BIOCLIP 2: Emergent Properties from Scaling Hierarchical Contrastive Learning

Jianyang Gu[1†], Samuel Stevens[1], Elizabeth G Campolongo[1], Matthew J Thompson[1], Net Zhang[1], Jiaman Wu[1], Andrei Kopanev[1], Zheda Mai[1], Alexander E. White[2], James Balhoff[3], Wasila Dahdul[4], Daniel Rubenstein[5], Hilmar Lapp[6], Tanya Berger-Wolf[1], Wei-Lun Chao[1], Yu Su[1†]
[1]The Ohio State University, [2]Smithsonian Institution, [3]UNC Chapel Hill,
[4]University of California, Irvine, [5]Princeton University, [6]Duke University

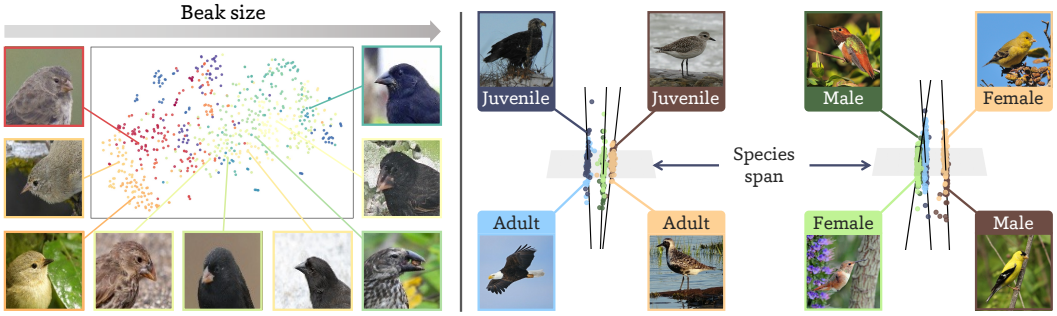


Figure 1: While BIOCLIP 2 is trained to distinguish species, it demonstrates emergent properties beyond the initial training objective. **Left**: At the *inter-species* level, the embedding distribution of different species aligns with ecological relationships; the embeddings of Darwin's finches arrange themselves by beak size from left to right. **Right**: Instead of collapsing, the *intra-species* variations are preserved in subspaces orthogonal to the inter-species variation (the black lines point from the mean embedding of one variant to that of the other variant). Orthogonality increases with scale (see Figure 2c).

Abstract

Foundation models trained at scale exhibit remarkable emergent behaviors, learning new capabilities beyond their initial training objectives. We find such emergent behaviors in biological vision models via large-scale contrastive vision-language training. To achieve this, we first curate TREEOFLIFE-200M, comprising 214 million images of living organisms, the largest and most diverse biological organism image dataset to date. We then train BIOCLIP 2 on TREEOFLIFE-200M to distinguish different species. Despite the narrow training objective, BIOCLIP 2 yields extraordinary accuracy when applied to various biological visual tasks such as habitat classification and trait prediction. We identify emergent properties in the learned embedding space of BIOCLIP 2. At the inter-species level, the embedding distribution of different species aligns closely with functional and ecological meanings (*e.g.* beak sizes and habitats). At the intra-species level, instead of being diminished, the intra-species variations (*e.g.* life stages and sexes) are preserved and better separated in subspaces orthogonal to inter-species distinctions. We provide formal proof and analyses to explain why hierarchical supervision and contrastive objectives encourage these emergent properties. Crucially, our results reveal that these properties become increasingly significant with larger-scale training data, leading to a biologically meaningful embedding space.

Models, data, and code available at imageomics.github.io/bioclip-2. [†]{gu.1220, su.809}@osu.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

1 Introduction

Recent advances in artificial intelligence (AI) are transforming core scientific workflows to become more efficient and automated [1, 2]. Tasks that once demanded overwhelming time and labor, like inferring atomic protein folds, designing functional materials, or producing global weather forecasts, can now be efficiently accomplished by large-scale predictive models [3, 4, 5, 6, 7, 8]. Particularly, a growing class of *domain-specific foundation models* demonstrate capabilities that arise without explicit definition during training [9, 10]. For example, language models trained purely on large-scale amino-acid strings unexpectedly develop an understanding of 3D chemistry to predict atomic folds with near-experimental accuracy [4, 11]. These scale-driven emergent abilities are reshaping scientific inference and opening new avenues for data-centric discovery.

In ecology and evolutionary biology, previous efforts leveraged hierarchical taxonomic labels and CLIP-style contrastive training [12] to achieve pronounced species classification accuracy across the tree of life [13, 14, 15]. This work asks a simple but intriguing question: *what properties emerge if we scale up hierarchical contrastive training?* To answer this question, we curate TreeOfLife-200M, comprising 214M organism images spanning 952K taxonomic classes, making it the largest and most diverse visual catalog of life to date. Through training at scale, our model BioCLIP 2 improves species classification accuracy by 18.1% over BioCLIP [13]. More importantly, we explore whether representations learned solely through species-level supervision can generalize to diverse biological questions **beyond species classification**.

To probe these capabilities, we evaluate BioCLIP 2 on a variety of existing biological visual tasks, including habitat classification [16], trait prediction [17, 18], new-species identification [19], and agricultural disease detection [20]. These applications push beyond simple species recognition and apply directly to biodiversity conservation, trait organization, and agricultural health. Despite being trained primarily with species-level supervision, BioCLIP 2 outperforms both vision-language (*e.g.* SigLIP [21]) and vision-only baselines (*e.g.* DINOv2 [22]) by an average margin of 10.9% on these tasks. We then look deeper into the embedding space of BioCLIP 2 and identify two *emergent properties* as the training scales up.

At the **inter-species** level, the embedding distribution of different species aligns with their ecological relationships. As shown on the left side of Figure 1, BioCLIP 2 embeddings of Darwin's finches demonstrate an increasing beak size from left to right, which is not observed in the original CLIP embedding space. We attribute the property to the adopted hierarchical taxonomic labels, which inherently encode functional and ecological information [23]. The hierarchical supervision at scale pushes related species to co-locate in functionally coherent "macro-clusters." In such a way, BioCLIP 2 acquires functional trait knowledge without using explicitly labeled traits.

At the **intra-species** level, contrary to the intuition that fine-grained differences collapse after extensive training [24, 25], BioCLIP 2 keeps the intra-species variations (*e.g.* life stages and sexes) distinct. On the right side of Figure 1, three species form tight clusters when projected onto the "species plane," while their intra-species variations fan out along axes orthogonal to the plane. Such variation cues are not encoded in taxonomic labels. We theoretically prove that when species prototypes are nearly orthogonal (*i.e.*, species are well separated), the contrastive objective prioritizes orthogonality between intra-species variations and inter-species differences over raw magnitude (Theorem D.1). Furthermore, these variations are observed to be increasingly separable as training data scales up. This microstructure preserves the intra-species representational diversity without interfering with inter-species distinctions, enabling various attribute recognition applications (§3.1).

We show in §3.2 through quantitative and qualitative analyses that larger-scale training improves both inter-species ecological alignment and separation of intra-species variants. These scale-amplified patterns make the embedding space more interpretable and biologically meaningful.

2 BioCLIP 2 and Species Classification

In addition to the significantly larger and more diverse dataset, we also scale model capacity by adopting a larger vision transformer (ViT-L/14 pre-trained on LAION-2B [12, 26, 27]). An auxiliary replay mechanism is introduced [28, 29] to maintain general-domain understanding for broader applications [30, 31]; a portion of CLIP training data (LAION-2B) is interleaved simultaneously with species contrastive learning.

Table 1: Zero-, one-, and five-shot species classification top-1 accuracy across 10 tasks for different models. **Bold** and underlined entries indicate the **best** and second best accuracies, respectively. BIOCLIP 2 outperforms both strong general- (CLIP, SigLIP, DINOv2) and domain-specific- (BIOCLIP) baselines. "Camera Trap" is mean performance across 5 camera-trap datasets; Appendix D.6 contains more details.

	Animals					Plants & Fungi					
Model	NABirds	Plankton	Insects	Insects 2	Camera Trap	PlantNet	Fungi	PlantVillage	Med. Leaf	Rare Species	Mean
Random Guessing	0.2	1.2	1.0	1.0	3.5	4.0	4.0	2.6	4.0	0.3	2.2
Zero-Shot Classification											
CLIP (ViT-L/14)	66.5	1.3	9.0	11.7	29.5	61.7	7.6	6.5	25.6	35.2	25.5
SigLIP	61.7	2.4	27.3	20.7	33.7	81.8	36.9	**28.5**	54.5	47.6	39.5
BIOCLIP	58.8	**6.1**	34.9	20.5	31.7	88.2	40.9	19.0	38.5	37.1	37.6
BIOCLIP 2	**74.9**	3.9	**55.3**	**27.7**	**53.9**	**96.8**	**83.8**	25.1	**57.8**	**76.8**	**55.6**
One-Shot Classification											
CLIP (ViT-L/14)	$42.7_{\pm 0.8}$	$28.9_{\pm 0.6}$	$29.0_{\pm 0.4}$	$17.0_{\pm 0.8}$	$36.0_{\pm 2.8}$	$58.7_{\pm 2.8}$	$20.7_{\pm 2.2}$	$56.7_{\pm 2.1}$	$74.4_{\pm 1.9}$	$34.3_{\pm 0.8}$	39.8
SigLIP	$39.9_{\pm 0.9}$	$28.4_{\pm 0.5}$	$32.3_{\pm 0.6}$	$20.6_{\pm 1.3}$	$37.8_{\pm 2.6}$	$66.3_{\pm 5.3}$	$28.7_{\pm 0.9}$	$64.1_{\pm 3.0}$	$81.7_{\pm 2.3}$	$38.8_{\pm 0.7}$	43.9
DINOv2	$55.4_{\pm 0.5}$	$29.8_{\pm 1.4}$	$8.7_{\pm 1.2}$	$22.0_{\pm 1.5}$	$47.7_{\pm 2.5}$	$73.2_{\pm 4.0}$	$32.8_{\pm 2.5}$	$61.8_{\pm 1.8}$	$\mathbf{94.4}_{\pm 1.6}$	$47.4_{\pm 0.4}$	47.3
BIOCLIP	$57.4_{\pm 1.2}$	$29.7_{\pm 1.1}$	$57.1_{\pm 1.0}$	$20.4_{\pm 0.9}$	$35.0_{\pm 2.8}$	$67.7_{\pm 3.9}$	$44.6_{\pm 2.0}$	$59.5_{\pm 2.5}$	$83.7_{\pm 1.8}$	$44.9_{\pm 0.7}$	50.0
BIOCLIP 2	$\mathbf{82.4}_{\pm 1.1}$	$32.0_{\pm 0.4}$	$\mathbf{74.6}_{\pm 0.4}$	$\mathbf{28.4}_{\pm 0.7}$	$\mathbf{48.1}_{\pm 2.2}$	$\mathbf{85.8}_{\pm 4.5}$	$\mathbf{70.3}_{\pm 2.6}$	$\mathbf{67.6}_{\pm 1.1}$	$92.0_{\pm 1.9}$	$\mathbf{59.5}_{\pm 0.9}$	**64.1**

Table 2: Biological visual tasks beyond species classification. **Bold** and underlined entries indicate the **best** and second best accuracies. See §F for task and evaluation methodology details.

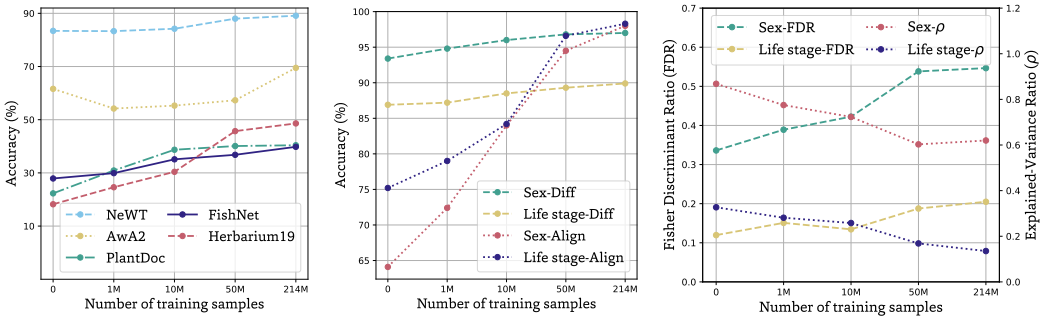
	Animals			Plants		
Model	FishNet	NeWT	AwA2	Herb. 19	PlantDoc	Mean
CLIP (ViT-L/14)	$27.9_{\pm 0.2}$	$83.4_{\pm 0.1}$	$61.6_{\pm 0.6}$	$18.2_{\pm 0.1}$	$22.3_{\pm 3.3}$	42.7
SigLIP	$31.9_{\pm 0.1}$	$83.2_{\pm 0.1}$	$67.3_{\pm 0.6}$	$18.6_{\pm 0.2}$	$28.2_{\pm 5.3}$	45.8
DINOv2	$37.4_{\pm 0.2}$	$83.7_{\pm 0.0}$	$48.6_{\pm 1.7}$	$28.1_{\pm 0.6}$	$38.6_{\pm 3.0}$	47.3
BIOCLIP	$30.1_{\pm 0.2}$	$82.7_{\pm 0.1}$	$65.9_{\pm 0.3}$	$26.8_{\pm 0.4}$	$39.5_{\pm 2.3}$	49.0
BIOCLIP 2	$\mathbf{39.8}_{\pm 0.4}$	$\mathbf{89.1}_{\pm 0.1}$	$\mathbf{69.5}_{\pm 1.1}$	$\mathbf{48.6}_{\pm 0.6}$	$\mathbf{40.4}_{\pm 3.7}$	**57.5**

Species classification performance. We evaluate BIOCLIP 2 on species classification tasks in Table 1. We use the same benchmarks as BioCLIP [13], including seven tasks from Meta-Album [32] and Rare Species [33]. We substitute Birds-525 [34] with NABirds [35] due to data inaccessibility. Additionally, we collect a test set of camera trap images from the Labeled Information Library of Alexandria: Biology and Conservation (LILA-BC) [36, 37, 38, 39, 40, 41] to illustrate more realistic species classification applications in the wild. The results suggest substantial improvements of BIOCLIP 2 over BIOCLIP. Particularly, attributed to more comprehensive species and image type coverage of TREEOFLIFE-200M, we observe over 20% zero-shot improvement on CameraTrap, Fungi, and Rare Species. On average, BIOCLIP 2 surpasses the second-best model by 18.1% and provides a 30.1% improvement over the original CLIP model. Please refer to Table 5 for full results.

