we used a curated dataset of Teleost fish images from five ichthyological research collections that participated in the Great Lakes Invasives Network Project (GLIN). After obtaining the raw images from these collections, we handpicked a subset of about 11,000 images and pre-processed them by resizing and appropriately padding each image to be of a 256 × 256 pixel resolution. Finally, we split the subset into a training set and a validation set of ratios 80% and 20%, respectively. Details of the data pre-processing are provided in Appendix A.

Our dataset includes images from 38 species of Teleost fishes with an average number of 40 images per species. We further discretized the phylogeny tree to have $n_\mathrm{l} = 4$ ancestor levels, where the last level is the species class. Details about the phylogeny tree selection and discretization are provided in the Appendix B.

5.2 Backbone Encoder and Decoder

Since *Phylo-NN* can operate on the feature space $\mathbf{x}$ of any backbone encoder E and produce reconstructed feature maps $\hat{\mathbf{x}}$ that can be decoded back to images by a corresponding backbone decoder D, we tried different encoder-decoder choices including pix2pix [17], ALAE [33], and StyleGAN [23]. However, we found VQGAN [13] feature maps to produce images of better visual quality than other encoder-decoder models. Hence, we used the embeddings of a base VQGAN encoder E as inputs in *Phylo-NN* for all our experiments. The reconstructed feature maps of *Phylo-NN* were then fed it into a base VQGAN quantizer serving as the backbone decoder D. Note that while training *Phylo-NN*, we kept the parameters of the backbone models fixed, thus saving training time and resources while benefiting from the modeling capabilities of the backbone models.

5.3 Baseline Methods

We considered the following baseline methods to compare the results for image generation and translation with *Phylo-NN*:

Vanilla VQGAN [13]: The first baseline that we consider is a vanilla VQGAN model trained on the fish dataset. By comparing the learned embeddings and generated images of *Phylo-NN* with vanilla VQGAN, we aim to demonstrate the importance of using biological knowledge to guide the extraction of evolutionary traits from images, rather than solely relying on the information contained in image data.

Concept whitening (CW) [5]: For this second baseline, we replaced the last normalization layer in the encoder block of vanilla VQGAN with the concept whitening (CW) module, where we used species class labels as concept definitions. This is intended to evaluate if CW is capable of disentangling the evolutionary traits of species automatically from images without using the structured knowledge of how species are related to one another in phylogenetic trees. The whitened embeddings z_{cw} produced by the CW module are fed into the quantizer module of vanilla VQGAN for converting the embeddings to images. While training the CW module, we optimized the whitening and rotation matrices for all concepts at every 30 batches. Similar to what we did in Section 4.1, we leveraged the VQGAN's transformer and conditioned on the species label to generate plausible sequences of z_{cw} that are then decoded into specimen images.

Latent Space Factorization (LSF) [26]: The third baseline that we considered is the LSF method, which is another method for feature disentanglement given concept attribute labels. Specifically, we introduced a variational autoencoder (VAE) model between the encoder and the quantization layer of the base VQGAN model. Similar to CW, for factorizing the latent space using LSF, we chose the species class of each image itself as the concept attribute. The LSF module was trained to optimize VAE's KL-divergence loss and recreation loss along with the attribute and non-attribute losses, as originally defined in the LSF method [26].

6 RESULTS

In the following, we analyze the results of *Phylo-NN* from multiple angles to assess the quality of its learned embeddings and generated images in comparison with baseline methods. The code and datasets for running all the analyses reported in this paper can be found at https://github.com/elhamod/phylonn.

6.1 Validating Species Distances in the Embedding Space

In order to evaluate the ability of *Phylo-NN* to extract evolutionary traits from images in an *unsupervised* manner (i.e., without using trait labels), we show that distances between species pairs in the embedding space of *Phylo-NN* are biologically meaningful and are correlated with ground-truth values better than baseline methods. In the following, we describe the two types of ground-truths used, the approach used for computing distances in the embedding space of comparative methods, and the comparison of correlations with ground-truth values.

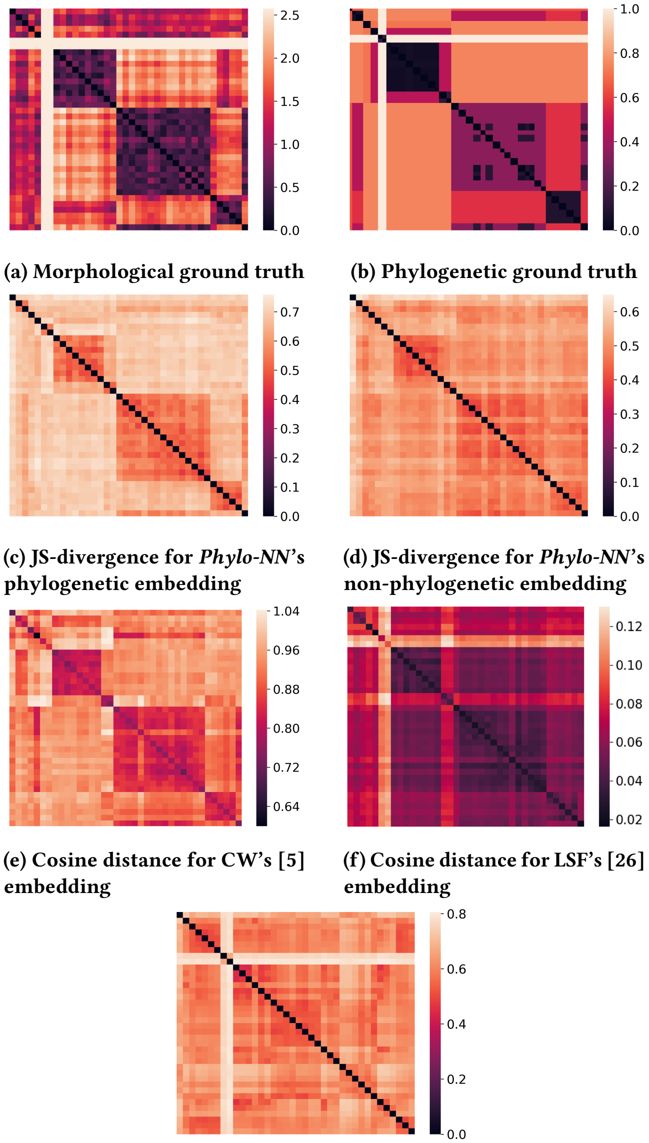
Phylogenetic Ground-truth (GT): The first type of ground-truth that we consider for evaluating the embedding distance between a pair of species is the *evolutionary distance* between their corresponding nodes in the phylogenetic tree. In particular, for any two species, we can calculate the sum total of edge lengths that is needed to traverse the path between their nodes in the phylogeny. The longer the path, the more distant the species are on the evolutionary scale. Hence, if *Phylo-NN* indeed captures evolutionary traits in its embedding space, we would expect it to show higher correlations with evolutionary distances computed from the phylogenetic tree as compared to baselines.

