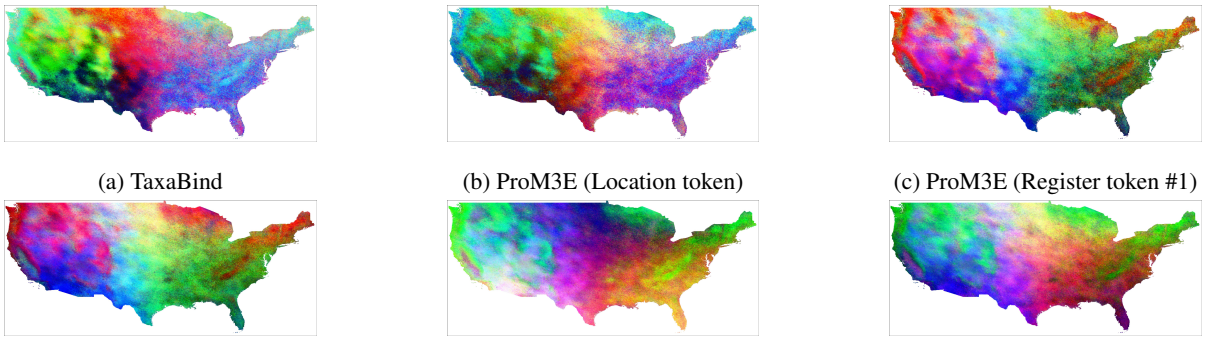
ProM3E: Probabilistic Masked MultiModal Embedding Model for Ecology

Supplementary Material

1. Mapping and Visualization

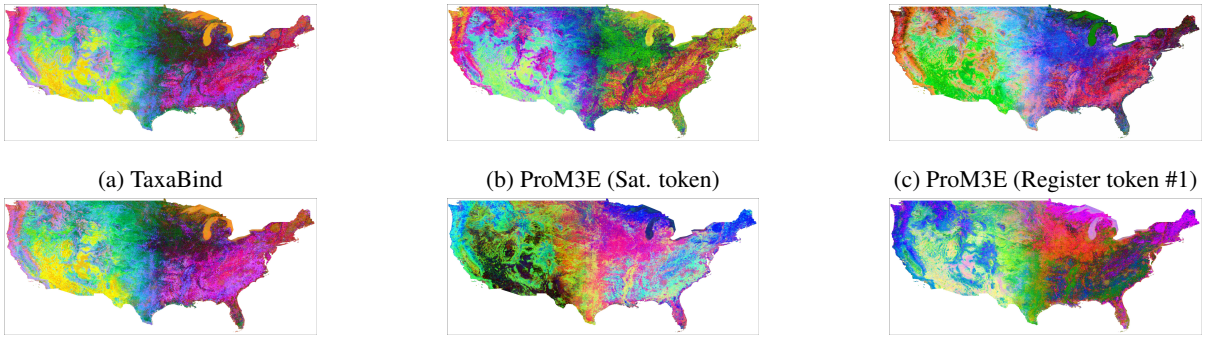
1.1. ICA Visualization of Geo-Embeddings



(a) TaxaBind (b) ProM3E (Location token) (c) ProM3E (Register token #1)

(d) ProM3E (Register token #2) (e) ProM3E (Register token #3) (f) ProM3E (Register token #4)

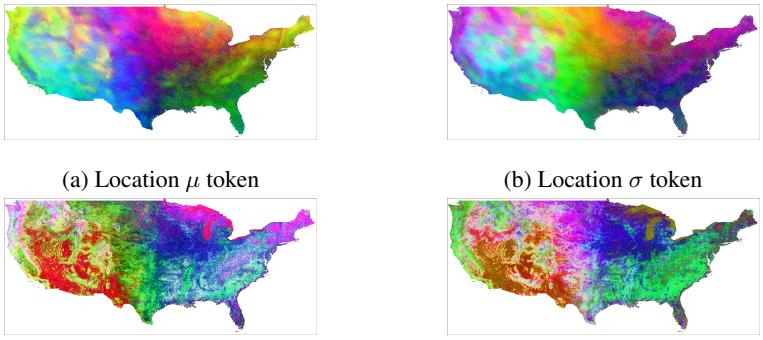
Figure 1. **ICA Plot of Location Embeddings**. We visually compare embeddings obtained from various tokens in the hidden representation of our model with the representation from TaxaBind. We notice that each register token captures different information.



(a) TaxaBind (b) ProM3E (Sat. token) (c) ProM3E (Register token #1)

(d) ProM3E (Register token #2) (e) ProM3E (Register token #3) (f) ProM3E (Register token #4)

Figure 2. **ICA Plot of Satellite Image Embeddings**. Similary, we compare satellite image embeddings with TaxaBind and notice register tokens capture diverse information.