3 Emergent Properties from Scaling Hierarchical Contrastive Learning

3.1 Beyond Species Classification

Biology's organization extends beyond species taxonomies; if scaling truly induces emergent behavior, model representations learned through species-level supervision should transfer to problems far removed from species classification. We collect and benchmark models on five visual benchmarks that push past species ID: habitat classification (ecological context) [16], trait prediction (evolutionary studies) [18, 17], new-species identification (biodiversity monitoring) [19], and agricultural disease detection [20]. For each task, we keep the evaluated models frozen and extract the corresponding sample embeddings. The embeddings are subsequently processed using machine-learning techniques (*e.g.* support vector classifiers). Detailed evaluation procedures are listed in §F.



(a) (b) (c)

Figure 2: **(a)** The model performance on five downstream tasks under different scales of training data. **(b)** The model performance on differentiating and aligning different life stages and sexes. **(c)** The separation and orthogonality evaluation of models trained with different amounts of data.

Table 2 presents the performance comparison among BIOCLIP 2, vision-language baselines, and vision-only models. Although no information on these tasks is explicitly described during training, BIOCLIP 2 yields an average performance improvement of 14.8% over the original CLIP baseline. DINOv2 is commonly believed to capture fine-grained visual features and is adopted for diverse visual tasks [22, 42, 43]. Nevertheless, BIOCLIP 2 yields a 10.2% performance gap over DINOv2.

3.2 Scaling Trends

Better species classification with more species-labeled data is expected, but its effect on other, non-species classification tasks is unexpected. To better understand the relationship between training data scale and non-species classification performance, we apply the same hierarchical training with varying sizes of data: 1M, 10M, 50M, and 214M samples. The smaller datasets are randomly sampled from the complete set without losing taxa representativeness. We compare the performance of the baseline CLIP ViT-L/14 and the four obtained models on five non-species classification tasks in Figure 2a. A consistent improvement is observed as the volume of training data increases from 1M to 214M. In AwA2, for example, the 1M model is worse than the baseline. However, the model gradually learns more generalized representations for different attributes across species and obtains improved performance as data scales up.

Next, we investigate how scale affects representations within species. We collect two groups of images with intra-species appearance variations: life stage variations from NeWT [18] and sex variations from NABirds [35]. We ask whether scaling hierarchical contrastive training collapses all images of one species onto a single prototype or still distinguishes juveniles from adults and males from females. We accordingly design two complementary tasks for each type of variation: (i) *alignment*, where a species classifier trained on one variant (*e.g.* juvenile images) is expected to recognize the species on the other variant (*e.g.* adult images), and (ii) *differentiation*, where the task is to tell the variants apart (*e.g.* juvenile *vs.* adult). As illustrated in Figure 2b, data scale steadily improves cross-variant species recognition. But at the same time, the model also becomes *better* at distinguishing the variants themselves.

3.3 Emergent Properties and Qualitative Analysis

Why does scaling data boost tasks that are never supervised during hierarchical contrastive training? We look deeper into BIOCLIP 2's embedding space and identify two emergent properties that generalize beyond species classification.

First, the embedding distribution of different species *aligns with their ecological and functional relationships*. Figure 3 shows t-SNE plots [44] of FishNet test set embeddings at four training scales, colored by whether the fish can live in freshwater or not. In the baseline CLIP plot (left), freshwater and non-freshwater fish have a large portion of overlap. Larger training sets progressively separate the two groups in the embedding space. We note that there is no explicit constraint to arrange meaningful distribution across species in contrastive loss, highlighting the emergence at scale.

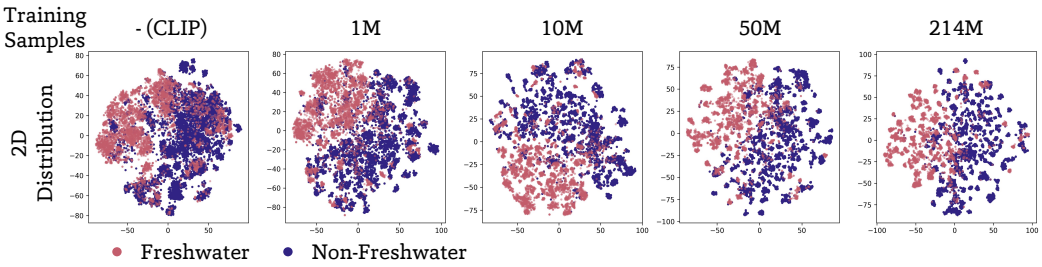


Figure 3: t-SNE embedding visualization of FishNet test set for models trained with different amounts of data. The leftmost plot is the original LAION-2B CLIP ViT-L/14. As the training data scales, freshwater fish become more distinct from saltwater fish and brackish fish, despite no explicit supervision, demonstrating that data scale contributes to emergent properties in model representations.

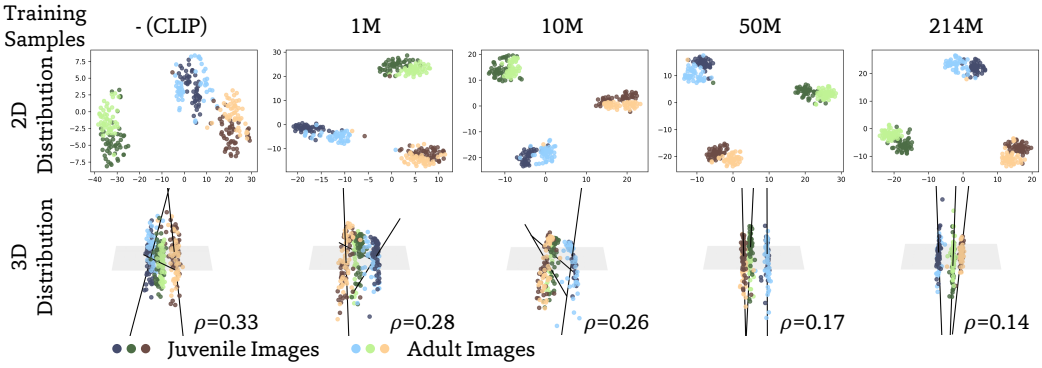


Figure 4: The embedding distribution of life stage variations under different scales of training data. The 2D distributions are obtained using t-SNE. For the 3D distributions, we first run SVD with the mean embedding of each species. The first two singular vectors are used to construct the gray plane that captures most inter-species differences. The embeddings are then projected into the 3D space with an additional orthogonal dimension. The straight lines point from the mean embedding of juvenile images to that of the adult images. As the training scales up, the intra-species variations are preserved in the subspace orthogonal to the inter-species differences. Orthogonality improves with data scale, as evidenced by the decreasing explained-variance ratio ρ.

Second, the intra-species variations are *preserved and separated*. Figure 4 shows t-SNE plots of BIOCLIP 2 embeddings of three species from NeWT exhibiting life-stage variation. In the 2D plots (top), different species (shown in different colors) tighten progressively from left to right, which is a direct consequence of scaled contrastive training. At the same time, the intra-species variations are preserved and better separated than the baseline CLIP model (leftmost sub-figure). We further project the embeddings onto 3D spaces created by singular value decomposition (SVD, bottom), which reveals that the intra-species variations lie in subspaces roughly orthogonal to the species span. Therefore, the existence of intra-species variation does not interfere with inter-species distinctions. We provide formal analyses of emergent properties and more empirical observations in §D.

4 Conclusion

In this work, we curate TREEOFLIFE-200M, the largest and most diverse biological organism dataset to date, and train BIOCLIP 2 with hierarchical taxonomic labels. BIOCLIP 2 achieves state-of-the-art accuracy on species classification. More importantly, large-scale training gives rise to two emergent properties not described during training. At the inter-species level, the embedding distribution of different species aligns with their ecological relationships. At the intra-species level, the appearance variations within species are preserved and well separated in the embedding space. We demonstrate that combining the effort of domain-specific scaling and structured supervision leads to effective generalization beyond the initial training objectives. BIOCLIP 2 serves as a strong foundation model for biological research and simultaneously evidences the effectiveness of scale-driven scientific discovery.

References

[1] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023. 2

[2] Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4), 2021. 2

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021. 2

[4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. 2

[5] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. 2

[6] Huan Tran, Rishi Gurnani, Chiho Kim, Ghanshyam Pilania, Ha-Kyung Kwon, Ryan P Lively, and Rampi Ramprasad. Design of functional and sustainable polymers assisted by artificial intelligence. *Nature Reviews Materials*, pages 1–21, 2024. 2

[7] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. 2, 15

[8] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023. 2

[9] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *TMLR*, pages 1–30, 2022. 2, 11

[10] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. 2

[11] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, 2021. 2

[12] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2, 12, 26

[13] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *CVPR*, pages 19412–19424, 2024. 2, 3, 11, 16, 22, 23

[14] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. In *NeurIPS*, volume 37, pages 102101–102120, 2024. 2, 11, 12, 13, 26

[15] José M Padial, Aurélien Miralles, Ignacio De la Riva, and Miguel Vences. The integrative future of taxonomy. *Frontiers in zoology*, 7:1–14, 2010. 2, 18

[16] Faizan Farooq Khan, Xiang Li, Andrew J Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *ICCV*, pages 20496–20506, 2023. 2, 3, 11, 25

[17] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 41(9):2251–2265, 2018. 2, 3, 20, 25

[18] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *CVPR*, pages 12884–12893, 2021. 2, 3, 4, 11, 17, 25

[19] Kiat Chuan Tan and Yulong Liu. Herbarium challenge 2019 - fgvc6. https://kaggle.com/competitions/herbarium-2019-fgvc6, 2019. Kaggle. 2, 3, 20, 25

[20] 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, 2020. 2, 3, 11, 20, 25

[21] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 2, 26

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, pages 1–32, 2024. 2, 4, 26

[23] Sangeet Lamichhaney, Jonas Berglund, Markus Sällman Almén, Khurram Maqbool, Manfred Grabherr, Alvaro Martinez-Barrio, Marta Promerová, Carl-Johan Rubin, Chao Wang, Neda Zamani, et al. Evolution of darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(7539):371–375, 2015. 2, 20

[24] Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In *NeurIPS*, volume 35, pages 31697–31710, 2022. 2, 23

[25] Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. 2, 23, 24

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, pages 1–21, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. 2, 26

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, pages 1–17, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY. 2, 26

[28] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 2

[29] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *NeurIPS*, volume 32, 2019. 2

[30] Julian D Santamaria, Claudia Isaza, and Jhony H Giraldo. Catalog: A camera trap language-guided contrastive learning model. In *WACV*, pages 1197–1206. IEEE, 2025. 2, 11

[31] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. In *NeurIPS*, volume 37, pages 126500–126514, 2024. 2, 20

[32] 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 *NeurIPS*, 2022. URL https://meta-album.github.io/. 3

[33] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Rare species, 2023. URL https://huggingface.co/datasets/imageomics/rare-species. 3, 17, 23

[34] Gerald Piosenka. Birds 525 species - image classification, 05 2023. URL https://www.kaggle.com/datasets/gpiosenka/100-bird-species. 3

[35] 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 *CVPR*, pages 595–604, 2015. 3, 4, 25

[36] Labeled information library of alexandria: Biology and conservation (LILA-BC). https://lila.science/datasets. 3, 22

[37] Island Conservation. Island conservation camera traps. https://lila.science/datasets/island-conservation-camera-traps/. 3, 23

[38] Desert Lion Conservation. Desert lion conservation camera traps. https://lila.science/datasets/desert-lion-conservation-camera-traps/, July 2024. 3, 23

[39] Juliana Vélez, Paula J Castiblanco-Camacho, Michael A Tabak, Carl Chalmers, Paul Fergus, and John Fieberg. Choosing an appropriate platform and workflow for processing camera trap data using artificial intelligence. *arXiv preprint arXiv:2202.02283*, 2022. 3, 23

[40] S. Balasubramaniam. Optimized classification in camera trap images: An approach with smart camera traps, machine learning, and human inference. Master's thesis, The Ohio State University, 2024. URL http://rave.ohiolink.edu/etdc/view?acc_num=osu1721417695430687. 3, 23

[41] Hayder Yousif, Roland Kays, and Zhihai He. Dynamic programming selection of object proposals for sequence-level animal species classification in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 3, 23

[42] Guangxing Han and Ser-Nam Lim. Few-shot object detection with foundation models. In *CVPR*, pages 28608–28618, 2024. 4

[43] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *WACV*, pages 1319–1329. IEEE, 2025. 4

[44] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html. 4

[45] Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. Emergent abilities in large language models: A survey. *arXiv preprint arXiv:2503.05788*, 2025. 11

[46] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 11

[47] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18134–18144, 2022. 11

[48] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations. In *ECCV*, pages 220–238. Springer, 2024. 11

[49] Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Deciphering the role of representation disentanglement: Investigating compositional generalization in clip models. In *ECCV*, pages 35–50. Springer, 2024. 11

[50] Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *IJCV*, 132(9):3770–3786, 2024. 11