Morphological Ground-truth (GT): Another approach that we use to measure the similarity in the traits of two species is by using ground-truth measurements of linear morphological traits obtained from the FishShapes v1.0 dataset [34]. This dataset contains expert-measured traits that are known to carry evolutionary signals, defined and collected using traditional methods that are subjective and labor-intensive. We specifically used 8 functionally relevant traits from the FishShapes dataset for every fish species. Some species were not available in the FishShapes dataset, so when possible, either the closest relative was substituted or the species was dropped (see supplementary materials for details). The species were then matched to a time-calibrated phylogeny of fishes [3, 35] and the log-transformed measurements were rotated with phylogenetically-aligned components analysis (PACA, [7], which rotates the traits to the axis with the highest level of phylogenetic signal. After correcting for overall size and allometry, the principal components of PACA were used to compute the Mahalonobis distance between every pair of species, using a covariance matrix proportional to the evolutionary rate matrix. See supplementary materials for more details of the complete pipeline of calculations.

Computing Embedding Distances: To compute pair-wise distances in the embedding space of *Phylo-NN*, we first compute the probability distributions of quantized codes at every position of the Phylo-sequence (i.e., $\mathbf{z}_{\text{p}}^{Q}$ and $\mathbf{z}_{\text{np}}^{Q}$) in the test images for every species. We then compute the Jensen-Shannon (JS) divergence [29] between the probability distributions of codes at a pair of a species to measure the similarity of learned *Phylo-NN* embeddings at the two species. We adopt a similar approach for computing the JS-divergence of species-pairs in the quantized feature space of vanilla VQGAN. For baseline methods that operate in continuous feature spaces (CW and LSF), we first calculate the mean feature vector for every species and then compute the cosine similarity in the mean vectors of a pair of species.

Comparing Correlations with Ground Truth: Figure 4(a) and Figure 4(b) show the pair-wise species distance matrices for morphological and phylogenetic GTs, respectively. We can see that both ground-truths show a similar clustering structure of species, indicating groups of species that share evolutionary traits. However, there are differences too; while phylogenetic GT is solely based on phylogeny, the morphological GT uses both the phylogeny and information about phenotypic traits. Figure 4(c) and Figure 4(d) show the JS-divergences among species computed separately for the two disentangled parts of PhyloNN's embeddings ($\mathbf{z}_{\text{p}}^{Q}$ and $\mathbf{z}_{\text{np}}^{Q}$). We can see that the embeddings containing phylogenetic information show a similar clustering structure of distances as the GT matrices, in contrast to the non-phylogenetic embeddings. This shows the ability of *Phylo-NN* to disentangle features related to phylogeny from other unrelated features. Figure 4 also shows the embedding distance matrices of the baseline methods, which are not as visually clean as *Phylo-NN* in terms of matching with the GT matrices.

To quantitatively evaluate the ability of *Phylo-NN* to match with GT distances compared to baselines, we compute the Spearman correlation between the GT distance matrices and embedding distance matrices for different methods as shown in Table 1. We can see that *Phylo-NN* shows higher correlations at the species level with both morphological and phylogenetic GTs as compared to other baselines. Furthermore, since *Phylo-NN* learns a different Phylo-descriptor for every ancestor level in contrast to baseline methods that learn a flat representation, we can also compute *Phylo-NN*'s distance matrix at any ancestor level and compare it with GT matrices at the same ancestor level. We can see that *Phylo-NN*



(a) Morphological ground truth

(b) Phylogenetic ground truth

(c) JS-divergence for *Phylo-NN*'s phylogenetic embedding

(d) JS-divergence for *Phylo-NN*'s non-phylogenetic embedding

(e) Cosine distance for CW's [5] embedding

(f) Cosine distance for LSF's [26] embedding

(g) JS-divergence for Vanilla VQ-GAN's [13] embedding

Figure 4: Comparing embedding distance matrices of methods with morphological and phylogenetic ground-truths

shows significantly higher correlations with GT matrices at higher ancestor levels than the species level.

Table 1: Correlations between GT distances and embedding distances

		Morphological	Phylogenetic
PhyloNN	level0	0.86	0.83
	level1	0.87	0.85
	level2	0.78	0.83
	species	0.70	0.78
LSF		0.33	0.53
CW		0.71	0.66
vanilla VQGAN		0.31	0.24

6.2 Evaluating Species-to-species Image Translations

To further assess how well *Phylo-NN*'s learned embedding captures phylogenetic traits, we investigate how altering the quantized embedding of an image specimen incrementally in a phylogenetically meaningful ordering affects the morphological traits when the altered embeddings are decoded back as an image. To do that, we set up the following experiment. We pick two specimen images from a pair of species. By encoding the two images using *Phylo-NN*, we obtain their corresponding encodings, $\mathbf{z}_1^Q$ and $\mathbf{z}_2^Q$. We then start to replace the codes in the Phylo-sequence $\mathbf{z}_1^Q$ with the corresponding codes in $\mathbf{z}_2^Q$ iteratively, until $\mathbf{z}_1^Q$ transforms completely into $\mathbf{z}_2^Q$. The order of this iterative replacement is by first replacing the codes representing the non-phylogenetic part of the embedding $\mathbf{z}^Q_{\text{np}_1}$, then the part representing the information in the earliest ancestral level (level 0), to the next ancestor level (level 1), till we eventually reach the last level of the phylogenetic tree, which is the species level. At the final point, the entire Phylo-sequence $\mathbf{z}_1^Q$ has been replaced with $\mathbf{z}_2^Q$. This phylogeny-driven ordering of code replacements helps us capture key "snapshots" of the species-to-species translation process that are biologically meaningful. In particular, by observing the traits that appear or disappear at every ancestor level of code replacement, we can infer and generate novel hypotheses about the biological timing of trait changes as they may have happened in the process of evolution.

