BioCLIP: A Vision Foundation Model for the Tree of Life

Samuel Stevens[1*†], Jiaman Wu[1*], Matthew J Thompson[1], Elizabeth G Campolongo[1], Chan Hee Song[1], David Edward Carlyn[1], Li Dong[2], Wasila M Dahdul[3], Charles Stewart[4], Tanya Berger-Wolf[1], Wei-Lun Chao[1], and Yu Su[1†]

[1]The Ohio State University, [2]Microsoft Research, [3]University of California, Irvine, [4]Rensselaer Polytechnic Institute

Abstract

Images of the natural world, collected by a variety of cameras, from drones to individual phones, are increasingly abundant sources of biological information. There is an explosion of computational methods and tools, particularly computer vision, for extracting biologically relevant information from images for science and conservation. Yet most of these are bespoke approaches designed for a specific task and are not easily adaptable or extendable to new questions, contexts, and datasets. A vision model for general organismal biology questions on images is of timely need. *To approach this, we curate and release* TreeOfLife-10M*, the largest and most diverse ML-ready dataset of biology images. We then develop* BioCLIP*, a foundation model for the tree of life, leveraging the unique properties of biology captured by* TreeOfLife-10M*, namely the abundance and variety of images of plants, animals, and fungi, together with the availability of rich structured biological knowledge. We rigorously benchmark our approach on diverse fine-grained biology classification tasks and find that* BioCLIP *consistently and substantially outperforms existing baselines (by 16% to 17% absolute). Intrinsic evaluation reveals that* BioCLIP *has learned a hierarchical representation conforming to the tree of life, shedding light on its strong generalizability.*[1]

1. Introduction

Digital images and computer vision are quickly becoming pervasively used tools to study the natural world, from evolutionary biology [13, 51] to ecology and biodiversity [5, 77, 83]. The capability to rapidly convert vast quantities of images from museums [64], camera traps [1, 6, 7, 59, 77], and citizen science platforms [2, 40, 54, 58, 60, 62, 75, 79–81, 87, 88] into actionable information (e.g., species classification, individual identification, and trait detection) has accelerated and enabled new advances in tasks such as species delineation [32], understanding mechanisms of adaptation [23, 39], abundance and population structure estimation [3, 40, 58, 82], and biodiversity monitoring and conservation [83].

However, applying computer vision to answer a biological question is still a laborious task requiring substantial machine learning expertise and effort—biologists must manually label sufficient data for the specific taxa and task of interest, and find and train a suitable model for the task. Meanwhile, foundation models [12] such as CLIP [69] and GPT-3 [14] are extraordinarily valuable by enabling zero-shot or few-shot learning for a wide range of tasks. An analogous vision foundation model for biology should be useful for tasks spanning the *entire* tree of life [37, 53] instead of just the taxa it was trained on. Such a model would significantly lower the barrier to apply AI to biology.

In this work, we aim to develop such a vision foundation model for the tree of life. To be broadly useful for real-world biology tasks, this model should meet the following criteria. First, it should **generalize to the entire tree of life**, where possible, to ensure it supports researchers studying many different clades rather than a niche. Furthermore, it is infeasible to collect training data that covers the millions of known taxa [38, 44], so the model must generalize to taxa not present in training data. Second, it should learn **fine-grained representations** of images of organisms as biology frequently engages with organisms that are visually similar, like closely related species within the same genus [67] or species mimicking others' appearances for a fitness advantage [39]. This fine-grained granularity is crucial because the tree of life organizes living things into both broad categories (animal, fungus, and plant) and very fine-grained ones (see Fig. 1). Finally, due to the high cost of data collection and labeling in biology, **strong performance in the**

*Equal contribution. †{stevens.994,su.809}@osu.edu

[1] imageomics.github.io/bioclip has models, data and code.

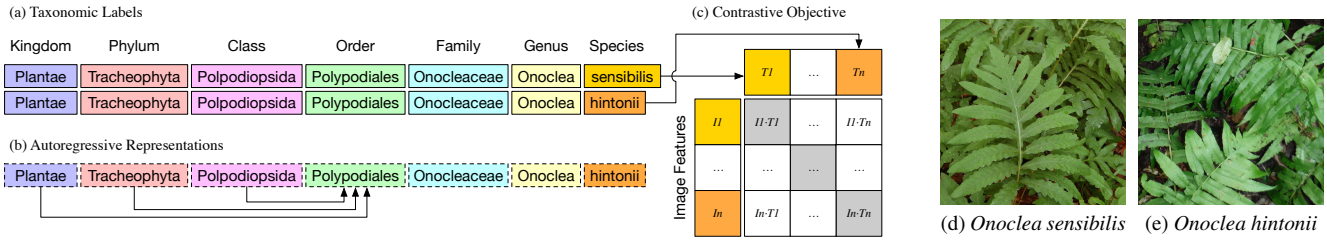

Figure 1. (a) Two taxa, or taxonomic labels, for two different plants, *Onoclea sensibilis* (d) and *Onoclea hintonii* (e). These taxa are identical except for the species. (b) The autoregressive text encoder naturally encodes the hierarchical structure of the taxonomy. See how the Order token(s) (Polypodiales) can incorporate information from the Kingdom, Phylum and Class tokens, but nothing later in the hierarchy. This helps align the visual representations to this same hierarchical structure (see §4.6). (c) These hierarchical representations of taxonomic labels are fed into the standard contrastive pre-training objective and are matched with image representations (d) and (e).

low-data regime (i.e., zero-shot or few-shot) is critical.

While the goals of **generalization**, **fine-grained classification**, and **data efficiency** are not new in computer vision, existing general-domain vision models [61, 69, 95] trained on hundreds of millions of images fall short when applied to evolutionary biology and ecology. Specifically, existing vision models produce *general* fine-grained representations, useful for comparing common organisms like dogs and wolves, but not for more fine-grained comparisons, e.g., *Onoclea sensibilis* and *Onoclea hintonii* (see Fig. 1).

We identify two major barriers to developing a vision foundation model for biology. First, there is a need for suitable **pre-training datasets**: existing datasets [28, 86, 88, 89] lack either scale, diversity, or fine-grained labels. Second, there is a need to investigate suitable **pre-training strategies** that leverage special properties of the biology domain to better achieve the three pivotal goals, e.g., the tree of life taxonomy, which is insufficiently considered in mainstream pre-training algorithms [48, 61, 69].

In light of these goals and challenges in achieving them, we introduce 1) **TREEOFLIFE-10M**, a large-scale ML-ready biology image dataset, and 2) **BIOCLIP**, a vision foundation model for the tree of life, trained with suitable use of taxa in TREEOFLIFE-10M. We outline the contributions, conceptual framework, and design decisions below:

TREEOFLIFE-10M: a large-scale, diverse ML-ready biology image dataset. We curate and release the largest ML-ready dataset to-date of biology images with associated taxonomic labels, containing over 10 million images covering 454 thousand taxa in the tree of life.[2] In comparison, the current largest ML-ready biology image dataset, iNat21 [86], contains only 2.7 million images covering 10 thousand taxa. TREEOFLIFE-10M integrates existing high-quality datasets like iNat21 and BIOSCAN-1M [28]. More importantly, it includes newly curated images from the Encyclopedia of Life (eol.org), which supplies most of TREEOFLIFE-10M's data diversity. Every image in TREEOFLIFE-10M is labeled with its taxonomic hierarchy to the finest level possible, as well as higher taxonomic ranks in the tree of life (see Fig. 1 and Tab. 3 for examples of taxonomic ranks and labels). TREEOFLIFE-10M enables training BIOCLIP and future biology foundation models.

BIOCLIP: a vision foundation model for the tree of life. With a large-scale labeled dataset like TREEOFLIFE-10M, a standard, intuitive training strategy (as adopted by other vision models like ResNet50 [33] and Swin Transformer [48]) is to use a supervised classification objective and learn to predict the taxonomic indices from an image. However, this fails to recognize and leverage the rich structure of taxonomic labels—taxa do not exist in isolation but are interconnected in a comprehensive taxonomy. Consequently, a model trained via plain supervised classification may not generalize well to taxa unseen in training, nor could it support zero-shot classification of unseen taxa.

Instead, we propose a novel strategy *combining CLIP-style multimodal contrastive learning [69] with the rich biological taxonomy* for BIOCLIP. We "flatten" the taxonomy from Kingdom to the distal-most taxon rank into a string called *taxonomic name*, and use the CLIP contrastive learning objective to learn to match images with their corresponding taxonomic names. Intuitively, this helps the model generalize to unseen taxa—even if the model has not seen a species, it has likely learned a reasonable representation for that species' genus or family (see Fig. 1). BIOCLIP also supports zero-shot classification with taxonomic names of unseen taxa. We further propose, and demonstrate the effectiveness of, a *mixed text type* training strategy; by mixing different text types (e.g., taxonomic vs. scientific vs. common names) during training, we retain the generalization from taxonomic names while being more flexibility at test time. For example, BIOCLIP still excels even if only common species names are offered by downstream users.

Comprehensive benchmarking. We comprehensively evaluate BIOCLIP on 10 fine-grained image classification datasets covering animals, plants, and fungi, including a newly curated RARE SPECIES dataset unseen in training. BIOCLIP achieves strong performance in both zero-shot

[2] By ML-ready, we mean the data is standardized in a format suitable for training ML models and is readily available for downloading.

and few-shot settings and substantially outperforms both CLIP [69] and OpenCLIP [42], leading to an average absolute improvement of **17%** (zero-shot) and **16%** (few-shot). Intrinsic analysis further reveals that BIOCLIP has learned a more fine-grained hierarchical representation conforming to the tree of life, explaining its superior generalization.