[51] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *WACV*, pages 1765–1774. IEEE, 2025. 11

[52] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. TreeOfLife-10M, 2023. URL https://huggingface.co/datasets/imageomics/TreeOfLife-10M. 12

[53] Grant Van Horn and Oisin Mac Aodha. inat challenge 2021 - fgvc8, 2021. URL https://kaggle.com/competitions/inaturalist-2021. 12, 17

[54] Z. Gharaee, Z. Gong, N. Pellegrino, I. Zarubiieva, J. B. Haurum, S. C. Lowe, J. T. A. McKeown, C. Y. Ho, J. McLeod, Y. C. Wei, J. Agda, S. Ratnasingham, D. Steinke, A. X. Chang, G. W. Taylor, and P. Fieguth. A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS) Datasets & Benchmarks Track*, volume 36, pages 43593–43619. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/87dbbdc3a685a97ad28489a1d57c45c1-Paper-Datasets_and_Benchmarks.pdf. 12

[55] GBIF.org. GBIF occurrence download, May 2024. URL https://doi.org/10.15468/dl.bfv433. 12, 14

[56] Encyclopedia of Life (EOL). URL https://eol.org. Accessed August 2024. 12

[57] Zahra Gharaee, Scott C. Lowe, ZeMing Gong, Pablo Millan Arias, Nicholas Pellegrino, Austin T. Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Lila Kari, Dirk Steinke, Graham W. Taylor, Paul Fieguth, and Angel X. Chang. BIOSCAN-5M: A multimodal dataset for insect biodiversity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *NeurIPS*, volume 37, pages 36285–36313. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3fdbb472813041c9ecef04c20c2b1e5a-Paper-Datasets_and_Benchmarks_Track.pdf. 12

[58] Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, and Katherine L. C. Bell. FathomNet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1):15914, September 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-19939-2. URL https://www.nature.com/articles/s41598-022-19939-2. 12

[59] iNaturalist contributors and iNaturalist. iNaturalist Research-grade Observations. Accessed via GBIF.org (Occurrence Snapshot: 10.15468/DL.BFV433) 2024-05-01. 12, 13, 14

[60] Smithsonian Institution. URL https://www.si.edu/. Accessed May 15, 2025. 12

[61] Flickr. URL https://www.flickr.com/. Accessed May 15, 2025. 12

[62] Dmitry Mozzherin. GNverifier – a reconciler and resolver of scientific names against more than 100 data sources, November 2024. URL https://github.com/gnames/gnverifier. 12

[63] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review, 2019. 12, 15

[64] Andres Hernandez, Zhongqi Miao, Luisa Vargas, Sara Beery, Rahul Dodhia, and Juan Lavista. Pytorch-wildlife: A collaborative deep learning framework for conservation, 2024. 12, 15

[65] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 12, 17

[66] Facebook, Inc. Threatexchance. https://github.com/facebook/ThreatExchange, 2019. 13, 17

[67] IUCN. IUCN Red List Summary Table 1a, March 2025. URL https://nc.iucnredlist.org/redlist/content/attachment_files/2025-1_RL_Table_1a.pdf. 13, 17, 27

[68] IUCN. The IUCN Red List of Threatened Species. version 2025-1. https://www.iucnredlist.org/search?dl=true&permalink=fd473452-551c-4561-a2b2-fb51fe51d7b7, 2025. Accessed on 14 May 2025. 13, 17

[69] GBIF Secretariat. GBIF Backbone Taxonomy, 2023. URL https://doi.org/10.15468/39omei. 14

[70] J. Doubt, R. Blades, and A. Tyrell. Canadian Museum of Nature Herbarium, 2025. Version 1.283. Canadian Museum of Nature. Occurrence dataset https://doi.org/10.15468/kowta4 accessed via GBIF.org on 2025-05-22. 15, 16

[71] MNHN and S. Chagnoux. The fossil collection (F) of the Muséum national d'Histoire naturelle (MNHN - Paris), 2025. Version 68.409. MNHN - Museum national d'Histoire naturelle. Occurrence dataset https://doi.org/10.15468/fjjmkb accessed via GBIF.org on 2025-05-14. https://www.gbif.org/occurrence/417149247. Collected in France (licensed under http://creativecommons.org/licenses/by/4.0/). 16

[72] Natural History Museum and Benjamin Drew. Natural History Museum (London) Collection Specimens, 2025. Occurrence dataset https://doi.org/10.5519/qd.x8gq8f5t accessed via GBIF.org on 2025-05-14. https://www.gbif.org/occurrence/1056329684. Collected in Suriname by The Trustees of the Natural History Museum, London (licensed under http://creativecommons.org/licenses/by/4.0/). 17

[73] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 19

[74] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4 edition, 2013. ISBN 978-1-4214-0794-4. 19

[75] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2): 179–188, 1936. 19

[76] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 22

[77] Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, pages 1–43, 2022. 23

[78] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 23, 24

[79] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, volume 33, pages 18661–18673, 2020. 23, 24

[80] Jaidev Gill, Vala Vakilian, and Christos Thrampoulidis. Engineering the neural collapse geometry of supervised-contrastive loss. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7115–7119. IEEE, 2024. 23

[81] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *NeurIPS*, volume 34, pages 29820–29834, 2021. 23

[82] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, pages 7492–7501, 2022. 25

[83] 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. 25

[84] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, pages 1–33, 2023. 26

[85] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 26

[86] 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. 27

[87] David Lindenmayer and Ben Scheele. Do not publish. *Science*, 356(6340):800–801, 2017. doi: 10.1126/science.aan1362. 27

Appendix

The remaining content of this document is organized as follows:

- In §A we present research related to this work.
- In §B we provide the overview of TreeOfLife-200M.
- In §C we provide the details of data processing for TreeOfLife-200M.
- In §D we demonstrate more empirical observations of BioCLIP 2.Theorem D.1.
- In §E we discuss the relationship between the adopted training scheme and neural collapse.
- In §F we describe both the domain-specific benchmarks and the implementation details.
- In §G we describe the training implementation for BioCLIP 2 in detail.
- In §H we introduce the baselines employed for the numerical experiments.
- In §I we discuss limitations of our work.
- In §J we discuss the broader impacts of our work.

A Related Work

Emergent properties in foundation models. Emergent properties refer to the capabilities implicitly acquired from the training process and generalized beyond the initial training objective. Large language models (LLMs) illustrate a variety of in-context learning skills after next-token-prediction pre-training [9, 45]. Such emergence also arises in computer vision. DINO learns semantic segmentation through purely visual self-supervision [46], whereas GroupViT learns semantic segmentation solely through text supervision [47]. The studies closest to this work are [48, 49]. Alper and Averbuch-Elor explore the visual-semantic hierarchies in CLIP models [48]. Although hierarchical semantic structures are never explicitly presented as the training supervision, CLIP models acquire the capability of matching images with varying levels of descriptions. Abbasi *et al*. discover that CLIP models possess disentangled representational sub-spaces for different factors of variations [49]. It allows CLIP models to generalize across compositional out-of-distribution concepts. This work leverages CLIP to distinguish different species with hierarchical labels, which is a different scenario from the above work. We investigate the emergent properties under this setting.

Computer vision for ecology & evolutionary biology. Ecology and evolutionary biology is a naturally challenging regime for computer vision systems due to long-tail distributions, extremely fine-grained classifications, and a wide variety of image distributions. Existing work formalizes these challenges into specific visual tasks such as attribute prediction [NeWT, 18], trait prediction [FishNet, 16], and plant disease detection [PlantDoc, 20].

Recent advancements in computer vision have led to the development of foundation models for biological applications. BioCLIP incorporates taxonomic labels in the vision-language contrastive training, yielding promising species classification accuracy [13]. Follow-up work scaled data to 162M images [BioTrove, 14], specialized the data to camera traps [CATALOG and WildCLIP, 50, 30], and added additional model modalities [TaxaBind, 51]. We investigate both data and model scaling, with a focus on both broad biological applications and any emergent properties after extensive training.

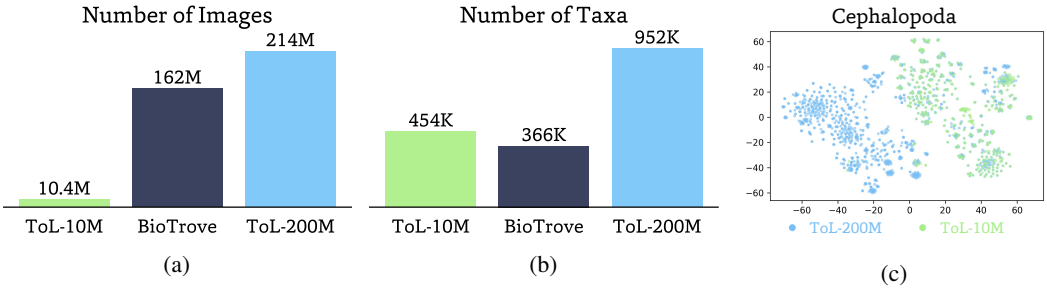


Figure 5: **(a)** Number of images across organismal biology datasets. **(b)** Biodiversity comparison across datasets (measured unique 7-tuples for TreeOfLife (ToL) datasets, species count provided by BioTrove). **(c)** The taxa distributional difference in the *Cephalopoda* class (octopuses, squids, *etc*.) between ToL-200M and ToL-10M.

B TreeOfLife-200M

Large-scale, clean, diverse data drives progress in machine learning. There have been efforts such as TreeOfLife-10M [52] and BioTrove [14] to create large-scale biological organismal datasets for machine learning (ML). As shown in Figure 5, TreeOfLife-10M [52] improves on prior work such as iNat21 [53] and BIOSCAN-1M [54] by increasing taxa diversity by a factor of 45. BioTrove [14] increases data scale to 162M but fails to match the biodiversity of TreeOfLife-10M. In this work, we combine the vast breadth of Global Biodiversity Information Facility (GBIF) [55] images with those of the Encyclopedia of Life project (EOL) [56], BIOSCAN-5M [57], and FathomNet Database [58]. With nearly 214 million images representing 952,257 taxa, TreeOfLife-200M is the largest *and* most diverse public ML-ready dataset for computer vision models in biology.

Unlike BioTrove [14], which relies solely on iNaturalist and contains 162M images but only 366K unique species, our use of museum, camera-trap, and citizen-science contributions expands the taxonomic coverage to 2.6× more taxa. Our curation efforts ensure this breadth does not come at the cost of data quality. We quantify image type diversity from GBIF: 51.8M museum specimen, 617.8K camera trap, and 151M citizen science images. Beyond the taxa-wise diversity, these images also provide more observing perspectives for the focal species. In such a way, the robustness of models trained on TreeOfLife-200M is significantly enhanced against a variety of use cases. Specifically, we demonstrate that BioCLIP 2 yields a 22.8% performance gap compared with BioCLIP on camera trap images (See Table 5).

B.1 Images

TreeOfLife-200M consists of images curated from four core data providers: GBIF, EOL, BIOSCAN-5M, and FathomNet. GBIF and EOL aggregate biological data from various sources, such as iNaturalist [59], the Smithsonian Institution [60], and Flickr [61]. BIOSCAN-5M and FathomNet are curated collections of expert-annotated images designed to improve cataloging and identification of species from highly diverse and under-represented branches of the tree of life. BIOSCAN-5M is part of the ongoing DNA Barcoding project to improve insect identification (one of the most diverse classes, *Insecta*). FathomNet focuses on a habitat rather than a clade: all animals that live in the ocean. Together, these comprise TreeOfLife-200M.

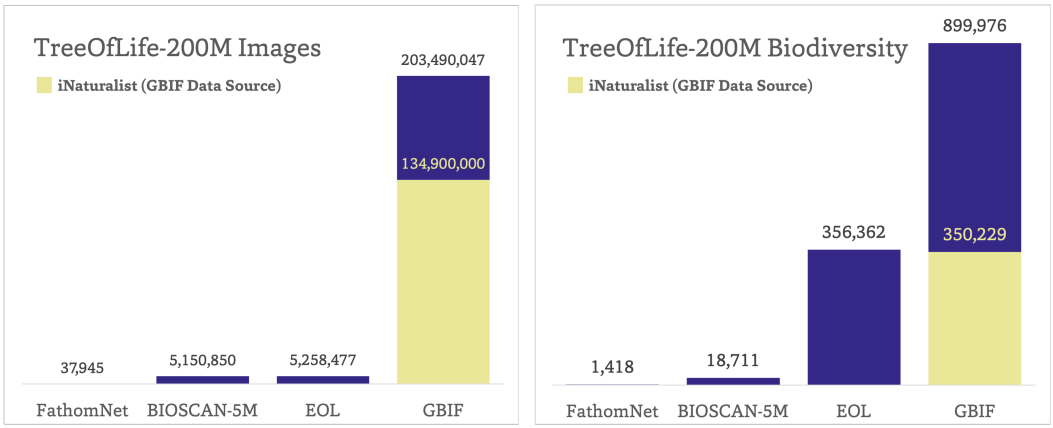
B.2 Data Curation & Filtering

Once data from GBIF, EOL, BIOSCAN-5M, and FathomNet were retrieved, we started with 222,065,140 images and 1,359,405 unique taxonomic hierarchies. We then cleaned the data, focusing on (1) aligning taxonomic labels, (2) image quality, and (3) eliminating duplication and data leakage.

