Jacob Halaweh, David Lund, Mario Lucido
Department of Computer Science
Sonoma State University

Gaia: A Spatial-Spectral Vision Transformer for Biodiversity Mapping from Airborne Hyperspectral Imagery

1. Abstract

Accurate mapping of terrestrial biodiversity is vital for proactive conservation planning, particularly within hyper-diverse ecosystems such as South Africa's Greater Cape Floristic Region (GCFR). In this work, we present Gaia, an autonomous Masked Spatial-Spectral Vision Transformer (SSVT) designed to predict multi-taxa animal species richness (encompassing birds, frogs, and insects) directly from airborne hyperspectral imagery. Fine-tuned on global EnMAP satellite foundation weights, Gaia bridges the gap between large-scale self-supervised chemical embeddings and localized ecological regression tasks. We leverage high-resolution aerospace imagery from the NASA BioSCape campaign (AVIRIS-NG resampled mosaics) paired with 521 ground-truth acoustic and observational richness monitoring sites. To address spatial resolution and patch size disparities, we introduce bicubic positional embedding interpolation, scaling input patches from the pre-trained 8×8 footprint up to 16×16 (at 30m resolution, representing a 480m × 480m area) and 64×64 (at 5m resolution, representing a 320m × 320m area). To ensure spatial generalizability and to combat geographic data leakage, models are evaluated using a rigorous Group K-Fold cross-validation. Our fine-tuned ViT-Micro architecture achieves a validation R^2 of 0.3036, a Root Mean Squared Error (RMSE) of 4.59 species, and a Mean Absolute Error (MAE) of 3.58 species at 5m resolution, representing an 18.7% improvement in RMSE and a 19.1% improvement in MAE over a standard mean-prediction baseline. At 30m resolution, a structured 10-fold cross-validation of the model yields a mean validation R^2 of 0.4324, demonstrating robust predictive capability across coarser landscape scales. Our findings demonstrate that transfer learning from spatial-spectral foundation models offers a robust, highly scalable pathway for regional biodiversity mapping, enabling high-resolution species richness forecasting across diverse vegetative landscapes.

2. Introduction

Global biodiversity loss presents one of the most pressing ecological crises of the Anthropocene. To mitigate these declines and design adaptive management strategies, ecologists require high-resolution, continuous monitoring of biodiversity indices. Traditional manual field surveys, while accurate, are labor-intensive, logistically constrained, and spatially localized—rendering them insufficient for tracking rapid ecological changes across expansive, hyper-diverse landscapes such as South Africa’s Greater Cape Floristic Region (GCFR). Consequently, remote sensing has emerged as a crucial tool for scaling biodiversity observations.

While multispectral satellites provide broad temporal coverage, they lack the spectral resolution necessary to capture the complex biochemical signatures and structural variations of diverse vegetation communities that support rich animal life. Hyperspectral imaging addresses this limitation by capturing hundreds of contiguous, narrow spectral bands, mapping chemical, structural, and taxonomic variation in fine detail. The NASA Biodiversity Survey of the Cape (BioSCape) campaign leverages airborne AVIRIS-NG (Airborne Visible/Infrared Imaging Spectrometer-Next Generation) sensor data to capture these spatial-spectral granules across the Western Cape. Concurrently, terrestrial animal species richness has been quantified via localized ground-truth acoustics and observational surveys, yielding a structured dataset of 521 unique richness monitoring sites (ranging from 2 to 39 species, with a mean of 23.66 ± 6.74 species).

Historically, mapping high-dimensional hyperspectral signatures to ecological properties has relied on traditional machine learning algorithms such as Convolutional Neural Networks. However, these methods frequently struggle to exploit spatial context or model complex spectral band co-dependencies simultaneously, and deep learning architectures trained from scratch suffer from severe overfitting on sparse ecological ground-truth datasets. Furthermore, training from-scratch models often leads to convergence plateaus, where networks take a safe prediction route, estimating the mathematical mean of agricultural or background soil signatures due to gradient starvation.

To overcome these limitations, we introduce Gaia, a deep learning framework centered on a Masked Spatial-Spectral Vision Transformer. Instead of training a high-parameter network from scratch, Gaia pivots to a foundation model approach, fine-tuning the MaskedSSVT model pre-trained on 2.1 million hyperspectral patches from the global EnMAP satellite dataset. This allows Gaia to leverage a pre-trained understanding of fundamental earth chemistry (capable of including chlorophyll absorption, red-edge gradients, and water absorption features) and transfer it directly to species richness regression.

We outline several technical breakthroughs designed to translate this satellite-scale classification foundation model into an airborne-scale regression system:

1. *Spatial Positional Interpolation*: Using bicubic interpolation, we expand the spatial context of the pre-trained model's positional embeddings from 8×8 patches to 16×16 (30m resolution) and 64×64 (5m resolution) to capture contiguous landscape-level characteristics.
2. *Regression Bridge*: We adapt the classification head of the pre-trained transformer into a global mean-pooled regression head mapping directly to multi-taxa species richness vectors.
3. *Data Stratification and Cleaning*: We address high variance in training folds by implementing custom stratification based on both ground-truth richness and vegetation groundcover classes, resolving initial k-fold instabilities.
4. *Computational Efficiency*: To facilitate un-frozen training of spatial-spectral transformers under limited VRAM budgets, we integrate Automatic Mixed Precision (AMP), Scaled

Dot-Product Attention (SDPA), and Gradient Accumulation, reducing peak VRAM requirements from over 80GB to under 40GB without compromising model accuracy.

Through systematic validation using Group K-Fold cross-validation grouped by flightlines, we verify Gaia's capacity to generalize to entirely unseen geographic regions. At 5m resolution, our model achieves a validation **R^2 of 0.3036**. At 30m resolution, where spatial average pooling operates over a wider 480m × 480m spatial footprint, a 10-fold cross-validation demonstrates even higher predictive alignment with a mean validation **R^2 of 0.4324**. This report outlines our methodology, dataset processing, and validation results, highlighting a path forward for regional-scale biodiversity mapping.