Figure 5 shows an example of such a translation process between a specimen of the species Carassius Auratus to a specimen of the species Lepomis Cyanellus. We can see that although the two specimens look similar on the surface, there are several subtle traits that are different in the two species that are biologically interesting. For example, the source species has a V-shaped tail fin (termed caudal fin), while the target species has a flatter tail fin. By looking at their place of occurrence in the translation process of *Phylo-NN*, we can generate novel biological hypotheses of whether they are driven by phylogeny or not, and whether they appeared earlier or later in the target species in the course of evolution. For example, we can see that the flat tail feature of the target species appears right after replacing the non-phylogenetic part of the embedding, indicating that this feature may not be capturing evolutionary signals and instead maybe be affected by the environment. On the other hand, if we observe another fin that appears in the middle of the body of the target species (termed pectoral fin), we can see that it seems to get more prominent only in the later levels (it is absent in level 0). This suggests the hypothesis that the presence of pectoral fin in the target species may have been added later in the course of evolution. Our work opens novel opportunities for generating such biological hypotheses, which can be further investigated by biologists to potentially accelerate scientific discoveries. Figure 5 also shows the translations obtained by baseline methods for the same pair of species specimens. We can see that the baselines are mostly performing a smooth interpolation between the source and target images. In particular, the transition points in the translation process of baseline methods do not correspond to biologically meaningful events, since they only rely on the information contained in data and do not use biological knowledge.

6.3 Generalization to Unseen Species

As the objective of *Phylo-NN* is to encode specimen images into their corresponding phylogenetic and non-phylogenetic quantized sequences of codes, we are interested in studying the distribution of codes used within different descriptors of phylogenetic concepts (e.g., species nodes or ancestor nodes). Naturally, we expect specimens that belong to the same species to largely share the same phylogenetic code in terms of the species descriptor D_{n_l}, while varying in terms of the non-phylogenetic codes. More generally, specimens belonging to species that share a common ancestor at a phylogenetic level i should largely share the codes with D_i while varying in terms of the rest of the phylogenetic and non-phylogenetic codes. This should also apply for specimens of *unseen* (or newly discovered) species that we have not yet observed in the training set. We posit that by looking at the similarity of codes generated for an unseen species during testing, we should be able to infer its ancestor lineage in terms of the species sampled during training.

To quantify this phenomenon, for a given phylogenetic concept of interest, we construct two sets of histograms, H_p and H_np of sizes $[n_\mathrm{l} \times n_\mathrm{p}]$ and $[n_\mathrm{np}]$, respectively. Each value in the histograms, ${H_\mathrm{p}}^{i,j}$ and ${H_\mathrm{np}}^{k}$, describes the distribution of codes of that corresponding location in the sequence across all the specimen in the dataset of interest. Take a look at Appendix F for some examples.

Once these histograms are constructed, we can qualitatively inspect them and extract the unique or most common codes for a certain sequence location. However, a quantitative way to evaluate the purity of *Phylo-NN* codes is by calculating the entropy of each histogram in H_p and H_np. If the entropy is low for a certain sequence location, it means only a few possible codes occur at that location, alluding to the fact that those specific codes at that location are key at characterizing the phylogenetic concept in question. On the other hand, higher entropy means a variety of codes occur at that location, implying that such a location is not discriminative to the phylogenetic concept of interest.

Finally, to compare the code distributions for two species, we use the JS-divergence metric for calculating the difference between two histograms of a sequence location. Similar to what is done in Section 6.1, such a metric can be aggregated to quantify the coding differences between the species in question.

To assess the *Phylo-NN*'s ability to generalize to unseen species, we train it on a subset of the species in our dataset and then evaluate the quality of the embedding space when the model is introduced to species it has never seen before during training. In our experiment, we chose to train on the same dataset as before while only excluding three species. Once the model is trained, we look at the average JS-divergence distance between each of the three missing species and three other species in the tree. These three other species were selected such that each missing species has one seen species that is close to it phylogenetically (i.e., both species share the same ancestor at the immediate ancestor level) while the others are relatively far from it.

Table 2: JS-diveregence of the phylogenetic codes at the species level between unseen and seen species

		Seen species		
		Notropis nubilus	Lepomis macrochirus	Noturus flavus
Unseen species	Notropis percobromus	0.47	0.71	0.62
	Lepomis megalotis	0.73	0.43	0.72
	Noturus miurus	0.62	0.71	0.48