Taxonomic alignment. Curating large biological image collections from distributed data providers requires the alignment of noisy and inconsistent taxonomic labels both between and within providers. We develop a taxonomic alignment package `TaxonoPy` in consultation with taxonomists that resolves entries to both a seven-rank Linnaean hierarchy and a common name. `TaxonoPy` queries GNVerifier [62] against the GBIF Backbone, Catalogue of Life, and OpenTree hierarchies (in that order). From the original 1.36M taxa, `TaxonoPy` filters 407K taxonomic hierarchies (such as synonyms, provisional names, etc.) and yields 952K unique taxa.[1]

Image-quality screens. Digital archives contain herbarium labels, empty camera-trap frames, and occasional people. None add biological signal, and faces raise privacy concerns, so we drop them to keep learning focused on organisms via three neural-network-based filters: (i) **Museum non-organism removal.** A pre-trained CLIP-L/14 [12] is used for a nearest-centroid classifier spanning 25 fine-grained subtypes (10 collection areas, each split into categories such as specimen, fossil, drawer-label, etc.). The classifier is fit to 8.5K manually-curated examples and predicts a subtype for all museum images; we drop all non-organismal images. (ii) **Camera-trap trimming.** We apply a pre-trained camera-trap model MegaDetector [63, 64] to filter for frames with visible animals. (iii) **Face removal.** We apply a pre-trained face-detection model MTCNN [65] to discard images containing human faces.

[1] See the anonymized repository for more details and `TaxonoPy` code.



(a) (b)

Figure 6: The number of images (a) and unique 7-rank taxa (b) by core data provider in TreeOfLife-200M. iNaturalist [59] is the largest GBIF data source (by image count) and the data provider used by BioTrove [14], so it is included here for reference.

Duplicate and leakage control. Exact duplicates in the training set are removed with MD5 hashes. GBIF includes images from iNaturalist, a popular source for computer vision ecology benchmarks. To prevent train-test leakage and inflated downstream scores, we compute both MD5 and perceptual PDQ [66] hashes for every test image and purge any near or exact duplicates from training.

B.3 Taxa Coverage

The International Union for Conservation of Nature (IUCN) estimates 2.14M species have been described [67]. Following curation, there are nearly 868K unique taxa labeled to the level of species in TreeOfLife-200M.[2] Based on the most recent IUCN Red List assessment [68], TreeOfLife-200M demonstrates a particularly strong representation of threatened species, with 77.1% coverage (36,370 species). This coverage establishes that the approach to integrating diverse data sources used in TreeOfLife-200M is a valuable resource for conservation research, providing representation for a substantial majority of species prioritized for global conservation action.

Notably, TreeOfLife-200M adds diverse clades that are extremely under-represented in prior work like TreeOfLife-10M and BioTrove. Figure 5c compares the distribution of taxonomic names in the class of *Cephalopoda* between TreeOfLife-10M and TreeOfLife-200M. While they share overlaps, there are clades almost completely absent in TreeOfLife-10M. These under-represented clades receive a substantial influx of samples in TreeOfLife-200M (1,102 new taxa). Another example is 55,085 taxa of *Fungi* in TreeOfLife-200M, close to $4\times$ of that in TreeOfLife-10M (14,793). The improved diversity facilitates accurate species classification of these clades, as evidenced in Table 5 (42.9% absolute improvement over BioCLIP on zero-shot Fungi benchmark).

C Data Processing Details

In this section, we provide more details on the data curation process.

C.1 Taxonomic Standardization

When GNVerifier returns a result from a query, our taxonomic alignment package combines it with the input taxonomic hierarchy along with the query parameters to form a resolution attempt. The query response, based on the most specific taxonomic term available in the input data, along with the remaining input hierarchy (entry's resolution attempt) is then matched against pre-defined profiles. The algorithm uses these three components to fit against the series of profiles to determine whether a

[2]Not all images contain full 7-rank Linnaean taxa; for instance, 93% of BIOSCAN-5M images are not labeled to the species level. Thus, the unique taxa with non-null species is a more appropriate comparison.

Table 3: Top taxa contributors in GBIF [55]. Here, "Distinct Taxa" is used to describe 7-rank taxonomic labels only provided to GBIF by that publisher, while "Total Taxa" is the total number of unique taxa labels within that publisher. Observe that, following iNaturalist [59], the next three most diverse sources are museums which, together, account for more distinct taxa than iNaturalist alone, despite having significantly fewer images overall.

Publisher (GBIF Data Source)	Distinct Taxa	Total Taxa	Images	% GBIF Taxa
iNaturalist.org	81,671	350,229	134,877,019	9.07%
Natural History Museum	38,535	148,637	3,603,256	4.28%
Museum national d'Histoire nat.	37,201	258,766	6,185,477	4.13%
Naturalis Biodiversity Center	25,421	244,697	5,282,428	2.82%

confident resolution is found or if an alternative query strategy is needed (such as using a different query term or data source), iterating until a match is made or alternative approaches are exhausted. If all attempts ultimately fail to match a profile, the original input data is used for the entry's label in the final output. After hierarchical resolution, common names are annotated. For each entry, the common name corresponding to its most specific resolved taxonomic term is selected, when available, using the preferred English vernacular names from the GBIF Backbone Taxonomy [69]. In the output, the proportion of records changed for each data source is EOL: 98.7%, FathomNet: 16.0%, BIOSCAN: 11.8%, and GBIF: 0.3%. The notably low modification rate for GBIF reflects our taxonomic alignment package's preference for the GBIF Backbone Taxonomy as its primary reference. After standardization, we summarize the image number and taxa number from each source in Figure 6. This part is included in the anonymized repository as `TaxonoPy`.

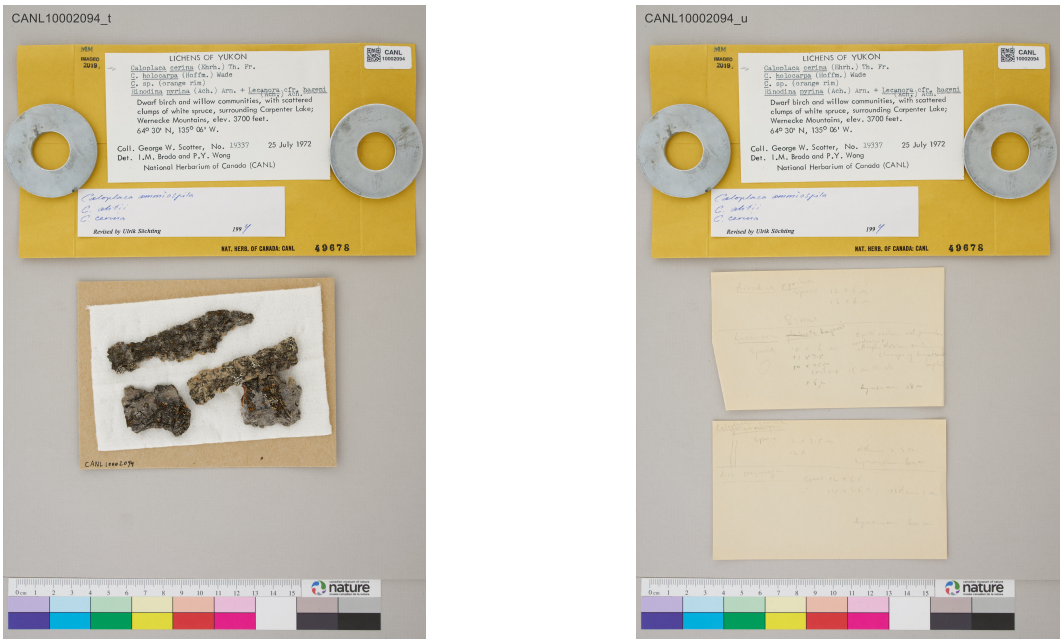
C.2 Image-quality Screens

At download, all images were checked to be sufficiently large (224 pixels or more on the shortest side), and resized, where needed, so that the largest side does not exceed 720 pixels. The code, support set class embeddings, and further details of the image-quality screens and the subsequent duplicate control are provided in our dataset repository, `TreeOfLife-toolbox` (included in the anonymized repository).

C.2.1 Museum Specimen Processing

Museums often consider multiple specimens of the same species to be connected instead of differentiating them (as they are often collected from the same or similar location). Thus, these occurrences often include images of their metadata, duplicates of the same or different specimens under one occurrence ID, or less informative images. Some common complications to consider when working with museum specimen images that influenced our processing are that

1. Plant and fungi specimens may be stored in envelopes or folders. These are often photographed and digitized under the same occurrence ID as the image of the specimen itself, thus creating extraneous images we do not want to include in training. They also sometimes pair this with living specimen images (which would be worth keeping as well). See Figure 7, which demonstrates variety within a single occurrence for fungi specimens. Plant specimens have a secondary confounding factor in that they are often pressed for preservation with their metadata on the page (see Figure 8).
2. Fish, worms, and similar will be stored in jars. A single jar is considered an occurrence, so only one specimen may be photographed—perhaps at multiple angles—the jar may be photographed, multiple specimens photographed, etc. In any of these cases, the images will all be labeled with the same metadata that does not include this context.
3. Fossilized specimens are often digitized with a photo of the fossil (appearing as a "rock", lacking a clear view of the specimen) and a close-up of the specimen itself, e.g., in the case of a bug in amber (see Figure 9).
4. For animal specimens, both of the above cases may occur: there may be simply a close-up of the tag (similar issue to envelopes with plants and fungi). There may also be multiple specimens photographed within the same occurrence (as noted with 2 about fish and worms). Museums often consider multiple specimens of the same species to be connected instead of differentiating



(a) (b)

Figure 7: These images all belong to GBIF Occurrence 2301948912 [70], and thus share metadata. (a) Contains the specimen (*Caloplaca cerina*), (b) contains the file metadata from the specimen folder; only (a) should be retained. Collected in Canada by ©Canadian Museum of Nature (licensed under CC BY-NC 4.0).

them. This is two-fold, in that they are generally, collected from the same or similar location at or about the same time, and a specimen is kept as a representative of its species, so there is not a clear need for distinguishing between them. We also see multiple views of the same specimen within an occurrence. Examples in Figure 10.

In order to appropriately separate these images, we treated them as museums do, specifically by first dividing them into 11 collection areas (Fungi, Insect, Invertebrate Zoology, Microbiology, Plant, Uncategorized, and five classes of Vertebrate Zoology: Amphibians, Birds, Fishes, Mammals, Others) inspired by the Smithsonian Institution's categorical subdivisions for their biological museum collections. From here, we further divided each category based on its image type (e.g., fossil or preserved specimen, as specified in GBIF metadata).

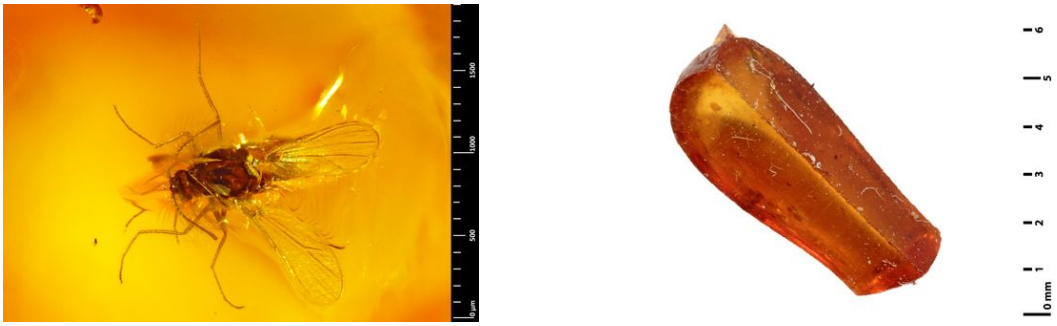
For each museum specimen category, we manually curated a small "support set" of representative specimen and non-specimen images. We embedded these examples using CLIP (ViT-L/14@336px) [7]; we chose not to use BioCLIP since museum specimen labels were not filtered from its training data (e.g., EOL contains many museum specimen images). To capture intra-class diversity, we ran K-Means on each support set and retained the resulting cluster centers. During classification, we L2-normalized both the input image's embedding and each center, then assigned the image to the nearest center in Euclidean space. This processing was applied to all museum specimen images identified within GBIF. The support set embeddings are included in `TreeOfLife-toolbox`.

C.2.2 Camera Trap Images

Some occurrences include large-volume camera-trap sequences, with up to 10,000 images. These occurrences have a single taxonomic label applied across all images, though there are different taxa (*e.g.*, the first few frames may have a duck, while later images have a swan, another a goose, but the label for all is a duck). To reduce the risk of introducing such noise while still capturing relevant biodiversity, we filter the dataset to include only occurrences with 15 images or fewer. We then use MegaDetector [63, 64] to filter "empty" frames. In creating a dataset to train a foundation model on the entire tree of life, it is prudent to avoid introducing too many plant images labeled as animals.

(a) (b)

Figure 8: Both images contain specimens included in the dataset, but their dominance in-frame is quite different. The pressed flower sheets are the same size, but (a) *Osmorhiza longistylis* is dominant in the frame, while (b) *Chara globularis* covers less pixel area than the text. These images belong to GBIF Occurrence 5132787320 and GBIF Occurrence 5135831095 [70], respectively, and were both collected in United States of America by The New York Botanical Garden (licensed under CC BY 4.0).



(a) (b)

Figure 9: These images belong to GBIF Occurrence 417149247 [71], and thus share metadata labeling them as *Corynoneura eocenica*. However, the specimen is not visible through the amber in (b). Additionally, this particular occurrence has many more distinct specimen and amber images.

C.2.3 Citizen Science Images

Similarly, citizen science occurrences may contain many images assigned a single label. For each occurrence, we thus embedded the images using BioCLIP [13], standardized them, calculated the pair-wise cosine similarity, and used the mean to indicate the occurrence's "distinctness". In these instances, we identified three broad categories into which to further subdivide them for reduction:

1. Mixed occurrences: multiple species all labeled as just one of them (low similarity). These were treated as noise and discarded.
2. Multiple image occurrences of same species: observations suggest these are creature close-ups and environment images under a single occurrence. They may also be larger groups with close-ups. For these, we randomly sample down to 5 images.