3. Dataset

3.1 Data sources

The primary remote sensing dataset used in this project was BioSCape: AVIRIS-NG L2B Enhanced Surface Reflectance, accessed through NASA Earthdata [1]. This dataset contains corrected surface reflectance data from the Airborne Visible/Infrared Imaging Spectrometer–Next Generation, or AVIRIS-NG, flown on a NASA Gulfstream III aircraft during a NASA led Biodiversity Survey of the Cape, or BioSCape project. AVIRIS-NG imagery was captured over the Greater Cape Floristic Region of South Africa from October 22 to November 26, 2023. The spatial extent of the AVIRIS-NG imagery used in this project is shown in Figure 1.

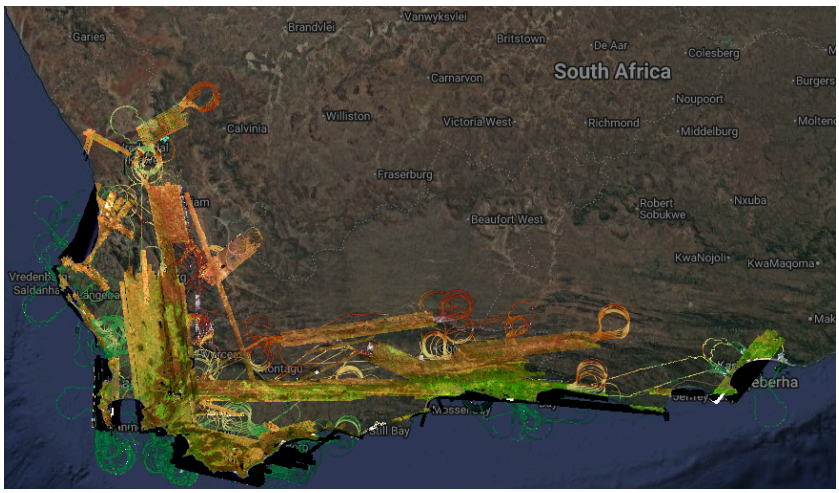


Figure 1 *Study region and AVIRIS-NG coverage used in the project. The dataset comes from NASA BioSCape AVIRIS-NG imagery collected over the Greater Cape Floristic Region of South Africa during October–November 2023.*

Figure 1 shows why targeted downloading was necessary: the full BioSCape archive covers a large region, but our model only needed granules overlapping labeled richness sites. This allowed the pipeline to reduce storage and bandwidth while preserving the imagery most relevant to supervised training. Standard RGB imagery, which contains only three color channels, AVIRIS-NG hyperspectral imagery records reflectance across hundreds of narrow spectral bands. This is useful for biodiversity modeling because vegetation, soil, water, and

surface structure can produce different spectral patterns that may relate to species richness. BioSCape was designed to connect field biodiversity observations with airborne, satellite and modeling datasets to improve the measurement and monitoring of biodiversity at large spatial scales [2]. The L2B data contains reflectance derived from L1B radiance data in 425 spectral bands. The files are georeferenced, projected into UTM coordinates, and distributed in NetCDF format. The full dataset contains 3,647 files/granules and is approximately 4.297 TB [1].

For this project, we did not download every BioSCape AVIRIS-NG granule. Instead, our pipeline used a targeted inventory of labeled granules, meaning granules that overlapped with sites that had available species richness labels. The script searched NASA Earthdata for the BioSCape_AVNG_L2B_BRDF_GCFR_2385 collection, downloaded only the target granules, applied the project bad-band list, and optionally downsampled the imagery to a lower spatial resolution for local training. This reduced storage and bandwidth while keeping the AVIRIS-NG data linked to the available richness labels.

The processed NetCDF files were then used as the source imagery for extracting hyperspectral patches around labeled site coordinates. Each patch represented the local spectral-spatial environment around a biodiversity sampling site and served as the input to the species richness prediction model. Although a later L3 resampled mosaic product was also discussed during the project, the original training pipeline shown here used the L2B enhanced surface reflectance granules and performed its own bad-band filtering and downsampling.

3.2 Species Richness Labels

Species richness was used as the prediction target for this project. Species richness refers to the number of distinct species observed or estimated at a given site. In our model, species richness was treated as a regression target, meaning the model predicted a continuous numerical value. The richness labels used in this project came from a project-provided site-level richness file from the western-cape-richness repository, based on prior BioSCape and BioSoundSCape work [3]. The broader BioSoundSCape dataset collected biodiversity-related acoustic data across the Greater Cape Floristic Region using Autonomous Recording Units [4]. These recorders were deployed at 521 sites during the wet and dry seasons of 2023 and captured sounds within approximately a 50 m radius of each site. These sites' biodiversity observations provide the biological context for the richness values used in our model. The distribution of species richness labels is shown in *Figure 2*.

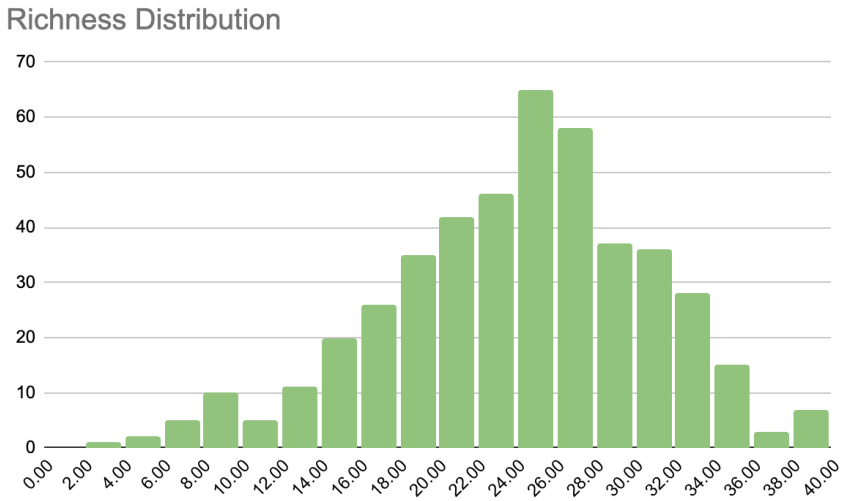


Figure 2. *Distribution of species richness labels used for training. The labels range from approximately 2 to 39 species, with an average richness of about 23.66 species. The distribution is approximately normal with a slight negative skew.*