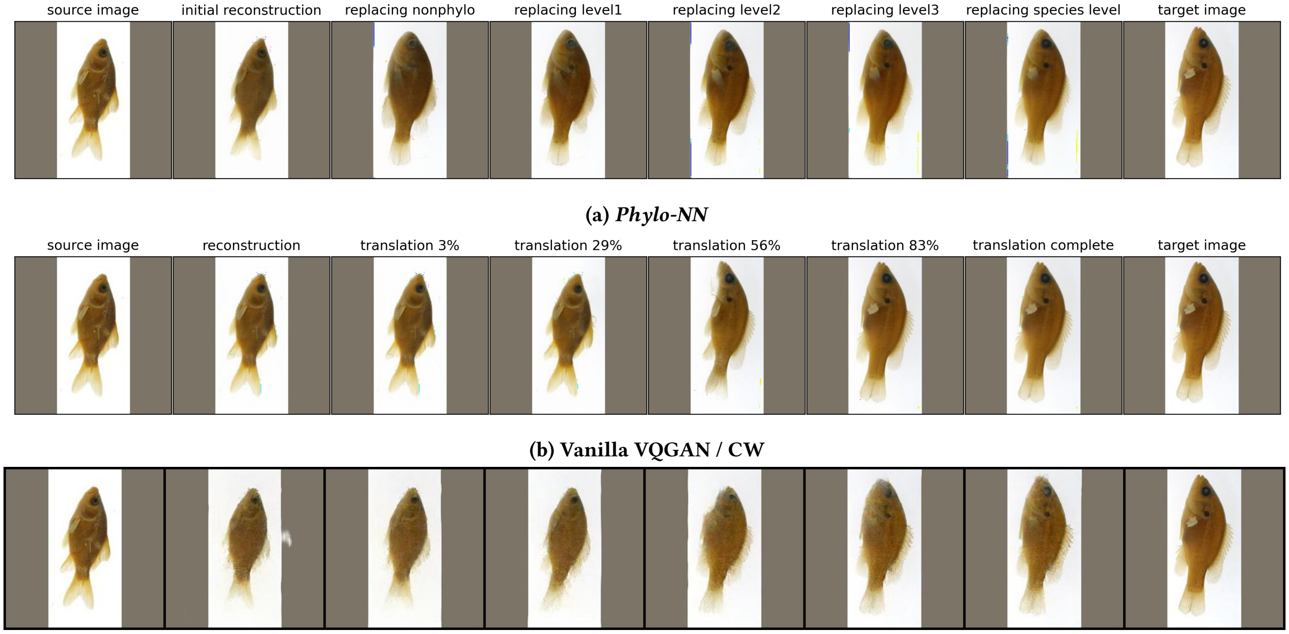
Table 2 shows the average distance of the phylogenetic codes among the six aforementioned species. Looking at the table, one can see that this distance is smallest for each unseen species and its counterpart that shares the same immediate ancestor (shown as the diagonal in the table). This confirms that even though the model has not seen the former species, it was able to characterize it using a coding sequence that is significantly closer to that of its seen counterpart than the other species'.

While Table 2 highlights the phylogenetic matching in the embedding space at the species descriptor level, D_{n_l}, Table 3 does the same but for the descriptor at a distant ancestor level (level 0), i.e., D_0. Based on the phylogeny tree we have used in this example, both the *Notropis* and *Noturus* species share the same distant ancestor at that descriptor level. On the other hand, *Lepomis* species does not share that ancestor. Hence, we find that the JS-divergences increase for the *Lepomis* unseen species with seen species that are not *Lepomis* as compared to Table 2. On the other hand, the JS-divergences decrease for the other two unseen species w.r.t. seen species that are on the off-diagonals of the table. This confirms that D_0 specifically captures the phylogenetic information of that distant ancestor that is common across *Notropis* and *Noturus* seen and unseen species.

Finally, to confirm that this phylogenetic correlation is mainly constrained only to the phylo-descriptors, we calculate the same distances but using the non-phylogenetic part of the sequences. The result is shown in Table 4. Here, we can see the distances are much closer to each other, implying that the non-phylogenetic embedding is not specialized at differentiating among the different species, and hence cannot be used to phylogenetically categorize the unseen species.

6.4 Assessing the Clustering Quality of the Embedding Space Using t-SNE Plots

In this section, we qualitatively assess the quality of generated images by visualizing their embedding space. Visualization tools such as loss landscape visualizations [25] and t-SNE plots [44], have been frequently used as investigative tools in deep learning in recent years as they help gauge a model's generalization power. To that end, we are interested in understanding how *Phylo-NN* clusters the



(a) ***Phylo-NN***

(b) **Vanilla VQGAN / CW**

(c) **LSF**

Figure 5: Comparing species-to-species image translations from a Carassius Auratus specimen to a Lepomis Cyanellus specimen

Table 3: JS-diveregence of the phylogenetic codes at the earliest ancestral level between unseen and seen species

		Seen species		
		Notropis nubilus	Lepomis macrochirus	Noturus flavus
Unseen species	Notropis percobromus	0.26	0.81	0.50
	Lepomis megalotis	0.81	0.27	0.81
	Noturus miurus	0.52	0.80	0.31

Table 4: JS-diveregence of the non-phylogenetic codes between unseen and seen species

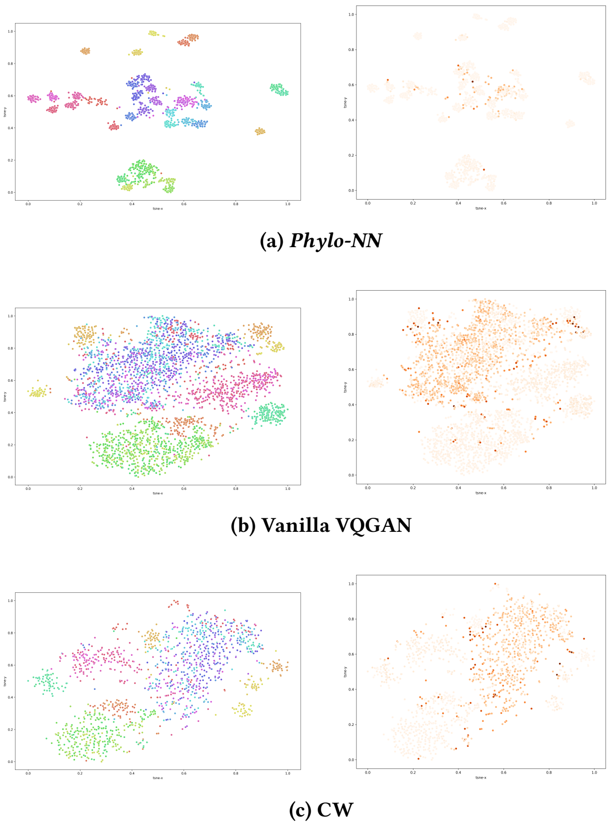
		Seen species		
		Notropis nubilus	Lepomis macrochirus	Noturus flavus
Unseen species	Notropis percobromus	0.39	0.45	0.39
	Lepomis megalotis	0.46	0.36	0.48
	Noturus miurus	0.40	0.42	0.36

embedding space compared to other baselines by analyzing these models' t-SNE plots.