2. TREEOFLIFE-10M

Recent work has shown that data quality and diversity are critical when training CLIP models [24, 26, 57]. We curate TREEOFLIFE-10M, the most diverse large-scale public ML-ready dataset for computer vision models in biology.

2.1. Images

The largest ML-ready biology image dataset is iNat21 [86] with 2.7M images of 10K species. Despite this class breadth compared to popular general-domain datasets like ImageNet-1K [70], 10K species is limited for biology. The International Union for Conservation of Nature (IUCN) reported over 2M total described species in 2022, with over 10K bird species and over 10K reptile species alone [44]. iNat21's species diversity limits its potential for training a foundation model for the entire tree of life.

Motivated to find high-quality biology images with a focus on species diversity, we turn to the Encyclopedia of Life project (EOL; eol.org). EOL collaborates with a variety of institutions to gather and label millions of images. We download 6.6M images from EOL and expand our dataset to cover an additional **440**K taxa.

Species are not evenly distributed among the different subtrees in the tree of life; insects (of the class *Insecta* with 1M+ species), birds (of the class *Aves* with 10K+ species) and reptiles (of the class *Reptilia* with 10K+ species) are examples of highly diverse subtrees with many more species. To help a foundation model learn extremely fine-grained visual representations for insects, we also incorporate BIOSCAN-1M [28], a recent dataset of 1M lab images of insects, covering 494 different families.[3] Furthermore, BIOSCAN-1M contains *lab* images, rather than in situ images like iNat21, diversifying the *image* distribution.

2.2. Metadata & Aggregation

The TREEOFLIFE-10M dataset integrates iNat21 (training split), our curated EOL dataset, and BIOSCAN-1M by aggregating the images and canonicalizing the labels. *This is a highly non-trivial task because taxonomic hierarchies are notoriously noisy and rarely consistent between sources* [4, 31, 36, 52, 63], likely contributing to the prior lack of

[3]We note that BIOSCAN-1M's label granularity may still be limited for insects. 98.6% of BIOSCAN-1M's images are labeled to the family level but only 22.5% and 7.5% of the images have genus or species indicated, respectively. Lack of label granularity is an inherent challenge.

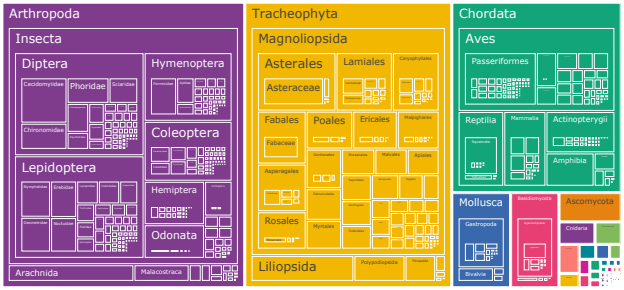

Figure 2. Treemap of the 108 phyla in TREEOFLIFE-10M. Different colors are different phyla; nested boxes represent classes, orders, and families. Box size, not number of inner boxes, represents relative number of samples.

image datasets large enough to train a foundation-scale vision model for the entire tree of life. We carefully unify and backfill taxonomic hierarchies from EOL, the Integrated Taxonomic Information System (ITIS) [43], and iNaturalist with special consideration for the existence of homonyms (genus-species labels shared among higher-order taxa). For more information on this process, the challenges, our solutions, and remaining issues, see Appendix C.

2.3. Release & Statistics

Tab. 1 presents dataset statistics: TREEOFLIFE-10M has over 10M images across more than 450K unique taxonomic names. Fig. 2 shows the distribution of images by phyla and the respective lower-rank taxa (order through family).

Our curated training and test datasets (TREEOFLIFE-10M and RARE SPECIES, described in §4.2) are available on Hugging Face (with DOIs) under a public domain waiver, to the extent primary source licenses allow. This includes CSVs with image metadata and links to the primary sources, accompanied by a GitHub repository with the scripts to generate the datasets.[4]

3. Modeling

BIOCLIP is initialized from OpenAI's public CLIP checkpoint and continually pre-trained on TREEOFLIFE-10M with CLIP's multimodal contrastive learning objective.

3.1. Why CLIP?

Compared with general domain computer vision tasks, one of the most salient differences for the biology domain is its rich label space. Not only are the taxon labels large in quantity (there are 2M+ recorded species as of 2022 [44]), but they are also connected with each other in a hierarchical taxonomy. This is a challenge for training a good foundation model that can achieve satisfactory coverage and generalization. Despite this, the intricate structure in the label

[4]We encourage future work to cite iNat21 [86], BIOSCAN-1M [28] and to appropriately attribute images from EOL based on their licenses if citing TREEOFLIFE-10M.

Dataset	Description	Images	Unique Classes
iNat21	Citizen scientist labeled image dataset from iNaturalist for fine-grained classification.	2.7M	10,000
BIOSCAN-1M	Expert labeled image dataset of insects for classification.	1.1M	7,831
EOL	A new dataset with citizen scientist images sourced from Encyclopedia of Life and taxonomic labels standardized by us.	6.6M	448,910
TREEOFLIFE-10M	**Largest-to-date ML-ready dataset of biology images with taxonomic labels.**	**10.4M**	**454,103**

Table 1. Training data sources used in TREEOFLIFE-10M. We integrate and canonicalize taxonomic labels across the sources (§2.2).

	Name	Description	Examples	Classes	Labels
Animals	Birds 525	Scraped dataset of bird images from web search. [68]	89,885	525	Taxonomic
	Plankton	Expert-labeled in situ images of plankton [35].	4,080	102	Mixed
	Insects	Expert and volunteer-labeled in-the-wild citizen science images of insects [74].	4,680	117	Scientific
	Insects 2	Mixed common and scientific name classification for insect pests [91].	4,080	102	Mixed
Plants & Fungi	PlantNet	Citizen science species-labeled plant images, some drawings [27].	1,000	25	Scientific
	Fungi	Expert-labeled images of Danish fungi [66].	1,000	25	Scientific
	PlantVillage	Museum-style leaf specimens labeled with common names [25].	1,520	38	Common
	Medicinal Leaf	Species classification of leaves from mature, healthy medicinal plants [71].	1,040	26	Scientific
	PlantDoc	17 diseases for 13 plant species [76].	1,080	27	Common
	RARE SPECIES	Subset of species in the IUCN Red List categories: Near Threatened through Extinct in the Wild (iucnredlist.org).	12,000	400	Taxonomic

Table 2. Datasets used for evaluation. All tasks are classification evaluated with Top-1 accuracy.

space, accumulated through centuries of biology research, provides very rich signal for learning better generalization. Intuitively, if the label space's structure is successfully encoded in a foundation model, even if the model has not seen a certain species, it will likely have learned a good representation for that species' corresponding genus or family. Such a hierarchical representation serves as a strong prior to enable few-shot or even zero-shot learning of new taxa.

Many vision foundation models, such as ResNet [33] and Swin Transformer [48], adopt a supervised classification objective and directly learn the mapping from input images to class indices. As a result, each class label is treated as a distinct symbol, and their relationships are neglected. *A key realization of our work is that the multimodal contrastive learning objective used in CLIP can be repurposed for leveraging the hierarchical structure of the label space.* This is not an obvious choice; after all, TREEOFLIFE-10M is largely labeled with class labels and not with free-form text like image captions. The autoregressive text encoder naturally embeds the taxonomic hierarchy into a dense label space by conditioning later taxonomic rank representations on higher ranks (Fig. 1). While hierarchical classification [9, 11, 96] can also leverage taxonomy, we empirically show that CLIP-style contrastive learning significantly improves generalization (§4.4). We note that *repurposing CLIP's multimodal contrastive learning objective for learning hierarchical representations conforming to a taxonomy* is a novel and non-trivial technical contribution.

CLIP trains two uni-modal embedding models, a vision encoder and a text encoder, to (1) maximize feature similarity between *positive* (image, text) pairs and (2) minimize feature similarity between *negative* (image, text) pairs, where positive pairs are from the training data and negative pairs are all other possible (image, text) pairings in a batch. After training, CLIP's encoder models embed individual instances of their respective modalities into a shared feature space. Next, we discuss formatting the text input to CLIP to incorporate the taxonomic structure.

Text Type	Example
Common	black-billed magpie
Scientific	*Pica hudsonia*
Taxonomic	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia*
Scientific + Common	*Pica hudsonia* with common name black-billed magpie
Taxonomic + Common	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia* with common name black-billed magpie

Table 3. Text types considered in the training of BIOCLIP.

3.2. Text Types

An advantage of CLIP is that the text encoder accepts free-form text. In biology, unlike other classification tasks, class names are diversely formatted. We consider the following:

Taxonomic name. A standard seven-level biology taxonomy from higher to lower level is kingdom, phylum, class, order, family, genus and species. For each species, we "flatten" the taxonomy by concatenating all labels from root to leaf into a single string, which we call the *taxonomic name*.

Scientific name. Scientific names are composed of genus and species (e.g., *Pica hudsonia*).

Common name. Taxonomy categories are usually Latin, which is not often seen in generalist image-text pre-training datasets. Instead, the common name, such as "black-billed magpie," is more widespread. Note that common names may not have a 1-to-1 mapping to taxa: A single species may have multiple common names, or the same common name may refer to multiple species.