Figure 2 shows that most labels fall near the middle of the richness range, with fewer examples at very low and very high richness values. This matters for training because the model sees more examples of moderate-richness sites than extreme sites, which can make high- and low-richness predictions more difficult. The geographic distribution of labeled richness sites is shown in *Figure 3.*

Figure 3. *Spatial distribution of species richness sites. Each point represents a labeled biodiversity sampling site, with color indicating the corresponding richness value. These site coordinates were matched to AVIRIS-NG granules to extract hyperspectral image patches.*

Figure 3 shows that richness sites are distributed across the BioSCape study area rather than being concentrated in only one location. In the training pipeline, these coordinates were used to

identify which AVIRIS-NG granule contained each site, allowing the model to extract a local hyperspectral patch around each label. In the training pipeline the species richness file is loaded in as a CSV that contains site coordinates and richness values. Each site is then matched with AVIRIS-NG NetCDF granules by checking whether the site coordinate fell within the geographic bounds of a granule. If a site is inside a AVIRIS-NG granule, the pipeline extracted a hyperspectral image patch centered around that site and paired it with the corresponding richness value. This process created the supervised learning pairs with AVIRIS-NG hyperspectral patch as the input and species richness value as the target.

3.3 AVIRIS-NG bad bands and EnMAP-Compatible Input

AVIRIS-NG files used in this project contain hyperspectral surface reflectance data with hundreds of spectral bands. While this high spectral resolution provides useful environmental information, not every band is reliable for modeling. Some AVIRIS-NG bands are affected by atmospheric absorption, sensor noise, or other quality issues that make them unsuitable for training. If these bands are left in the dataset, the model may learn from noisy or physically unreliable signals rather than meaningful vegetation and surface reflectance patterns.

To address this, the preprocessing pipeline applied a bad-band list, provided through project guidance [5]. In this list, each spectral band is marked as either usable or unusable, where 1 indicates a good band and 0 indicates a bad band. The pipeline used this list to keep only the usable bands from the NetCDF reflectance cube. After bad-band filtering, the data still contained more spectral channels than the pretrained model could accept. The baseline model used pretrained weights from a Masked Spatial-Spectral Transformer trained on EnMAP-style hyperspectral data [6,7]. This model expected a 200-band input, so the AVIRIS-NG data had to be represented in a compatible 200-channel format before being passed into the Vision Transformer. The spectral transformation is shown in *Figure 4.*

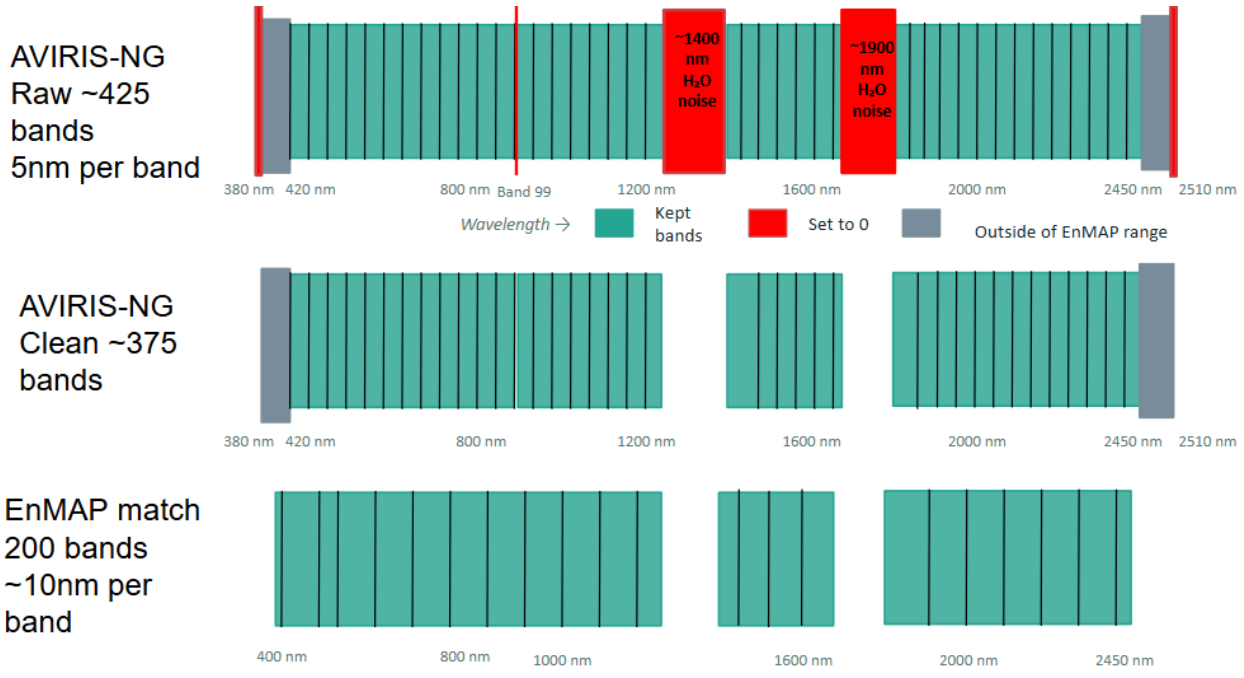


***Figure 4**. Spectral preprocessing workflow. The original AVIRIS-NG reflectance data contain 425 spectral bands. Bands marked as unreliable in the bad-band list were removed before the data were converted into a 200-band representation compatible with the pretrained EnMAP transformer.*

Figure 4 shows the main spectral transformation used in the baseline model. The original AVIRIS-NG reflectance cube begins with 425 spectral bands. The bad-band list removes unreliable bands, and the remaining spectral information is converted into a 200-band EnMAP-compatible representation. This step was necessary because the pretrained MaskedSST model expected 200 spectral channels. The 200-band representation allowed the project to reuse pretrained EnMAP hyperspectral embeddings instead of training a transformer completely from scratch. This was useful because the number of available species richness labels was limited compared with the amount of data normally needed to train a large transformer model. Using pretrained spatial-spectral features gave the model a stronger starting point for learning relationships between hyperspectral image patches and species richness.

However, converting AVIRIS-NG data into a 200-band EnMAP-compatible format also introduced a tradeoff. AVIRIS-NG contains more spectral information than the baseline model could use, so reducing the data to 200 channels may remove some information that could be useful for biodiversity prediction.

