[{'title': 'BIOCLIP: A Vision Foundation Model for the Tree of Life',

'authors': [{'name': 'Samuel Stevens',

'affiliation': 'The Ohio State University'},

{'name': 'Jiaman Wu', 'affiliation': 'The Ohio State University'},

{'name': 'Matthew J Thompson', 'affiliation': 'The Ohio State University'},

{'name': 'Elizabeth G Campolongo',

'affiliation': 'The Ohio State University'},

{'name': 'Chan Hee Song', 'affiliation': 'The Ohio State University'},

{'name': 'David Edward Carlyn', 'affiliation': 'The Ohio State University'},

{'name': 'Li Dong', 'affiliation': 'Microsoft Research'},

{'name': 'Wasila M Dahdul',

'affiliation': 'University of California, Irvine'},

{'name': 'Charles Stewart',

'affiliation': 'Rensselaer Polytechnic Institute'},

{'name': 'Tanya Berger-Wolf', 'affiliation': 'The Ohio State University'},

{'name': 'Wei-Lun Chao', 'affiliation': 'The Ohio State University'},

{'name': 'Yu Su', 'affiliation': 'The Ohio State University'}],

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

'key\_findings': 'BIOCLIP consistently and substantially outperforms existing baselines (by 16% to 17% absolute) on diverse fine-grained biology classification tasks. Intrinsic evaluation reveals that BIOCLIP has learned a hierarchical representation conforming to the tree of life.',

'limitation\_of\_sota': 'Most computational methods and tools for extracting biologically relevant information from images are bespoke approaches designed for a specific task and are not easily adaptable or extendable to new questions, contexts, and datasets.',

'proposed\_solution': 'The development of BIOCLIP, a foundation model for the tree of life, leveraging the unique properties of biology captured by TREEOFLIFE-10M, namely the abundance and variety of images of plants, animals, and fungi, together with the availability of rich structured biological knowledge.',

'paper\_limitations': '',

'source': 'BIOCLIP: A Vision Foundation Model for the Tree of Life\nSamuel Stevens1\*†, Jiaman Wu1\*, Matthew J Thompson1, Elizabeth G Campolongo1, Chan Hee Song1,\nDavid Edward Carlyn1, Li Dong2, Wasila M Dahdul3, Charles Stewart4, Tanya Berger-Wolf1,\nWei-Lun Chao1, and Yu Su1†\n1The Ohio State University,2Microsoft Research,3University of California, Irvine,\n4Rensselaer Polytechnic Institute\nAbstract\nImages of the natural world, collected by a variety of\ncameras, from drones to individual phones, are increasingly\nabundant sources of biological information. There is an ex-\nplosion of computational methods and tools, particularly\ncomputer vision, for extracting biologically relevant infor-\nmation from images for science and conservation. Yet most\nof these are bespoke approaches designed for a specific task\nand are not easily adaptable or extendable to new ques-\ntions, contexts, and datasets. A vision model for general or-\nganismal biology questions on images is of timely need. To\napproach this, we curate and release TREEOFLIFE-10M ,\nthe largest and most diverse ML-ready dataset of biology\nimages. We then develop BIOCLIP , a foundation model\nfor the tree of life, leveraging the unique properties of bi-\nology captured by TREEOFLIFE-10M , namely the abun-\ndance and variety of images of plants, animals, and fungi,\ntogether with the availability of rich structured biological\nknowledge. We rigorously benchmark our approach on di-\nverse fine-grained biology classification tasks and find that\nBIOCLIP consistently and substantially outperforms exist-\ning baselines (by 16% to 17% absolute). Intrinsic evalua-\ntion reveals that BIOCLIP has learned a hierarchical rep-\nresentation conforming to the tree of life, shedding light on\nits strong generalizability.1\n1. Introduction\nDigital images and computer vision are quickly becoming\npervasively used tools to study the natural world, from evo-\nlutionary biology [13, 51] to ecology and biodiversity [5,\n77, 83]. The capability to rapidly convert vast quantities of\nimages from museums [64], camera traps [1, 6, 7, 59, 77],\nand citizen science platforms [2, 40, 54, 58, 60, 62, 75,\n\*Equal contribution.†{stevens.994,su.809 }@osu.edu\n1imageomics.github.io/bioclip has models, data and code.79–81, 87, 88] into actionable information (e.g., species\nclassification, individual identification, and trait detection)\nhas accelerated and enabled new advances in tasks such\nas species delineation [32], understanding mechanisms of\nadaptation [23, 39], abundance and population structure es-\ntimation [3, 40, 58, 82], and biodiversity monitoring and\nconservation [83].\nHowever, applying computer vision to answer a biologi-\ncal question is still a laborious task requiring substantial ma-\nchine learning expertise and effort—biologists must manu-\nally label sufficient data for the specific taxa and task of\ninterest, and find and train a suitable model for the task.\nMeanwhile, foundation models [12] such as CLIP [69] and\nGPT-3 [14] are extraordinarily valuable by enabling zero-\nshot or few-shot learning for a wide range of tasks. An\nanalogous vision foundation model for biology should be\nuseful for tasks spanning the entire tree of life [37, 53] in-\nstead of just the taxa it was trained on. Such a model would\nsignificantly lower the barrier to apply AI to biology.\nIn this work, we aim to develop such a vision foundation\nmodel for the tree of life. To be broadly useful for real-\nworld biology tasks, this model should meet the following\ncriteria. First, it should generalize to the entire tree of\nlife, where possible, to ensure it supports researchers study-\ning many different clades rather than a niche. Furthermore,\nit is infeasible to collect training data that covers the mil-\nlions of known taxa [38, 44], so the model must generalize\nto taxa not present in training data. Second, it should learn\nfine-grained representations of images of organisms as bi-\nology frequently engages with organisms that are visually'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': '',

'proposed\_solution': "The proposed solution emphasizes the importance of a model's ability to generalize to taxa not present in training data, learn fine-grained representations of images of organisms, and perform strongly despite the high cost of data collection and labeling in biology. This approach is crucial for dealing with the vast diversity of life, which includes organisms that are visually similar or mimic others for survival advantages.",

'paper\_limitations': '',

'source': 'to taxa not present in training data. Second, it should learn\nfine-grained representations of images of organisms as bi-\nology frequently engages with organisms that are visually\nsimilar, like closely related species within the same genus\n[67] or species mimicking others’ appearances for a fitness\nadvantage [39]. This fine-grained granularity is crucial be-\ncause the tree of life organizes living things into both broad\ncategories (animal, fungus, and plant) and very fine-grained\nones (see Fig. 1). Finally, due to the high cost of data col-\nlection and labeling in biology, strong performance in thearXiv:2311.18803v3 [cs.CV] 14 May 2024'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'The introduction of TREEOFLIFE-10M, a large-scale, diverse ML-ready biology image dataset with over 10 million images covering 454 thousand taxa, and BIOCLIP, a vision foundation model for the tree of life, trained with suitable use of taxa in TREEOFLIFE-10M. These tools address the need for suitable pre-training datasets and strategies in the field of evolutionary biology and ecology, where existing general-domain vision models fall short.',

'limitation\_of\_sota': 'Existing general-domain vision models trained on hundreds of millions of images fall short when applied to evolutionary biology and ecology, particularly for fine-grained comparisons such as between Onoclea sensibilis and Onoclea hintonii. Current datasets lack either scale, diversity, or fine-grained labels, and mainstream pre-training algorithms insufficiently consider the tree of life taxonomy.',

'proposed\_solution': 'The proposed solution includes the creation of TREEOFLIFE-10M, the largest ML-ready dataset to-date of biology images with associated taxonomic labels, and BIOCLIP, a vision foundation model for the tree of life. BIOCLIP is trained using a suitable pre-training strategy that leverages the special properties of the biology domain, such as the tree of life taxonomy, to better achieve generalization, fine-grained classification, and data efficiency.',