To construct the t-SNE plot for each model, we iterate through its generated images, encode them, obtain the quantized embedding vector for each image ($\mathbf{z}_{\mathrm{p}}^{Q}$ and $\mathbf{z}^{Q}$ for *Phylo-NN* and vanilla VQGAN, respectively), and finally create the t-SNE plots. For CW, we use the whitened embeddings z_{cw} instead.

Figure 6 shows these constructed t-SNE plots with two different color-coding schemes. The first one (left column) color-codes the data-points based on the grouping of species at the second phylogenetic level (i.e., the direct ancestor of the specimen's species). Using this color-coding scheme allows us to inspect how different species cluster in the embedding space. The second color-coding (right column) is the average phylogenetic distance between the data-point and its k-nearest neighbors (KNN), where $k = 5$ in this setup. The higher the average distance (i.e., the darker the data-point's color), the more distant the specimen is from those k specimen's that are closest to it in the quantized embedding space. This color-coding helps us spot how well the different species are separated from each other in the embedding space, which generally characterizes the quality of the encoding and its propensity for downstream tasks, such as classification.

Looking at Figure 6, we can see that *Phylo-NN* (top row) clearly clusters the generated images better than vanilla VQGAN and CW as evident from its hierarchical clustering where the specimens belonging to the same species clump into small clusters and these clusters in turn clump into larger clusters (representing ancestor nodes) that have a singular color. This demonstrates that *Phylo-NN* is able to learn a phylogenetically-meaningful encoding, whereas the other base models' clustering is quite fuzzy and poorly characterizes any biological knowledge. Also, by looking at the right column, we can see that *Phylo-NN* commits very little clustering error in terms of its phylogenetic constraints because the average phylogenetic distance is low (almost zero) for the majority of the data points. This is in contrast to the other baselines where there is quite a high clustering error as seen from the "heat" of its scatter plot.



(a) *Phylo-NN*

(b) Vanilla VQGAN

(c) CW

Figure 6: t-SNE plots of the images generated by *Phylo-NN* and other baselines.

7 CONCLUSIONS AND FUTURE WORK

In this work, we presented a novel approach of *Phylo-NN* for discovering biological traits related to evolution automatically from images in an unsupervised manner without requiring any trait labels. The key novelty of our approach is to leverage the biological knowledge of phylogenetic trees to structure the quantized embedding space of *Phylo-NN*, where different parts of the embedding capture phylogenetic information at different ancestor levels of the phylogeny. This enables our method to perform a variety of tasks in a biologically meaningful way such as species-to-species image translation and identifying the ancestral lineage of newly discovered unseen species.

In the future, our work can be extended to include a larger number of embedding dimensions to improve the visual quality of generated images and can be applied to other image datasets beyond the fish dataset. Future work can explore extensions of *Phylo-NN* to generate images of ancestor species or to predict images of species that are yet to be evolved. Future work can also focus on making the discovered Phylo-sequence embeddings explainable by understanding the correspondence of each quantized code with a region in the image space. Our work opens a novel area of research in grounding image representations using biological knowledge available in the form of phylogenetic trees, which can lead to new research paradigms in other fields of science where images are abundant but labels are scarce.

ACKNOWLEDGMENTS

This work was supported, in part, by NSF awards for the HDR Imageomics Institute (Award # 2118240) and the Biology-guided Neural Network (BGNN) projectss (Award # 1940247, # 1940322, # 1940233, # 2022042, # 1939505).

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[2] Brandon Anderson, Truong Son Hy, and Risi Kondor. 2019. Cormorant: Covariant Molecular Neural Networks. *Advances in Neural Information Processing Systems* 32 (2019), 14537–14546.

[3] Jonathan Chang, Daniel L Rabosky, Stephen A Smith, and Michael E Alfaro. 2019. An R package and online resource for macroevolutionary studies using the ray-finned fish tree of life. *Methods in Ecology and Evolution* 10, 7 (2019), 1118–1124.

[4] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf

[5] Zhi Chen, Yijie Bei, and Cynthia Rudin. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence* 2, 12 (2020), 772–782.

[6] Julien Clavel, Gilles Escarguel, and Gildas Merceron. 2015. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution* 6, 11 (2015), 1311–1319.

[7] Michael L. Collyer and Dean C. Adams. 2021. Phylogenetically aligned component analysis. *Methods in Ecology and Evolution* 12, 2 (2021), 359–372. https://doi.org/10.1111/2041-210X.13515 arXiv:https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13515

[8] Arka Daw, Anuj Karpatne, William D Watkins, Jordan S Read, and Vipin Kumar. 2017. Physics-guided neural networks (pgnn): An application in lake temperature modeling. In *Knowledge-Guided Machine Learning*. Chapman and Hall/CRC, 353–372.

[9] Anderson Aparecido dos Santos and Wesley Nunes Gonçalves. 2019. Improving Pantanal fish species recognition through taxonomic ranks in convolutional neural networks. *Ecological Informatics* 53 (2019), 100977.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[12] Mohannad Elhamod, Kelly M. Diamond, A. Murat Maga, Yasin Bakis, Henry L. Bart Jr., Paula Mabee, Wasila Dahdul, Jeremy Leipzig, Jane Greenberg, Brian Avants, and Anuj Karpatne. 2022. Hierarchy-guided neural network for species classification. *Methods in Ecology and Evolution* 13, 3 (2022), 642–652. https://doi.org/10.1111/2041-210X.13768 arXiv:https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13768

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

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16000–16009.

[16] David Houle and Daniela M Rossoni. 2022. Complexity, Evolvability, and the Process of Adaptation. *Annual Review of Ecology, Evolution, and Systematics* 53 (2022).