In addition to spectral preprocessing, the pipeline could optionally downsample the spatial resolution of the imagery. The synchronization script supported a target resolution parameter, allowing the native 5 m data to be averaged into lower-resolution versions such as 30 m for faster local training. This made it possible to test the model on smaller, more manageable data before scaling up to full-resolution experiments.

4. Approach and Algorithm:

In order to ensure our dataset was properly cleaned we created a data visualization tool that displayed the different possible images for each site. It made it clear that although we already ignored any images that were all zeros, we weren’t always using the best image.

When making our test/train splits originally we left it fully random. However this caused the issue of the test set sometimes not being representative of the training set. This was especially clear in our initial k-fold validation, where we would frequently have one fold perform much worse than the overall fold average. Switching to stratification based on species richness provided a mild improvement, and stratification based on richness and groundcover label improved it even more. The groundcover labels were collected from the same dataset where we got the richness labels from, compiled by a previous student.

Given the small dataset size we tried to artificially increase the size by multisampling the training set with data augmentation. However this overall didn’t have much effect on the performance for the test set. The augmentation we tried was primarily spatial, with minor rotations, flips, and cropping changes. It's possible that spectral augmentation would have a better stronger effect.

4.1 Model Architecture

To predict animal species richness (Birds, Frogs, and Insects) across diverse ecological zones, our primary research models (Gaia 30m and Gaia 5m) leverage a deep transfer-learning approach built upon a pre-trained Masked Spatial-Spectral Vision Transformer foundation model.

The Core Foundation Encoder

The underlying backbone is a ViT architecture characterized by the following parameters:

- Transformer Dimension: 96
- Encoder Depth: 4 Layers
- Attention Heads: 8
- Spectral Patch Size: 10 bands
- Spatial Patch Size: 1 pixel
- Attention Mechanism: Factorized attention (spatial attention over the grid tokens, followed sequentially by spectral attention across spectral groups).

Initially pre-trained on the global EnMAP satellite dataset (comprising 2.1 million hyperspectral patches), the encoder has acquired a rich self-supervised representation of terrestrial chemistry, such as chlorophyll absorption curves, red-edge transition zones, and water absorption bands.

Bicubic Positional Embedding Interpolation

The baseline EnMAP foundation model was natively pre-trained on 8×8 spatial blocks. To capture landscape-level ecological contexts, our target datasets are extracted in larger dimensions:

- Gaia 30m model: Processes 16×16 spatial grids (representing a 480m×480m area).
- Gaia 5m model: Processes 64×64 spatial grids (representing a 320m×320m area).

Because the transformer requires matching positional encodings for these expanded coordinate spaces, we implement Bicubic Spatial Positional Interpolation. This mathematical scaling maps the learned 8×8 positional coordinate grid to the target 16×16 and 64×64 matrices prior to token injection, preserving the relative spatial relationship of the pixels without requiring from-scratch positional learning.

The Regression Bridge and MLP Head

The foundation model's original classification head is discarded and replaced with a multi-taxa regression bridge. This MLP head processes the global mean-pooled spatial-spectral tokens to output continuous multi-dimensional richness estimates (*Figure 5*).

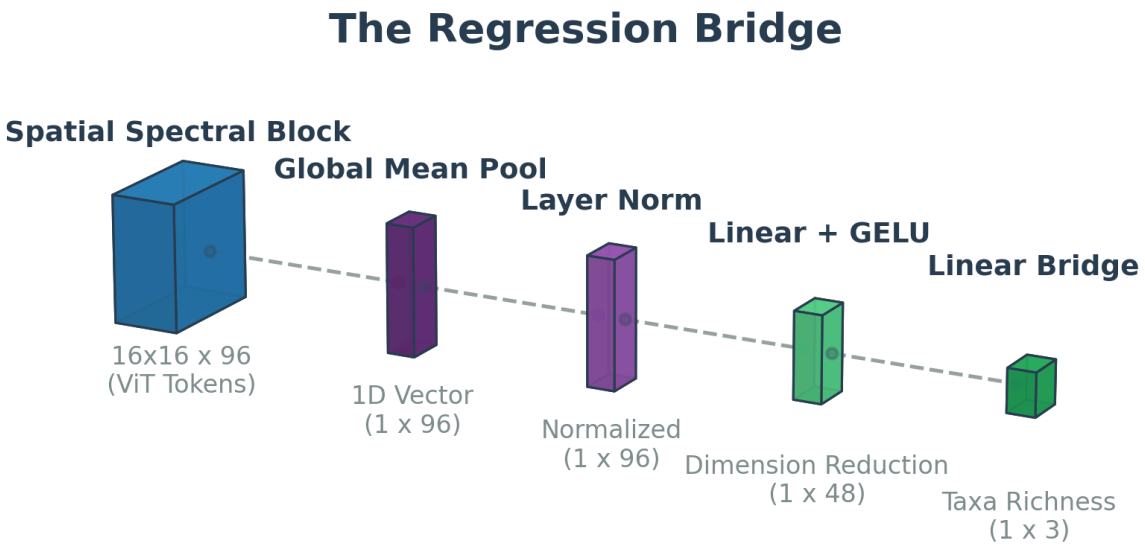


Figure 5. *The process and dimensionality of a Spatial Spectral Block of tokens being transformed into a richness output.*

During downstream fine-tuning, the regression MLP head is fully optimized while the pre-trained EnMAP encoder weights are either frozen (for quick verification) or unfrozen (for native adaptation on the BioSCape dataset) depending on the validation strategy.

4.2 Custom Embedding Generation

In addition to utilizing satellite-scale EnMAP foundation models, we implemented a custom, native self-supervised pre-training pipeline designed specifically for airborne hyperspectral sensors (such as NASA's AVIRIS-NG instrument).

AVIRIS-NG Native Foundation Model Pipeline

While satellite sensors are subject to spectral binning (e.g., EnMAP's 200 resampled bands), AVIRIS-NG records high-fidelity imagery across 370 native spectral bands (after removing bad band list / BBL ranges corresponding to atmospheric water vapor absorption). To exploit this rich spectral resolution directly, we constructed an AVIRIS-specific embedding space using self-supervised SimMIM (Simple Masked Image Modeling) pre-training.