(a) Location μ token (b) Location σ token

(c) Satellite μ token (d) Satellite σ token

Figure 3. **ICA Plot of *[μ]* and *[σ]* Tokens**. We plot the representations obtained from *[μ]* and *[σ]* tokens for geographic location and satellite images across the USA.

1.2. Species Distribution Mapping

In Figure 4, we show species distribution maps generated using our model given a query ground-level image depicting a species. We create a dense grid of satellite imagery over the USA and compute ProM3E embeddings for each location on the grid. We then compute cosine similarity of the query image and the embeddings of each location.

1.3. Generating an iNaturalist Species Diversity Map

We create species diversity and richness maps of the USA using iNaturalist observations. To create the maps, we first filtered observations to include only those within the contiguous United States (excluding Alaska and Hawaii). We then employed a spatial analysis technique that divided the US territory into a 250×500 grid based on the geographic bounding coordinates of the USA. We then filtered out all cells falling outside the USA. For each grid cell, we identified and counted the number of unique species observed by mapping latitude and longitude coordinates to their corresponding grid indices. For calculating species diversity, we used the Shannon index, which computes the entropy in the species distribution. The species richness is calculated as the number of unique species present within each grid cell. To each of the maps, we applied a kernel density estimation (KDE) based Gaussian smoothing with a sigma parameter of 2.0, which smoothed the discrete data across neighboring cells.

Additionally, we generate an uncertainty map of the USA by computing the $||\sigma||_1$ value at each grid cell. We then compare all the generated maps visually. We visualize the maps in Figure 5. Remember, in section, we conducted a quantitative comparison between $||\sigma||_1$ and Shannon diversity index and found a significant positive correlation between them. The maps in the figure show similarities visually. This is in agreement with the quantitative analyses conducted in the previous sections.

2. Dataset Details

2.1. Training Datasets

ProM3E has a flexible two-stage framework that can be trained independently. The first stage allows for training on large-scale image-paired datasets while the second stage requires an all paired dataset of all modalities for training. The first stage involves training modality-specific encoders using TaxaBind recipe. Below we present the details for the pretraining datasets used in stage one.

TreeofLife-10M. This dataset is composed of 10M pairs of species images and their corresponding taxonomic labels derived from open databases such as . It was introduced by Stevens et al. [14], which was used to train BioCLIP. Here, we utilize BioCLIP's image-text frozen embedding space and project all other modalities to this space.

iNaturalist-2021. We use the iNaturalist-2021 dataset primarily to train for aligning geographic location and species images. This dataset consists of 2.7M images across 10k species categories captured around the globe. Each image is associated with metadata including geographic location, timestamp, etc.

iSatNat. Sastry et al. [13] curated a paired dataset of satellite and species images using the iNaturalist-2021 dataset. For each ground-level image, they download a 256x256 Sentinel-2 imagery. This dataset is used to align satellite image with species images. There are 2.55M samples for training, 134k samples for validation and 100k samples for testing.

iSoundNat. This dataset [13] consists of paired species images and audio downloaded from the iNaturalist platform. There are 74k samples for training, 4k samples for validation and 8k samples for testing.

WorldClim-2. Climatic variables derived from WorldClim-2 are used to align environmental covariates and species images. These are environmental covariates are curated for each species in the iNaturalis-2021 dataset.

2.2. Evaluation Datasets

Below we provide details on the evaluation datasets used in the paper.

Taxabench-8k. This dataset consists of 8813 observations from the iNaturalist platform including all modalities paired for each observation. This dataset is used primarily for evaluating the models for the task of cross-modal retrieval.

BirdClef series. These datasets, released annually as part of the LifeClef [7] competition, contain geographically confined audio recordings of rare bird species. These datasets are used to identify bird species based on their soundscapes. We use the training, validation and testing split from TaxaBind for this task of bird species audio classification.

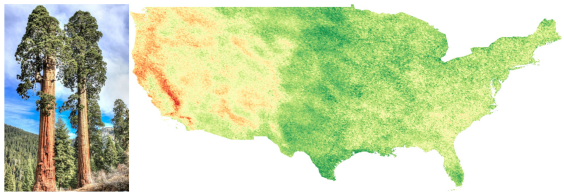
EcoRegions & Biome. We follow Range [2] and use their curated dataset for ecoregion and biome classification of given geographic locations. The dataset was curated by randomly sampling 100k geographic locations across the globe. Each geographic location was assigned a ecoregion label and a biome label. In total, there exist 846 ecoregions and 14 biomes.