For certain downstream use cases of BIOCLIP, it might be the case that only one type of label, e.g., scientific names, is available. To improve the flexibility at inference time, we propose a *mixed text type* training strategy: at each training step, we pair each input image with a text randomly sampled from all of its available text types (shown in Tab. 3). We empirically show that this simple strategy retains the generalization benefits of taxonomic names while providing more flexibility in using other names at inference time (§4.3). The final text input to CLIP is the name in the standard CLIP template, e.g., "a photo of *Pica hudsonia*".

4. Experiments

We train BIOCLIP on TREEOFLIFE-10M, compare BIOCLIP to general vision models and investigate how our modeling choices affect BIOCLIP's performance.

4.1. Training and Evaluation Details

To train BIOCLIP, we initialize from OpenAI's CLIP weights [69] with a ViT-B/16 vision transformer [22] image encoder and a 77-token causal autoregressive transformer text encoder. We continue pre-training on TREEOFLIFE-10M for 100 epochs with a cosine learning rate schedule [49]. We train on 8 NVIDIA A100-80GB GPUs over 2 nodes with a global batch size of 32,768. We also train a baseline model on only the iNat21 dataset and multiple ablation models on 1M examples randomly sampled from TREEOFLIFE-10M (Secs. 4.3 and 4.4), following the same procedure for BIOCLIP except with a smaller global batch size of 16,384 on 4 NVIDIA A100 GPUs on 1 node. All hyperparameters and training details are in Appendix D and training and evaluation code is publicly available.

We evaluate on **10 different classification tasks**: the 8 biologically-relevant tasks from **Meta-Album** [84], **Birds 525** [68] and our new **RARE SPECIES** task (described in §4.2). Meta-Album is a dataset collection for meta-learning, encompassing various subjects. Specifically, we use the Plankton, Insects, Insects 2, PlantNet, Fungi, PlantVillage, Medicinal Leaf, and PlantDoc datasets. Our classification tasks cover all four multi-celled kingdoms in the tree of life (animals, plants, fungi, and protists) and have a diverse image distribution (photographs, microscope images, drawings, and museum specimens). Tab. 2 summarizes the datasets; they comprise a variety of label types from full taxonomic names to only scientific or common name.

For **zero-shot learning**, we follow the same procedure as CLIP. For **few-shot learning**, we follow SimpleShot [90] and use a nearest-centroid classifier. For k-shot learning, we first randomly sample k examples for each class and obtain the image embedding from the visual encoder of the pre-trained models. We then compute the average feature vector of the k embeddings as the centroid for each class. All the examples left in the dataset are used for testing. After applying mean subtraction and L2-normalization to each centroid and test feature vector, we choose the class with the nearest centroid to the test vector as the prediction. We repeat each few-shot experiment 5 times with different random seeds and report the mean accuracy in Tab. 4. Results with standard deviations are reported in Appendix E.

We compare BIOCLIP with the original OpenAI CLIP [69] and OpenCLIP [42] trained on LAION-400M [73]. Intuitively, common names of organisms are most pervasive in the training data of CLIP and OpenCLIP and these models work best with common names. This is also confirmed in our preliminary tests. Therefore, we use common names as class labels for CLIP and OpenCLIP by default unless unavailable for a dataset. BIOCLIP can leverage taxonomic names, so we use taxonomic+common names by default. However, as noted in Tab. 2, the test datasets come in a variety of labels. Whenever the preferred label type is not available, we use labels that come with the dataset. We also compare to an ImageNet-21K [21] pre-trained model [78] and DINO [15] for few-shot classification.

4.2. Can BIOCLIP Generalize to Unseen Taxa?

Taxonomic names are added, removed, and changed as biologists discover and classify new and existing species. BIOCLIP should generalize to unseen taxonomic names to avoid re-training for every new species. To empirically answer whether BIOCLIP generalizes well to unseen taxa, we introduce a new evaluation task that is both biologically and empirically motivated: **RARE SPECIES.**

Classifying "rare" species is an important and challenging computer vision application in biology, particularly in the context of global conservation efforts [83]. To the best of our knowledge, there is no diverse, publicly available rare species classification dataset. Recently published work [47, 56] lack species diversity with only a dozen classes. We aim to fill this gap, collecting all $\approx$ 25K species on the IUCN Red List (iucnredlist.org) classified[5] as Near Threatened, Vulnerable, Endangered, Critically Endangered, or Extinct in the Wild. We select 400 such species represented by at least 30 images in our EOL dataset, then remove

[5]IUCN has classified 150,388 species and generally updates their list twice per year (IUCN Update Schedule). The classifications used for this dataset are current as of July 13, 2023.

Model	Animals				Plants & Fungi							
	Birds 525	Plankton	Insects	Insects 2	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc	Rare Species	Mean	Mean (Δ)
Random Guessing	0.2	1.2	1.0	1.0	4.0	4.0	2.6	4.0	3.7	0.3	2.2	
Zero-Shot Classification												
CLIP	49.9	3.2	9.1	9.8	58.5	10.2	5.4	15.9	26.1	31.8	21.9	–
OpenCLIP	54.7	2.2	6.5	9.6	50.2	5.7	8.0	12.4	25.8	29.8	20.4	−1.5
BioCLIP	**72.1**	**6.1**	**34.8**	**20.4**	**91.4**	40.7	**24.4**	**38.6**	**28.4**	**38.0**	**39.4**	+17.5
– iNat21 Only	56.1	2.6	30.7	11.5	88.2	**43.0**	18.4	25.6	20.5	21.3	31.7	+9.8
One-Shot Classification												
CLIP	43.7	25.1	21.6	13.7	42.1	17.2	49.7	70.1	24.8	28.5	33.6	–
OpenCLIP	53.7	32.3	23.2	14.3	45.1	18.4	53.6	71.2	26.8	29.2	36.7	+3.1
Supervised-IN21K	60.2	22.9	14.7	14.4	46.7	16.9	**62.3**	58.6	27.7	28.0	35.2	+1.6
DINO	40.5	**37.0**	23.5	16.4	30.7	20.0	60.0	79.2	23.7	31.0	36.2	+2.6
BioCLIP	71.8	30.6	**57.4**	**20.4**	64.5	**40.3**	58.8	**84.3**	**30.7**	**44.9**	**50.3**	+16.7
– iNat21 Only	**74.8**	29.6	53.9	19.7	**67.4**	35.5	55.2	75.1	27.8	36.9	47.5	+13.9
Five-Shot Classification												
CLIP	73.5	41.2	39.9	24.6	65.2	27.9	71.8	89.7	35.2	46.0	51.5	–
OpenCLIP	81.9	52.5	42.6	25.0	68.0	30.6	77.8	91.3	42.0	47.4	55.9	+4.4
Supervised-IN21K	83.9	39.2	32.0	25.4	70.9	30.9	**82.4**	82.3	44.7	47.3	53.9	+2.4
DINO	70.8	**56.9**	46.3	28.6	50.3	34.1	82.1	94.9	40.3	50.1	55.4	+3.9
BioCLIP	90.0	49.3	**77.8**	**33.6**	**85.6**	**62.3**	80.9	**95.9**	**47.5**	**65.7**	**68.8**	+17.3
– iNat21 Only	**90.1**	48.2	73.7	32.1	84.7	55.6	77.2	93.5	41.0	55.6	65.1	+13.6

Table 4. Zero-, one- and five-shot classification top-1 accuracy for different models. **Bold** indicates best accuracy. All models use the same ViT-B/16 architecture. "iNat21 Only" follows the same procedure as BioCLIP but uses iNat21 instead of TreeOfLife-10M. Δ denotes the difference in mean accuracy with CLIP. Supervised-IN21K [78] and DINO [15] are vision-only models and cannot do zero-shot classification.

Dataset	Train↓Test→	Com	Sci	Tax	Sci+Com	Tax+Com
ToL-1M	Com	24.9	9.5	10.8	22.3	21.0
	Sci	11.0	22.3	4.5	21.5	8.0
	Tax	11.8	10.1	26.6	16.0	24.8
	Sci+Com	24.5	12.9	12.6	28.0	24.9
	Tax+Com	20.5	8.0	19.7	24.0	30.4
	Mixture	**26.1**	**24.9**	**26.7**	**29.5**	**30.9**
iNat21-2.7M	Mixture	20.4	14.7	15.6	20.9	21.3
ToL-10M	Mixture	31.6	30.1	34.1	37.0	38.0

Table 5. Zero-shot accuracy on species not seen during training (Rare Species task). Different rows and columns indicate different text types during training and test time, respectively. **Blue** indicates best accuracy and Orange indicates second-best. Using the taxonomic name over the scientific name always improves accuracy (22.3→26.6 and 28.0→30.4). The final rows use the full iNat21 dataset and TreeOfLife-10M for reference.

them from TreeOfLife-10M, creating an *unseen* Rare Species test set with 30 images per species. This dataset demonstrates (1) BioCLIP's out-of-distribution generalization to unseen taxa, (2) BioCLIP's potential applications, and (3) provides a crucial dataset for the community to address the ongoing biodiversity crisis.

Results. Tab. 4 shows that BioCLIP substantially outperforms both baseline CLIP models, as well as the iNat21-trained CLIP model, at zero-shot classification, especially on unseen taxa (see the "Rare Species" column). We attribute BioCLIP's strong zero-shot performance on this broad and diverse set of tasks to the broad and diverse classes in TreeOfLife-10M. We explore how data diversity leads to broadly useful image representations in §4.3.

4.3. How Do Text Types Affect Generalization?

We investigate how different text types affect zero-shot generalization by training BioCLIP on a 10% subset of TreeOfLife-10M (10% due to computational constraints). We use our Rare Species dataset because the test classes have every text type, and all species are excluded from training, making it ideal for testing generalization to unseen taxa. Prior works find that the diversity of captions makes stronger vision models [57] and randomly use one of five different captions for each image during training rather than a single fixed caption [72]. Similarly, we use a mixed text type strategy (§3.2). How does that affect performance?

