Research Overview

AI foundation models are neural networks that learn to represent large datasets [1]. These neural networks can become state-of-the-art scientific modeling tools [2, 3]. Cutting-edge foundation models for vision [4], language [5], chemistry [6], climate [7], and many other scientific fields have been developed. Foundation models could potentially be developed for biogeography and ecophysiology [8, 9] as well. They could be trained to geospatially simulate biophysics of ecosystems across the planet. Fine-tuning such a foundation model is expected to unlock state-of-the-art inference systems for biology, ecology, farming, landscaping, and other important fields.

Development of Artificial Intelligence for Science: 2006 to 2024

In 2006, Geoffrey Hinton published a neural network "autoencoder" in *Science* [10], capable of generating masked parts of images. Andrew Ng and Jeff Dean then showed in 2012 that larger neural networks could generate image data more accurately [11]. In 2018, OpenAI generated masked text data using a GPT (Generative Pre-trained Transformer) neural network [12]. By 2022, GPT-3.5 was trained to generate 100s of billions of words, and then "fine-tuned" from human feedback into ChatGPT [13]. Neural networks have since been trained to generate masked data from models of physics, chemistry, biology, and geosciences [6, 14, 15, 16, 17, 18, 19, 20, 21]. ClimaX from Microsoft in 2023 [22] trained a vision transformer (ViT) neural network [23] to generate parts of climate and weather models [24, 25]. ClimaX neural network parameters were then fine-tuned to generate more precise data from a regional-scale climate model, suddenly unlocking global inference capabilities at higher resolution. Fine-tuned foundation models can learn new tasks with 99% less labeled data [26].

Research Hypothesis: Foundation Model Generalization

Perhaps foundation models can generalize functional capabilities learned during fine-tuning on small-scale datasets back into their original training distribution. If this is true, then an ecology foundation model could learn to generate masked data from scientific models of ecophysiological responses and biogeographic distributions at a global scale across all species; and then potentially learn mappings back and forth from its original knowledge to smaller datasets. For example, as ClimaX learned to reconstruct higher resolution datasets globally, from limited regional data, a global ecology foundation model could potentially generate 1x1 meter community-scale distributions of species globally, from fine-tuning on a small sample of labeled data.

Overview of Research Methods

Following summarizes key steps in the development of an ecology foundation model:

1. *Training* neural networks to generate masked data ("masked autoencoding"):
   > Data gathered from global ecophysiological and biogeographic models
   > AI learns to minimize difference between generated and ground truth data
   > Trained ecology foundation model can simulate statistics similar to real data
2. *Fine-tuning* foundation model into state-of-the-art domain inference models:
   > Data gathered from small-scale datasets (*e.g.* genomic sequences, plant traits)
   > Foundation model fine-tuned to generate masked parts of small-scale data
   > Foundation model can generalize fine-tuning to global scale across species

Phase I: Training Foundation Model to Simulate Scientific Data

Development of an ecology foundation model can begin on small datasets with limited computing resources. A neural network could first be trained to generate masked plant habitat models, *e.g.* a snapshot in time of native habitat ranges [**27**, **28**]. For example, a ViT neural network could learn to generate masked native plant habitat distributions, by jointly encoding geographic and biological data representations. Encoding genetic and evolutionary history from recently developed phylogenetic trees [**29**, **30**, **31**], *e.g.* via graph neural networks [**32**], will likely improve ecology foundation models. The same is true for other abiotic (*e.g.* climatic) variables: neural networks that learn to encode these variables, perhaps through a similar approach taken by ClimaX, will yield better performance in their generation of masked ground truth data.

Phase II: Fine-Tuning Foundation Model for State-of-the-Art Inference

Geospatial encoders of habitat, phylogenetic, and climatic variables could be enough to demonstrate state-of-the-art scientific inferences from fine-tuning on small datasets. A first experiment for fine-tuning could be to geospatially predict distributions of plant traits (e.g. [**33**, **34**]). This could work by training and testing a foundation model's ability to generate masked plant traits that are currently known and validated, including geographic information to the extent available. This fine-tuned model could then potentially be able to infer global distributions of most traits for most plant species. This demonstration could inspire project funding (*e.g.* [**35**]). Additional fine-tuning possibilities: geospatial generation of microclimates, community ecology, genomic sequences, agricultural yield, and high-performance ecological planting designs.