3. Implementation Details

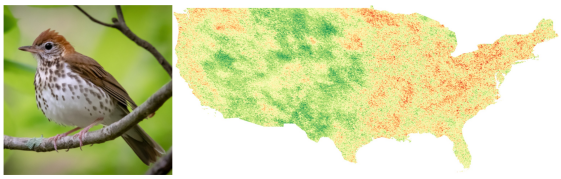
Below we provide all the implementation details that were used to train our model.

(a) Northern Cardinal

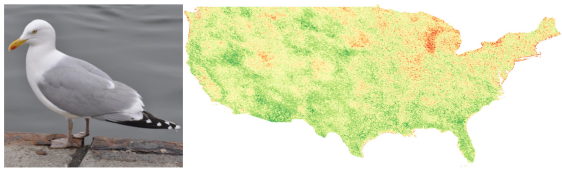
(b) Bighorn Sheep



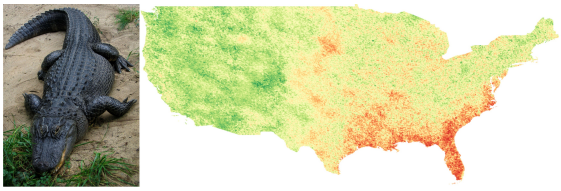
(c) Giant Sequoia Trees



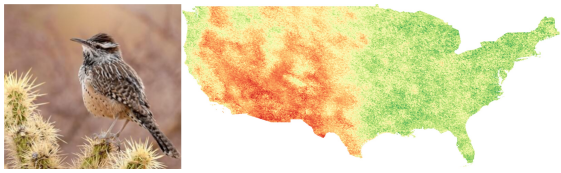
(d) Wood Thrush



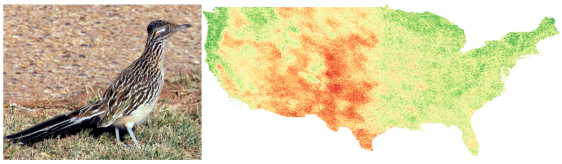
(e) European Herring Gull



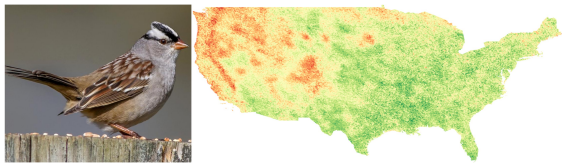
(f) American Alligator



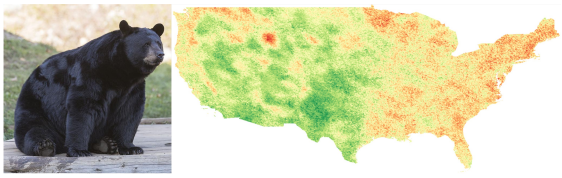
(g) Cactus Wren



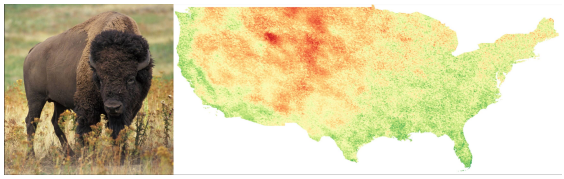
(h) Greater Roadrunner



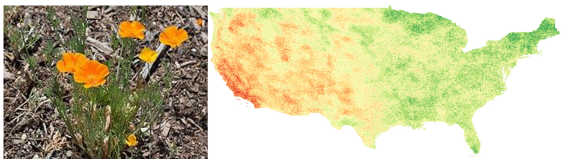
(i) White Crowned Sparrow



(j) American Black Bear

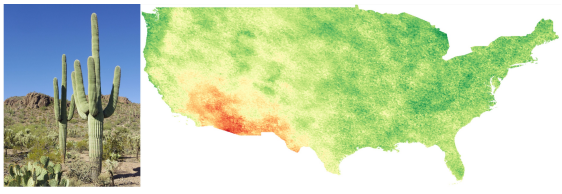


(k) American Bison



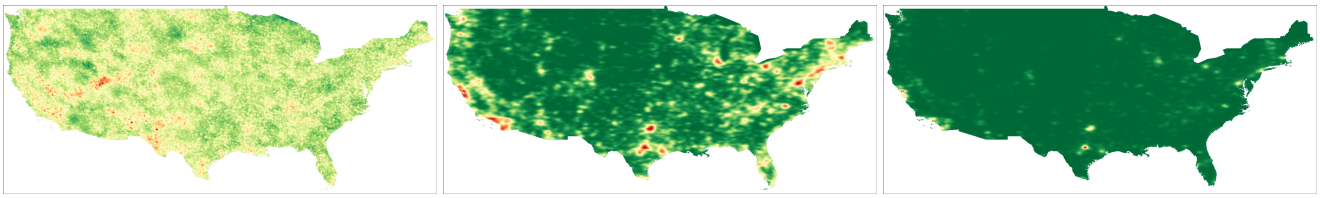
(l) California Poppy

(m) Bald Cypress



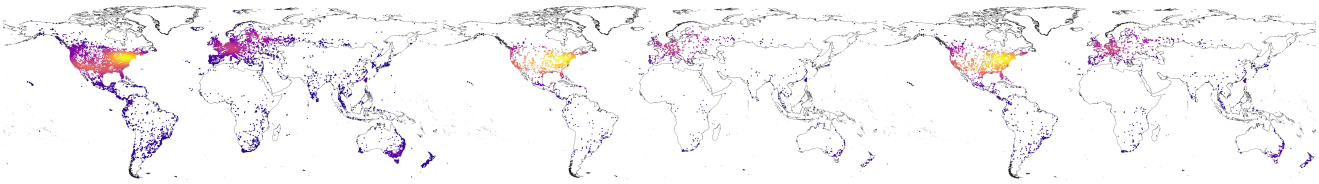
(n) Saguaro Cactus

Figure 4. **High Resolution Species Distribution Mapping using Ground-Level Imagery.** We create species distribution maps by computing the similarity between query ground-level image and geo-locations sampled uniformly across USA.



(a) $||\sigma||_1$ (b) Shannon Diversity Index (c) Species Richness

Figure 5. **Species Biodiversity Maps**. We plot $||\sigma||_1$ values predicted by our model and compare it with shannon diversity index and species richness maps derived from iNaturalist observations. The maps are plotted using a rectangular grid of 250x500 points over USA.



(a) Train (MultiNat) (b) Validation (MultiNat) (c) Test (Taxabench-8k)

Figure 6. **Spatial Distribution of Data**. The spatial distribution of our MultiNat dataset covering the globe.

Hyperparameter	Value
batch size	1024
max training epochs	500
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.98$
learning rate	1e-4
scheduler	cosine decay
weight decay	0.01
gpu type	NVIDIA H-100
num gpus	1

Table 1. Configuration used for training.

Configuration	Value
VIB term weight (λ)	0.001
base scale parameter (α)	-5.0
base shift parameter (β)	5.0

Table 2. Base configuration for our loss function for MVAE.

Configuration	Value
input embedding dim	512
num projection layers	2
encoder dim	1024
feedforward activation function	GeLU
num encoder layers	1
decoder dim	1024
num decoder layers	1

Table 3. Base architecture configuration for MVAE.

Modality	Model Architecture
ground-level image	OpenCLIP [6] (ViT-B/16)
satellite image	CLIP [12] (ViT-B/16)
geographic location	GeoCLIP [15]