(a) (b) (c) (d)

Figure 10: These images all belong to GBIF Occurrence 1056329684 [72], and thus share metadata. (a) through (c) all contain the specimen (*Pristimantis zeuctotylus*), though (c) is primarily focused on its label or tag. Meanwhile, (d) is the label for another specimen of the same species from the collection; there are multiple views of this specimen as well. (c) provides an alternate view and can be retained, while (d) should be removed.

3. Camera trap images: likely images collected in backyards (high similarity). Occasionally people upload images from camera traps to citizen science platforms like iNaturualist (also ex: MammalWeb, a citizen science platform for camera trap images). These are processed under camera trap protocols described above.

The final step in our quality control processing pipeline was to filter out identifiable images of people from the training data. This was done by running MTCNN [65] on all images downloaded from GBIF and EOL.

C.3 Duplicate and Leakage Control

GBIF and EOL are large biodiversity data aggregators, sourcing images that are also used in biological benchmarks such as iNat21 [53] and NeWT [18]. Due to metadata provenance complications, the images in these test sets cannot easily be matched to their sources or copies downloaded from other sources. To prevent the introduction of test data sourced from iNaturalist or "other sources online", we perform two content deduplication steps. The first is to run MD5 hashes on all downloaded images. During our download, these hash sums are used to distinguish images that have already been downloaded. We record the hash of both the original image and our resized image. These were used to ensure that Rare Species [33] images (sourced from EOL) were not included in the training data.

Traditional hash sums are highly sensitive to small changes–a one-pixel difference between two images will produce a different hash. Hence, we applied perceptual hashing [PDQ 66], with distance less than 10, to identify training images that may be in our desired test sets that could not otherwise be filtered out (i.e., by MD5 hash sum or through metadata). Note that PDQ hash evaluation was only run on GBIF citizen science images and EOL images sourced from Flickr since they are not reliable for museum specimens.

IUCN Red List coverage. According to the most recent IUCN Red List assessment[68], TreeOfLife-200M contains images of 69.5% (55,512) of all IUCN-assessed species in categories characterized as rare species or data deficient. This coverage was determined by applying our taxonomic alignment package to the IUCN taxonomic data to enable direct comparison with our standardized dataset. TreeOfLife-200M demonstrates particularly strong representation of threatened species, with 77.1% coverage across threatened categories (36,370 species), including 79.7% of Vulnerable species (14,038), 79.4% of Endangered species (15,190), and 68.4% of Critically Endangered species (7,142). Coverage extends to 81.7% of Near Threatened species (8,073) and 82.7% of species classified as Extinct in the Wild (67). Data Deficient species have a lower representation at 48.4% (11,002), likely reflecting the challenges in imaging and identifying the species in this group. Notably, these species are designated in this category because there is not sufficient information about them for IUCN to evaluate their status; only 8% of the 2.14M described species have been evaluated [67]. Thus, including 48% of these species in TreeOfLife-200M, along with the threatened species coverage, establishes the approach to integrating diverse data sources used in TreeOfLife-200M as a valuable resource for conservation research, providing visual representation for a substantial majority of species prioritized for global conservation action.

D Formal Analyses and More Empirical Observations

In this section, we present formal analyses and more empirical observations of the training design, detailed numerical results, and qualitative analyses.

D.1 Formal Analyses

We next ask *why* these two properties of inter-species ecological alignment and intra-species variation separation emerge with scale.

Inter-species ecological alignment. As training data increases, species that share proximal taxonomic labels are pulled toward common textual prototypes at multiple levels. As related taxa typically share morphology, behavior, and ecological characteristics, this multi-level supervision aligns visual similarity with functional similarity [15]. With more samples per species providing supervision at scale, embeddings of species in the same family or genus form coherent macro-clusters. In effect, the embeddings extracted by BIOCLIP 2 are more separable for different ecological groups.

Intra-species variation separation. While the inter-species ecological alignment can be explained by hierarchical supervision, the intra-species variations are *not* encoded in taxonomic labels. The preservation of intra-species variations also contradicts the common intuition of contrastive training effects. Therefore, we investigate the optimization of contrastive loss. We propose the following theorem to suggest that subspace orthogonal to inter-species differences is allowed after extensive training to accommodate intra-species variations.

Theorem D.1. *Let $\boldsymbol{\mu}$ be the prototypes of species, with $\boldsymbol{\mu}_s$ as the prototype of species s. Let τ be temperature. If different $\boldsymbol{\mu}_k$ are nearly orthogonal (*i.e.*, species are well separated), the intra-species variation δ for species s is constrained by*

$$\boldsymbol{\delta}^\top \left[\frac{1}{2\tau^2} \left(\sum_k w_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top - \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top \right) \right] \boldsymbol{\delta}, \quad \textit{where} \quad w_k = \frac{\exp(\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_k)/\tau}{\sum_k (\exp(\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_k)/\tau)}.$$

Proof. The contrastive loss for one visual embedding $\boldsymbol{z}$ belonging to the class s and the corresponding text embedding $\boldsymbol{c}_s$ is:

$$l(\boldsymbol{z}, \boldsymbol{c}_s) = -\log \frac{\exp(\boldsymbol{z}^\top \boldsymbol{c}_s/\tau)}{\sum_k \exp(\boldsymbol{z}^\top \boldsymbol{c}_k/\tau)}. \tag{1}$$

Let $h_\phi(L(\cdot))$ be the text encoder. Assume the representation is already close to the species prototype $\boldsymbol{\mu}_s = h_\phi(L(s)) = \boldsymbol{c}_s$, and there is a residual $\boldsymbol{\delta}$ representing the intra-species variance:

$$\boldsymbol{\delta} := \boldsymbol{z} - \boldsymbol{\mu}_s, \quad \|\boldsymbol{\delta}\| \ll \|\boldsymbol{\mu}_s - \boldsymbol{\mu}_{k \neq s}\|.$$

Define:

$$a_k := \frac{\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_k}{\tau}, \quad Z = \sum_k \exp(a_k), \quad w_k := \frac{\exp(a_k)}{Z}$$

Substituting $\boldsymbol{z} = \boldsymbol{\mu}_s + \boldsymbol{\delta}$ into Equation 1:

$$l(\boldsymbol{\mu}_s + \boldsymbol{\delta}, \boldsymbol{c}_s) = -\frac{(\boldsymbol{\mu}_s + \boldsymbol{\delta})^\top \boldsymbol{\mu}_s}{\tau} + \log \left[\sum_k \exp\left(\frac{(\boldsymbol{\mu}_s + \boldsymbol{\delta})^\top \boldsymbol{\mu}_k}{\tau} \right) \right] \tag{2}$$

$$= -\frac{\boldsymbol{\mu}_s^\top (\boldsymbol{\mu}_s + \boldsymbol{\delta})}{\tau} + \log \left[\sum_k \exp\left(\frac{\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_k}{\tau} \right) \cdot \exp\left(\frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau} \right) \right] \tag{3}$$

$$= -\frac{\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_s}{\tau} - \frac{\boldsymbol{\mu}_s^\top \boldsymbol{\delta}}{\tau} + \log \left[\sum_k \exp\left(\frac{\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_k}{\tau} \right) \sum_k \frac{\exp(\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_k/\tau)}{\sum_k \exp(\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_k/\tau)} \cdot \exp\left(\frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau} \right) \right] \tag{4}$$

$$= -\frac{\boldsymbol{\mu}_s^\top \boldsymbol{\mu}_s}{\tau} - \frac{\boldsymbol{\mu}_s^\top \boldsymbol{\delta}}{\tau} + \log Z + \log \left[\sum_k w_k \exp\left(\frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau} \right) \right]. \tag{5}$$

Use the Taylor expansion to the second order for the argument of the logarithm of the last term $\Psi(\boldsymbol{\delta}) = \log\left[\sum_k w_k \exp(\boldsymbol{\delta}^\top \boldsymbol{\mu}_k/\tau)\right]$:

$$\sum_k w_k \exp\left(\frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau}\right) = 1 + \sum_k w_k \frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau} + \frac{1}{2}\sum_k w_k \frac{(\boldsymbol{\delta}^\top \boldsymbol{\mu}_k)^2}{\tau^2} + O(\|\boldsymbol{\delta}\|^3), \tag{6}$$

$$\text{obtaining} \quad \Psi(\boldsymbol{\delta}) = \sum_k w_k \frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau} + \frac{1}{2}\left[\sum_k w_k \frac{(\boldsymbol{\delta}^\top \boldsymbol{\mu}_k)^2}{\tau^2} - \left(\sum_k w_k \frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau}\right)^2\right] + O(\|\boldsymbol{\delta}\|^3). \tag{7}$$

Insert $\Psi(\boldsymbol{\delta})$ back into Eq.2:

$$\begin{aligned} l = l(\boldsymbol{\mu}_s, \boldsymbol{\mu}_s) - \frac{\boldsymbol{\mu}_s^\top \boldsymbol{\delta}}{\tau} + \sum_k w_k \frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau} \\ + \frac{1}{2}\left[\sum_k w_k \frac{(\boldsymbol{\delta}^\top \boldsymbol{\mu}_k)^2}{\tau^2} - \left(\sum_k w_k \frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau}\right)^2\right] + O(\|\boldsymbol{\delta}\|^3). \end{aligned}$$

Define $\boldsymbol{m} := \sum_k w_k \boldsymbol{\mu}_k$. Then collecting first order terms around $\boldsymbol{\delta}$ results in the following:

$$\frac{(\sum_k w_k \boldsymbol{\mu}_k - \boldsymbol{\mu}_s)^\top \boldsymbol{\delta}}{\tau} = \frac{(\boldsymbol{m} - \boldsymbol{\mu}_s)^\top \boldsymbol{\delta}}{\tau}.$$

Suppose the training resulted in $w_s \approx 1$, leading to $\boldsymbol{m} \approx \boldsymbol{\mu}_s$. Then the first-order terms will vanish as the embeddings of different species are better separated.

We further rewrite the second-order term of Eq.6 as:

$$\begin{aligned} \frac{1}{2}\left[\sum_k w_k \frac{(\boldsymbol{\delta}^\top \boldsymbol{\mu}_k)^2}{\tau^2} - \left(\sum_k w_k \frac{\boldsymbol{\delta}^\top \boldsymbol{\mu}_k}{\tau}\right)^2\right] &= \frac{1}{2}\left[\sum_k w_k \frac{(\boldsymbol{\delta}^\top \boldsymbol{\mu}_k)^2}{\tau^2} - \frac{(\boldsymbol{\delta}^\top \boldsymbol{m})^2}{\tau^2}\right] \\ &= \boldsymbol{\delta}^\top \left[\frac{1}{2\tau^2}\left(\sum_k w_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top - \boldsymbol{m}\boldsymbol{m}^\top\right)\right] \boldsymbol{\delta}. \end{aligned}$$

Substituting $\boldsymbol{m} \approx \boldsymbol{\mu}_s$, we have an approximation as

$$\boldsymbol{\delta}^\top \left[\frac{1}{2\tau^2}\left(\sum_k w_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top - \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top\right)\right] \boldsymbol{\delta}.$$

The Hessian matrix lies in the span of species as each term is an outer product of a prototype. Therefore, as long as the residual $\boldsymbol{\delta}$ is distributed in a subspace orthogonal to the species differences, the scale of $\boldsymbol{\delta}$ will not interfere with the overall contrastive optimization. □

Thus, as long as the variation $\boldsymbol{\delta}$ is distributed in a subspace orthogonal to the inter-species distinctions, the scale of $\boldsymbol{\delta}$ won't interfere with the overall contrastive optimization. The orthogonality is qualitatively supported by Figure 4. To further quantify it, we calculate the explained-variance ratio, *i.e.*, the ratio in the intra-species variation that is captured by the species span [73]. We first obtain an orthonormal basis for the species prototypes $\boldsymbol{U}$ using QR decomposition. Let $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n] \in \mathbb{R}^{d \times n}$ be the matrix stacking n intra-species variation difference vectors with dimension d. The explained-variance ratio calculates the energy fraction inside species span by $\rho = \|\boldsymbol{U}^\top \boldsymbol{D}\|_F^2 / \|\boldsymbol{D}\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm [74]. We show the ratio change as the data scales up in Figure 2c. The results suggest that the intra-species variations are increasingly orthogonal to the species differences. Due to the orthogonality, the existence of intra-species variations will not interfere with the inter-species distinctions. The observation of smaller projection areas in Figure 4 at the species span also indicates better species classification accuracy after extensive contrastive training.

Furthermore, we look at the separation between these variations using the Fisher Discriminant Ratio (FDR) metric [75]. Given two variant classes A and B, FDR is defined by their embedding mean $\boldsymbol{\mu}$ and standard deviation σ: FDR $= \|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|^2 / (\sigma_A^2 + \sigma_B^2)$. The increasing trends in Figure 2c

Table 4: Performance comparison between different replay designs of CLIP training data. All the models in the bottom three rows are initialized with CLIP (the first row) and trained with TreeOfLife-10M data. The CLIP model is pre-trained with LAION-2B data, from which we randomly select 2M samples for this experiment. Δ represents the performance gap over the CLIP baseline. **Bold** entries indicate the best accuracy.