'paper\_limitations': '',

'source': 'PlantaeTracheophytaPolpodiopsidaPolypodialesOnocleaceaeOnocleahintoniiPlantaePlantaeTracheophytaTracheophytaPolpodiopsidaPolpodiopsidaPolypodialesPolypodialesOnocleaceaeOnocleaceaeOnocleaOnocleasensibilishintoniiKingdomPhylumClassOrderFamilyGenusSpecies(b) Autoregressive Representations(a) Taxonomic LabelsT1TnI1I1·T1…I1·Tn……In·Tn…\nInIn·T1………(c) Contrastive ObjectiveImage Features\n(d)Onoclea sensibilis\n (e)Onoclea hintonii\nFigure 1. (a) Two taxa, or taxonomic labels, for two different plants, Onoclea sensibilis (d) and Onoclea hintonii (e). These taxa are\nidentical except for the species. (b) The autoregressive text encoder naturally encodes the hierarchical structure of the taxonomy. See\nhow the Order token(s) (Polypodiales) can incorporate information from the Kingdom, Phylum and Class tokens, but nothing later in the\nhierarchy. This helps align the visual representations to this same hierarchical structure (see §4.6). (c) These hierarchical representations\nof taxonomic labels are fed into the standard contrastive pre-training objective and are matched with image representations (d) and (e).\nlow-data regime (i.e., zero-shot or few-shot) is critical.\nWhile the goals of generalization ,fine-grained classifi-\ncation , and data efficiency are not new in computer vision,\nexisting general-domain vision models [61, 69, 95] trained\non hundreds of millions of images fall short when applied\nto evolutionary biology and ecology. Specifically, exist-\ning vision models produce general fine-grained represen-\ntations, useful for comparing common organisms like dogs\nand wolves, but not for more fine-grained comparisons, e.g.,\nOnoclea sensibilis andOnoclea hintonii (see Fig. 1).\nWe identify two major barriers to developing a vision\nfoundation model for biology. First, there is a need for\nsuitable pre-training datasets : existing datasets [28, 86,\n88, 89] lack either scale, diversity, or fine-grained labels.\nSecond, there is a need to investigate suitable pre-training\nstrategies that leverage special properties of the biology do-\nmain to better achieve the three pivotal goals, e.g., the tree\nof life taxonomy, which is insufficiently considered in main-\nstream pre-training algorithms [48, 61, 69].\nIn light of these goals and challenges in achieving them,\nwe introduce 1) TREEOFLIFE-10M , a large-scale ML-\nready biology image dataset, and 2) BIOCLIP , a vision\nfoundation model for the tree of life, trained with suitable\nuse of taxa in T REEOFLIFE-10M. We outline the contribu-\ntions, conceptual framework, and design decisions below:\nTREEOFLIFE-10M: a large-scale, diverse ML-ready bi-\nology image dataset . We curate and release the largest\nML-ready dataset to-date of biology images with associ-\nated taxonomic labels, containing over 10million images\ncovering 454thousand taxa in the tree of life.2In compar-\nison, the current largest ML-ready biology image dataset,\niNat21 [86], contains only 2.7million images covering\n10thousand taxa. T REEOFLIFE-10M integrates existing\nhigh-quality datasets like iNat21 and B IOSCAN -1M [28].\nMore importantly, it includes newly curated images from\nthe Encyclopedia of Life (eol.org), which supplies most\nof T REEOFLIFE-10M’s data diversity. Every image in\n2By ML-ready, we mean the data is standardized in a format suitable\nfor training ML models and is readily available for downloading.TREEOFLIFE-10M is labeled with its taxonomic hierar-\nchy to the finest level possible, as well as higher taxonomic\nranks in the tree of life (see Fig. 1 and Tab. 3 for examples of\ntaxonomic ranks and labels). T REEOFLIFE-10M enables\ntraining B IOCLIP and future biology foundation models.\nBIOCLIP: a vision foundation model for the tree of life .\nWith a large-scale labeled dataset like T REEOFLIFE-10M,\na standard, intuitive training strategy (as adopted by other\nvision models like ResNet50 [33] and Swin Transformer\n[48]) is to use a supervised classification objective and learn\nto predict the taxonomic indices from an image. However,'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': 'This fails to recognize and leverage the rich structure of taxonomic labels—taxa do not exist in isolation but are interconnected in a comprehensive taxonomy. Consequently, a model trained via plain supervised classification may not generalize well to taxa unseen in training, nor could it support zero-shot classification of unseen taxa.',

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

'paper\_limitations': '',

'source': 'vision models like ResNet50 [33] and Swin Transformer\n[48]) is to use a supervised classification objective and learn\nto predict the taxonomic indices from an image. However,\nthis fails to recognize and leverage the rich structure of tax-\nonomic labels—taxa do not exist in isolation but are inter-\nconnected in a comprehensive taxonomy. Consequently, a\nmodel trained via plain supervised classification may not\ngeneralize well to taxa unseen in training, nor could it sup-\nport zero-shot classification of unseen taxa.\nInstead, we propose a novel strategy combining CLIP-\nstyle multimodal contrastive learning [69] with the rich bi-\nological taxonomy for B IOCLIP. We “flatten” the taxon-\nomy from Kingdom to the distal-most taxon rank into a\nstring called taxonomic name , and use the CLIP contrastive\nlearning objective to learn to match images with their cor-\nresponding taxonomic names. Intuitively, this helps the\nmodel generalize to unseen taxa—even if the model has not\nseen a species, it has likely learned a reasonable represen-\ntation for that species’ genus or family (see Fig. 1). B IO-\nCLIP also supports zero-shot classification with taxonomic\nnames of unseen taxa. We further propose, and demonstrate\nthe effectiveness of, a mixed text type training strategy; by\nmixing different text types (e.g., taxonomic vs. scientific vs.\ncommon names) during training, we retain the generaliza-\ntion from taxonomic names while being more flexibility at\ntest time. For example, B IOCLIP still excels even if only\ncommon species names are offered by downstream users.\nComprehensive benchmarking . We comprehensively\nevaluate B IOCLIP on 10fine-grained image classification\ndatasets covering animals, plants, and fungi, including a\nnewly curated R ARE SPECIES dataset unseen in training.\nBIOCLIP achieves strong performance in both zero-shot'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'BIOCLIP substantially outperforms both CLIP and OpenCLIP in zero-shot and few-shot settings with an average absolute improvement of 17% (zero-shot) and 16% (few-shot). BIOCLIP has learned a more fine-grained hierarchical representation conforming to the tree of life, explaining its superior generalization.',

'limitation\_of\_sota': 'Existing CLIP models like CLIP and OpenCLIP are limited in their ability to generalize across the diverse and fine-grained categories found in biological data.',

'proposed\_solution': 'The TREEOFLIFE-10M dataset was curated, integrating iNat21, EOL dataset, and BIOSCAN-1M, to create the most diverse large-scale public ML-ready dataset for computer vision models in biology. This dataset aims to address the limitations of existing datasets by providing a more diverse and comprehensive collection of biology images, focusing on species diversity and including a significant number of images from underrepresented categories such as insects.',

'paper\_limitations': 'The label granularity in BIOSCAN-1M may still be limited for insects, with a significant percentage of images labeled only to the family level and not to the more specific genus or species levels. Taxonomic hierarchies are notoriously noisy and rarely consistent between sources, which poses challenges for label aggregation and standardization.',