Results. The zero-shot ablation results are in Tab. 5; there are several salient observations. First, using taxo-

Objective	Mean 1-Shot	Mean 5-shot
Cross-entropy	16.5	26.2
Hier. cross-entropy	19.3	30.5
CLIP	**44.7**	**63.8**

Table 6. One- and five-shot classification top-1 accuracy for different pre-training objectives on TREEOFLIFE-1M. Results are macro-averaged over all the test sets in Tab. 4.

nomic+common names yields the strongest performance, showing the importance of incorporating the taxonomic structure for generalization. Second, when only using a single text type for training, performance degrades substantially when a different text type is used at test time. Using mixed text types for training yields consistently strong performance across all text types during testing. These results indicate that mixed text type pre-training largely retains the generalization benefits of using taxonomic names while also providing flexibility of different text types for inference, an important property for a foundation model that may be used for diverse downstream tasks. Finally, using 1M examples from TREEOFLIFE-10M outperforms using 2.7M examples from iNat21, further confirming the importance of the added data diversity from TREEOFLIFE-10M.

4.4. Is the CLIP Objective Necessary?

Using the CLIP objective to pre-train on a labeled image dataset is an unintuitive decision (Goyal et al. [29] fine-tune using the CLIP objective, but do not pretrain). We justify our choice by training two ViT-B/16 models on TREEOFLIFE-1M using a cross-entropy classification loss and a multitask hierarchical variant, then compare them against the CLIP objective under the few-shot setting. The multitask hierarchical training objective is to predict the labels for kingdom, phylum, etc., down to species, using cross entropy for each level of the taxonomy, then summing those losses [11]. Pseudo-code is provided in Listing 1.

Results. We evaluate each model on the same set of 10 tasks but only in the one-shot and five-shot settings because non-CLIP models cannot do zero-shot classification. We report mean accuracy in Tab. 6. The hierarchical classification model outperforms simple classification and is comparable to the CLIP baseline (see Tab. 4). However, the CLIP objective massively outperforms both baselines and strongly justifies our repurposing of the CLIP objective.

4.5. Can BIOCLIP Classify More Than Species?

BIOCLIP is trained on a (contrastive) species-classification objective, which might limit its use beyond species classification. We consider plant diagnosis with the PlantVillage and PlantDoc datasets, which require classifying both species and disease (if any). Large-scale data labeling is expensive, but biologists always label several instances for field guides or museum collections. Few-shot classification is thus an ideal setting for this sort of task transfer.

Results. BIOCLIP outperforms baselines at classifying plant diseases based on visual symptoms, in both zero-shot and few-shot settings (see PlantVillage and PlantDoc in Tab. 4). While Radford et al. [69] find that CLIP one-shot and two-shot classification is often worse than zero-shot (because few-shot settings cannot use the semantic information in the class name), BIOCLIP has learned useful visual representations that are useful even with only one labeled example: BIOCLIP's mean one-shot accuracy is 9.1% higher than zero-shot accuracy.

4.6. Does BIOCLIP Learn the Hierarchy?

BIOCLIP demonstrates strong performance in a low-data regime on our extrinsic evaluation, but why? We further conduct an intrinsic evaluation and visualize BIOCLIP's learned image representations to shed light on this question (Fig. 3). We embed image representations from iNat21's validation set (unseen during training) using t-SNE [85] and color the points by the image's taxonomic label. For each plot, we run t-SNE independently on the subset of examples under the labeled taxonomical rank. Each plot visualizes one taxonomic hierarchy rank and the top six categories of the next rank, e.g., the top left plot visualizes the six most common phyla in the Animalia kingdom. At higher ranks like kingdom (omitted for space) and phylum, both CLIP and BIOCLIP have good separation, but BIOCLIP's representations are more fine-grained and contain a richer clustering structure. At lower ranks, BIOCLIP produces evidently more separable features, while CLIP's features are cluttered and lack a clear structure. Appendix F has more qualitative results and visuals.

5. Related Work

Multimodal foundation model training data. CLIP [69] trained state-of-the-art vision models from noisy, web-scale (100M+) image-text datasets using a contrastive objective that is optimized for image retrieval. ALIGN [45] and BASIC [65] further scaled the number of training examples from 400M to 6.6B, improving vision representation quality. However, further work [24, 26, 57, 93, 94] all find that *dataset diversity and better alignment between the image and caption semantics* are more important than dataset size and lead to stronger performance on downstream tasks. TREEOFLIFE-10M *emphasizes the importance of diversity*, adding over 440K classes to iNat21's 10K and leading to stronger zero-shot performance.

Domain-specific CLIPs. Domain-specific training often beats general training [18, 30], but subject-matter experts are often too expensive to hire to label large-scael domain-specific datasets. Image-text training is thus particularly potent because models can learn from noisy image-text pairs. Ikezogwo et al. [41] and Lu et al. [50] gathered 1M+ image-

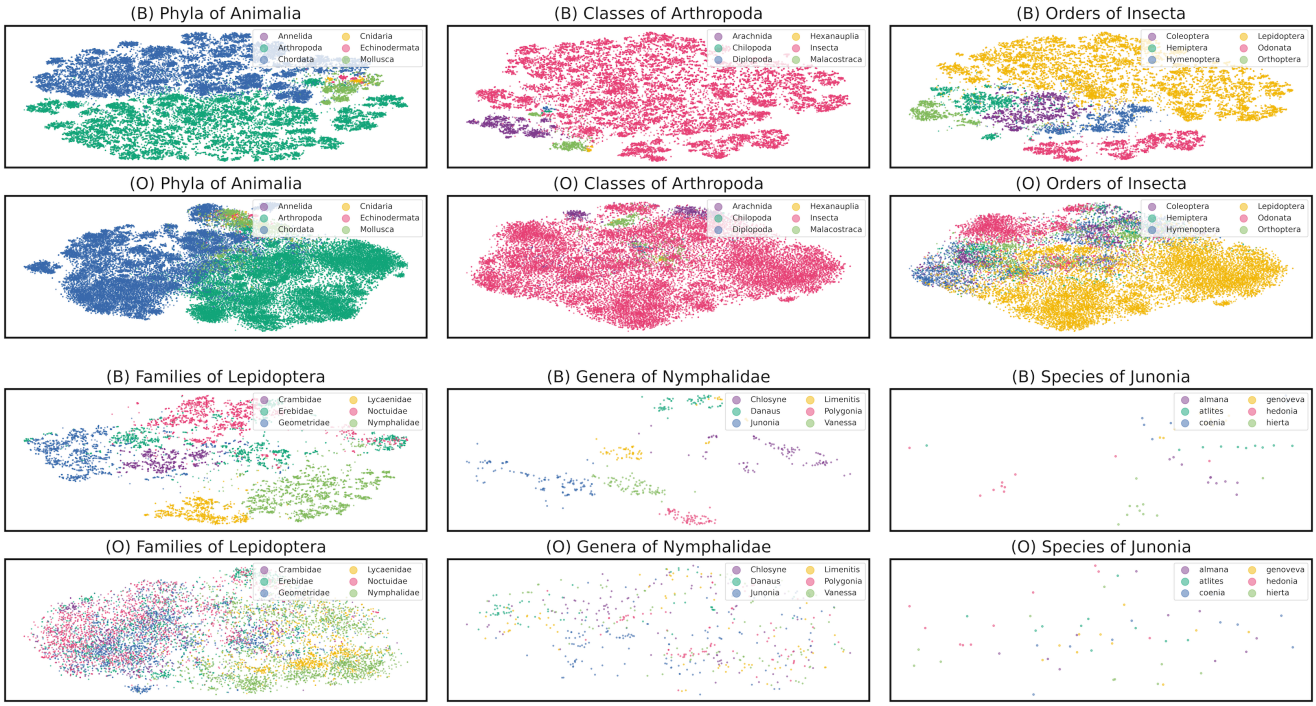

Figure 3. T-SNE visualization of image features, colored by taxonomic labels. BIOCLIP (B) is visualized in the first and third row and OpenAI's CLIP (O) is visualized in the second and fourth rows. BIOCLIP's features better preserve the hierarchical structure: while both BIOCLIP and CLIP cleanly separate the phyla in the Animalia Kingdom (top left), only BIOCLIP successfully separates the orders in the Insecta Class (top right) and the families in the Lepidoptera Order (bottom left).

text pairs for computational pathology. We gather $10\times$ the images, emphasizing class diversity.

Hierarchy in computer vision. Hierarchy in computer vision is well-studied, in part because ImageNet [70] classes are from the hierarchical WordNet [55]. Bilal et al. [10] study model predictions on ImageNet and find that model confusion patterns follow the hierarchical class structure. They incorporate hierarchy into AlexNet's architecture [46] and improve ImageNet top-1 error by 8% absolute. Bertinetto et al. [9] measure image classifiers' mistake severity and propose alternative objectives that incorporate hierarchy, reducing mistake severity at the expense of worsening top-1 accuracy. Zhang et al. [96] propose a contrastive objective where the hierarchical distance between labels corresponds to the desired distance in the embedding space, and outperform cross-entropy on ImageNet and iNat17 [88]. We apply hierarchical classification to 454K unique classes through a repurposed CLIP objective, while prior work applied hierarchies to smaller label spaces.