Pre-Training Methodology (SimMIM)

1. Spectral-Spatial Tokenization: The 3D input volume is sliced into patches of size (10 bands×1 pixel×1 pixel) . This maps the 370 bands into 37 spectral groups across the spatial grid.
2. Masking: A random masking ratio of 60% is applied to the spatial-spectral tokens. Masked positions are replaced with a single, learnable token parameter vector.
3. Backbone Processing: The partially masked token sequence is passed through the spatial-spectral transformer to generate contextualized latent embeddings.

4. Reconstruction & Decoder: A lightweight 2-layer MLP decoder is trained to reconstruct the original raw reflectance values of the masked patches. Optimization is driven by an L1 or L2 loss computed exclusively on the masked regions.

Downstream Fine-Tuning

Once self-supervised pre-training on the unlabeled BioSCape flightlines converges, the reconstructed decoder is removed. The pre-trained AVIRIS-specific encoder is then frozen, and the regression MLP head is attached and fine-tuned on the labeled western cape biodiversity dataset (521 monitoring sites), producing custom high-performance embeddings highly sensitive to regional vegetative and structural nuances.

5. Results:

5.1 30 Meter Per Pixel Results

Our primary experiments evaluated the Gaia framework at a 30m spatial resolution using 16×16 pixel patches and a 200-band EnMAP-compatible spectral representation. This configuration served as the primary benchmark because it aligns naturally with the spatial scale of the pre-trained EnMAP satellite foundation weights, and proved computationally efficient, enabling comprehensive hyperparameter tuning, convergence validation, and stable cross-validation.

To evaluate spatial generalizability and robust performance across diverse vegetative landscapes, we conducted a rigorous Group K-Fold cross-validation structured by flightline (10-fold split). The 30m model achieved a mean validation R^2 of 0.4324 across the ten folds. Individual fold R^2 values demonstrated robust predictive capability under geographic dispersion, yielding: 0.4322, 0.4500, 0.5597, 0.2691, 0.7232, 0.3384, 0.4637, 0.4639, 0.3495, and 0.2745, as seen in *Figure 6*. Compared with the mean-prediction baseline, the 30m model consistently outperformed simple average-based guessing, demonstrating that the network successfully learned non-trivial ecological patterns.

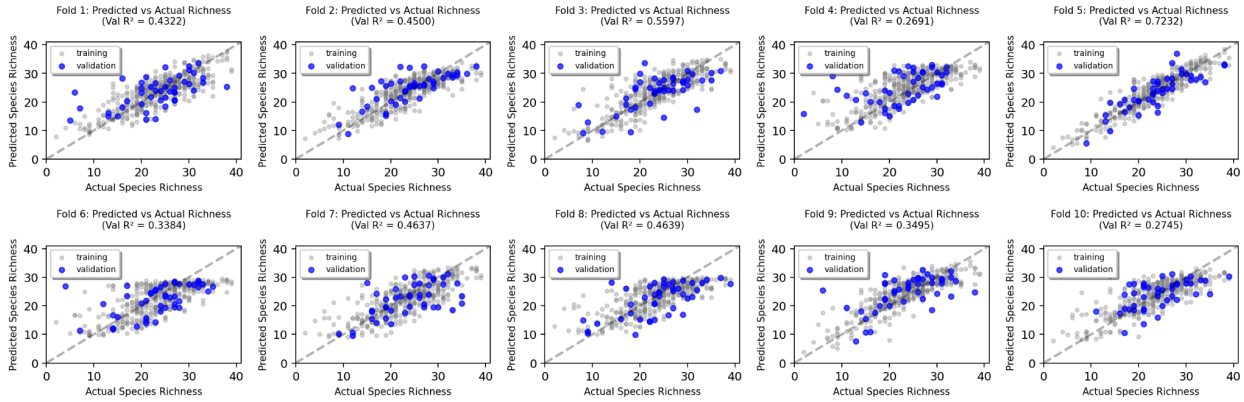


Figure 6. *The results of a 10 - Fold validation for the 30m Gaia Model.*

The strong performance of the 30m configuration over the exploratory 5m test is likely due to two key factors. First, the 30m model's 16×16 pixel patches cover a larger total spatial footprint of 480 m × 480 m, compared to the 320 m × 320 m area of the 5m model. This broader context

allows the model to capture macro-level landscape characteristics and habitat heterogeneity that heavily influence multi-taxa species richness (such as bird, frog, and insect distributions). Second, the spatial scale of the 30m resampled imagery matches the native footprint of the EnMAP satellite data on which the foundation weights were pre-trained, facilitating a highly effective transfer of learned spatial-spectral embeddings.

Furthermore, the 30m model was computationally efficient, requiring substantially less memory and training time than the 5m model. This allowed for extensive hyperparameter sweeps, repeated runs to confirm optimization stability, and robust convergence checks. The resulting stability and superior alignment make the 30m configuration our primary recommended setup for regional biodiversity mapping. *Figure 6* shows the evaluation dashboard for the primary 30m configuration.

5.2 5 Meter Per Pixel Results

In addition to the main 30 m experiments, we ran an exploratory test using 5 m spatial resolution, 64×64 pixel patches, and the same 200-band EnMAP-compatible spectral representation. This experiment was not treated as the primary result because it was more computationally expensive and was tested less extensively than the 30 m configuration. However, it was useful for evaluating whether higher spatial resolution could improve species richness prediction.

The 5 m experiment achieved a validation R^2 of 0.3036, with an RMSE of 4.59 species and an MAE of 3.58 species across 46 validation sites. Compared with a mean-prediction baseline, this represented an 18.7% improvement in RMSE and a 19.1% improvement in MAE. Although this performance was weaker than our stronger 30 m experiments, it still showed that the model learned useful patterns beyond simply predicting the average richness value.

One reason the 5 m result may not have improved over the 30 m configuration is the difference in spatial footprint. The 30 m model used 16×16 pixel patches, covering 16 × 16 pixels × 30 m = 480 m × 480 m. While the 5 m model used 64×64 pixel patches, covering 64 × 64 pixels × 5 m = 320 m × 320 m. Although the 5 m patches contain more fine-grained spatial detail, they cover a smaller total area around each site. Species richness may depend on broader landscape context.