'source': 'and few-shot settings and substantially outperforms both\nCLIP [69] and OpenCLIP [42], leading to an average abso-\nlute improvement of 17% (zero-shot) and 16% (few-shot).\nIntrinsic analysis further reveals that B IOCLIP has learned\na more fine-grained hierarchical representation conforming\nto the tree of life, explaining its superior generalization.\n2. T REEOFLIFE-10M\nRecent work has shown that data quality and diversity are\ncritical when training CLIP models [24, 26, 57]. We cu-\nrate T REEOFLIFE-10M, the most diverse large-scale pub-\nlic ML-ready dataset for computer vision models in biology.\n2.1. Images\nThe largest ML-ready biology image dataset is iNat21 [86]\nwith 2.7M images of 10K species. Despite this class\nbreadth compared to popular general-domain datasets like\nImageNet-1K [70], 10K species is limited for biology. The\nInternational Union for Conservation of Nature (IUCN) re-\nported over 2M total described species in 2022, with over\n10K bird species and over 10K reptile species alone [44].\niNat21’s species diversity limits its potential for training a\nfoundation model for the entire tree of life.\nMotivated to find high-quality biology images with a fo-\ncus on species diversity, we turn to the Encyclopedia of Life\nproject (EOL; eol.org). EOL collaborates with a variety\nof institutions to gather and label millions of images. We\ndownload 6.6M images from EOL and expand our dataset\nto cover an additional 440K taxa.\nSpecies are not evenly distributed among the differ-\nent subtrees in the tree of life; insects (of the class In-\nsecta with 1M+ species), birds (of the class Aves with\n10K+ species) and reptiles (of the class Reptilia with 10K+\nspecies) are examples of highly diverse subtrees with many\nmore species. To help a foundation model learn extremely\nfine-grained visual representations for insects, we also in-\ncorporate B IOSCAN -1M [28], a recent dataset of 1M lab\nimages of insects, covering 494 different families.3Fur-\nthermore, B IOSCAN -1M contains labimages, rather than in\nsitu images like iNat21, diversifying the image distribution.\n2.2. Metadata & Aggregation\nThe T REEOFLIFE-10M dataset integrates iNat21 (training\nsplit), our curated EOL dataset, and B IOSCAN -1M by ag-\ngregating the images and canonicalizing the labels. This\nis a highly non-trivial task because taxonomic hierarchies\nare notoriously noisy and rarely consistent between sources\n[4, 31, 36, 52, 63], likely contributing to the prior lack of\n3We note that B IOSCAN -1M’s label granularity may still be limited for\ninsects. 98.6% of B IOSCAN -1M’s images are labeled to the family level\nbut only 22.5% and 7.5% of the images have genus or species indicated,\nrespectively. Lack of label granularity is an inherent challenge.\n \nArthropoda Tracheophyta Chordata\nMolluscaB a s i d i o m y c o t aAs c o m y c o t a\nC n i d a r i aEc h i n o d er m at a\nB r y o p h y t a\nAn n e l i d aPr o t o z o aR h o d o p h y t aInsecta\nAr a c h n i d aM al ac o s t r ac aMagnoliopsida\nLiliopsidaP o l y p o d i o p s i d aP i n o p s i d aAves\nR e p t i l i aM a m m a l i aAc t i n o p t e r y g i i\nAm p h i b i aT el eo s t ei\nG a s t r o p o d a\nB i v a l v i aAg ar i c o m y c et es\nP u c c i n i o m y c e t e sDiptera\nLepidopteraH y m e n o p t e r a\nColeoptera\nHe m i p t e r a\nO d o n a t aO r t h o p t e r a\nMantodea P s o c o d e a\nT r i c h o p t e r a N e u r o p t e r aAsteralesL a m i a l e sC a r y o p h y l l a l e s\nF a b a l e s\nAs p a r a g a l e s\nR o s a l e sPo al es E r i c a l e sM a l p i g h i a l e s\nGentianales\nR a n u n c u l a l e s\nM y r t al esB r a s s i c a l e s M a l v a l e s Ap i a l e s\nSa p i n d a l e s\nSa x i f r a g a l e s\nS o l a n a l e sB o r a g i n a l e s F a g a l e s\nP r o t e a l e s\nL i l i a l e s\nAl i s m a t a l e sG e r a n i a l e s\nO x a l i d a l e s\nA r e c a l e s\nSa n t a l a l e sPas s er i fo r m esC h a r a d r i i f o r m e s\nAn s e r i f o r m e sA c c i p i t r i f o r m es\nSq u a m a t a'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': '',

'proposed\_solution': '',

'paper\_limitations': '',

'source': 'O x a l i d a l e s\nA r e c a l e s\nSa n t a l a l e sPas s er i fo r m esC h a r a d r i i f o r m e s\nAn s e r i f o r m e sA c c i p i t r i f o r m es\nSq u a m a t a\nT e s t u d i n e sC a r n i v o r aP e r c i f o r m e s\nA g a r i c a l e sC e c i d o m y i i d a e\nC h i r o n o m i d a eP h o r i d a eSc i a r i d a e\nC e r a t o p o g o n i d a e\nP s y c h o d i d a eD o l i c h o p o d i d a e\nS y r p h i d a e\nSp h aer o c er i d aeD r o s o p h i l i d a e T a c h i n i d a e As i l i d ae\nM u s c i d a eHybotidae\nT i p u l i d a e\nAn t h o m y i i d a e\nE m p i d i d a e\nE p h y d r i d a e\nSc a t o p s i d a eu n k n o w nC al l i p h o r i d ae M i l i c h i i d a eB i b i o n i d a e\nU l i d i i d a e\nN y m p h a l i d a e\nG e o m e t r i d a eE r e b i d a e\nN o c t u i d a eL y c a e n i d a e C r a m b i d a eH e s p e r i i d a e\nT o r t r i c i d ae\nP i e r i d a e\nSp h i n g i d a eP a p i l i o n i d a e Py r al i d aeN o t o d o n t i d a e\nG e l e c h i i d a e\nSa t u r n i i d a e\nO e c o p h o r i d a eT i n e i d a e\nN o l i d a eC o s m o p t e r i g i d a e Pt er o p h o r i d ae\nZ y g a e n i d a e\nN e p t i c u l i d a e\nSe s i i d a e\nCo l eo p h o r i d aeP s y c h i d a e C o s s i d a eH e p i a l i d a e U r an i i d ae\nAd e l i d a eF o r m i c i d a e Ap i d a e\nI c h n e u m o n i d a eB r ac o n i d aeSc el i o n i d ae Ve s p i d a e\nC r a b r o n i d a e\nH a l i c t i d a e\nE u m e n i d a eP l a t y g a s t r i d a e B e t h y l i d a e D i a p r i i d a e Po m p i l i d ae\nSphecidae\nAn d r en i d ae\nC y n i p i d a eE u l o p h i d a e M y m a r i d a e F i g i t i d a e\nC o l l et i d ae\nM u t i l l i d a eu n k n o w nSc o l i i d a e\nC e r a m b y c i d a e\nCar ab i d aeC h r y s o m e l i d a e\nCu r c u l i o n i d aeStaphylinidae\nC o c c i n e l l i d a e\nC e t o n i i d a e\nD y t i s c i d aeM e l o i d a eM e l o l o n t h i d a e L u c a n i d a eL a m p y r i d a e\nE l a t e r i d a e Si l p h i d a e R u t e l i d a e\nN i t i d u l i d a e\nC l e r i d a e\nM o r d e l l i d a e\nPt i n i d aeH i s t e r i d a e\nM i r i d a e\nP e n t a t o m i d a eC i c a d i d a e\nC o r e i d a eAp h i d i d a e\nA l ey r o d i d aeR e d u v i i d a e\nM e m b r a c i d a eAc r i d i d a eT e t t i g o n i i d a eAs t e r a c e a eC a m p a n u l a c e a e\nG o o d e n i a c e a eL a m i a c e a e\nPl an t ag i n ac eaeO r o b a n c h a c e a e\nP h r y m a c e a e\nL e n t i b u l a r i a c e a eG e s n e r i a c e a eC a c t a c e a e\nP o l y g o n a c e a eC a r y o p h y l l a c e a e\nA i z o a c e a e\nD r o s e r a c e a eN y c t a g i n a c e a e M o n t i a c e a e\nF ab ac eae\nO r c h i d a c e a eAs p a r a g a c e a e\nI r i d a c e a eAs p h o d e l a c e a e\nR o s a c e a eR h a m n a c e a e U r t i c a c e a eB r o m e l i a c e a e P r i m u l a c e a e P o l e m o n i a c e a e V i o l a c e a e Sa l i c a c e a e\nAp o c y n a c e a e R u b i a c e a eT y r a n n i d a e\nP a r u l i d a eF r i n g i l l i d a eH i r u n d i n i d a e\nT u r d i d a e\nCar d i n al i d aeM o t a c i l l i d a e\nP a r i d a e\nM e l i p h a g i d a e\nT r o g l o d y t i d a eSt u r n i d a e\nN e c t a r i n i i d a e\nP y c n o n o t i d a e\nF u r n ar i i d ae\nE s t r i l d i d a eLaniidae P l o c e i d a e Vi r e o n i d a e M i m i d a e\nC o l u b r i d a eFigure 2. Treemap of the 108phyla in T REEOFLIFE-10M. Dif-\nferent colors are different phyla; nested boxes represent classes,\norders, and families. Box size, not number of inner boxes, repre-\nsents relative number of samples.\nimage datasets large enough to train a foundation-scale vi-\nsion model for the entire tree of life. We carefully unify and\nbackfill taxonomic hierarchies from EOL, the Integrated\nTaxonomic Information System (ITIS) [43], and iNaturalist\nwith special consideration for the existence of homonyms\n(genus-species labels shared among higher-order taxa). For\nmore information on this process, the challenges, our solu-'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'TREEOFLIFE-10M dataset contains over 10M images across more than 450K unique taxonomic names. The datasets TREEOFLIFE-10M and RARE SPECIES are available on Hugging Face under a public domain waiver, including CSVs with image metadata and links to the primary sources, accompanied by a GitHub repository with scripts to generate the datasets. BIOCLIP is initialized from OpenAI’s public CLIP checkpoint and continually pre-trained on TREEOFLIFE-10M with CLIP’s multimodal contrastive learning objective.',