environment covariates	SINR [1]
text	OpenCLIP [6]
audio	CLAP [4]

Table 4. Architecture configuration for modality-specific encoders.

4. Additional Experiments & Ablations

4.1. Linear Probing

4.1.1. Embedding Generation Ablation

In this experiment, we evaluate several design choices for curating embedding effective for linear probing tasks. The first choice is to utilized all the reconstructed embeddings. This is done by concatenating all embeddings at the output of our MVAE decoder. The rest of the choices involve using the hidden representations of our MVAE encoder. We could use the *[μ]*, *[m_i]* or register tokens. Additionally, we can concatenate all the tokens from the hidden representations of our MVAE encoder. Table 5 presents results of all these choices. We perform linear probing on the BirdClef-2023, EcoRegions and Biome datasets. Except the experiment with *[μ]* token, all the choices outperform TaxaBind. Using all the hidden tokens results in the best performance. We find that register tokens are essential and result in a significant gain in performance.

4.1.2. Probing Geo-Location Embeddings

Our models can serve as general purpose ecological predictors over space. Generating insights about habitat and climatic conditions of various geographic locations around the world is crucial in understanding global ecological trends. In this experiment, we compare the performance of several pretrained geographic location encoders in predicting various ecological indicators over space. In Table 6, we show the performance of our location encoder in predicting Biome, EcoRegion, Temperature and Elevation at a given geographic location. Climplicit is considered the absolute SOTA since it is specifically training on rich spatio-temporal climate data. We find that our model has the best performance beating TaxaBind, SINR and SatCLIP on all the tasks. We also conduct linear probing for the task of predicting several climatic variables in the ERA5 dataset. The results are presented in Table 7. We find that our model beats TaxaBind by a large margin and achives the second best performance on average after SINR. We believe that high-frequency geographic location features may not be necessary for these tasks. Climatic variables are typically low-frequency and often do not vary significantly across large regions. SINR is a simple feedforward-based model that outputs low-frequency geo-location embeddings. As a result, it achieves superior performance over other location encoding frameworks.