	Species		Non-Species							
Model	NABirds	Rarespecies	FishNet	NeWT	AWA2	Herb. 19	PlantDoc	INQUIRE	Mean (Δ)	
CLIP (ViT-L/14)	66.5	35.2	$27.9_{\pm 0.2}$	$83.4_{\pm 0.1}$	$61.6_{\pm 0.6}$	$18.2_{\pm 0.1}$	$22.3_{\pm 3.3}$	35.0	43.8	–
No Replay	68.9	46.1	$\mathbf{35.3}_{\pm 0.1}$	$83.8_{\pm 0.1}$	$58.0_{\pm 2.8}$	$29.4_{\pm 0.4}$	$36.0_{\pm 2.8}$	34.4	49.0	$\uparrow 5.2$
Single-proj	68.8	44.8	$34.4_{\pm 0.2}$	$\mathbf{84.5}_{\pm 0.2}$	$\mathbf{58.1}_{\pm 1.4}$	$\mathbf{31.0}_{\pm 0.2}$	$38.1_{\pm 2.4}$	34.7	49.3	$\uparrow 5.5$
Separate-proj (Ours)	**71.2**	**47.2**	$35.1_{\pm 0.1}$	$84.2_{\pm 0.1}$	$57.4_{\pm 0.9}$	$30.4_{\pm 0.3}$	$\mathbf{38.7}_{\pm 3.7}$	**37.1**	**50.2**	$\uparrow 6.4$

support that in addition to orthogonality, the intra-species variations are more separable as the training scales up. This trend is also qualitatively evidenced by the top row of Figure 4. These theoretical and empirical insights validate that BioCLIP 2 learns to preserve and separate intra-species variations without explicit training constraints. We provide more qualitative analyses in §D.3.

D.2 CLIP Training Data Replay

Together with the contrastive supervision of hierarchical labels, we also introduce the replay of CLIP training data to retain the understanding of general knowledge. For each experiment of different training scales, we randomly select a subset from LAION-2B with 10%-20% of the corresponding biological image total. Specifically, for the largest run on 214M biological images, we select 26M samples from LAION-2B. Each training batch consists of 69,312 biological images and 8,192 replay samples. However, the text labels in TreeOfLife-200M are primarily different forms of taxonomic names, which have a distributional gap from the CLIP training data. Therefore, we apply a separate visual projector specifically for the replay data to avoid the optimization conflict. Other than this difference, the biological data and replay data share the same visual backbone and text encoder.

We evaluate different replay settings quantitatively in Table 4. The "No Replay" row shows the baseline performance applying biological contrastive training on top of the pre-trained CLIP model, where a 5.2% performance improvement is achieved. When the replay data is added and share the visual projector with biological images, we observe conflicted performance change. On Herb. 19 [19] and PlantDoc [20], more than 2% improvement is acquired. However, the species classification accuracy on Rare Species drops by 1.3%. We attribute the inconsistent performance change to the distribution gap between biological and replay text labels. After applying a separate visual projector for the replay data, we observe overall improved performance across multiple benchmarks. At the same time, we also notice that replay has limited influence on benchmarks like AwA2 [17]. Additionally, we evaluate the preservation of general knowledge understanding on INQUIRE-Rerank [31]. Applying contrastive training with taxonomic labels slightly hurts the performance, while the single-projector replay fails to retain the understanding. Comparatively, the adopted separate-projector replay improves the performance by 2.1%. Given that there is still a distribution gap between the taxonomic labels and natural language, we do not treat INQUIRE as one of our main focuses in this work.

D.3 Qualitative Analyses

Embedding distribution of Darwin's finches. In Figure 1 we show that the embedding distribution of Darwin's finches aligns with the beak size. Here we further visualize the distribution under different training scales in Figure 11. Based on the genome-based phylogeny, warbler finches were the most ancient branches, while tree finches and ground finches form the recent branches [23]. Among these species, warbler finches have the smallest beak, convenient for extracting tiny arthropods from leaves. Comparatively, ground finches have larger beaks, which are more suitable for cracking seeds and nuts. In the original CLIP embedding space, the warbler finches and tree finches are mixed. As the training data scales up, these two groups are separated, and the relative geometric relationship of all the finches aligns with their phylogeny tree. While the species separation is induced by taxonomic supervision,

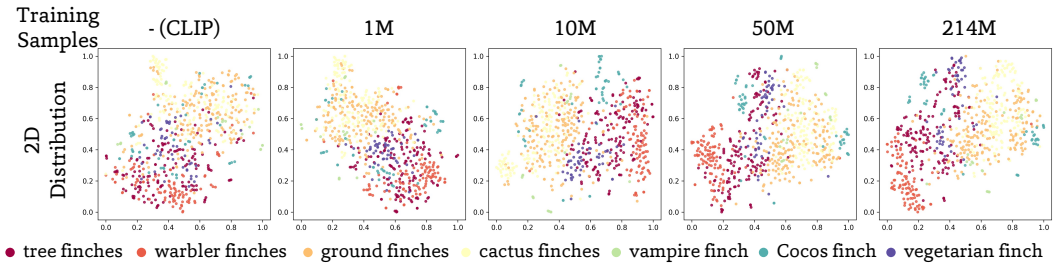


Figure 11: The t-SNE distribution of Darwin's finches under different scales of training data. Different colors represent different groups of finch species. As the training data scales up, the embeddings form biologically meaningful clusters that align with the phylogenetic tree and their functional traits.

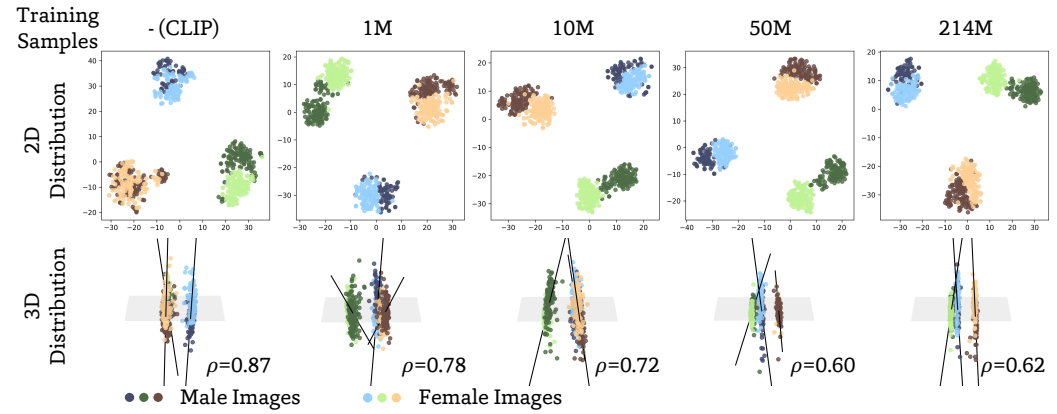


Figure 12: The embedding distribution of sex variations under different scales of training data. The 2D distributions are obtained using t-SNE. For the 3D distributions, we first run SVD with the mean embedding of each species. The first two singular vectors are used to construct the gray plane that captures most inter-species differences. The embeddings are then projected into the 3D space with an additional orthogonal dimension. The straight lines point from the mean embedding of male images to that of the female images. As the training data scales up, the intra-species variations are preserved in the subspace orthogonal to the inter-species differences. Orthogonality improves with data scale, as evidenced by the decreasing explained-variance ratio ρ.

BIOCLIP 2's embedding space again illustrates emergent higher-level biological meaningfulness after extensive training.

Embedding distribution of Sex data. Similar to Figure 4, we visualize the 2D and 3D distributions of embeddings from 3 species of the Sex data in Figure 12. We can draw conclusions similar to that obtained from Figure 4. When no training data is incorporated, the embeddings extracted by the original CLIP visual encoder (leftmost sub-figure) demonstrate large portions of overlap between male and female images. After extensive vision-language contrastive training, the embeddings present clear decision boundaries between the two variants. Furthermore, as evidenced in the 3D distribution, the embeddings within each species form more compact clusters when projected onto the species span (the gray plane). Comparatively, instead of being eliminated after contrastive training, the intra-species variations are embedded in the subspace orthogonal to the inter-species differences. The extensive training facilitates BIOCLIP 2 to acquire a biologically meaningful embedding space, highlighting its value in serving as a biology foundation model.

Embedding distribution of PlantDoc data. We further visualize the distribution of 6 classes from the PlantDoc dataset in Figure 13. When no training is processed (leftmost sub-figure), embeddings of different species, as well as diseases, are mixed. As the training scale increases, we observe two trends. First, the margin between different species is enlarged, and the embeddings belonging to the same species are clustered together. Second, the diseased leaves are easier to separate within each species, although not explicitly constrained during training. More interestingly, the embeddings of

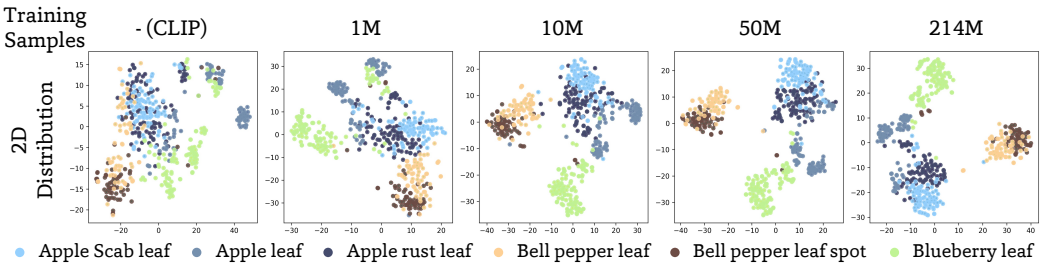
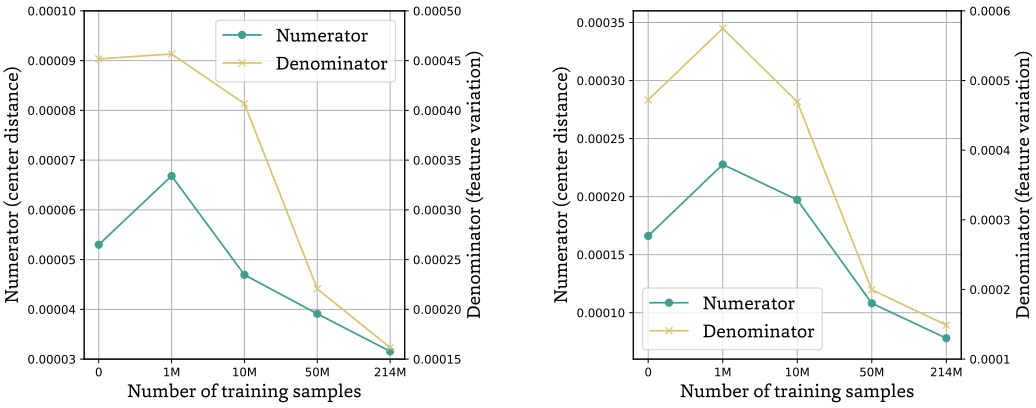


Figure 13: The t-SNE distribution of 6 classes from the PlantDoc dataset including three species and three different diseases. As the training data scales up, not only are the species better separated, but the intra-species variations also form clusters, making them easier to separate.



(a) (b)

Figure 14: The decay curves for the numerator (difference between two class centers) and the denominator (feature variation) terms of the FDR metric along the increasing data scale. (a) The curves of life stage data. (b) The curves of sex data.

healthy apple leaves are distributed close to the healthy blueberry leaves. These observations again highlight the biologically meaningful embedding space of BIOCLIP 2 after extensive training.

D.4 The decay of FDR numerator/denominator along the data scale.

We observe that FDR between two intra-species variation classes is increasing as the training data scales up. More specifically, we look into the numerator term (difference between two class centers) and the denominator term (the variation of features). We visualize the curve of the numerator and denominator terms in Figure 14a and Figure 14b for life stage data and sex data, respectively. The y-axes are scaled to the same maximum ratio to the minimum value. As the number of training samples scales up from 1M to 214M, the denominator term goes through a larger decay than the numerator term. The numerical results further support the increasing FDR trend.

D.5 Full Species Classification Results

We present the full species classification results in Table 5. BIOCLIP 2 demonstrates superior species recognition capability across both zero-shot and few-shot scenarios, over a broad range of foundation models (including the recently released DINOv3 [76]).

D.6 Camera Trap Results

In addition to the species classification benchmarks adopted in [13], we further introduce a set of camera trap images from LILA-BC [36] to construct a more realistic application scenario. Specifically, we select five datasets from LILA-BC that are labeled to the image level to avoid testing on noisy

Table 5: Zero-, one-, and five-shot species classification top-1 accuracy across 10 tasks for different models. **Bold** and underlined entries indicate the **best** and second best accuracies, respectively. BIOCLIP 2 outperforms both strong general- (CLIP, SigLIP, DINOv2) and domain-specific- (BIOCLIP, BioTrove-CLIP) baselines. "Camera Trap" is mean performance across 5 camera-trap datasets; Appendix D.6 contains more details.