'limitation\_of\_sota': 'One of the most salient differences for the biology domain is its rich label space, with 2M+ recorded species as of 2022, connected in a hierarchical taxonomy, posing a challenge for training a foundation model that can achieve satisfactory coverage and generalization.',

'proposed\_solution': 'The proposed solution involves using BIOCLIP, initialized from OpenAI’s public CLIP checkpoint and continually pre-trained on TREEOFLIFE-10M with CLIP’s multimodal contrastive learning objective, to address the challenge of a rich label space in the biology domain.',

'paper\_limitations': '',

'source': 'with special consideration for the existence of homonyms\n(genus-species labels shared among higher-order taxa). For\nmore information on this process, the challenges, our solu-\ntions, and remaining issues, see Appendix C.\n2.3. Release & Statistics\nTab. 1 presents dataset statistics: T REEOFLIFE-10M has\nover 10M images across more than 450K unique taxonomic\nnames. Fig. 2 shows the distribution of images by phyla and\nthe respective lower-rank taxa (order through family).\nOur curated training and test datasets (T REEOFLIFE-\n10M and R ARE SPECIES , described in §4.2) are avail-\nable on Hugging Face (with DOIs) under a public domain\nwaiver, to the extent primary source licenses allow. This\nincludes CSVs with image metadata and links to the pri-\nmary sources, accompanied by a GitHub repository with the\nscripts to generate the datasets.4\n3. Modeling\nBIOCLIP is initialized from OpenAI’s public CLIP check-\npoint and continually pre-trained on T REEOFLIFE-10M\nwith CLIP’s multimodal contrastive learning objective.\n3.1. Why CLIP?\nCompared with general domain computer vision tasks, one\nof the most salient differences for the biology domain is\nits rich label space. Not only are the taxon labels large in\nquantity (there are 2M+ recorded species as of 2022 [44]),\nbut they are also connected with each other in a hierarchical\ntaxonomy. This is a challenge for training a good founda-\ntion model that can achieve satisfactory coverage and gen-\neralization. Despite this, the intricate structure in the label\n4We encourage future work to cite iNat21 [86], B IOSCAN -1M [28] and\nto appropriately attribute images from EOL based on their licenses if citing\nTREEOFLIFE-10M.'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'The key realization of this work is that the multimodal contrastive learning objective used in CLIP can be repurposed for leveraging the hierarchical structure of the label space, significantly improving generalization for learning hierarchical representations conforming to a taxonomy.',

'limitation\_of\_sota': 'Many vision foundation models, such as ResNet and Swin Transformer, adopt a supervised classification objective and directly learn the mapping from input images to class indices, treating each class label as a distinct symbol and neglecting their relationships.',

'proposed\_solution': 'This work proposes repurposing CLIP’s multimodal contrastive learning objective for learning hierarchical representations conforming to a taxonomy. The approach involves training two uni-modal embedding models, a vision encoder and a text encoder, to maximize feature similarity between positive (image, text) pairs and minimize feature similarity between negative (image, text) pairs, embedding the taxonomic hierarchy into a dense label space.',

'paper\_limitations': '',

'source': 'Dataset Description Images Unique Classes\niNat21 Citizen scientist labeled image dataset from iNaturalist for fine-grained classification. 2.7M 10,000\nBIOSCAN -1M Expert labeled image dataset of insects for classification. 1.1M 7,831\nEOLA new dataset with citizen scientist images sourced from Encyclopedia of Life and\ntaxonomic labels standardized by us.6.6M 448 ,910\nTREEOFLIFE-10M Largest-to-date ML-ready dataset of biology images with taxonomic labels. 10.4M 454 ,103\nTable 1. Training data sources used in T REEOFLIFE-10M. We integrate and canonicalize taxonomic labels across the sources (§2.2).\nName Description Examples Classes LabelsAnimalsBirds 525 Scraped dataset of bird images from web search. [68] 89,885 525 Taxonomic\nPlankton Expert-labeled in situ images of plankton [35]. 4,080 102 Mixed\nInsects Expert and volunteer-labeled in-the-wild citizen science images of insects [74]. 4,680 117 Scientific\nInsects 2 Mixed common and scientific name classification for insect pests [91]. 4,080 102 MixedPlants & FungiPlantNet Citizen science species-labeled plant images, some drawings [27]. 1,000 25 Scientific\nFungi Expert-labeled images of Danish fungi [66]. 1,000 25 Scientific\nPlantVillage Museum-style leaf specimens labeled with common names [25]. 1,520 38 Common\nMedicinal Leaf Species classification of leaves from mature, healthy medicinal plants [71]. 1,040 26 Scientific\nPlantDoc 17 diseases for 13 plant species [76]. 1,080 27 Common\nRARE SPECIESSubset of species in the IUCN Red List categories: Near Threatened through\nExtinct in the Wild (iucnredlist.org).12,000 400 Taxonomic\nTable 2. Datasets used for evaluation. All tasks are classification evaluated with Top-1 accuracy.\nspace, accumulated through centuries of biology research,\nprovides very rich signal for learning better generalization.\nIntuitively, if the label space’s structure is successfully en-\ncoded in a foundation model, even if the model has not seen\na certain species, it will likely have learned a good repre-\nsentation for that species’ corresponding genus or family.\nSuch a hierarchical representation serves as a strong prior\nto enable few-shot or even zero-shot learning of new taxa.\nMany vision foundation models, such as ResNet [33]\nand Swin Transformer [48], adopt a supervised classifica-\ntion objective and directly learn the mapping from input im-\nages to class indices. As a result, each class label is treated\nas a distinct symbol, and their relationships are neglected.\nA key realization of our work is that the multimodal con-\ntrastive learning objective used in CLIP can be repurposed\nfor leveraging the hierarchical structure of the label space.\nThis is not an obvious choice; after all, T REEOFLIFE-10M\nis largely labeled with class labels and not with free-form\ntext like image captions. The autoregressive text encoder\nnaturally embeds the taxonomic hierarchy into a dense la-\nbel space by conditioning later taxonomic rank representa-\ntions on higher ranks (Fig. 1). While hierarchical classi-\nfication [9, 11, 96] can also leverage taxonomy, we empiri-\ncally show that CLIP-style contrastive learning significantly\nimproves generalization (§4.4). We note that repurposing\nCLIP’s multimodal contrastive learning objective for learn-\ning hierarchical representations conforming to a taxonomy\nis a novel and non-trivial technical contribution.\nCLIP trains two uni-modal embedding models, a vision\nencoder and a text encoder, to (1) maximize feature sim-Text Type Example\nCommon black-billed magpie\nScientific Pica hudsonia\nTaxonomicAnimalia Chordata Aves Passeriformes\nCorvidae Pica hudsonia\nScientific + CommonPica hudsonia with common name\nblack-billed magpie\nTaxonomic + CommonAnimalia Chordata Aves Passeriformes\nCorvidae Pica hudsonia with common\nname black-billed magpie\nTable 3. Text types considered in the training of B IOCLIP.\nilarity between positive (image, text) pairs and (2) mini-\nmize feature similarity between negative (image, text) pairs,'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': '',