Another possible issue is that the pretrained EnMAP foundation weights were originally designed around EnMAP-style hyperspectral data, which is closer to a 30 m spatial scale than the native 5 m AVIRIS-NG imagery. Because of this mismatch, the pretrained spatial-spectral features may transfer more naturally to the 30 m configuration than to the 5 m configuration. The 5 m model also required substantially more memory and training time, making it harder to run repeated experiments, tune hyperparameters, and verify convergence. With more time, compute, and testing, the 5 m approach may improve, especially if paired with custom AVIRIS-NG embeddings or a larger spatial patch size.

Overall, the 5 m experiment should be viewed as a promising but unfinished direction. It demonstrated improvement over the mean baseline, but it did not yet outperform the more stable 30 m configuration. For this reason, we report it as a secondary exploratory result rather than the main performance claim. *Figure 7* shows the evaluation dashboard for the exploratory 5 m experiment.

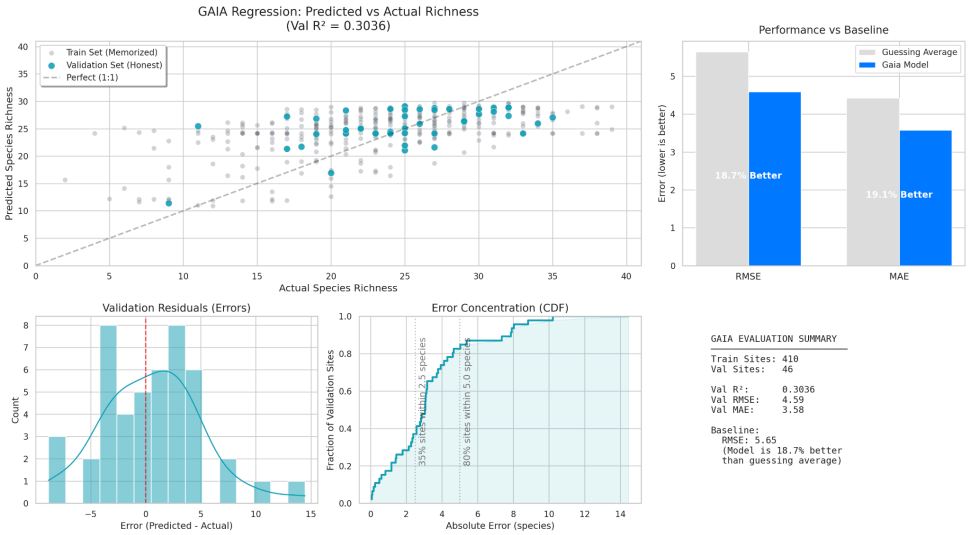


Figure 7*. Evaluation dashboard for the exploratory 5 m, 64×64, 200-band experiment. The model achieved a validation R² of 0.3036, RMSE of 4.59 species, and MAE of 3.58 species across 46 validation sites.*

5.3 Qualitative Heatmap Results

To complement the numerical evaluation, we generated predicted richness heatmaps using the 200-band EnMAP-compatible model. These heatmaps provide a qualitative way to inspect where the model predicts higher or lower species richness across larger AVIRIS-NG regions. In the 30 m configuration, each 16×16 prediction patch covers approximately 480 m × 480 m, so the heatmap shows broader landscape-level patterns rather than individual plants or small objects. Warmer colors indicate higher predicted richness, while cooler colors indicate lower predicted richness. The regional prediction map is shown in *Figure 8*.

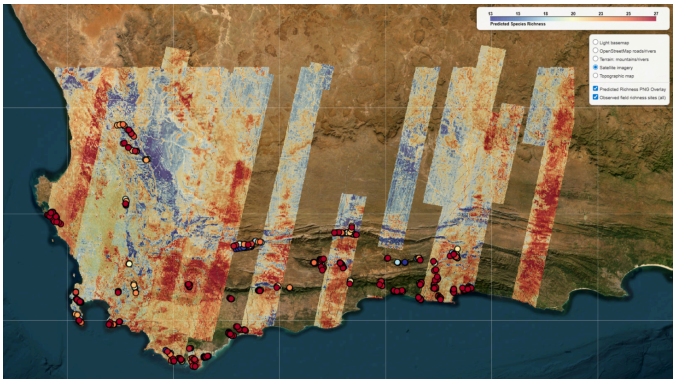


Figure 8. *Regional predicted species richness heatmap generated using the 200-band EnMAP-compatible model. Warmer colors indicate higher predicted richness, while cooler colors indicate lower predicted richness.*

The regional heatmap *Figure 8* shows that the model does not assign uniform richness across the landscape. Instead, the predictions form spatial patterns that appear to follow differences in vegetation, moisture, and land cover. This suggests that the model is responding to spectral-spatial variation in the AVIRIS-NG imagery rather than only predicting a constant average value.

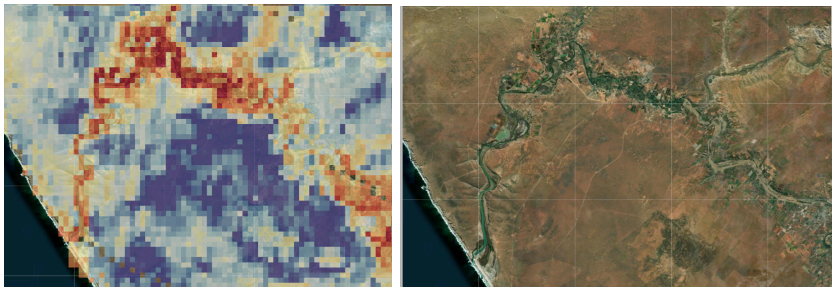


Figure 9*. Lake close-up comparing predicted species richness with the corresponding visual landscape context.*

In the first close-up *Figure 9*, higher predicted richness follows a lake or wetland system, while nearby barren or drier regions are predicted as lower richness. This pattern is visually reasonable because water-adjacent and greener areas often support more habitat variation, which may be associated with higher biodiversity.