Computer vision for biology. Fine-grained classification is a classic challenge in computer vision, and biological images are often used to benchmark models. Berg et al. [8], Piosenka [68], Wah et al. [89] all use bird species classification to evaluate fine-grained classification ability. Biology tasks are used for contrastive learning frameworks [20, 92], weakly supervised object detection [19] and semi-supervised learning methods [34].

6. Conclusion

We introduce TREEOFLIFE-10M and BIOCLIP, a large-scale diverse biology image dataset and a foundation model for the tree of life, respectively. Through extensive evaluation, we show that BIOCLIP is a strong fine-grained classifier for biology in both zero- and few-shot settings. We corroborate our hypothesis that using the entire taxonomic name leads to stronger generalization than other caption types through an ablation on unseen species and by visualizing BIOCLIP's representations, finding that BIOCLIP-embedded images better match the taxonomic hierarchy.

Although the CLIP objective efficiently learns visual representations over 450K taxa, BIOCLIP is fundamentally trained to do classification. Future work will further scale up the data, e.g., incorporating more than 100M research-grade images from iNaturalist, and collect richer textual descriptions of species' appearances such that BIOCLIP can extract fine-grained trait-level representations.

Acknowledgements

We thank the Imageomics team (including Josef Uyeda, Jim Balhoff, Dan Rubenstein, Hank Bart, Hilmar Lapp, Sara Beery and Dipanjyoti Paul) and the OSU NLP group for their valuable feedback, the BIOSCAN-1M and iNaturalist teams for sharing their data, and Jennifer Hammock at EOL for her help accessing EOL's images. Our research is sup-

ported by NSF OAC 2118240 and resources from the Ohio Supercomputer Center [16].

References

[1] Jorge A Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y Zhao, Walter Jetz, Margaret Kinnaird, Sayali Kulkarni, Arnaud Lyet, David Thau, Michelle Duong, Ruth Oliver, and Anthony Dancer. Wildlife Insights: A Platform to Maximize the Potential of Camera Trap and Other Passive Sensor Wildlife Data for the Planet. *Environmental Conservation*, 47(1):1–6, 2020. Edition: 2019/09/26 ISBN: 0376-8929 Publisher: Cambridge University Press.

[2] Alexandre Antonelli, Kiran L. Dhanjal-Adams, and Daniele Silvestro. Integrating machine learning, remote sensing and citizen science to create an early warning system for biodiversity. *PLANTS, PEOPLE, PLANET*, 5(3):307–316, 2023.

[3] Gonzalo Araujo, Ariana Agustines, Steffen S. Bach, Jesse E. M. Cochran, Emilio De La Parra-Galván, Rafael De La Parra-Venegas, Stella Diamant, Alistair Dove, Steve Fox, Rachel T. Graham, Sofia M. Green, Jonathan R. Green, Royale S. Hardenstine, Alex Hearn, Mahardika R. Himawan, Rhys Hobbs, Jason Holmberg, Ibrahim Shameel, Mohammed Y. Jaidah, Jessica Labaja, Savi Leblond, Christine G. Legaspi, Rossana Maguiño, Kirsty Magson, Stacia D. Marcoux, Travis M. Marcoux, Sarah Anne Marley, Meynard Matalobos, Alejandra Mendoza, Joni A. Miranda, Brad M. Norman, Cameron T. Perry, Simon J. Pierce, Alessandro Ponzo, Clare E. M. Prebble, Dení Ramírez-Macías, Richard Rees, Katie E. Reeve-Arnold, Samantha D. Reynolds, David P. Robinson, Christoph A. Rohner, David Rowat, Sally Snow, Abraham Vázquez-Haikin, and Alex M. Watts. Improving sightings-derived residency estimation for whale shark aggregations: A novel metric applied to a global data set. *Frontiers in Marine Science*, 9:775691, 2022.

[4] Jonathan A. Rees and Karen Cranston. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*, 5:e12581, 2017.

[5] Sara Beery. Scaling Biodiversity Monitoring for the Data Age. *XRDS: Crossroads, The ACM Magazine for Students*, 27(4):14–18, 2021.

[6] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

[7] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

[8] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2011–2018, 2014.

[9] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12506–12515, 2020.

[10] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):152–162, 2018.

[11] Kim Bjerge, Quentin Geissmann, Jamie Alison, Hjalte MR Mann, Toke T Høye, Mads Dyrmann, and Henrik Karstoft. Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics*, 77:102278, 2023.

[12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[13] Marek L Borowiec, Rebecca B Dikow, Paul B Frandsen, Alexander McKeeken, Gabriele Valentini, and Alexander E White. Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13(8):1640–1660, 2022.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9650–9660, 2021.

[16] Ohio Supercomputer Center. Ohio supercomputer center, 1987.

[17] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 52–68. Springer, 2016.

[18] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022.

[19] Elijah Cole, Kimberly Wilber, Grant Van Horn, Xuan Yang, Marco Fornoni, Pietro Perona, Serge Belongie, Andrew Howard, and Oisin Mac Aodha. On label granularity and object localization. In *European Conference on Computer Vision*, pages 604–620, 2022.

[20] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14755–14764, 2022.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Briana D. Ezray, Drew C. Wham, Carrie E. Hill, and Heather M. Hines. Unsupervised machine learning reveals mimicry complexes in bumblebees occur along a perceptual continuum. *Proceedings of the Royal Society B: Biological Sciences*, 286(1910):20191501, 2019.

[24] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (CLIP). In *International Conference on Machine Learning*, pages 6216–6234, 2022.

[25] Geetharamani G. and Arun Pandian J. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Computers & Electrical Engineering*, 76:323–338, 2019.

[26] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[27] Camille Garcin, alexis joly, Pierre Bonnet, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, and Joseph Salmon. Pl@ntnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2021.

[28] 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.

[29] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023.

[30] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

[31] Robert P. Guralnick, Nico Cellinese, John Deck, Richard L. Pyle, John Kunze, Lyubomir Penev, Ramona Walls, Gregor Hagedorn, Donat Agosti, John Wieczorek, Terry Catapano, and Roderic D. M. Page. Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys*, 494:133–154, 2015.

[32] Oskar L. P. Hansen, Jens-Christian Svenning, Kent Olsen, Steen Dupont, Beulah H. Garner, Alexandros Iosifidis, Benjamin W. Price, and Toke T. Høye. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and Evolution*, 10(2):737–747, 2020.

[33] 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.

[34] Wei He, Kai Han, Ying Nie, Chengcheng Wang, and Yunhe Wang. Species196: A one-million semi-supervised dataset for fine-grained species recognition. In *Advances in Neural Information Processing Systems*, 2024.

[35] Emily F. Brownlee Heidi M. Sosik, Emily E. Peacock. Annotated plankton images - data set for developing and evaluating classification methods, 2015.

[36] Cody E. Hinchliff, Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, Karl Gude, David S. Hibbett, Laura A. Katz, H. Dail Laughinghouse, Emily Jane McTavish, Peter E. Midford, Christopher L. Owen, Richard H. Ree, Jonathan A. Rees, Douglas E. Soltis, Tiffani Williams, and Karen A. Cranston. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):12764–12769, 2015.

[37] Cody E Hinchliff, Stephen A Smith, James F Allman, J Gordon Burleigh, Ruchi Chaudhary, Lyndon M Coghill, Keith A Crandall, Jiabin Deng, Bryan T Drew, Romina Gazis, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):12764–12769, 2015.

[38] Donald Hobern, Saroj K Barik, Les Christidis, Stephen T. Garnett, Paul Kirk, Thomas M Orrell, Thomas Pape, Richard L Pyle, Kevin R Thiele, Frank E Zachos, et al. Towards a global list of accepted species vi: The catalogue of life checklist. *Organisms Diversity & Evolution*, 21(4):677–690, 2021.

[39] Jennifer F Hoyal Cuthill, Nicholas Guttenberg, Sophie Ledger, Robyn Crowther, and Blanca Huertas. Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. *Science advances*, 5(8):eaaw4967, 2019.

[40] Toke T Høye, Johanna Ärje, Kim Bjerge, Oskar LP Hansen, Alexandros Iosifidis, Florian Leese, Hjalte MR Mann, Kristian Meissner, Claus Melvad, and Jenni Raitoharju. Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences*, 118(2): e2002545117, 2021.

[41] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1M: One million image-text pairs for histopathology. In *Advances in Neural Information Processing Systems*, pages 37995–38017, 2023.

[42] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021.

[43] ITIS. Integrated taxonomic information system (ITIS) on-

line database. www.itis.gov, 2023. Retrieved July 21, 2023.

[44] IUCN. IUCN Red List Summary Table 1a, 2022.

[45] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916, 2021.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[47] Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang. LoTE-Animal: A long time-span dataset for endangered animal behavior understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 20064–20075, 2023.

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021.

[49] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[50] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023.

[51] 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.

[52] Richard L. Pyle. Towards a global names architecture: The future of indexing scientific names. *ZooKeys*, 550:261–281, 2016.

[53] David R. Maddison and K.-S. Schultz. The Tree of Life Web Project. 2007.

[54] Duncan C. McKinley, Abe J. Miller-Rushing, Heidi L. Ballard, Rick Bonney, Hutch Brown, Susan C. Cook-Patton, Daniel M. Evans, Rebecca A. French, Julia K. Parrish, Tina B. Phillips, Sean F. Ryan, Lea A. Shanley, Jennifer L. Shirk, Kristine F. Stepenuck, Jake F. Weltzin, Andrea Wiggins, Owen D. Boyle, Russell D. Briggs, Stuart F. Chapin, David A. Hewitt, Peter W. Preuss, and Michael A. Soukup. Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, 208:15–28, 2017.