'limitation\_of\_sota': '',

'proposed\_solution': 'In the context of training B IOCLIP, the proposed solution involves two main objectives: (1) maximize feature similarity between positive (image, text) pairs and (2) minimize feature similarity between negative (image, text) pairs. Positive pairs are from the training data, and negative pairs are all other possible (image, text) pairings in a batch. After training, CLIP’s encoder models embed individual instances of their respective modalities into a shared feature space. The text input to CLIP is formatted to incorporate the taxonomic structure by considering the taxonomic name, which concatenates all labels from root to leaf into a single string for each species.',

'paper\_limitations': '',

'source': 'Table 3. Text types considered in the training of B IOCLIP.\nilarity between positive (image, text) pairs and (2) mini-\nmize feature similarity between negative (image, text) pairs,\nwhere positive pairs are from the training data and negative\npairs are all other possible (image, text) pairings in a batch.\nAfter training, CLIP’s encoder models embed individual in-\nstances of their respective modalities into a shared feature\nspace. Next, we discuss formatting the text input to CLIP\nto incorporate the taxonomic structure.\n3.2. Text Types\nAn advantage of CLIP is that the text encoder accepts free-\nform text. In biology, unlike other classification tasks, class\nnames are diversely formatted. We consider the following:\nTaxonomic name. A standard seven-level biology taxon-\nomy from higher to lower level is kingdom, phylum, class,\norder, family, genus and species. For each species, we “flat-\nten” the taxonomy by concatenating all labels from root to\nleaf into a single string, which we call the taxonomic name .'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'The mixed text type training strategy retains the generalization benefits of taxonomic names while providing more flexibility in using other names at inference time. BIOCLIP, trained on TREEOFLIFE-10M, shows improved performance over general vision models and demonstrates effectiveness in a variety of biologically-relevant classification tasks across all four multi-celled kingdoms in the tree of life.',

'limitation\_of\_sota': "Existing models like OpenAI's CLIP and OpenCLIP, though effective with common names due to their pervasive presence in training data, lack the flexibility and specificity required for scientific and taxonomic name-based inference in biological classification tasks.",

'proposed\_solution': "BIOCLIP is proposed with a mixed text type training strategy, utilizing a combination of scientific names, common names, and other text types for training. This approach is aimed at improving flexibility and performance in biological image classification tasks, especially when only one type of label is available. The model is pre-trained on the TREEOFLIFE-10M dataset with OpenAI's CLIP weights and further evaluated on various classification tasks to demonstrate its efficacy.",

'paper\_limitations': 'The paper does not explicitly mention its limitations, but potential limitations could include the reliance on the availability and quality of taxonomic and common name data, the generalizability of the model to unseen species, and the computational resources required for training on large datasets like TREEOFLIFE-10M.',

'source': 'Scientific name. Scientific names are composed of genus\nand species (e.g., Pica hudsonia ).\nCommon name. Taxonomy categories are usually Latin,\nwhich is not often seen in generalist image-text pre-training\ndatasets. Instead, the common name, such as “black-billed\nmagpie,” is more widespread. Note that common names\nmay not have a 1-to-1 mapping to taxa: A single species\nmay have multiple common names, or the same common\nname may refer to multiple species.\nFor certain downstream use cases of B IOCLIP, it might\nbe the case that only one type of label, e.g., scientific names,\nis available. To improve the flexibility at inference time, we\npropose a mixed text type training strategy: at each training\nstep, we pair each input image with a text randomly sam-\npled from all of its available text types (shown in Tab. 3).\nWe empirically show that this simple strategy retains the\ngeneralization benefits of taxonomic names while provid-\ning more flexibility in using other names at inference time\n(§4.3). The final text input to CLIP is the name in the stan-\ndard CLIP template, e.g., “a photo of Pica hudsonia ”.\n4. Experiments\nWe train B IOCLIP on T REEOFLIFE-10M, compare B IO-\nCLIP to general vision models and investigate how our\nmodeling choices affect B IOCLIP’s performance.\n4.1. Training and Evaluation Details\nTo train B IOCLIP, we initialize from OpenAI’s CLIP\nweights [69] with a ViT-B/16 vision transformer [22] image\nencoder and a 77-token causal autoregressive transformer\ntext encoder. We continue pre-training on T REEOFLIFE-\n10M for 100 epochs with a cosine learning rate schedule\n[49]. We train on 8NVIDIA A100-80GB GPUs over 2\nnodes with a global batch size of 32,768. We also train\na baseline model on only the iNat21 dataset and multiple\nablation models on 1M examples randomly sampled from\nTREEOFLIFE-10M (Secs. 4.3 and 4.4), following the same\nprocedure for B IOCLIP except with a smaller global batch\nsize of 16,384on4NVIDIA A100 GPUs on 1node. All\nhyperparameters and training details are in Appendix D and\ntraining and evaluation code is publicly available.\nWe evaluate on 10 different classification tasks : the 8\nbiologically-relevant tasks from Meta-Album [84], Birds\n525[68] and our new RARE SPECIES task (described in\n§4.2). Meta-Album is a dataset collection for meta-learning,\nencompassing various subjects. Specifically, we use the\nPlankton, Insects, Insects 2, PlantNet, Fungi, PlantVillage,\nMedicinal Leaf, and PlantDoc datasets. Our classification\ntasks cover all four multi-celled kingdoms in the tree of\nlife (animals, plants, fungi, and protists) and have a di-\nverse image distribution (photographs, microscope images,\ndrawings, and museum specimens). Tab. 2 summarizes thedatasets; they comprise a variety of label types from full\ntaxonomic names to only scientific or common name.\nForzero-shot learning , we follow the same procedure as\nCLIP. For few-shot learning , we follow SimpleShot [90]\nand use a nearest-centroid classifier. For k-shot learning,\nwe first randomly sample kexamples for each class and ob-\ntain the image embedding from the visual encoder of the\npre-trained models. We then compute the average feature\nvector of the kembeddings as the centroid for each class.\nAll the examples left in the dataset are used for testing. Af-\nter applying mean subtraction and L2-normalization to each\ncentroid and test feature vector, we choose the class with\nthe nearest centroid to the test vector as the prediction. We\nrepeat each few-shot experiment 5times with different ran-\ndom seeds and report the mean accuracy in Tab. 4. Results\nwith standard deviations are reported in Appendix E.\nWe compare B IOCLIP with the original OpenAI CLIP\n[69] and OpenCLIP [42] trained on LAION-400M [73]. In-\ntuitively, common names of organisms are most pervasive in\nthe training data of CLIP and OpenCLIP and these models\nwork best with common names. This is also confirmed in\nour preliminary tests. Therefore, we use common names as'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': "BIOCLIP can leverage taxonomic names and is tested for generalization to unseen taxa, specifically in the context of classifying rare species. A new evaluation task, RARE SPECIES, is introduced to assess BIOCLIP's performance in classifying species classified as Near Threatened, Vulnerable, Endangered, Critically Endangered, or Extinct in the Wild by the IUCN Red List. Approximately 25K species on the IUCN Red List are considered, and 400 species represented by at least 30 images in the EOL dataset are selected for this task.",