4.1.3. Habitat Classification

In this experiment, our aim is to classify the habitat of species represented using a given ground-level image. To achieve this, we use the iNat-2021 dataset that includes over 2.7M images of species with corresponding geographic location information. For each sample, we extract the Biome and EcoRegion label. We then obtain the image embeddings for each sample using our model and train a single layer linear classification model to predict the Biome/EcoRegion label given the image embedding. We note that this is a single positive multi label (SPML) problem. For training, we use the assume negative loss which is a common loss used in SPML problems. We evaluate the trained model on the testing split of iNat-2021 dataset.

4.2. Cross-Modal Retrieval

4.2.1. Embedding Generation Ablation

In this section, we investigate an optimal procedure to generate embeddings for effective cross-modal retrieval. There are several design choices one could use. We compare these design choices in Table 9. We find that using the representations from the hidden *[m_i]* leads to poor performance. We suspect that the representations useful for reconstruction are not necessarily useful for retrieval. We compare the reconstructed embeddings alone for retrieval and find that its performance is better than simply using the TaxaBind representations. We get the best performance using our proposed hybrid approach.

5. Uncertainty & Modality Gap

6. Broader Impact

6.1. Limitations

We acknowledge that the datasets used for training and evaluation in this paper suffer from various biases including geographic, socio-economic and human biases. The aim of this paper is to demonstrate the benefits of fusing multiple

Task	Dataset	Modality	TaxaBind [13]	Recons.	$[\mu]$	$[m_i]$	Reg. Tokens	All
Loc. Classification	EcoRegions		73.75	75.96	74.06	74.38	79.44	81.35
Loc. Classification	Biome		71.73	76.45	71.19	73.80	78.81	82.30
Audio Classification	BirdCLEF-2023		42.19	41.17	43.20	45.30	51.65	52.30
Audio Classification	BirdCLEF-2023	+	46.97	49.25	42.55	54.09	58.10	59.06
Average			58.66	60.71	57.75	61.89	67.00	**68.73**

Table 5. **Embedding Generation Ablation**. We investigate different choices for generating embeddings for linear probing. We find using all hidden tokens to achieve the best performance on average.

	Modality	Biome	EcoRegions	Temperature	Elevation
Direct		29.1	0.6	0.381	0.025
Cartesian_3D		30.2	1.8	0.362	0.030
Wrap [9]		34.4	1.1	0.861	0.085
Theory [5]		33.5	1.0	0.849	0.093
SphereM [11]		36.4	27.3	0.629	0.139
SphereM$^+$ [11]		58.7	50.1	0.886	0.294
SphereC [11]		36.3	52.9	0.461	0.185
SphereC$^+$ [11]		53.2	61.6	0.842	0.260
CSP-INat [10]		61.1	57.1	0.717	0.388
CSP-FMoW [10]		61.4	58.0	0.865	0.399
SINR [1]		67.9	54.9	**0.942**	0.644
GeoCLIP [15]		70.2	71.6	0.916	0.604
SatCLIP [8]		68.9	69.3	0.825	0.666
TaxaBind [13]		71.7	73.7	0.915	0.601
ProM3E **(ours)**		**82.3** +10.6	**81.3** +7.6	0.918 +0.003	**0.772** +0.171
Climplicit† [3] (Absolute SOTA)		83.3	78.4	0.985	0.898

Table 6. Comparison of various pretrained location encoders on predicting four ecological indicators. †Note that climplicit is pretrained on rich spatio-temporal climate data.

Models	temp_mean	temp_min	temp_max	dew_temp	precipitation	pressure	u_wind	v_wind	Avg
CSP	0.944	0.933	0.940	0.918	0.610	0.427	0.499	0.550	0.727
CSP-INat	0.987	0.897	0.886	0.857	0.534	0.307	0.413	0.386	0.658
SINR	**0.982**	**0.975**	**0.976**	**0.977**	0.758	0.706	**0.726**	**0.694**	**0.849**
GeoCLIP	0.960	0.953	0.948	0.954	0.591	0.651	0.502	0.529	0.761
SatCLIP	0.904	0.900	0.887	0.894	0.497	0.743	0.488	0.455	0.721
TaxaBind	0.965	0.954	0.955	0.957	0.637	0.662	0.525	0.560	0.777
ProM3E **(Ours)**	0.978 +0.013	0.971 +0.017	0.970 +0.015	0.972 +0.015	**0.730** +0.093	**0.758** +0.096	0.630 +0.105	0.638 +0.078	0.830 +0.058

Table 7. We show the linear probe results on real-world climate data from ERA5. Our model consistently beats TaxaBind and achieves the second best performance on average after SINR.