[55] George A. Miller. WordNet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[56] Chao Mou, Aokang Liang, Chunying Hu, Fanyu Meng, Baixun Han, and Fu Xu. Monitoring endangered and rare wildlife in the field: A foundation deep learning model integrating human knowledge for incremental recognition with few data and low cost. *Animals*, 13(20):3168, 2023.

[57] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of CLIP. In *Advances in Neural Information Processing Systems*, pages 21455–21469, 2022.

[58] Bradley M. Norman, Jason A. Holmberg, Zaven Arzoumanian, Samantha D. Reynolds, Rory P. Wilson, Dani Rob, Simon J. Pierce, Adrian C. Gleiss, Rafael De La Parra, Beatriz Galvan, Deni Ramirez-Macias, David Robinson, Steve Fox, Rachel Graham, David Rowat, Matthew Potenski, Marie Levine, Jennifer A. Mckinney, Eric Hoffmayer, Alistair D. M. Dove, Robert Hueter, Alessandro Ponzo, Gonzalo Araujo, Elson Aca, David David, Richard Rees, Alan Duncan, Christoph A. Rohner, Clare E. M. Prebble, Alex Hearn, David Acuna, Michael L. Berumen, Abraham Vázquez, Jonathan Green, Steffen S. Bach, Jennifer V. Schmidt, Stephen J. Beatty, and David L. Morgan. Undersea Constellations: The Global Biology of an Endangered Marine Megavertebrate Further Informed through Citizen Science. *BioScience*, 67(12):1029–1043, 2017.

[59] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12 (1):150–161, 2021.

[60] Jill Nugent. Inaturalist. *Science Scope*, 41(7):12–13, 2018.

[61] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: learning robust visual features without supervision. *arXiv preprint 2304.07193*, 2023.

[62] Jason Parham, Jonathan Crall, Charles Stewart, Tanya Berger-Wolf, and Daniel Rubenstein. Animal Population Censusing at Scale with Citizen Science and Photographic Identification. In *AAAI 2017 Spring Symposium on AISOC*, 2017.

[63] David Patterson, Dmitry Mozzherin, David Peter Shorthouse, and Anne Thessen. Challenges with using names to link digital biodiversity information. *Biodiversity Data Journal*, 4:e8080, 2016.

[64] Katelin D Pearson, Gil Nelson, Myla FJ Aronson, Pierre Bonnet, Laura Brenskelle, Charles C Davis, Ellen G Denny, Elizabeth R Ellwood, Hervé Goëau, J Mason Heberling, et al. Machine learning using digitized herbarium specimens to advance phenological research. *BioScience*, 70(7):610–620, 2020.

[65] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555: 126658, 2023.

[66] Lukáš Picek, Milan Šulc, Jiří Matas, Thomas S Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias

Frøslev. Danish fungi 2020-not just another image recognition dataset. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1525–1535, 2022.

[67] Catarina Pinho, Antigoni Kaliontzopoulou, Carlos A Ferreira, and João Gama. Identification of morphologically cryptic species with computer vision models: wall lizards (Squamata: Lacertidae: Podarcis) as a case study. *Zoological Journal of the Linnean Society*, 198(1):184–201, 2022.

[68] Gerald Piosenka. Birds 525 species - image classification, 2023.

[69] 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, 2021.

[70] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[71] Roopashree S and Anitha J. Medicinal leaf dataset, 2020.

[72] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a Caption Worth a Thousand Images? A Controlled Study for Representation Learning. *arXiv preprint arXiv:2207.07635*, 2022.

[73] Cristoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In *Proceedings of NeurIPS Data-Centric AI Workshop*, 2021.

[74] Hortense Serret, Nicolas Deguines, Yikweon Jang, Gregoire Lois, and Romain Julliard. Data quality and participant engagement in citizen science: comparing two approaches for monitoring pollinators in france and south korea. *Citizen Science: Theory and Practice*, 4(1):22, 2019.

[75] Robert Simpson, Kevin R. Page, and David De Roure. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1049–1054, Seoul, Korea, 2014. Association for Computing Machinery. Type: 10.1145/2567948.2579215.

[76] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 249–253, New York, NY, USA, 2020. Association for Computing Machinery.

[77] Robin Steenweg, Mark Hebblewhite, Roland Kays, Jorge Ahumada, Jason T Fisher, Cole Burton, Susan E Townsend, Chris Carbone, J Marcus Rowcliffe, Jesse Whittington, Jedediah Brodie, J Andrew Royle, Adam Switalski, Anthony P Clevenger, Nicole Heim, and Lindsey N Rich. Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, 15(1):26–34, 2017. ISBN: 1540-9295 Publisher: John Wiley & Sons, Ltd Type: https://doi.org/10.1002/fee.1448.

[78] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[79] Brian L. Sullivan, Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theo Damoulas, André A. Dhondt, Tom Dietterich, Andrew Farnsworth, Daniel Fink, John W. Fitzpatrick, Thomas Fredericks, Jeff Gerbracht, Carla Gomes, Wesley M. Hochachka, Marshall J. Iliff, Carl Lagoze, Frank A. La Sorte, Matthew Merrifield, Will Morris, Tina B. Phillips, Mark Reynolds, Amanda D. Rodewald, Kenneth V. Rosenberg, Nancy M. Trautmann, Andrea Wiggins, David W. Winkler, Weng-Keen Wong, Christopher L. Wood, Jun Yu, and Steve Kelling. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, 169: 31–40, 2014.

[80] Brian L Sullivan, Jocelyn L Aycrigg, Jessie H Barry, Rick E Bonney, Nicholas Bruns, Caren B Cooper, Theo Damoulas, André A Dhondt, Tom Dietterich, Andrew Farnsworth, et al. The ebird enterprise: An integrated approach to development and application of citizen science. *Biological conservation*, 169:31–40, 2014.

[81] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, 2(150026):1–14, 2015.

[82] Mélisande Teng, Amna Elmustafa, Benjamin Akera, Hugo Larochelle, and David Rolnick. Bird distribution modelling using remote sensing and citizen science data. *arXiv preprint arXiv:2305.01079*, 2023.

[83] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

[84] Ihsan Ullah, Dustin Carrion, Sergio Escalera, Isabelle M Guyon, Mike Huisman, Felix Mohr, Jan N van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. Meta-Album: multi-domain meta-dataset for few-shot image classification. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, pages 3232–3247, 2022.

[85] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (11), 2008.

[86] Grant Van Horn and Oisin Mac Aodha. iNat Challenge 2021 - FGVC8, 2021.

[87] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[88] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[89] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[90] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens Van Der Maaten. SimpleShot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

[91] Xiaoping Wu, Chi Zhan, Yukun Lai, Ming-Ming Cheng, and Jufeng Yang. IP102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8787–8796, 2019.

[92] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021.

[93] Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. CiT: Curation in training for effective vision-language data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15180–15189, 2023.

[94] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. *arXiv preprint arXiv:2309.16671*, 2023.

[95] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[96] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, 2022.

Appendices

Many details are omitted in the main text because of space concerns; we present relevant details here.

1. Appendix A: Reproducibility statement
2. Appendix B: Ethics statement
3. Appendix C: Details of training data aggregation
4. Appendix D: Training details and hyperparameters
5. Appendix E: Standard deviations for few-shot results
6. Appendix F: Example zero-shot predictions on our evaluation tasks.
7. Appendix G: Additional text-type results
8. Appendix H: Generalized zero-shot learning setting

A. Reproducibility Statement

We ensure reproducibility of our results by releasing our datasets (TREEOFLIFE-10M and RARE SPECIES), data pre-processing code, training code, evaluation code, code to generate all figures (Figs. 2 and 3), and pre-trained model weights. With these resources, anyone with sufficient compute resources can download the original data, then reproduce the pre-processing, training, and evaluation. For those with limited compute, the pre-trained model weights enable full reproducibility of our evaluation results (§4).

We provide DOIs as permanent references to our new digital assets:

- TREEOFLIFE-10M: doi:10.57967/hf/1972
- RARE SPECIES: doi:10.57967/hf/1981
- BIOCLIP: doi:10.57967/hf/1511
- Code: doi:10.5281/zenodo.10895871

B. Ethics Statement

We are not aware of any major ethical issues that arise from our work. BIOCLIP is further pre-trained from the original CLIP model; many of the same concerns (class design, surveillance, etc.) apply; however, these concerns are discussed in great detail in Radford et al. [69], so we will focus on addressing these concerns as they relate to the biological addition provided in BIOCLIP.

How could BIOCLIP affect endangered species–does BIOCLIP or TREEOFLIFE-10M pose a threat by aiding poachers? Though BIOCLIP leads to improved automatic species classification, it does not include specific geographic information such as GPS coordinates. Furthermore, animal conservation status is not included during training.

Could BIOCLIP have a negative impact on biologists? BIOCLIP is designed to combine visual cues with an established taxonomic hierarchy to aid in scientific discovery. Concerns regarding over-reliance on model predictions is a warning that accompanies many–if not all–contemporary models and is not unique to BIOCLIP. The goal is for BIOCLIP to aid biologists in their work, not to replace them. As such, it is important for users to retain that understanding/context when applying BIOCLIP to downstream tasks.

C. Training Data Aggregation

We aggregate images and labels from the iNat21 training data, BIOSCAN-1M's, and data downloaded from EOL. While every image has at least one taxonomic rank labeled, full taxonomic hierarchies and common names are scraped on a best-effort basis from metadata providers, including iNaturalist (iNaturalist Taxonomy DarwinCore Archive), Encyclopedia of Life (eol.org) and Integrated Taxonomic Information System (ITIS) (itis.gov).