	Animals					Plants & Fungi					
Model	NABirds	Plankton	Insects	Insects 2	Camera Trap	PlantNet	Fungi	PlantVillage	Med. Leaf	Rare Species	Mean
Random Guessing	0.2	1.2	1.0	1.0	3.5	4.0	4.0	2.6	4.0	0.3	2.2
Zero-Shot Classification											
CLIP (ViT-L/14)	66.5	1.3	9.0	11.7	29.5	61.7	7.6	6.5	25.6	35.2	25.5
SigLIP	61.7	2.4	27.3	20.7	33.7	81.8	36.9	**28.5**	54.5	47.6	39.5
BioTrove-CLIP	39.4	1.0	20.5	15.7	10.7	64.4	38.2	15.7	31.6	24.6	26.2
BIOCLIP	58.8	**6.1**	34.9	20.5	31.7	88.2	40.9	19.0	38.5	37.1	37.6
BIOCLIP 2	**74.9**	3.9	**55.3**	**27.7**	**53.9**	**96.8**	**83.8**	25.1	**57.8**	**76.8**	**55.6**
One-Shot Classification											
CLIP (ViT-L/14)	$42.7_{\pm 0.8}$	$28.9_{\pm 0.6}$	$29.0_{\pm 0.4}$	$17.0_{\pm 0.8}$	$36.0_{\pm 2.8}$	$58.7_{\pm 2.8}$	$20.7_{\pm 2.2}$	$56.7_{\pm 2.1}$	$74.4_{\pm 1.9}$	$34.3_{\pm 0.8}$	39.8
SigLIP	$39.9_{\pm 0.9}$	$28.4_{\pm 0.5}$	$32.3_{\pm 0.6}$	$20.6_{\pm 1.3}$	$37.8_{\pm 2.6}$	$66.3_{\pm 5.3}$	$28.7_{\pm 0.9}$	$64.1_{\pm 3.0}$	$81.7_{\pm 2.3}$	$38.8_{\pm 0.7}$	43.9
Supervised-IN21K	$43.8_{\pm 0.8}$	$23.5_{\pm 1.2}$	$15.2_{\pm 1.1}$	$18.2_{\pm 1.5}$	$30.6_{\pm 2.4}$	$63.8_{\pm 4.4}$	$26.4_{\pm 1.4}$	$52.8_{\pm 3.5}$	$75.2_{\pm 4.4}$	$31.6_{\pm 0.7}$	38.1
DINOv2	$55.4_{\pm 0.5}$	$29.8_{\pm 1.4}$	$8.7_{\pm 1.2}$	$22.0_{\pm 1.5}$	$47.7_{\pm 2.5}$	$73.2_{\pm 4.0}$	$32.8_{\pm 2.5}$	$61.8_{\pm 1.8}$	$\mathbf{94.4}_{\pm 1.6}$	$47.4_{\pm 0.4}$	47.3
DINOv3	$48.3_{\pm 1.1}$	$\mathbf{36.5}_{\pm 0.8}$	$8.8_{\pm 0.7}$	$18.8_{\pm 1.6}$	$42.6_{\pm 2.4}$	$66.8_{\pm 4.7}$	$27.5_{\pm 1.8}$	$64.8_{\pm 1.2}$	$92.1_{\pm 2.2}$	$41.5_{\pm 0.2}$	44.8
BioTrove-CLIP	$61.9_{\pm 0.6}$	$26.4_{\pm 0.5}$	$57.1_{\pm 1.4}$	$20.9_{\pm 0.7}$	$31.2_{\pm 2.3}$	$69.7_{\pm 3.4}$	$47.3_{\pm 2.1}$	$55.8_{\pm 3.4}$	$83.5_{\pm 1.1}$	$34.9_{\pm 0.4}$	48.9
BIOCLIP	$57.4_{\pm 1.2}$	$29.7_{\pm 1.1}$	$57.1_{\pm 1.0}$	$20.4_{\pm 0.9}$	$35.0_{\pm 2.8}$	$67.7_{\pm 3.9}$	$44.6_{\pm 2.0}$	$59.5_{\pm 2.5}$	$83.7_{\pm 1.8}$	$44.9_{\pm 0.7}$	50.0
BIOCLIP 2	$\mathbf{82.4}_{\pm 1.1}$	$32.0_{\pm 0.4}$	$\mathbf{74.6}_{\pm 0.4}$	$\mathbf{28.4}_{\pm 0.7}$	$\mathbf{48.1}_{\pm 2.2}$	$\mathbf{85.8}_{\pm 4.5}$	$\mathbf{70.3}_{\pm 2.6}$	$\mathbf{67.6}_{\pm 1.1}$	$92.0_{\pm 1.9}$	$\mathbf{59.5}_{\pm 0.9}$	**64.1**
Five-Shot Classification											
CLIP (ViT-L/14)	$68.2_{\pm 0.3}$	$48.2_{\pm 1.5}$	$50.6_{\pm 0.7}$	$30.1_{\pm 0.7}$	$53.9_{\pm 2.2}$	$75.9_{\pm 1.2}$	$31.4_{\pm 2.5}$	$78.3_{\pm 1.4}$	$92.6_{\pm 0.7}$	$53.3_{\pm 0.4}$	58.3
SigLIP	$64.2_{\pm 0.3}$	$47.4_{\pm 1.1}$	$54.9_{\pm 0.7}$	$35.2_{\pm 0.5}$	$56.9_{\pm 2.0}$	$81.6_{\pm 1.4}$	$45.5_{\pm 1.6}$	$81.1_{\pm 0.7}$	$94.1_{\pm 0.7}$	$57.8_{\pm 0.6}$	61.9
Supervised-IN21K	$57.5_{\pm 0.4}$	$40.1_{\pm 0.6}$	$30.1_{\pm 0.7}$	$30.3_{\pm 0.2}$	$48.0_{\pm 2.2}$	$77.2_{\pm 1.4}$	$39.6_{\pm 1.9}$	$78.0_{\pm 1.1}$	$92.8_{\pm 0.9}$	$48.8_{\pm 0.4}$	54.2
DINOv2	$73.7_{\pm 0.4}$	$48.7_{\pm 0.5}$	$14.7_{\pm 0.6}$	$31.7_{\pm 0.7}$	$66.8_{\pm 1.5}$	$87.6_{\pm 0.3}$	$49.6_{\pm 2.1}$	$78.8_{\pm 1.3}$	$\mathbf{99.2}_{\pm 0.1}$	$61.4_{\pm 0.5}$	61.2
DINOv3	$72.3_{\pm 0.4}$	$\mathbf{57.8}_{\pm 1.6}$	$19.7_{\pm 0.7}$	$34.2_{\pm 0.6}$	$63.3_{\pm 2.8}$	$83.3_{\pm 1.4}$	$44.1_{\pm 1.6}$	$82.4_{\pm 1.1}$	$98.8_{\pm 0.5}$	$62.6_{\pm 0.5}$	61.8
BioTrove-CLIP	$78.5_{\pm 0.2}$	$44.6_{\pm 0.6}$	$77.0_{\pm 0.8}$	$34.2_{\pm 0.6}$	$47.9_{\pm 2.0}$	$86.0_{\pm 1.0}$	$65.2_{\pm 0.8}$	$75.1_{\pm 0.8}$	$96.2_{\pm 0.7}$	$51.3_{\pm 0.2}$	65.6
BIOCLIP	$78.2_{\pm 0.3}$	$49.2_{\pm 1.1}$	$78.0_{\pm 0.6}$	$33.9_{\pm 0.6}$	$54.3_{\pm 2.2}$	$85.7_{\pm 1.7}$	$61.6_{\pm 1.9}$	$81.7_{\pm 1.1}$	$96.7_{\pm 0.6}$	$65.7_{\pm 0.4}$	68.5
BIOCLIP 2	$\mathbf{92.4}_{\pm 0.2}$	$50.5_{\pm 1.1}$	$\mathbf{89.3}_{\pm 0.4}$	$\mathbf{44.3}_{\pm 1.1}$	$\mathbf{67.7}_{\pm 1.9}$	$\mathbf{94.4}_{\pm 0.8}$	$\mathbf{85.0}_{\pm 1.1}$	$\mathbf{83.9}_{\pm 0.9}$	$98.4_{\pm 0.4}$	$\mathbf{77.2}_{\pm 0.4}$	**78.3**

images—those labeled as containing an animal when it is simply the animal's habitat. The Island Conservation Camera Traps [37] were of particular interest for their stated purpose of assisting in the prevention of endangered island species' extinction and the varied ecosystems represented. This provides a fine-grained complement to the Rare Species test set [13, 33]. The Desert Lion Conservation Camera Traps dataset [38] is similarly intended to advance conservation efforts. With the Orinoquía Camera Traps [39], Ohio Small Animals [40], and ENA24 [41] camera trap datasets, we can test on camera trap images across varied settings. The Island Conservation set uses common names, while the remaining datasets are reduced to just those that are labeled to the species level, and then sampled to create a balanced test set across the remaining classes (these are evaluated on scientific names). The sampling and image manifests are provided in the anonymized repository. We report the detailed accuracy on each camera trap dataset in Table 6.

E Discussion on the Relationship with Neural Collapse

Neural collapse is a status where the intra-class variations collapse to zero, and the embeddings collapse to their corresponding class prototypes [77, 78]. The class prototypes collapse to the vertices of a simplex Equiangular Tight Frame (ETF). It has been empirically observed and theoretically proved that the commonly used loss functions—including cross-entropy loss and supervised contrastive loss [79]—lead to neural collapse at the terminal phase of training [24, 25, 80, 81]. If neural collapse happens, the intra-species appearance variations will be hardly separable. However, we reveal that after extensive training of BIOCLIP 2, the intra-species variations are preserved in the subspace orthogonal to inter-species differences, indicating that BIOCLIP 2 does not suffer neural collapse.

Why BIOCLIP 2 does not lead to Neural Collapse. We summarize two key reasons that allow the existence of the intra-species variation subspace in BIOCLIP 2.

Table 6: Zero-, one-, and five-shot species classification accuracy of camera trap images for different models. **Bold** and underlined entries indicate the **best** and second best accuracies, respectively.

	LILA-BC					
Model	Desert Lion	ENA24	Island	Orinoquia	OH Small Animals	Mean
Random Guessing	3.1	5.0	3.0	3.6	2.6	3.5
Zero-Shot Classification						
CLIP (ViT-L/14)	35.2	38.2	27.1	25.6	21.6	29.5
SigLIP	46.9	41.0	28.1	31.9	20.5	33.7
BioTrove-CLIP	9.7	10.4	13.2	8.0	12.4	10.7
BIOCLIP	47.2	42.5	18.4	27.1	23.1	31.7
BIOCLIP 2	**58.8**	**68.5**	**42.6**	**47.9**	**51.5**	**53.9**
One-Shot Classification						
CLIP (ViT-L/14)	$37.6_{\pm 2.5}$	$35.9_{\pm 3.8}$	$44.7_{\pm 3.0}$	$31.3_{\pm 2.7}$	$30.7_{\pm 2.0}$	36.0
SigLIP	$41.1_{\pm 1.8}$	$39.1_{\pm 4.0}$	$48.5_{\pm 2.9}$	$32.7_{\pm 1.6}$	$27.6_{\pm 2.5}$	37.8
Supervised-IN21K	$32.4_{\pm 2.0}$	$28.4_{\pm 3.2}$	$40.1_{\pm 2.2}$	$26.3_{\pm 2.4}$	$26.0_{\pm 2.4}$	30.6
DINOv2	$\mathbf{56.3}_{\pm 0.8}$	$\mathbf{48.9}_{\pm 2.7}$	$48.2_{\pm 2.4}$	$\mathbf{43.1}_{\pm 3.4}$	$41.9_{\pm 3.3}$	47.7
DINOv3	$48.1_{\pm 1.3}$	$38.5_{\pm 3.2}$	$\mathbf{50.5}_{\pm 2.8}$	$40.0_{\pm 2.6}$	$36.1_{\pm 2.3}$	42.6
BioTrove-CLIP	$30.5_{\pm 1.7}$	$30.5_{\pm 2.2}$	$37.9_{\pm 2.1}$	$29.0_{\pm 2.6}$	$28.3_{\pm 3.1}$	31.2
BIOCLIP	$39.9_{\pm 2.2}$	$34.4_{\pm 2.7}$	$45.7_{\pm 3.0}$	$27.5_{\pm 3.7}$	$27.5_{\pm 2.2}$	35.0
BIOCLIP 2	$54.3_{\pm 1.1}$	$48.6_{\pm 2.3}$	$49.5_{\pm 2.5}$	$\mathbf{43.1}_{\pm 2.1}$	$\mathbf{45.0}_{\pm 3.0}$	**48.1**
Five-Shot Classification						
CLIP (ViT-L/14)	$58.3_{\pm 1.7}$	$57.8_{\pm 2.6}$	$66.1_{\pm 2.4}$	$43.9_{\pm 2.9}$	$43.5_{\pm 1.6}$	53.9
SigLIP	$64.3_{\pm 2.2}$	$60.0_{\pm 2.7}$	$71.6_{\pm 1.4}$	$46.2_{\pm 1.8}$	$42.6_{\pm 1.7}$	56.9
Supervised-IN21K	$51.3_{\pm 2.3}$	$48.5_{\pm 2.6}$	$59.0_{\pm 1.3}$	$39.3_{\pm 3.4}$	$41.9_{\pm 1.6}$	48.0
DINOv2	$71.9_{\pm 2.1}$	$\mathbf{73.5}_{\pm 1.4}$	$69.2_{\pm 0.5}$	$\mathbf{61.2}_{\pm 1.6}$	$58.3_{\pm 1.9}$	66.8
DINOv3	$66.3_{\pm 3.3}$	$63.2_{\pm 3.1}$	$\mathbf{73.4}_{\pm 1.3}$	$59.9_{\pm 3.7}$	$53.8_{\pm 2.4}$	63.3
BioTrove-CLIP	$47.6_{\pm 3.0}$	$46.7_{\pm 2.1}$	$62.2_{\pm 1.1}$	$41.1_{\pm 1.6}$	$41.8_{\pm 2.4}$	47.9
BIOCLIP	$62.5_{\pm 1.7}$	$57.0_{\pm 3.2}$	$63.2_{\pm 3.8}$	$44.6_{\pm 1.7}$	$44.0_{\pm 0.5}$	54.3
BIOCLIP 2	$\mathbf{73.4}_{\pm 2.9}$	$73.4_{\pm 0.9}$	$70.7_{\pm 1.6}$	$59.4_{\pm 2.8}$	$\mathbf{61.8}_{\pm 1.3}$	**67.7**

First, the class prototypes in standard classification tasks and supervised contrastive learning (SCL) can both be treated as fully trainable parameters [79]. In standard classification, the prototypes are weights of the linear classifier. In SCL, they are simply the mean embedding of the corresponding classes. The class prototypes will form a simplex ETF after extensive training. Comparatively, the adopted contrastive training scheme in BIOCLIP 2 employs text embeddings of the taxonomic labels as class prototypes. During the creation of taxonomic labels, hierarchical structures have been naturally embedded into them based on ecological and functional evidence. Different species can share higher taxonomic levels, which makes them hard negatives. Therefore, even if the text encoder is also being trained, the generated prototypes will not become ETF.