'limitation\_of\_sota': 'Existing models like CLIP and OpenCLIP work best with common names and have limitations in leveraging taxonomic names for classification tasks.',

'proposed\_solution': 'BIOCLIP is designed to leverage both taxonomic and common names for classification tasks, aiming to generalize well to unseen taxa without the need for re-training for every new species. This approach is particularly useful for classifying rare species, an important application in the context of global conservation efforts.',

'paper\_limitations': 'The paper does not specify the limitations of the proposed BIOCLIP solution directly, but it implies that the challenge of classifying rare species is significant due to the lack of a diverse, publicly available dataset for rare species classification.',

'source': 'the training data of CLIP and OpenCLIP and these models\nwork best with common names. This is also confirmed in\nour preliminary tests. Therefore, we use common names as\nclass labels for CLIP and OpenCLIP by default unless un-\navailable for a dataset. B IOCLIP can leverage taxonomic\nnames, so we use taxonomic+common names by default.\nHowever, as noted in Tab. 2, the test datasets come in a\nvariety of labels. Whenever the preferred label type is not\navailable, we use labels that come with the dataset. We also\ncompare to an ImageNet-21K [21] pre-trained model [78]\nand DINO [15] for few-shot classification.\n4.2. Can BIOCLIP Generalize to Unseen Taxa?\nTaxonomic names are added, removed, and changed as\nbiologists discover and classify new and existing species.\nBIOCLIP should generalize to unseen taxonomic names to\navoid re-training for every new species. To empirically an-\nswer whether B IOCLIP generalizes well to unseen taxa, we\nintroduce a new evaluation task that is both biologically and\nempirically motivated: RARE SPECIES .\nClassifying “rare” species is an important and challeng-\ning computer vision application in biology, particularly in\nthe context of global conservation efforts [83]. To the best\nof our knowledge, there is no diverse, publicly available\nrare species classification dataset. Recently published work\n[47, 56] lack species diversity with only a dozen classes.\nWe aim to fill this gap, collecting all ≈25K species on the\nIUCN Red List (iucnredlist.org) classified5as Near Threat-\nened, Vulnerable, Endangered, Critically Endangered, or\nExtinct in the Wild. We select 400such species represented\nby at least 30images in our EOL dataset, then remove\n5IUCN has classified 150 ,388 species and generally updates their list\ntwice per year (IUCN Update Schedule). The classifications used for this\ndataset are current as of July 13, 2023.'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': "BIOCLIP substantially outperforms both baseline CLIP models and the iNat21-trained CLIP model at zero-shot classification, especially on unseen taxa. BIOCLIP's strong zero-shot performance on a broad and diverse set of tasks is attributed to the broad and diverse classes in TREEOFLIFE-10M. The study demonstrates BIOCLIP's out-of-distribution generalization to unseen taxa and its potential applications in addressing the ongoing biodiversity crisis.",

'limitation\_of\_sota': 'Existing models struggle with zero-shot classification, particularly in accurately identifying rare or unseen species.',

'proposed\_solution': 'BIOCLIP, trained on a diverse dataset from TREEOFLIFE-10M, shows significant improvement in zero-shot classification accuracy across a variety of taxa, including rare species. This model leverages the diversity of the training data to achieve broad and useful image representations for classification tasks.',

'paper\_limitations': 'The study was conducted with computational constraints, limiting the training to only a 10% subset of TREEOFLIFE-10M. Additionally, the impact of different text types on generalization was explored, but the diversity of captions and their specific contributions to the strength of the vision models could be further detailed.',

'source': 'Animals Plants & Fungi\nModel\nBirds 525\nPlankton\nInsects\nInsects 2\nPlantNet\nFungi\nPlantVillage\nMed. Leaf\nPlantDoc\nRare Species Mean ( ∆)\nRandom Guessing 0.2 1 .2 1 .0 1 .0 4 .0 4 .0 2 .6 4 .0 3 .7 0 .3 2 .2\nZero-Shot Classification\nCLIP 49.9 3 .2 9 .1 9 .8 58 .5 10 .2 5 .4 15 .9 26 .1 31 .8 21 .9 –\nOpenCLIP 54.7 2 .2 6 .5 9 .6 50 .2 5 .7 8 .0 12 .4 25 .8 29 .8 20 .4−1.5\nBIOCLIP 72.1 6 .1 34 .8 20 .4 91 .440.724.4 38 .6 28 .4 38 .0 39 .4+17 .5\n– iNat21 Only 56.1 2 .6 30 .7 11 .5 88 .243.018.4 25 .6 20 .5 21 .3 31 .7 +9 .8\nOne-Shot Classification\nCLIP 43.7 25 .1 21 .6 13 .7 42 .1 17 .2 49 .7 70 .1 24 .8 28 .5 33 .6 –\nOpenCLIP 53.7 32 .3 23 .2 14 .3 45 .1 18 .4 53 .6 71 .2 26 .8 29 .2 36 .7 +3 .1\nSupervised-IN21K 60.2 22 .9 14 .7 14 .4 46 .7 16 .962.358.6 27 .7 28 .0 35 .2 +1 .6\nDINO 40.537.023.5 16 .4 30 .7 20 .0 60 .0 79 .2 23 .7 31 .0 36 .2 +2 .6\nBIOCLIP 71.8 30 .657.4 20 .464.540.358.884.3 30 .7 44 .9 50 .3+16 .7\n– iNat21 Only 74.829.6 53 .9 19 .767.435.5 55 .2 75 .1 27 .8 36 .9 47 .5 +13 .9\nFive-Shot Classification\nCLIP 73.5 41 .2 39 .9 24 .6 65 .2 27 .9 71 .8 89 .7 35 .2 46 .0 51 .5 –\nOpenCLIP 81.9 52 .5 42 .6 25 .0 68 .0 30 .6 77 .8 91 .3 42 .0 47 .4 55 .9 +4 .4\nSupervised-IN21K 83.9 39 .2 32 .0 25 .4 70 .9 30 .982.482.3 44 .7 47 .3 53 .9 +2 .4\nDINO 70.856.946.3 28 .6 50 .3 34 .1 82 .1 94 .9 40 .3 50 .1 55 .4 +3 .9\nBIOCLIP 90.0 49 .377.8 33 .6 85 .6 62 .380.995.9 47 .5 65 .7 68 .8+17 .3\n– iNat21 Only 90.148.2 73 .7 32 .1 84 .7 55 .6 77 .2 93 .5 41 .0 55 .6 65 .1 +13 .6\nTable 4. Zero-, one- and five-shot classification top-1 accuracy for different models. Bold indicates best accuracy. All models use the\nsame ViT-B/16 architecture. “iNat21 Only” follows the same procedure as B IOCLIP but uses iNat21 instead of T REEOFLIFE-10M. ∆\ndenotes the difference in mean accuracy with CLIP. Supervised-IN21K [78] and DINO [15] are vision-only models and cannot do zero-shot\nclassification.\nDataset Train ↓Test→Com Sci Tax Sci+Com Tax+Com\nTOL-1MCom 24.9 9 .5 10 .8 22 .3 21 .0\nSci 11.0 22 .3 4 .5 21 .5 8 .0\nTax 11.8 10 .1 26 .6 16 .0 24 .8\nSci+Com 24.5 12 .9 12 .6 28 .0 24 .9\nTax+Com 20.5 8 .0 19 .7 24 .0 30 .4\nMixture 26.1 24 .9 26 .7 29 .5 30 .9\niNat21-2.7M Mixture 20.4 14 .7 15 .6 20 .9 21 .3\nTOL-10M Mixture 31.6 30 .1 34 .1 37 .0 38 .0\nTable 5. Zero-shot accuracy on species not seen during training\n(RARE SPECIES task). Different rows and columns indicate dif-\nferent text types during training and test time, respectively. Blue\nindicates best accuracy and Orange indicates second-best. Using\nthe taxonomic name over the scientific name always improves ac-\ncuracy ( 22.3→26.6and28.0→30.4). The final rows use the full\niNat21 dataset and T REEOFLIFE-10M for reference.\nthem from T REEOFLIFE-10M, creating an unseen RARE\nSPECIES test set with 30images per species. This dataset\ndemonstrates (1) B IOCLIP’s out-of-distribution generaliza-\ntion to unseen taxa, (2) B IOCLIP’s potential applications,\nand (3) provides a crucial dataset for the community to ad-\ndress the ongoing biodiversity crisis.Results . Tab. 4 shows that B IOCLIP substantially outper-\nforms both baseline CLIP models, as well as the iNat21-\ntrained CLIP model, at zero-shot classification, especially\non unseen taxa (see the “Rare Species” column). We at-\ntribute B IOCLIP’s strong zero-shot performance on this\nbroad and diverse set of tasks to the broad and diverse\nclasses in T REEOFLIFE-10M. We explore how data diver-\nsity leads to broadly useful image representations in §4.3.\n4.3. How Do Text Types Affect Generalization?\nWe investigate how different text types affect zero-shot\ngeneralization by training B IOCLIP on a 10% subset\nof T REEOFLIFE-10M (10% due to computational con-\nstraints). We use our Rare Species dataset because the test\nclasses have every text type, and all species are excluded\nfrom training, making it ideal for testing generalization to\nunseen taxa. Prior works find that the diversity of captions\nmakes stronger vision models [57] and randomly use one of'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'The zero-shot ablation results show several salient observations. First, using a mixed text type strategy for training, which involves using different captions for each image, significantly affects the performance.',