We create a lookup between scientific name and taxonomic hierarchy and a lookup between scientific name and common name. We populate these lookups using the following sources in order of descending prioritization, as earlier sources are considered more authoritative. That is, if a duplicate appears in a later source, it is superseded by the higher priority source: BIOSCAN-1M metadata, EOL aggregate datasets: information retrieved using EOL page IDs with the pages API, which checks for a match in the ITIS hierarchy for higher-level taxa standardization (setting aside homonyms for proper linkage). The full list of taxa and vernacular names provided by iNaturalist and the iNat21 training set class names were maintained. From here, any taxa that could not be resolved using these sources were fed through the Global Names Resolver (GNR) API. Overall we were able to achieve 84% full taxa labeling for images in TREEOFLIFE-10M, for context, 10% of TREEOFLIFE-10M is only labeled down to the family rank (BIOSCAN-1M), thus, genus-species information is not available.

Despite our efforts, we discovered after training that some hemihomonyms were mislabeled at higher-level taxa (family up to kingdom). This impacts approximately $0.1 - 0.2\%$ of our data. We are in the process of developing a more robust solution to taxonomic labeling which will also account for re-naming (as is currently in process for many bird species). We intend to release a patch alongside the solution.

D. Hyperparameters & Training Details

Tabs. D1 and D2 contain our training hyperparameters for the different models. Tab. D2 notes the different epochs at which we had the lowest validation loss, as evaluated using the CLIP objective on the validation split of TREEOFLIFE-10M (even for the TREEOFLIFE-1M models). We will release our training code upon acceptance.

We trained a hierarchical classification model in §4.4. Python pseudocode for the training objective is in Listing 1. We will publicly release full training code upon acceptance.

Hyperparameter	Value
Architecture	ViT-B/16
Max learning rate	1×10^{-4}
Warm-up steps	1,000
Weight Decay	0.2
Input Res.	224×224

Table D1. Common hyperparameters among all models we train.

Dataset	Text Type	Batch Size	Epoch
TREEOFLIFE-10M	Mixture	32K	100
iNat21 Only	Mixture	16K	65
TREEOFLIFE-1M	Common	16K	86
	Scientific		87
	Taxonomy		87
	Sci+Com		87
	Tax+Com		86
	Mixture		91

Table D2. Hyperparameters that differ between the various models we train. We use a smaller batch size and only 1M examples for our text-type ablation because of limited compute.

```python
import torch.nn.functional as F

def forward(vit, heads, images, h_labels):
  """
  vit: vision transformer.
  heads: linear layers, one for each taxonomic
         rank.
  images: batch of input images
  h_labels: hierarchical labels; each image has
            7 labels
  """
  img_feats = vit(images)
  h_logits = [head(img_feats) for head in heads]
  losses = [F.cross_entropy(logits, label)
    for logits, labels in zip(h_logits, h_labels)]
  return sum(losses)
```

Listing 1. Python code to calculate the hierarchical multitask objective. Each image has 7 class labels: one for each taxonomic rank. The ViT calculates dense features for each image, then each taxonomic rank has its own linear layer that produces logits. By summing the losses, the ViT learns to produce image features that are useful for classifying images at multiple taxonomic ranks.

E. Standard Deviation of Main Results

Tabs. E3 and E4 show the accuracy with standard deviation over five runs on the test sets presented in Tab. 2. Since we randomly select the training examples from the datasets for few-shot, accuracies vary based on which examples are train examples and which are test examples. However, the variation is small enough that our conclusions in §4.5 still hold. Zero-shot results are deterministic and have no variation.

F. Example Predictions

Figs. F1 and F2 show BIOCLIP and CLIP zero-shot predictions on all ten evaluation tasks. We randomly pick examples from each dataset that BIOCLIP correctly labels and example that CLIP incorrect labels but BIOCLIP correctly labels. BIOCLIP performs well on a variety of tasks, including out-of-distribution images (Plankton, Medicinal Leaf) and mixes of scientific and common names (PlantVillage, PlantDoc).

G. More Results of Text-Type

We investigated the effects of text-type during training and testing in §4.3 using the RARE SPECIES task. We present zero-shot results for all text-types on all tasks using the same procedure as in §4.2, where we use whatever taxonomic+common if available, otherwise whatever text-type is available.

H. Generalized Zero-Shot Learning

Chao et al. [17] introduced *generalized zero-shot learning*, where a model must label images of unseen classes from a set of both seen and unseen labels. We pick out a set of 400 seen species from TREEOFLIFE-10M using the same methodology as we used for the RARE SPECIES task. We classify the same images from the RARE SPECIES task using this set of 800 labels (a mix of seen and unseen). CLIP and OpenCLIP achieve 23.0% and 18.2% top-1 accuracy, while BIOCLIP achieves 26.0% top-1 accuracy in this challenging GZSL setting.

Model	Birds 525	Plankton	Insects	Insects 2	Rare Species
One-Shot Classification					
CLIP	43.7 ± 0.26	25.1 ± 0.71	21.6 ± 1.05	13.7 ± 1.09	28.5 ± 0.65
OpenCLIP	53.7 ± 0.52	32.3 ± 0.63	23.2 ± 1.58	14.3 ± 0.67	29.2 ± 0.64
Supervised-IN21K	60.2 ± 1.02	22.9 ± 0.84	14.7 ± 1.38	14.4 ± 0.90	28.0 ± 0.77
DINO	40.5 ± 0.96	$\mathbf{37.0 \pm 1.39}$	23.5 ± 1.49	16.4 ± 0.78	31.0 ± 0.89
BIOCLIP	71.8 ± 0.47	30.6 ± 0.77	$\mathbf{57.4 \pm 2.4}$	$\mathbf{20.4 \pm 1.28}$	$\mathbf{44.9 \pm 0.73}$
– iNat21 Only	$\mathbf{74.8 \pm 0.89}$	29.6 ± 0.82	53.9 ± 0.97	19.7 ± 0.80	36.9 ± 1.02
Five-Shot Classification					
CLIP	73.5 ± 0.37	41.2 ± 1.01	39.9 ± 0.86	24.6 ± 0.90	46.0 ± 0.33
OpenCLIP	81.9 ± 0.25	52.5 ± 0.83	42.6 ± 0.82	25.0 ± 0.83	47.4 ± 0.34
Supervised-IN21K	83.9 ± 0.15	39.2 ± 1.66	32.0 ± 1.90	25.4 ± 2.13	47.3 ± 0.41
DINO	70.9 ± 0.34	$\mathbf{56.9 \pm 1.61}$	46.3 ± 1.37	28.6 ± 1.59	50.1 ± 0.47
BIOCLIP	90.0 ± 0.12	49.3 ± 1.14	$\mathbf{77.8 \pm 0.81}$	$\mathbf{33.6 \pm 0.74}$	$\mathbf{65.7 \pm 0.43}$
– iNat21 Only	$\mathbf{90.1 \pm 0.08}$	48.2 ± 1.24	73.7 ± 0.65	32.1 ± 1.97	55.6 ± 0.16

Table E3. Accuracy with standard deviation of five runs on animals and rare species in Tab. 4

Model	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc
One-Shot Classification					
CLIP	42.1 ± 3.40	17.2 ± 0.78	49.7 ± 2.53	70.1 ± 2.83	24.8 ± 1.61
OpenCLIP	45.1 ± 3.40	18.4 ± 1.26	53.6 ± 0.79	71.2 ± 3.58	26.8 ± 1.45
Supervised-IN21K	46.7 ± 6.30	16.9 ± 2.32	$\mathbf{62.3 \pm 2.28}$	58.6 ± 4.45	27.7 ± 2.86
DINO	30.7 ± 3.79	20.0 ± 1.53	60.0 ± 2.15	79.2 ± 2.74	23.7 ± 2.48
BIOCLIP	64.5 ± 2.15	$\mathbf{40.3 \pm 3.00}$	58.8 ± 2.83	$\mathbf{84.3 \pm 1.90}$	$\mathbf{30.7 \pm 1.75}$
– iNat21 Only	$\mathbf{67.4 \pm 4.54}$	35.5 ± 2.93	55.2 ± 1.58	75.1 ± 1.16	27.8 ± 1.31
Five-Shot Classification					
CLIP	65.2 ± 1.25	27.9 ± 2.54	71.8 ± 1.46	89.7 ± 1.45	35.2 ± 1.59
OpenCLIP	68.0 ± 0.86	30.6 ± 1.26	77.8 ± 1.28	91.3 ± 0.85	42.0 ± 1.32
Supervised-IN21K	70.9 ± 2.45	30.9 ± 2.64	$\mathbf{82.4 \pm 1.53}$	82.3 ± 3.81	44.7 ± 2.26
DINO	50.3 ± 3.20	34.1 ± 2.87	82.1 ± 1.31	94.9 ± 1.30	40.3 ± 2.32
BIOCLIP	$\mathbf{85.6 \pm 1.79}$	$\mathbf{62.3 \pm 1.82}$	80.9 ± 1.04	$\mathbf{95.9 \pm 1.07}$	$\mathbf{47.5 \pm 1.35}$
– iNat21 Only	84.7 ± 1.24	55.6 ± 2.61	77.2 ± 0.68	93.5 ± 1.13	41.0 ± 1.75

Table E4. Accuracy with standard deviation of five runs on plants and fungi in Tab. 4