Second, TREEOFLIFE-200M poses an enormous label space, where 952K classes are involved in training. Typical SCL is performed upon CIFAR or ImageNet-level datasets, consisting of $100-1,000$ classes. Previous works analyzing neural collapse usually assume the dimension of embeddings is larger than the class number and the sample number is balanced across different classes [25, 78]. In contrast, the class number involved in BIOCLIP 2 is much larger than the embedding dimension (768 channels). These conditions restrict the possibility of prototypes forming ETF.

F Biological Visual Evaluation Details

Instead of training specialist models, we treat the evaluated models as frozen visual embedding extractors. Standard machine learning algorithms are applied on top of the acquired visual embeddings to predict the corresponding labels. Such a design is adopted to evaluate the quality of the embeddings while avoiding the influence of complicated optimization loops. In the following, we introduce the details of adopted benchmarks and the evaluation algorithms. All the experiments are conducted with 1 NVIDIA A100 GPU, and the running time for each task is within 30 minutes.

FishNet. FishNet focuses on recognizing, locating, and predicting species and their functional traits [16]. Specifically, 94,532 images are collected with annotations of habitat, ecological role, and nutritional value. In this work, we mainly focus on the prediction of habitats and ecological roles, involving 9 groups of binary labels (*e.g.* whether the fish can live in freshwater). Following the practice in the original paper, we train a two-layer linear classifier with binary cross-entropy loss to predict the 9 labels. We count a correct prediction only if all the 9 labels are predicted correctly for the sample. The original train-test split is adopted, where 75,631 images are used in training, and the remaining 18,901 images are used for testing. This task evaluates whether the embedding distribution of different species is aligned with their ecological relationships.

NeWT. NeWT comprises 164 binary classification tasks in the natural world [18]. The tasks include appearance, gestalt, context, counting, and behavior concepts. In each task, 50-100 images are assigned per class per train-test split. After extracting visual embeddings, we apply a support vector classifier for each of the tasks. The average accuracy is reported across all the evaluated tasks.

AwA2. AwA2 consists of 37,322 images of 50 animal classes, with annotations of 85 numeric attribute values for each class [17]. The dataset can be used for testing the capability of attribute-based classification and zero-shot transfer learning of trait prediction. In this work, we mainly focus on the transfer learning scheme. The training set of 45 classes is used to train an attribute classifier, and the remaining 5 unseen classes are used for testing. Similar to FishNet, we incorporate a linear classifier on top of the extracted features to predict the binary labels for the 85 attributes. The average accuracy over all the attributes is reported for this benchmark.

Herb. 19. Herbarium 19 is a task for discovering new species [19]. Specifically, given the images of known and unknown species, the model is required to predict the labels for both of them. As there are no fixed labels for unknown species, the task is implemented with a form of semi-supervised clustering [82]. Given the total number of species, a semi-supervised K-means algorithm is conducted on top of the extracted embeddings to cluster images. Clustering accuracy is calculated for the predictions following the original practice [82].

PlantDoc. PlantDoc is a dataset targeting the incorporation of computer vision for scalable and early plant disease detection [20]. 2,598 images of 13 plant species and up to 17 classes of diseases are collected in uncontrolled natural settings. We evaluate the model on PlantDoc in a few-shot learning style. One image per class is randomly selected from the original training split as the support set, and SimpleShot [83] is employed to predict the class labels for the test set. Accuracy over all the testing samples is reported as the performance.

Life stage-Diff/Align. We use the data of Age tasks from NeWT [18] to construct the Life stage-Diff/Align benchmark, where images are labeled with juvenile and adult classes. Specifically, the differentiation tasks aim to separate the binary appearance variations within each species. Conversely, in the alignment task, we train a species classifier with juvenile images while testing it using adult images. It requires the embeddings of two variations within one species to be closer than the embeddings of different species. For both of the tasks, we incorporate a support vector classifier to predict the corresponding labels. Ideally, after extensive contrastive training, the embeddings of different variation classes are expected to collapse to the species prototype. However, we demonstrate that the intra-species variations are well preserved in the embedding space of BIOCLIP 2.

Sex-Diff/Align. NABirds consists of 48,000 images from 400 bird species [35]. Sex and life stage labels are provided for those species with large appearance variations. We manually examine the images and select 81 species with male-female differences to construct the Sex-Diff/Align benchmark, where 13,624 images are used in total. For the differentiation task, we filter out 20 male images and

20 female images per species as the test set. Among the remaining 10,384 images, we filter out at most 20 male and 20 female images for training. For the alignment setting, we use the images of female birds to train a species classifier and use the male images for testing. Similar to the life stage benchmark, we use a support vector classifier for both of the sub-tasks.

G Training Implementation Details

Table 7: The adopted hyper-parameter setting in training BIOCLIP 2.

Hyper-parameter	Value
Architecture	ViT-L/14
Optimizer	Adam
Batch size/GPU (organism)	2,816
Batch size/GPU (replay)	320
GPUs	32 H100s
Epochs	30
Max learning rate	1×10^{-4}
Warm-up steps	1,875
Weight decay	0.2
Input resolution	224

Table 8: The adopted hyper-parameter setting in ablation study.

Hyper-parameter	Value
Architecture	ViT-L/14
Optimizer	Adam
Batch size/GPU (organism)	2,816
Batch size/GPU (replay)	320
GPUs	8 H100s
Epochs	100
Max learning rate	1×10^{-4}
Warm-up steps	1,500
Weight decay	0.2
Input resolution	224

We list the adopted hyper-parameters in Table 7 and Table 8 for the BIOCLIP 2 training and ablation study, respectively. TREEOFLIFE-10M is used for the ablation study. The batch size presented in both tables is the size per GPU. In addition to the larger GPU number and smaller batch size compared with the training in BIOCLIP due to the larger model size, another important modification is the introduction of experience replay. An additional visual projector is introduced as described in the Appendix. Beyond the visual projector, the model architecture is kept the same as that of CLIP [12]. During the evaluation, all the embeddings are extracted with the visual projector for hierarchical label matching to avoid extra influence.

H Baseline Details

In the quantitative experiments, we compare BIOCLIP 2 with the following baseline vision-language models and vision-only models. Without specification, the input image size is 224.

- **CLIP (ViT-L/14).** We compare BIOCLIP 2 with the CLIP model pre-trained on LAION-2B [27], which has the same architecture and patch size. The CLIP model is also used as the weight initialization of the BIOCLIP 2 training. It uses a ViT-large visual encoder with a patch size of 14. We load the weight from the OpenCLIP repository.
- **SigLIP.** In addition to the standard CLIP model, we also evaluate the performance of SigLIP [21]. We adopt the SigLIP model pre-trained on WebLI data [84]. The adopted visual encoder is ViT-large [26], the patch size is 16, and the input image resolution is 256. The model weight is also loaded from the OpenCLIP repository.
- **Supervised-IN21K.** For vision-only models, we first select a ViT-large model [26] trained in a supervised way on ImageNet-21K dataset [85]. As it is a vision-only model, we only run a few-shot and non-species classification tasks with it. The patch size is 32. The model is publicly downloadable in Hugging Face.
- **DINOv2.** Besides supervised pre-training, we also evaluate the performance of DINOv2, which is pre-trained in an unsupervised way [22]. The backbone architecture is ViT-large and the patch size is 14. The model can be downloaded from Hugging Face. Similarly, we only run few-shot and non-species classification evaluations with DINOv2.
- **BioTrove-CLIP.** The above four models are trained on general knowledge covering a variety of topics. We also compare BIOCLIP 2 with domain-specific models. BioTrove-CLIP is trained with BioTrove [14]. The model weights can be downloaded from Hugging Face. Among the provided three models, we use the model initialized with OpenAI CLIP [12] that yields the best average accuracy [14]. The visual backbone is ViT-base with a patch size of 16.

- **BIOCLIP.** BIOCLIP is trained on TREEOFLIFE-10M. It adopts ViT-base as the visual encoder, with a patch size of 16. The model weight is publicly available in Hugging Face.

I Limitations

Theoretical limitation. In this work, we have proved that the intra-species variations are preserved in subspaces orthogonal to the inter-species difference. We also have empirical observations that the variations are more separable as the training data scales up (See Figure 2c). However, we haven't formally proved this. It will be our future work to have deeper theoretical analyses of the separation of intra-species variations to better understand BIOCLIP 2's emergent properties.

Data limitation. TREEOFLIFE-200M is an imbalanced dataset in both taxonomic coverage and image type. Specifically, the dataset exhibits a long-tailed distribution across taxa. This is to be expected when working with biological data—not all taxonomic ranks are represented evenly across the tree of life. For instance, though TREEOFLIFE-200M has a balanced representation (at the kingdom level) between plants and animals, animals represent a larger proportion of described species [67].

Some of this is due to the nature of the image type distribution, which we provide for GBIF (Camera-trap, Citizen Science, and eleven Museum Specimen types: Fungi, Insect, Invertebrate Zoology, Microbiology, Plant, Vertebrate Zoology - [Amphibians, Birds, Fishes, Mammals, Others], as well as Unidentified). EOL contains the same categories, but we do not have precise numbers; BIOSCAN-5M are essentially all insect museum specimens, though the images are taken by researchers, so will skew toward their area of study; FathomNet contains a mix. Citizen Science images are the vast majority (151M from GBIF alone); these will skew toward more charismatic species and plants. Our next largest category is museum specimen images (51.8M from GBIF), which are limited more to representatives of the species, so may not contain as many images per taxa, though more taxa are represented. Finally, camera trap images make up the smallest portion of the dataset (617.8K in GBIF), and when filtered, these are only images of animals and generally those large enough to trigger a motion sensor or be detected in the primary provider's post-processing.

Further emphasis on the impact of citizen science images in amassing larger representations of species: the most prevalent taxonomic classes, flowering plants, insects, birds, mushrooms, and mammals, have millions of representative images, while the least-represented, microscopic organisms (e.g., bacteria, viruses), have a dozen or less.

J Broader Impacts

BIOCLIP 2 and TREEOFLIFE-200M provide great potential to improve and enhance existing conservation efforts, in particular by facilitating recognition of threatened species. As noted in §B.3, TREEOFLIFE-200M has expansive coverage of threatened species, as classified by IUCN. It additionally builds on the coverage of species considered to be Data Deficient. Based on the emergent properties displayed by BIOCLIP 2, there is potential to add in the effort to understand the risks facing these species that cannot currently be classified by IUCN due to lack of available information, as suggested in [86]. These designations are crucial to the international effort to protect biodiversity across the planet.

Unfortunately, as with many open-source efforts to further conservation goals, there is potential for bad actors to make use of these tools for malicious purposes. Though the improvement on threatened species *could* make it easier for poachers to identify protected species, these types of tools are a force-multiplier to monitor illicit trade and sales of these same species. The primary risk to endangered species comes from disclosure of precise location information rather than improved classification capability [87]. Our data does not provide geo-tagged information of the organisms included, minimizing the vulnerabilities that could be used in poaching.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: This paper focuses on emergent properties in large-scale hierarchical contrastive learning for biological organismal data. We have theoretical and empirical results to support the claims.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: We have a limitation section in §I including the discussion of theoretical and data limitations.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: We propose a theorem to support the orthogonality between the intra-species variations and inter-species differences. The proof is provided in §**??**.
   Guidelines:
   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: We have provided all the required documentation, metadata, and resources used to produce the results in the supplementary material.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: Data and code are being provided as supplementary material to maintain anonymity, and will all be published open-source under permissive licenses. TreeOfLife-200M will be published on Hugging Face under the least restrictive licensing allowed by its constituent parts, accompanied by a well-documented dataset card, and archived with a DOI. The Camera Trap test set will also be published on Hugging Face, under the Community Data License Agreement (permissive variant) used by LILA-BC datasets. The reconstruction of TreeOfLife-200M requires downloading the images from their original sources, instructions for which are included with the dataset and will be published alongside it.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: We provide implementation details of the BIOCLIP 2 model training in the supplementary material. The evaluation details for biological visual tasks are listed in Sec. F.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: Among the experiments presented in the paper, zero-shot classification is deterministic, so we only report the accuracy for those experiments. Other than zero-shot classification, we conduct the other experiments—few-shot classification and other biological visual tasks—for 5 runs and report the mean and standard deviation.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
Answer: [Yes]
Justification: We have reported the required compute resources for the BIOCLIP 2 model training and the evaluation of other biological visual tasks.
Guidelines:
- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
Answer: [Yes]
Justification: The submission conforms to the NeurIPS Code of Ethics, where individual relevant adherence to components of the code are discussed in detail in other portions of the submission. For example, human faces are removed from the dataset to preserve privacy and data licenses and provenance are preserved.
Guidelines:
- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [Yes]
Justification: We have included a Broader Impacts section in §J.
Guidelines:
- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
Answer: [NA] .
Justification: Images were gathered from established, publicly accessible biodiversity repositories and we used established procedures to ensure removal of images with human faces. Additionally, there are no GPS locations associated with the threatened species whose images are included.
Guidelines:
- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**
Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes] .
Justification: The core data providers through whom we accessed all images and their metadata for TreeOfLife-200M are directly cited in the paper. We provide and document provenance for all images through a dedicated Parquet file connecting each image to license, creator, etc. and include an explanation in our dataset card. All remaining data, code, and models are cited appropriately, as indicated by their creators.
Guidelines:
- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [Yes] .
Justification: We have the detailed dataset cards for TreeOfLife-200M and CameraTrap images attached in the supplementary material. Documentation for code is also provided as it will be in the published GitHub repositories (anonymized for this submission).
Guidelines:
- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [NA]
Justification: We do not involve crowdsourcing or research with human subjects.
Guidelines:
- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [NA]
Justification: We do not involve crowdsourcing or research with human subjects.
Guidelines:
- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**
Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.
Answer: [NA]
Justification: LLMs were not involved in developing the core method of this paper.
Guidelines:
- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.