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

[18] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. *Nature Reviews Physics* 3, 6 (2021), 422–440.

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

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

[21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021), 852–863.

[22] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[24] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[25] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the Loss Landscape of Neural Nets. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf

[26] Xiao Li, Chenghua Lin, Ruizhe Li, Chaozheng Wang, and Frank Guerin. 2020. Latent space factorisation and manipulation via matrix subspace projection. In *International Conference on Machine Learning*. PMLR, 5916–5926.

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

[28] Michael Lynch. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45, 5 (1991), 1065–1080.

[29] ML Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute* 334, 2 (1997), 307–318.

[30] Meike Nauta, Ron van Bree, and Christin Seifert. 2021. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14933–14943.

[31] NSF HDR Imageomics Institute. 2021. Imageomics: A new frontier of biological information powered by knowledge-guided machine learning. https://imageomics.osu.edu/.

[32] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. https://doi.org/10.48550/ARXIV.1711.00937

[33] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14104–14113.

[34] Samantha A Price, Sarah T Friedman, Katherine A Corn, Olivier Larouche, Kasey Brockelsby, Anna J Lee, Maya Nagaraj, Nick G Bertrand, Mailee Danao, Megan C Coyne, et al. 2022. FishShapes v1: Functionally relevant measurements of teleost shape and size on three dimensions.

[35] Daniel L Rabosky, Jonathan Chang, Peter F Cowman, Lauren Sallan, Matt Friedman, Kristin Kaschner, Cristina Garilao, Thomas J Near, Marta Coll, Michael E Alfaro, et al. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* 559, 7714 (2018), 392–395.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[37] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* 378 (2019), 686–707.

[38] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2287–2296.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

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

[41] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[42] Randal A Singer, Kevin J Love, and Lawrence M Page. 2018. A survey of digitized data from US fish collections in the iDigBio data aggregator. *PloS one* 13, 12 (2018), e0207636.

[43] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems* 29 (2016).

[44] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[45] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8769–8778.

[46] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[47] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X. Yu. 2020. Orthogonal Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Rui Wang, Robin Walters, and Rose Yu. 2020. Incorporating symmetry into deep dynamics models for improved generalization. *arXiv preprint arXiv:2002.03061* (2020).

A DATASET

As mentioned in Section 5.1, the images we use in this work come from a 38-species semi-balanced subset of a larger collection that participated in the Great Lakes Invasives Network Project (GLIN). After further splitting this subset into training and test sets, we apply some pre-processing that is necessary for the neural network to yield best results. This pre-processing includes padding the images with the ImageNet mean RGB color. We also use data augmentation when training the base VQGAN model. This includes random horizontal flips, spatial shifts and rotations, and brightness and contrast changes.

B PHYLOGENY PREPROCESSING

As mentioned in Section 2, we use a phylogeny tree to characterize the evolutionary distances between the species in our dataset. The phylogeny tree corresponding to the dataset described in Section 5.1 was obtained using *opentree* (https://opentree.readthedocs.io/en/latest/) python package. Phylogeny processing and manipulation were done using *ete3* (http://etetoolkit.org/) python package.

In our application, we quantize our 38-species tree into $n_{\mathrm{l}} = 4$ distinct phylogenetic levels. Each level groups the 38 species based on their common ancestry within that level. Tables 5 and 6 outline each level with its corresponding species groupings.

C HYPER-PARAMETER SELECTION

In terms of hyper-parameter tuning, we used the following settings for each of the trained models:

Vanilla VQGAN: We trained a VQGAN with a codebook of 1024 possible codes and an embedding sequence of 256 codes. We trained the model for 836 epochs with a learning rate of 4.5×10^{-6}. We used this VQGAN as the base model for the rest of the models. A batch size of 32 was used.

Phylo-NN Taking the base VQGAN model described above, we trained a *Phylo-NN* that has $\mathbf{z}_{\mathrm{p}}^{Q}$ of dimensions ($n_{\mathrm{l}} = 4, n_{\mathrm{p}} = 8$). $\mathbf{z}_{\mathrm{np}}^{Q}$ also has the same dimensionality. The dimensionality of the embedding itself is $d = 16$, and the size of the codebook $n_{\mathrm{q}} = 64$. A batch size of 32 was used.

Concept Whitening (CW) Taking the base VQGAN model described above, we trained CW for 10 epochs using the same hyper-parameters as Vanilla VQGAN. We used a batch size of 20 for the concepts.

Latent Space Factorization (LSF): An embedding dimension of size 1024 was used.

D DETAILS OF MORPHOLOGICAL DISTANCE PROCESSING

The 8 functionally relevant traits that we used from the FishShapes dataset include: standard length, maximum body depth, maximum fish width, lower jaw length, mouth width, head depth, minimum caudal peduncle depth, and minimum caudal peduncle. Some species were not available in the FishShapes dataset, so when possible, the closest relative was substituted. (*Notropis percobromus* was replaced with *Notropis rubellus*, and *Carassius auratus* was replaced with *Carassius carassius*). Also, two species of *Lepisosteus* had no close relatives and were thus removed fro the Spearman correlation analysis. To correct for overall size and allometry, each measurement was log transformed and regressed against Standard Length (SL) using a phylogenetic regression in the R package *phylolm*, with the residuals from the regression being the inputs into PACA. Distances in the principal components of PACA were measured as the Mahalonobis distance between the multivariate means using a covariance matrix proportional to the evolutionary rate matrix from the multivariate Brownian Motion fit in the R package mvMORPH [6].