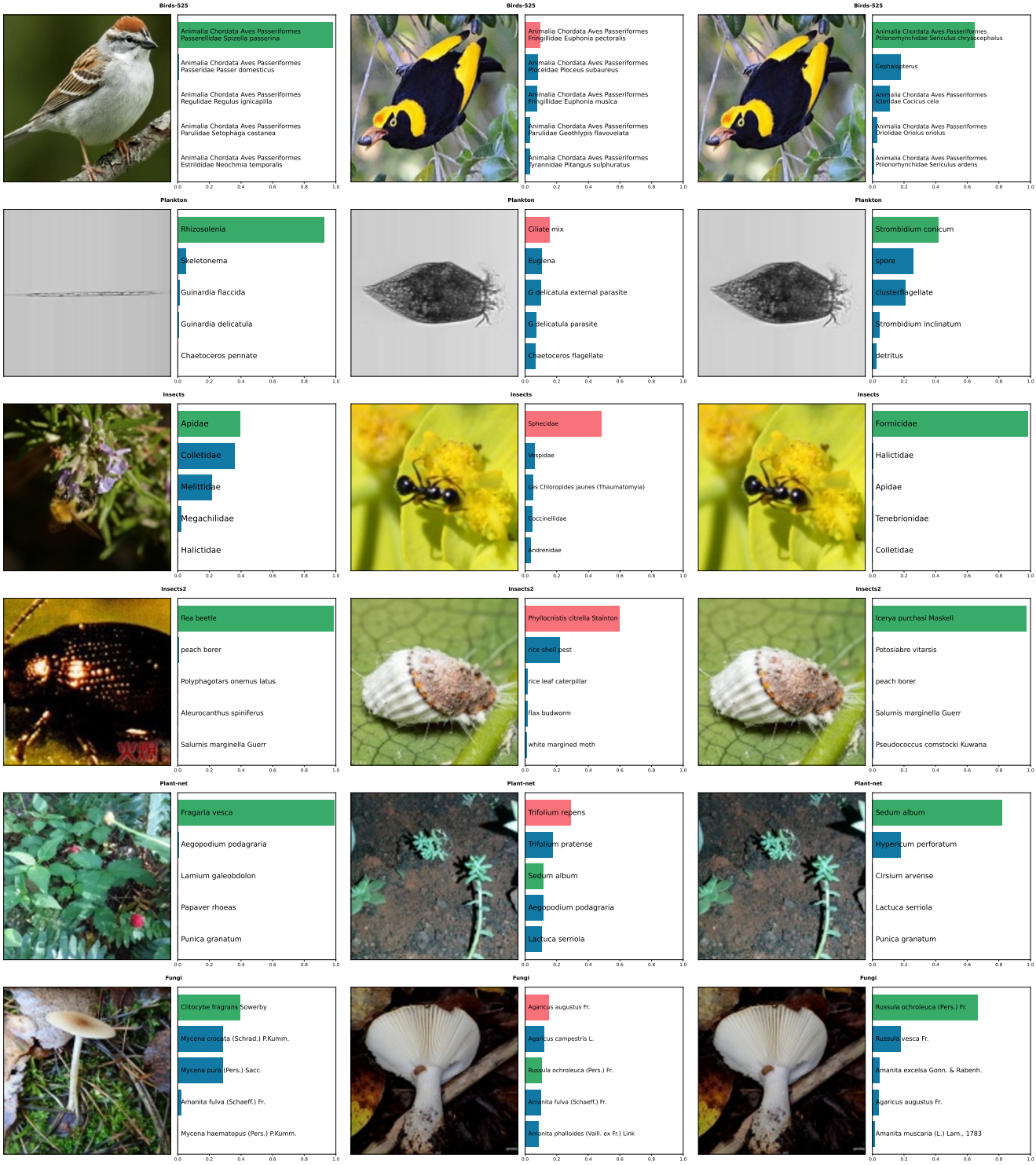

Figure F1. Example predictions for BIOCLIP and CLIP on Birds 525, Plankton, Insects, Insects2, PlantNet and Fungi tasks. Ground truth labels are green; incorrect predictions are red. Left: Correct BIOCLIP predictions. Center, Right: Images that CLIP incorrectly labels, but BIOCLIP correctly labels.

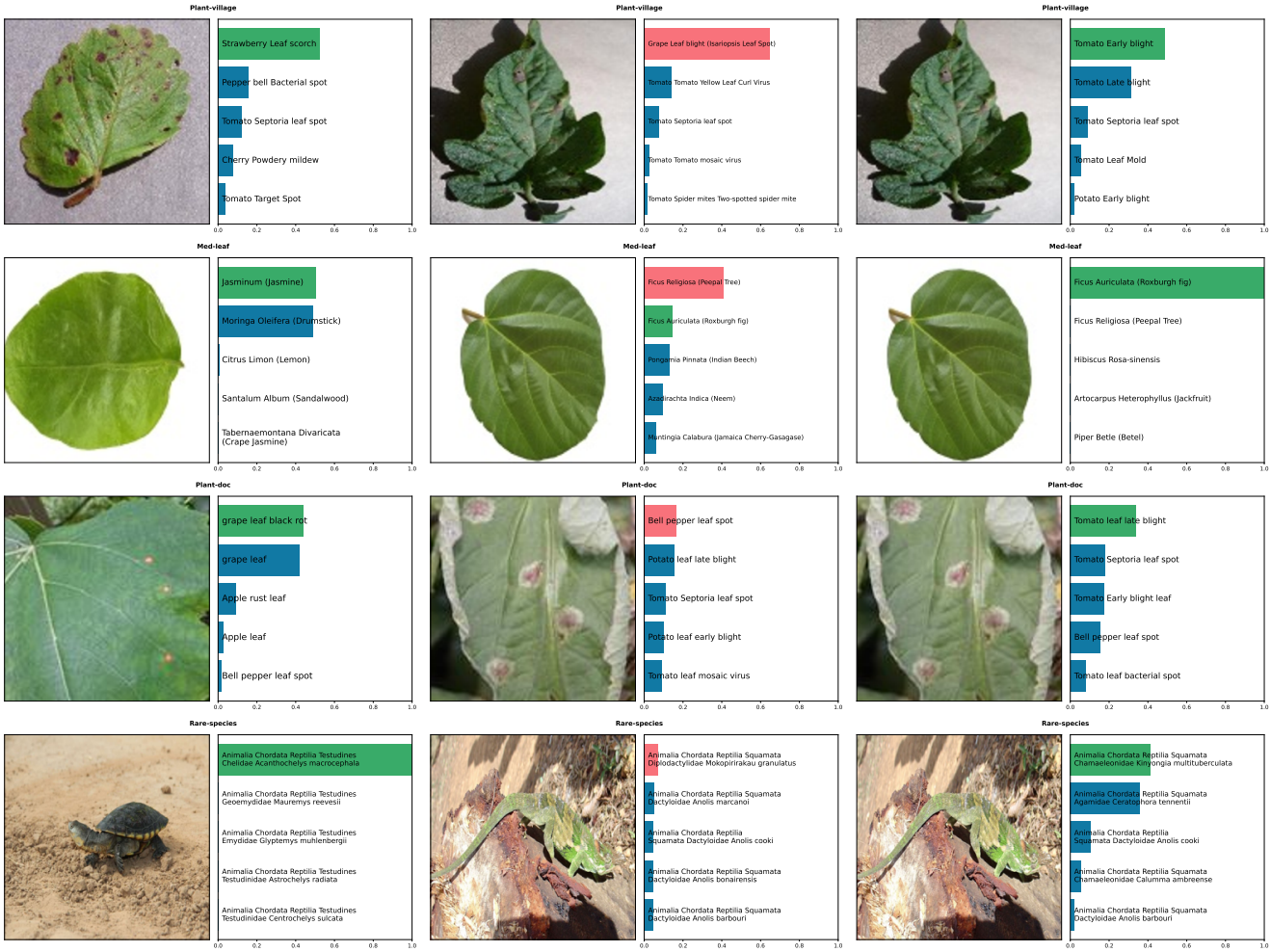

Figure F2. Example predictions for BIOCLIP and CLIP on PlantVillage, Medicinal Leaf, PlantDoc and RARE SPECIES. Ground truth labels are green; incorrect predictions are red. Left: Correct BIOCLIP predictions. Center, Right: Images that CLIP incorrectly labels, but BIOCLIP correctly labels.

	Animals				Plants & Fungi							
Training Text Type	Birds 525	Plankton	Insects	Insects 2	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc	Rare Species	Mean (Δ)	
Random Guessing	0.2	1.2	1.0	1.0	4.0	4.0	2.6	4.0	3.7	0.3	2.2	
Common	58.5	**4.4**	15.8	13.3	45.2	20.7	10.7	15.4	19.6	24.9	22.8	−10.1
Scientific	59.7	3.8	18.7	11.0	84.8	**35.3**	12.5	20.3	13.9	22.3	28.2	−4.7
Taxonomic	62.7	2.2	25.1	8.7	70.4	29.0	8.8	18.4	12.8	26.6	26.4	−6.5
Sci+Com	60.2	2.2	19.2	12.6	71.5	24.8	17.6	**21.5**	20.0	28.0	27.7	−5.2
Tax+Com	60.2	2.0	27.4	11.6	68.4	19.2	10.4	19.5	15.8	30.4	26.4	−6.5
Mixture	**65.1**	3.5	**30.6**	**17.3**	**86.3**	32.8	**19.9**	18.7	**24.5**	**30.9**	**32.9**	–

Table G5. Zero–shot classification top-1 accuracy for different text-types used during training. **Bold** indicates best accuracy. All models use the same architecture (ViT-B/16 vision encoders, 77-token text encoder) and are trained on the same dataset (TREEOFLIFE-1M). Δ denotes the difference in mean accuracy with "Mixture", which is the text-type we used for BIOCLIP.