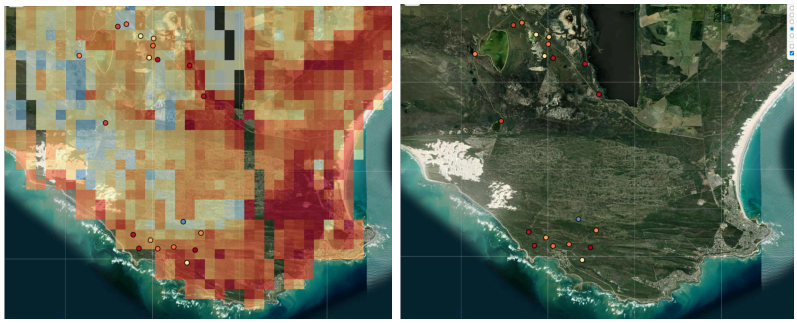


Figure 10. *Coastal close-up showing a more complex predicted richness pattern.*

In the second example *Figure 10*, the model gives a more complex pattern. Some areas near water are predicted as high richness, while nearby inland regions are lower. We cannot fully explain every prediction visually, but this is useful because it shows the model may be detecting patterns that are not obvious from RGB imagery alone.

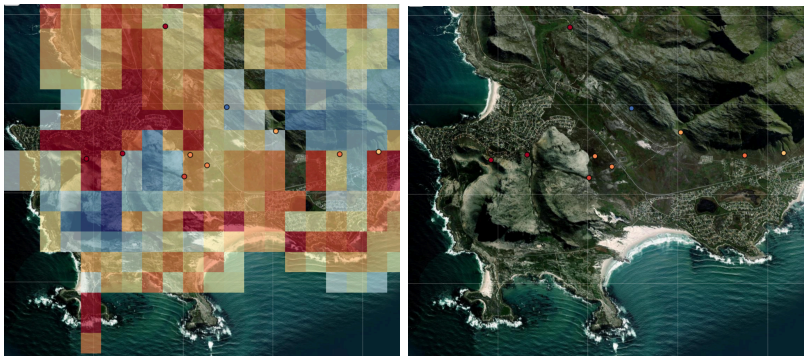


Figure 11. *Developed, rocky, and vegetated area close-up comparing predicted richness with visible landscape features.*

The final close-up *Figure 11* shows the model distinguishing between nearby developed, rocky, and vegetated surfaces. Some high predicted richness values appear near developed areas, so these predictions should be interpreted cautiously. They may reflect irrigated vegetation, parks, water proximity, spectral artifacts, or correlations in the training data rather than true species richness. Overall, the heatmaps are useful because they reveal both intuitive predictions and uncertain areas that require further ecological validation.

6. Conclusion

6.1 Reflection

The development of Gaia demonstrates that remote sensing foundation models can be successfully adapted to map terrestrial biodiversity. By fine-tuning a pre-trained spatial-spectral ViT, we bypassed the need for massive supervised datasets and enabled direct species richness prediction from hyperspectral imagery.

Our methodology highlighted critical lessons about data quality and preparation. An early bottleneck was attempting to ingest raw, uncalibrated AVIRIS-NG flightlines. Transitioning to the pre-cleaned, scale-normalized, and tiled mosaic version dramatically simplified our workflow. We also addressed the small size of the labeled dataset (521 sites) by leveraging pre-trained EnMAP foundation weights, which stabilized learning and prevented overfitting. Finally, implementing a visual patch inspection tool helped us identify and exclude corrupted patches, directly improving optimization stability and validation metrics.

From a larger perspective, the success of Gaia 30m on the extremely small dataset of 521 sites suggests that this method is extremely effective, and would likely see great improvement with more labeled data. While species richness is a metric that is hard to gather labels for, many other well labeled geological and biological phenomena would likely be easy to predict using the same architecture as Gaia.

Ultimately, Gaia provides a practical, scalable framework for continuous biodiversity monitoring, demonstrating the power of transfer learning in resource-constrained ecological research.

6.2 Challenges

One issue we ran into was not investigating our dataset earlier. It turns out there are two versions of the AVIRIS dataset, an uncleaned version of every flightline image, and a cleaned mosaic version. This means they did the work for us in terms of selecting which flightline was best for every region. They also normalized the scale, which we hadn't known was an issue, and split it into tiles, which makes for an overall simpler, smaller dataset.

7. References

1. *Kovach, K.R.; Ye, Z.; Frye, H.; Townsend, P.A. BioSCape: AVIRIS-NG L2B Enhanced Surface Reflectance, Version 1; ORNL Distributed Active Archive Center: Oak Ridge, TN, USA, 2025. https://doi.org/10.3334/ORNLDAAC/2385.*

2. *Cardoso, A.W., Hestir, E.L., Slingsby, J.A. et al. The biodiversity survey of the Cape (BioSCape), integrating remote sensing with biodiversity science. npj biodivers 4, 2 (2025). https://doi.org/10.1038/s44185-024-00071-5*

3. *Nguyen, R. western-cape-richness. Available online: https://github.com/nguyenry172/western-cape-richness (accessed on 21 May 2026).*

4. *Turner, A.A., Clark, M.L., Salas, L. et al. BioSoundSCape: A bioacoustic dataset for the Fynbos Biome. Sci Data 12, 1432 (2025). https://doi.org/10.1038/s41597-025-05685-3*

5. *Clark, M.L. (Sonoma State University, Rohnert Park, CA, USA). Personal communication, 2026.*