Method	Modality	Biome			EcoRegions		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
TaxaBind [13]		45.07	83.00	93.03	8.05	29.16	43.10
ProM3E **(ours)**		**57.75**	**89.51**	**93.70**	**25.37**	**54.06**	**61.72**
TaxaBind [13]		65.37	90.75	93.56	35.33	65.43	**74.93**
ProM3E **(ours)**		**76.57**	**93.64**	**93.96**	**54.79**	**69.63**	70.18
TaxaBind [13]		60.10	90.49	93.36	26.19	55.13	60.90
ProM3E **(ours)**		**83.05**	**93.90**	**94.00**	**64.95**	**73.76**	**73.79**

Table 8. **Habitat Classification.** We perform Biome and EcoRegion classification given species images as input. This is a challenging task and requires robust alignment of species images with geographic location and satellite images. We also test other inputs such as satellite images and environmental covariates.

Task	Dataset	Modality	TaxaBind [13]	Recons.	Hybrid
Image Classification	TaxaBench-8k	→	34.45	33.23	39.45
Image Classification	TaxaBench-8k	+ →	37.54	42.34	47.05
Retrieval	TaxaBench-8k	→	8.43	17.19	17.87
Retrieval	TaxaBench-8k	→	9.62	12.64	13.18
Average			58.66	60.71	**68.73**

Table 9. **Embedding Generation Ablation**. Here we investigate choices for embeddings useful for cross-modal retrieval.

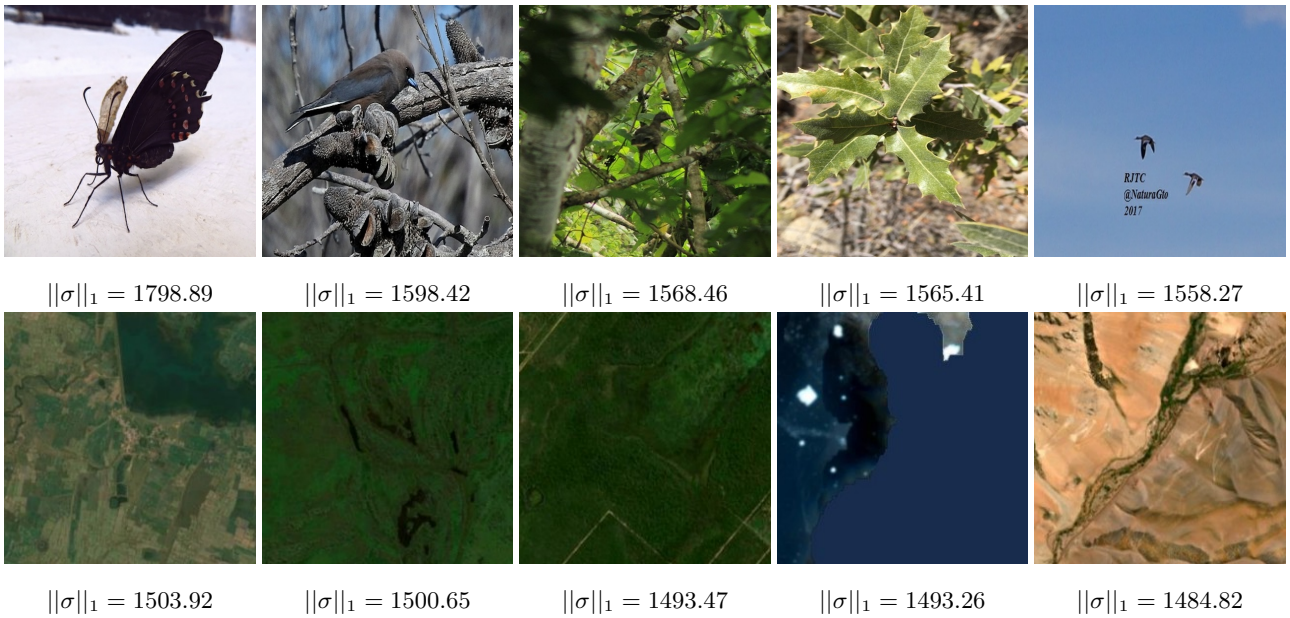


Figure 7. **Most Uncertain Images**. Most uncertain ground-level and satellite images.

modalities to improve performance of models on community accepted benchmark datasets. At present our model is limited to accept and process six modalities. However, given the simplicity of our approach, we believe it is trivial to incorporate additional modalities into the framework.