'limitation\_of\_sota': '',

'proposed\_solution': 'The proposed solution involves using a mixed text type strategy for training, which includes using different captions for each image rather than a single fixed caption. This approach is based on prior works that find the diversity of captions makes stronger vision models.',

'paper\_limitations': '',

'source': 'from training, making it ideal for testing generalization to\nunseen taxa. Prior works find that the diversity of captions\nmakes stronger vision models [57] and randomly use one of\nfive different captions for each image during training rather\nthan a single fixed caption [72]. Similarly, we use a mixed\ntext type strategy (§3.2). How does that affect performance?\nResults . The zero-shot ablation results are in Tab. 5;\nthere are several salient observations. First, using taxo-'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': "The study demonstrates the effectiveness of the CLIP objective for pre-training on a labeled image dataset, particularly in few-shot settings, outperforming both cross-entropy and hierarchical cross-entropy baselines significantly. It also highlights the importance of incorporating taxonomic structure for generalization and shows that using mixed text types for training enhances performance across different text types at test time. Additionally, BIOCLIP's ability to classify beyond species, such as plant diseases, in both zero-shot and few-shot settings is confirmed, indicating its utility for tasks requiring classification of both species and disease. The intrinsic evaluation reveals that BIOCLIP learns useful visual representations that align with taxonomic hierarchies.",

'limitation\_of\_sota': 'Existing models struggle with generalization across different text types and may not effectively incorporate taxonomic structures for classification tasks. Additionally, they may not be optimized for few-shot learning scenarios or capable of classifying beyond species level, such as identifying diseases based on visual symptoms.',

'proposed\_solution': 'The study proposes using the CLIP objective for pre-training on the TREEOFLIFE-1M dataset, which outperforms traditional cross-entropy and hierarchical cross-entropy methods in few-shot classification tasks. It emphasizes the importance of mixed text type pre-training for maintaining generalization benefits and flexibility across different text types. Furthermore, the study introduces BIOCLIP, which leverages the CLIP objective to classify both species and diseases, demonstrating its effectiveness in few-shot settings and its ability to learn taxonomic hierarchies.',

'paper\_limitations': 'The paper does not specify its limitations directly, but potential limitations could include the reliance on the specific dataset (TREEOFLIFE-1M) for training, which may not be representative of all biodiversity. Additionally, the effectiveness of the CLIP objective and BIOCLIP in broader or more complex classification tasks beyond the scope of the study remains to be fully explored.',

'source': 'Objective Mean 1-Shot Mean 5-shot\nCross-entropy 16.5 26 .2\nHier. cross-entropy 19.3 30 .5\nCLIP 44.7 63 .8\nTable 6. One- and five-shot classification top-1 accuracy for dif-\nferent pre-training objectives on T REEOFLIFE-1M. Results are\nmacro-averaged over all the test sets in Tab. 4.\nnomic+common names yields the strongest performance,\nshowing the importance of incorporating the taxonomic\nstructure for generalization. Second, when only using a sin-\ngle text type for training, performance degrades substan-\ntially when a different text type is used at test time. Us-\ning mixed text types for training yields consistently strong\nperformance across all text types during testing. These re-\nsults indicate that mixed text type pre-training largely re-\ntains the generalization benefits of using taxonomic names\nwhile also providing flexibility of different text types for in-\nference, an important property for a foundation model that\nmay be used for diverse downstream tasks. Finally, using\n1M examples from T REEOFLIFE-10M outperforms using\n2.7M examples from iNat21, further confirming the impor-\ntance of the added data diversity from T REEOFLIFE-10M.\n4.4. Is the CLIP Objective Necessary?\nUsing the CLIP objective to pre-train on a labeled image\ndataset is an unintuitive decision (Goyal et al. [29] fine-\ntune using the CLIP objective, but do not pretrain). We\njustify our choice by training two ViT-B/16 models on\nTREEOFLIFE-1M using a cross-entropy classification loss\nand a multitask hierarchical variant, then compare them\nagainst the CLIP objective under the few-shot setting. The\nmultitask hierarchical training objective is to predict the la-\nbels for kingdom, phylum, etc., down to species, using cross\nentropy for each level of the taxonomy, then summing those\nlosses [11]. Pseudo-code is provided in Listing 1.\nResults . We evaluate each model on the same set of 10\ntasks but only in the one-shot and five-shot settings because\nnon-CLIP models cannot do zero-shot classification. We re-\nport mean accuracy in Tab. 6. The hierarchical classification\nmodel outperforms simple classification and is comparable\nto the CLIP baseline (see Tab. 4). However, the CLIP ob-\njective massively outperforms both baselines and strongly\njustifies our repurposing of the CLIP objective.\n4.5. Can BIOCLIP Classify More Than Species?\nBIOCLIP is trained on a (contrastive) species-classification\nobjective, which might limit its use beyond species clas-\nsification. We consider plant diagnosis with the PlantVil-\nlage and PlantDoc datasets, which require classifying both\nspecies and disease (if any). Large-scale data labeling is\nexpensive, but biologists always label several instances for\nfield guides or museum collections. Few-shot classificationis thus an ideal setting for this sort of task transfer.\nResults . B IOCLIP outperforms baselines at classifying\nplant diseases based on visual symptoms, in both zero-\nshot and few-shot settings (see PlantVillage and PlantDoc\nin Tab. 4). While Radford et al. [69] find that CLIP one-\nshot and two-shot classification is often worse than zero-\nshot (because few-shot settings cannot use the semantic in-\nformation in the class name), B IOCLIP has learned useful\nvisual representations that are useful even with only one\nlabeled example: B IOCLIP’s mean one-shot accuracy is\n9.1% higher than zero-shot accuracy.\n4.6. Does BIOCLIP Learn the Hierarchy?\nBIOCLIP demonstrates strong performance in a low-data\nregime on our extrinsic evaluation, but why? We further\nconduct an intrinsic evaluation and visualize B IOCLIP’s\nlearned image representations to shed light on this question\n(Fig. 3). We embed image representations from iNat21’s\nvalidation set (unseen during training) using t-SNE [85] and\ncolor the points by the image’s taxonomic label. For each\nplot, we run t-SNE independently on the subset of examples\nunder the labeled taxonomical rank. Each plot visualizes\none taxonomic hierarchy rank and the top six categories of'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'At higher ranks like kingdom and phylum, both CLIP and BIOCLIP have good separation, but BIOCLIP’s representations are more fine-grained and contain a richer clustering structure. At lower ranks, BIOCLIP produces evidently more separable features, while CLIP’s features are cluttered and lack a clear structure.',