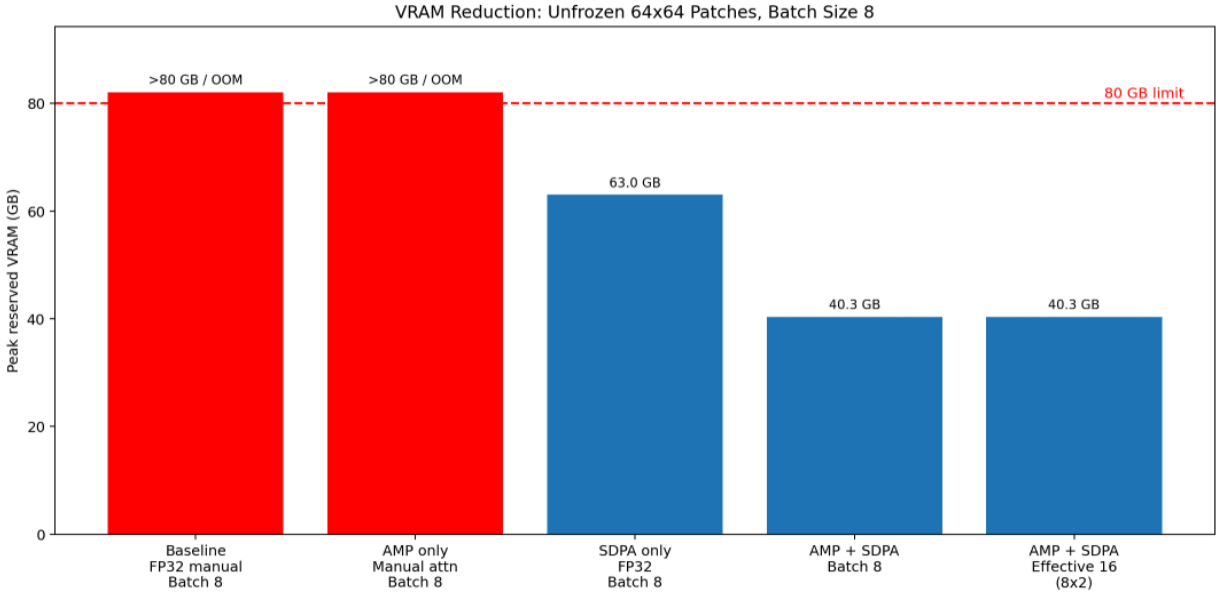
6. *L. Scheibenreif, M. Mommert and D. Borth, "Masked Vision Transformers for Hyperspectral Image Classification," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023, pp. 2166-2176, doi: 10.1109/CVPRW59228.2023.00210.*

7. *HSG-AIML. MaskedSST: Masked Vision Transformers for Hyperspectral Image Classification. Available online: https://github.com/HSG-AIML/MaskedSST/ (accessed on 21 May 2026).*

8. *Brodrick, P. G., Chlus, A. M., Eckert, R., Chapman, J. W., Eastwood, M., Geier, S., Helmlinger, M., Lundeen, S. R., Olson-Duvall, W., Pavlick, R., Rios, L. M., Thompson, D. R., & Green, R. O. (2025). BioSCape: AVIRIS-NG L3 Resampled Reflectance Mosaics, V2 (Version 2). ORNL Distributed Active Archive Center. https://doi.org/10.3334/ORNLDAAC/2427 Date Accessed: 2026-05-24*

Appendix

Appendix A. VRAM Optimization



We used PyTorch for the training experiments and tested three main changes to reduce VRAM usage on the unfrozen 64x64 model.

First, we enabled automatic mixed precision (AMP). AMP was easy to implement using PyTorch's mixed precision utilities, but it was not enough by itself. The AMP-only run still exceeded the 80GB GPU limit and failed with an out-of-memory error. AMP reduced memory usage but was insufficient by itself to prevent out-of-memory errors.

Second, we enabled PyTorch's scaled dot product attention (SDPA). This was more effective than AMP alone. The SDPA-only run completed successfully, reducing peak reserved VRAM to about 62.99GB. This was also the easy to change it was replacing the two lines in src/vit_spatial_spectral.py.

Third, we used gradient accumulation to simulate a larger effective batch size. By using physical batch size 8 with 2 accumulation steps, we achieved an effective batch size of 16 without increasing VRAM usage. Combined with AMP and SDPA, this run completed successfully with about 38.94GB peak allocated and 40.31GB peak reserved. This was harder to change but not bad and it allowed batch size to not become an issue did have the most downside but did not effect r^2 noticeably.

There is more that we can do but this allowed us to run everything that we wanted so I stopped optimizing.

Appendix B. Fine-Tuning Sweep Parameters

Figure B1. Fine-tuning sweep dashboard from W&B.
During fine-tuning, we used Weights & Biases (W&B) to track validation performance across multiple training runs. The goal of these sweeps was not to completely redesign the model, but to compare nearby hyperparameter settings and confirm that our manually selected values were reasonable. This was useful because the model had several sensitive training choices, including learning rate, whether the encoder was frozen or unfrozen, patience for early stopping, and the number of epochs before unfreezing.

The original MaskedSST repository includes its own fine-tuning workflow, but we created a modified finetune_sweep.py script to give us more control over regression-specific training and easier experiment visualization. W&B made it easier to compare validation R^2, validation RMSE, validation MAE, training loss, head learning rate, backbone learning rate, and encoder-freezing behavior across runs.

Most runs followed similar performance trends, suggesting that our original hand-selected training settings were already close to reasonable. The sweeps helped confirm that only minor parameter tuning was needed, rather than a complete change in the training strategy.

Appendix C. NRP Notes

It was pretty easy overall working with NRP and JupyterHub. There are some helpful user guides at https://nrp.ai/documentation/. For example choosing GPU type, and learning about JupyterHub, especially the Extending Images section. Day to day different GPUs will be available. Some days we had easy access to A100s and A10s, other days we had to pivot. In general don't assume you can run your tests last minute.

The default permanent memory size is only 5 GB, however you are able to request more in the NRP matrix chat. Based on our experience with these requests we recommend you mention 1) that you are a student 2) that you are using NRP's JupyterHub hosted at https://jupyterhub-west.nrp-nautilus.io/ 3) the namespace you're under and your username and 4) (possibly optional) some explanation of why you need this amount of storage. The administrators are overall very helpful and fairly quick at responding to requests. There are a lot of resources available, though best practice is to ask for only as much as you need (with some rounding/extra margin). You can always ask them to increase your storage again later.

Example message we have sent: "Hi admins, I'm a student using the NRP-managed JupyterHub (at jupyterhub-west.nrp-nautilus.io) and I'd like to request an increase in persistent storage to 128 GB to store some additional data for my ML project. My user is lundev@sonoma.edu under the namespace ssu-ggill."

When installing python packages that you want to persist after this session run `pip install --user <package_name>`. Other useful commands we have noted include `nvidia-smi` which allows you to see how much vRAM your GPU(s) have, and `df -h` to see your persistent local memory usage.