The species diversity and richness maps generated from iNaturalist observations might not accurately represent the Earth's biodiversity. As noted above, crowdsourcing and citizen science often lead to biased observations, favoring densely populated regions and documenting a limited num-

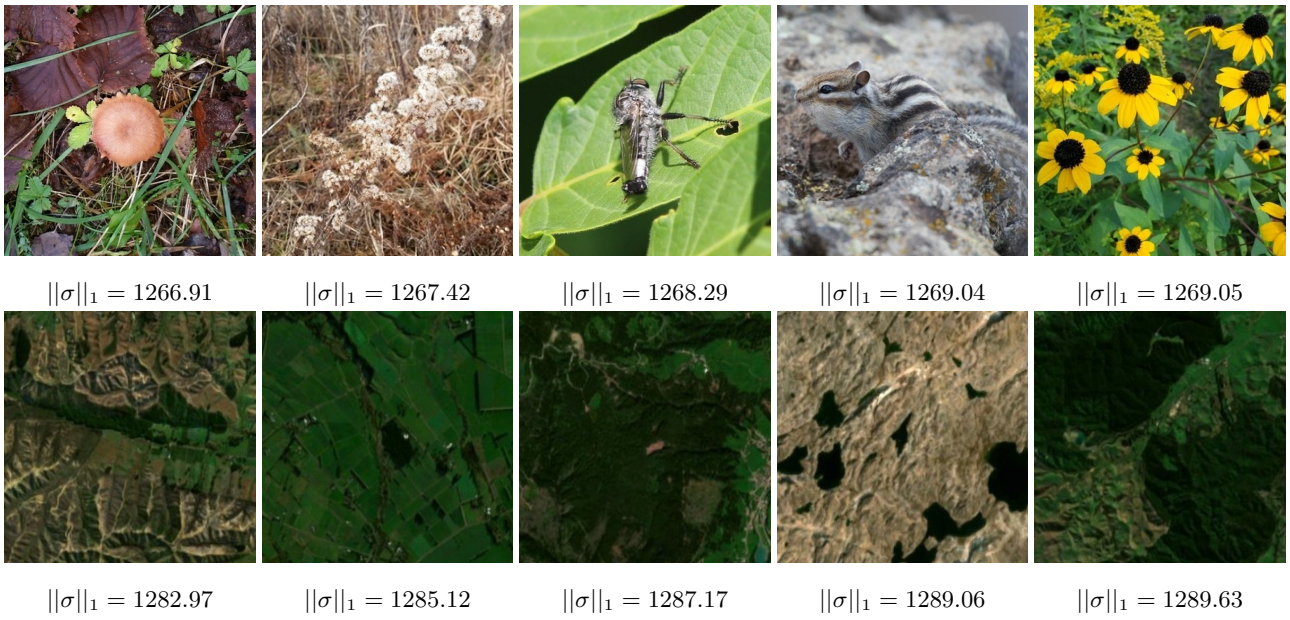
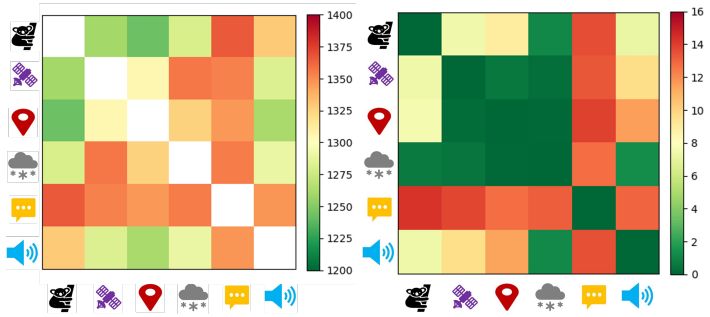


Figure 8. **Least Uncertain Images**. Least uncertain ground-level and satellite images.



(a) $||\sigma||_1$ (b) modality gap

Figure 9. **Pairwise $||\sigma||_1$ and Modality Gap**. We compare mean $||\sigma||_1$ values and modality gap when pairs of modality are provided as input. We find a spearman correlation of 0.32 between uncertainty and modality gap captured by our model.