Table 5: Phylogenetic groupings of the species included in this study at different ancestral levels

Level	Species groupings
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
	Esox americanus, Gambusia affinis, Lepomis auritus, Lepomis cyanellus, Lepomis gibbosus, Lepomis gulosus, Lepomis humilis, Lepomis macrochirus, Lepomis megalotis, Lepomis microlophus, Morone chrysops, Morone mississippiensis
	Lepisosteus osseus, Lepisosteus platostomus
1	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, Phenacobius mirabilis
	Esox americanus
	Gambusia affinis, Lepomis auritus, Lepomis cyanellus, Lepomis gibbosus, Lepomis gulosus, Lepomis humilis, Lepomis macrochirus, Lepomis megalotis, Lepomis microlophus, Morone chrysops, Morone mississippiensi
	Lepisosteus osseus, Lepisosteus platostomus
	Noturus exilis, Noturus flavus, Noturus gyrinus, Noturus miurus, Noturus nocturnus

Table 6: Phylogenetic groupings of the species included in this study at different ancestral levels (continued)

Level	Species groupings
2	Alosa chrysochloris
	Carassius auratus, Cyprinus carpi
	Esox americanus
	Gambusia affinis
	Lepisosteus osseus, Lepisosteus platostomus
	Lepomis auritus, Lepomis cyanellus, Lepomis gibbosus, Lepomis gulosus, Lepomis humilis, Lepomis macrochirus, Lepomis megalotis, Lepomis microlophus
	Morone chrysops, Morone mississippiensis
	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
	Noturus exilis, Noturus flavus, Noturus gyrinus, Noturus miurus, Noturus nocturnus

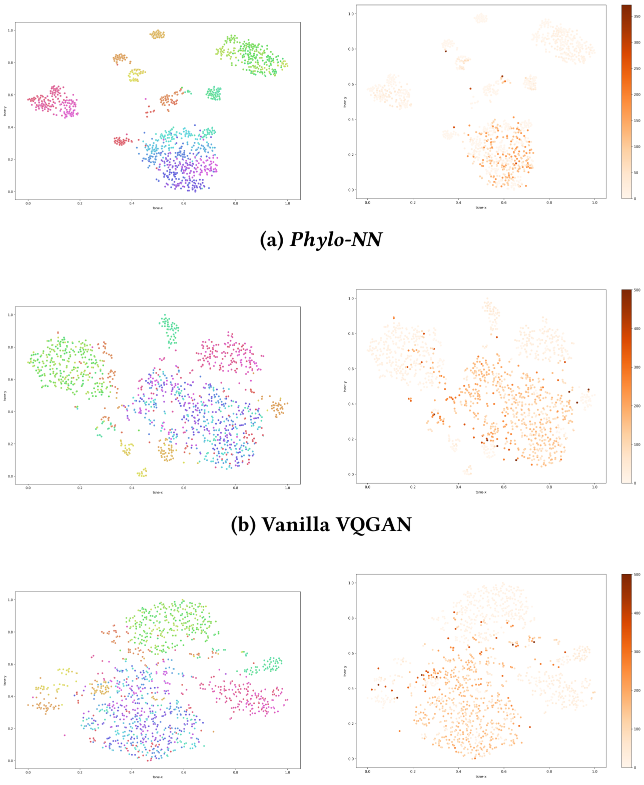
E T-SNE PLOTS FOR TEST IMAGES

As we have shown in Section 6.4, the embedding of the images generated by our *Phylo-NN* model exhibit much better qualities than those generated by a vanilla VQGAN. In this section, we run the same analysis on the test images. Clearly, a similar case can be made here, namely that the encoding of the images using *Phylo-NN* is superior to other models' in terms of its clustering. Figure 7 illustrates this point clearly when comparing *Phylo-NN*'s (top row) t-SNE plots with those of the other models'. Both vanilla VQGAN and CW perform worse at clustering the dataset compared to *Phylo-NN*.

F EXAMPLES OF PHYLO HISTOGRAMS

In Section 6.3, by means of calculating the average JS-divergence of sequence histograms, we investigated how well the Phylo-sequences match for species that share a common ancestor, as opposed to those that don't. In this section, we show some of these histogram plots to illustrate their value and the insight they could provide.

In figures 8 to 10, each histogram represents a code location in the phylogenetic sequence. Each column represents one of the $n_\text{l} = 4$ phylogenetic levels into which the phylogeny tree was quantized, starting with the species level from right and climbing the phylogenetic tree all the way to the earliest ancestral level on the left. Each column has $n_\text{p} = 8$ codes. Each histogram shows the relative frequency of each code of the learned $n_\text{q} = 64$ codes for its corresponding sequence location. The lower a histogram's entropy (i.e., when there is only one or a couple of codes that dominate the histogram's frequency spectrum), the more important that code location hypothetically is for characterizing the species at its corresponding phylogenetic level.


(a) *Phylo-NN*

(b) Vanilla VQGAN

(c) CW

Figure 7: t-SNE plots of the test set images using different models

As can be seen, both *Notropis* species share many codes at many sequence locations up to, but not including, the species level. This is because these species share an immediate ancestor. In contrast, We can see that the *Lepomis* species has a distinct histogram signature and does share almost no codes with the *Notropis* species, except for the earlier ancestral level (i.e., the left column).