'limitation\_of\_sota': 'Further work finds that dataset diversity and better alignment between the image and caption semantics are more important than dataset size and lead to stronger performance on downstream tasks.',

'proposed\_solution': 'TREEOFLIFE-10M emphasizes the importance of diversity, adding over 440K classes to iNat21’s 10K and leading to stronger zero-shot performance. Domain-specific training often beats general training, but subject-matter experts are often too expensive to hire to label large-scale domain-specific datasets. Image-text training is thus particularly potent because models can learn from noisy image-text pairs.',

'paper\_limitations': '',

'source': 'plot, we run t-SNE independently on the subset of examples\nunder the labeled taxonomical rank. Each plot visualizes\none taxonomic hierarchy rank and the top six categories of\nthe next rank, e.g., the top left plot visualizes the six most\ncommon phyla in the Animalia kingdom. At higher ranks\nlike kingdom (omitted for space) and phylum, both CLIP\nand B IOCLIP have good separation, but B IOCLIP’s repre-\nsentations are more fine-grained and contain a richer clus-\ntering structure. At lower ranks, B IOCLIP produces evi-\ndently more separable features, while CLIP’s features are\ncluttered and lack a clear structure. Appendix F has more\nqualitative results and visuals.\n5. Related Work\nMultimodal foundation model training data . CLIP [69]\ntrained state-of-the-art vision models from noisy, web-scale\n(100M+) image-text datasets using a contrastive objective\nthat is optimized for image retrieval. ALIGN [45] and BA-\nSIC [65] further scaled the number of training examples\nfrom 400M to 6.6B, improving vision representation qual-\nity. However, further work [24, 26, 57, 93, 94] all find\nthatdataset diversity and better alignment between the im-\nage and caption semantics are more important than dataset\nsize and lead to stronger performance on downstream tasks.\nTREEOFLIFE-10M emphasizes the importance of diver-\nsity, adding over 440K classes to iNat21’s 10K and leading\nto stronger zero-shot performance.\nDomain-specific CLIPs . Domain-specific training often\nbeats general training [18, 30], but subject-matter experts\nare often too expensive to hire to label large-scael domain-\nspecific datasets. Image-text training is thus particularly po-\ntent because models can learn from noisy image-text pairs.\nIkezogwo et al. [41] and Lu et al. [50] gathered 1M+ image-'},

{'title': '',

'authors': [],

'abstract': '',

'key\_findings': 'BIOCLIP is a strong fine-grained classifier for biology in both zero- and few-shot settings. Using the entire taxonomic name leads to stronger generalization than other caption types. BIOCLIP-embedded images better match the taxonomic hierarchy.',

'limitation\_of\_sota': 'Prior work applied hierarchies to smaller label spaces and did not efficiently learn visual representations over extensive taxa.',

'proposed\_solution': 'Introducing TREEOFLIFE-10M and BIOCLIP, a large-scale diverse biology image dataset and a foundation model for the tree of life, respectively, which applies hierarchical classification to 454K unique classes through a repurposed CLIP objective.',

'paper\_limitations': "BIOCLIP is fundamentally trained to do classification. Future work will scale up the data and collect richer textual descriptions of species' appearances for extracting fine-grained trait-level representations.",

'source': 'Figure 3. T-SNE visualization of image features, colored by taxonomic labels. B IOCLIP (B) is visualized in the first and third row and\nOpenAI’s CLIP (O) is visualized in the second and fourth rows. B IOCLIP’s features better preserve the hierarchical structure: while both\nBIOCLIP and CLIP cleanly separate the phyla in the Animalia Kingdom (top left), only B IOCLIP successfully separates the orders in the\nInsecta Class (top right) and the families in the Lepidoptera Order (bottom left).\ntext pairs for computational pathology. We gather 10×the\nimages, emphasizing class diversity.\nHierarchy in computer vision . Hierarchy in computer vi-\nsion is well-studied, in part because ImageNet [70] classes\nare from the hierarchical WordNet [55]. Bilal et al.\n[10] study model predictions on ImageNet and find that\nmodel confusion patterns follow the hierarchical class struc-\nture. They incorporate hierarchy into AlexNet’s architec-\nture [46] and improve ImageNet top-1 error by 8% abso-\nlute. Bertinetto et al. [9] measure image classifiers’ mis-\ntake severity and propose alternative objectives that incor-\nporate hierarchy, reducing mistake severity at the expense of\nworsening top-1 accuracy. Zhang et al. [96] propose a con-\ntrastive objective where the hierarchical distance between\nlabels corresponds to the desired distance in the embed-\nding space, and outperform cross-entropy on ImageNet and\niNat17 [88]. We apply hierarchical classification to 454K\nunique classes through a repurposed CLIP objective, while\nprior work applied hierarchies to smaller label spaces.\nComputer vision for biology . Fine-grained classification\nis a classic challenge in computer vision, and biological\nimages are often used to benchmark models. Berg et al.\n[8], Piosenka [68], Wah et al. [89] all use bird species clas-\nsification to evaluate fine-grained classification ability. Bi-\nology tasks are used for contrastive learning frameworks\n[20, 92], weakly supervised object detection [19] and semi-\nsupervised learning methods [34].6. Conclusion\nWe introduce T REEOFLIFE-10M and B IOCLIP, a large-\nscale diverse biology image dataset and a foundation model\nfor the tree of life, respectively. Through extensive evalua-\ntion, we show that B IOCLIP is a strong fine-grained clas-\nsifier for biology in both zero- and few-shot settings. We\ncorroborate our hypothesis that using the entire taxonomic\nname leads to stronger generalization than other caption\ntypes through an ablation on unseen species and by visu-\nalizing B IOCLIP’s representations, finding that B IOCLIP-\nembedded images better match the taxonomic hierarchy.\nAlthough the CLIP objective efficiently learns visual\nrepresentations over 450K taxa, B IOCLIP is fundamentally\ntrained to do classification. Future work will further scale\nup the data, e.g., incorporating more than 100M research-\ngrade images from iNaturalist, and collect richer textual de-\nscriptions of species’ appearances such that B IOCLIP can\nextract fine-grained trait-level representations.\nAcknowledgements\nWe thank the Imageomics team (including Josef Uyeda, Jim\nBalhoff, Dan Rubenstein, Hank Bart, Hilmar Lapp, Sara\nBeery and Dipanjyoti Paul) and the OSU NLP group for\ntheir valuable feedback, the B IOSCAN -1M and iNaturalist\nteams for sharing their data, and Jennifer Hammock at EOL\nfor her help accessing EOL’s images. Our research is sup-'}]