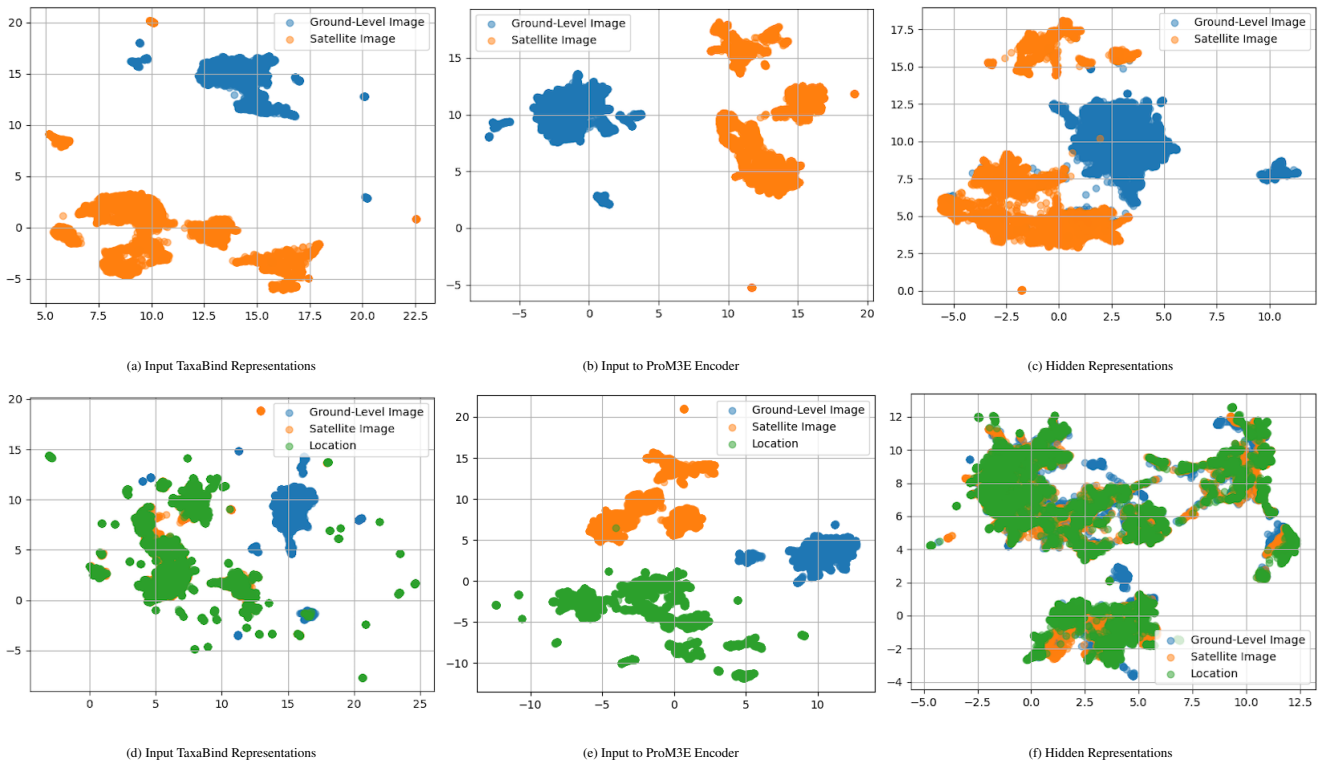
ber of species. Our study aimed to investigate whether the uncertainty captured by our model at different geographic locations correlates with the diversity of species observations in those areas. We found a significant positive correlation between these two factors. This is a promising result which we believe can form basis for future research.

6.2. Social Impact

Our models can be effectively adapted to address several remote sensing and ecological challenges. This might mean fine-tuning on additional datasets to adapt our models for specific applications. Our models can serve as a starting point from which interesting applications can emerge. However, utmost care must be taken before deploying them in the real world as is. They might need additional validation before they can be utilized for real world applications. The inherent biases present in the training datasets could potentially lead to inaccurate predictions in certain cases. Consequently, the application of our models in real-world scenarios can benefit from domain expertise. Our model was trained only on openly available species observation data and does not necessarily include information about sensitive species. Yet, care must be taken when using our models for such species.

References

[1] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *International Conference on Machine Learning*, pages 6320–6342. PMLR, 2023. 5, 6



(a) Input TaxaBind Representations (b) Input to ProM3E Encoder (c) Hidden Representations

(d) Input TaxaBind Representations (e) Input to ProM3E Encoder (f) Hidden Representations

Figure 10. **Adding Modalities Reduces the Modality Gap**. UMAP visualization of embeddings describing the reduction in modality gap between two modalities in presence of a third modality. Top row presents ground-level and satellite image embeddings while the bottom row presents the embeddings when location is additionally provided as input.

[2] Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. RANGE: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[3] Johannes Dollinger, Damien Robert, Elena Plekhanova, Lukas Drees, and Jan Dirk Wegner. Climplicit: Climatic implicit embeddings for global ecological tasks. *International Conference on Learning Representations (ICLR) Workshops*, 2025. 6

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 5

[5] Ruiqi Gao, Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. In *International Conference on Learning Representations*, 2019. 6

[6] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 5

[7] Alexis Joly, Lukáš Picek, Stefan Kahl, Hervé Goëau, Vincent Espitalier, Christophe Botella, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, Cesar Leblanc, et al. Lifeclef 2024 teaser: Challenges on species distribution prediction and identification. In *European Conference on Information Retrieval*, pages 19–27. Springer, 2024. 2

[8] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4347–4355, 2025. 6

[9] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019. 6

[10] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*, pages 23498–23515. PMLR, 2023. 6

[11] Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023. 6

[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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[13] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1765–1774. IEEE, 2025. 2, 6, 7

[14] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 2

[15] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2023. 5, 6