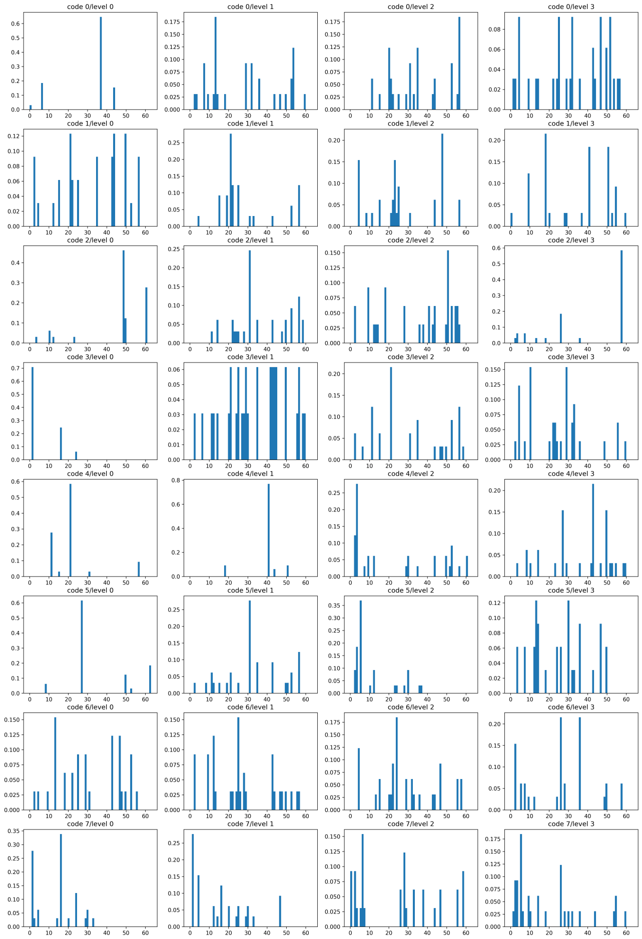
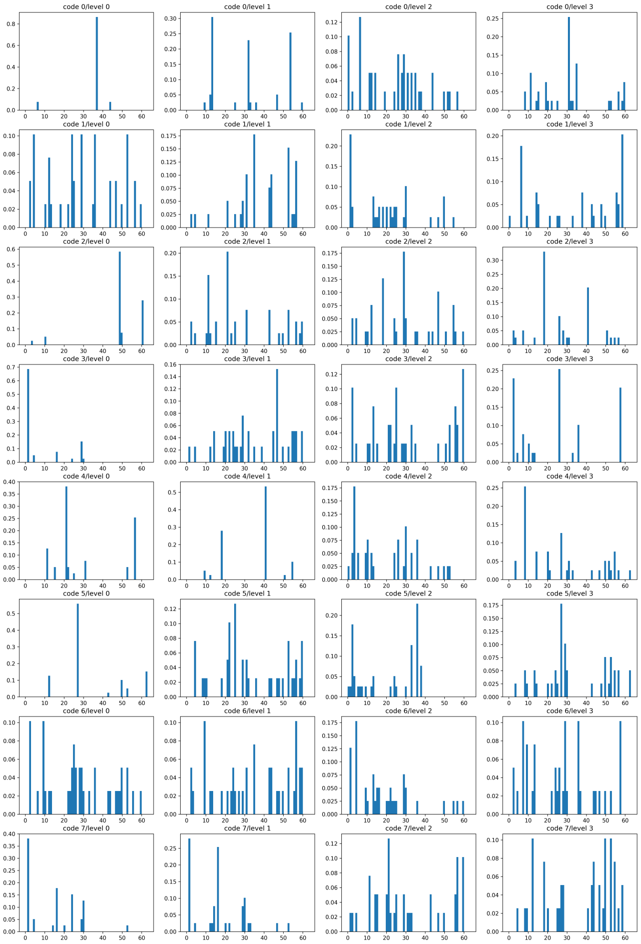
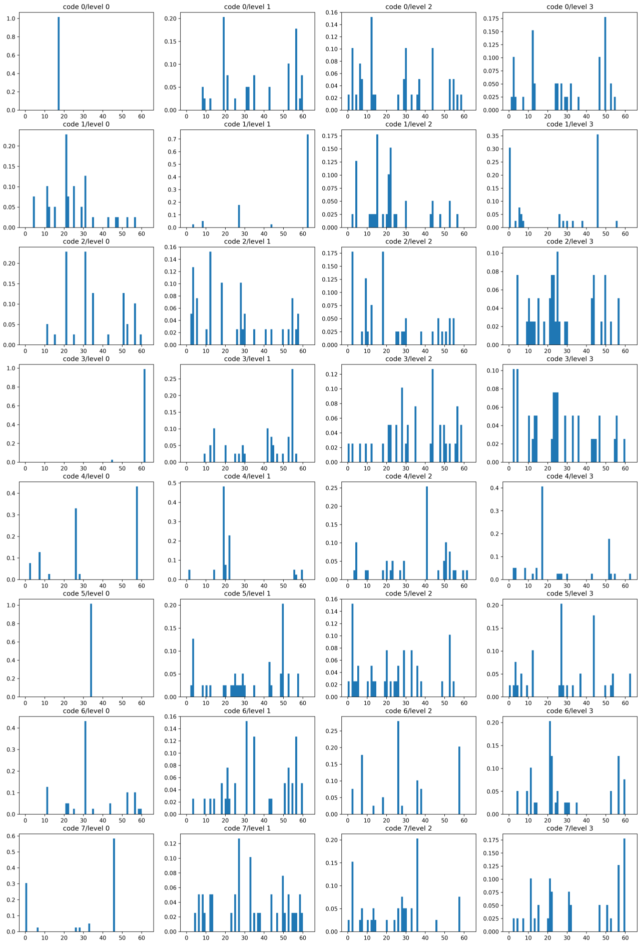


Figure 8: ***Notropis nubilus***



code 0/level 0
code 0/level 1
code 0/level 2
code 0/level 3
code 1/level 0
code 1/level 1
code 1/level 2
code 1/level 3
code 2/level 0
code 2/level 1
code 2/level 2
code 2/level 3
code 3/level 0
code 3/level 1
code 3/level 2
code 3/level 3
code 4/level 0
code 4/level 1
code 4/level 2
code 4/level 3
code 5/level 0
code 5/level 1
code 5/level 2
code 5/level 3
code 6/level 0
code 6/level 1
code 6/level 2
code 6/level 3
code 7/level 0
code 7/level 1
code 7/level 2
code 7/level 3

Figure 9: *Notropis percobromus*



code 0/level 0
code 0/level 1
code 0/level 2
code 0/level 3
code 1/level 0
code 1/level 1
code 1/level 2
code 1/level 3
code 2/level 0
code 2/level 1
code 2/level 2
code 2/level 3
code 3/level 0
code 3/level 1
code 3/level 2
code 3/level 3
code 4/level 0
code 4/level 1
code 4/level 2
code 4/level 3
code 5/level 0
code 5/level 1
code 5/level 2
code 5/level 3
code 6/level 0
code 6/level 1
code 6/level 2
code 6/level 3
code 7/level 0
code 7/level 1
code 7/level 2
code 7/level 3

Figure 10: *Lepomis macrochirus*

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009