BioSpace25 - Biodiversity insight from Space
10 - 14 February 2025 | ESA-ESRIN | Frascati - Italy

How do Earth Observation Foundation Models Help to Predict Multi-Trophic Soil Biodiversity

CERNA Selene, SI-MOUSSI Sara, MIELE Vincent, THUILLER Wilfried

Laboratoire d'Ecologie Alpine
Grenoble, France

Introduction and Objectives

Why predicting soil biodiversity?

To address critical threats such as:

- Land use change and intensification.
- Desertification
- Increased levels of pollution.
- Climate change.

LAND USE for NET ZERO HUB | About the Hub | About Land Use & Net Zero | Resources

3. Soil biodiversity is likely in decline, but it is key to successful above ground biodiversity

REVIEW

Received 25 Apr 2015 | Accepted 9 Oct 2015 | Published 23 Nov 2015

Extinction risk of soil biota

Stavros D. Veresoglou[1,2], John M. Halley[3] & Matthias C. Rillig[1,2]

Identifying potential threats to soil biodiversity

Mark Tibbett, Tandra D. Fraser and Sarah Duddigan

ARTICLE

https://doi.org/10.1038/s43247-023-01047-2 OPEN

Climate change and cropland management compromise soil integrity and multifunctionality

Marie Sünnemann[1,2✉], Remy Beugnon[1,3,4], Claudia Breitkreuz[5,6], François Buscot[5,1], Simone Cesarz[1,2], Arwyn Jones[7], Anika Lehmann[8,9], Alfred Lochner[1,2], Alberto Orgiazzi[7], Thomas Reitz[1,5], Matthias C. Rillig[8,9], Martin Schädler[1,10], Linnea C. Smith[1,11], Anja Zeuner[1,2], Carlos A. Guerra[1,2,11,12] & Nico Eisenhauer[1,2,12]

Introduction and Objectives

Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

Introduction and Objectives

Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

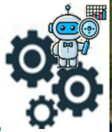
Challenges when building predictive models:

The largely **unknown** diversity of **soil organisms.**

→

How to overcome them in this study?

Built **efficient ML models** for predicting soil trophic abundance.



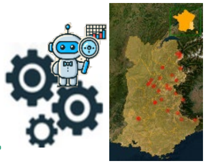
Introduction and Objectives

Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

Challenges when building predictive models:

How to overcome them in this study?

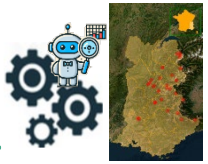
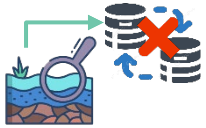
The largely **unknown** diversity of **soil organisms.** → Built **efficient ML models** for predicting soil trophic abundance.



Taxonomic and **technical limitations** in species identification. → Use **eDNA metabarcoding** to estimate soil trophic group abundances.

Introduction and Objectives

Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

Challenges when building predictive models:	How to overcome them in this study?
 The largely **unknown** diversity of **soil organisms.**	Built **efficient ML models** for predicting soil trophic abundance. 
 Taxonomic and **technical limitations** in species identification.	Use **eDNA metabarcoding** to estimate soil trophic group abundances. 
The **lack** of **large** and high-resolution environmental (soil) **datasets.** 	 Leverage pretrained **Earth Observation Foundation models** to extract belowground features.

Introduction and Objectives

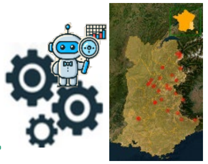
Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

Challenges when building predictive models:

How to overcome them in this study?

The largely **unknown** diversity of **soil organisms.**

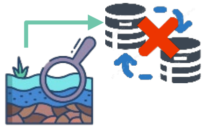
→ Built **efficient ML models** for predicting soil trophic abundance.



Taxonomic and **technical limitations** in species identification.

→ Use **eDNA metabarcoding** to estimate soil trophic group abundances.

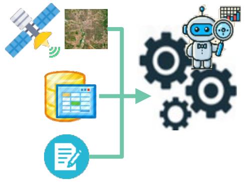
The **lack** of **large** and high-resolution environmental (soil) **datasets.**



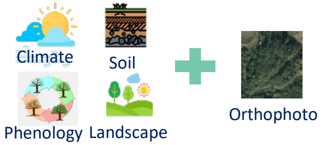
→

Leverage pretrained **Earth Observation Foundation models** to extract belowground features.

The **difficulty** of **integrating** diverse **data types** for soil prediction.



→



Integrate **tabular** data with remote sensing **features.**

Introduction and Objectives

Main goal of our study: To predict the abundance of 51 soil trophic groups in the French Alps

Challenges when building predictive models:		How to overcome them in this study?
The largely **unknown** diversity of **soil organisms.**	→	Built **efficient ML models** for predicting soil trophic abundance.
Taxonomic and **technical limitations** in species identification.	→	Use **eDNA metabarcoding** to estimate soil trophic group abundances.
The **lack** of **large** and high-resolution environmental (soil) **datasets**.	→	Leverage pretrained **Earth Observation Foundation models** to extract belowground features.
The **difficulty** of **integrating** diverse **data types** for soil prediction.	→	Integrate **tabular** data with remote sensing **features**.

Our specific objectives:

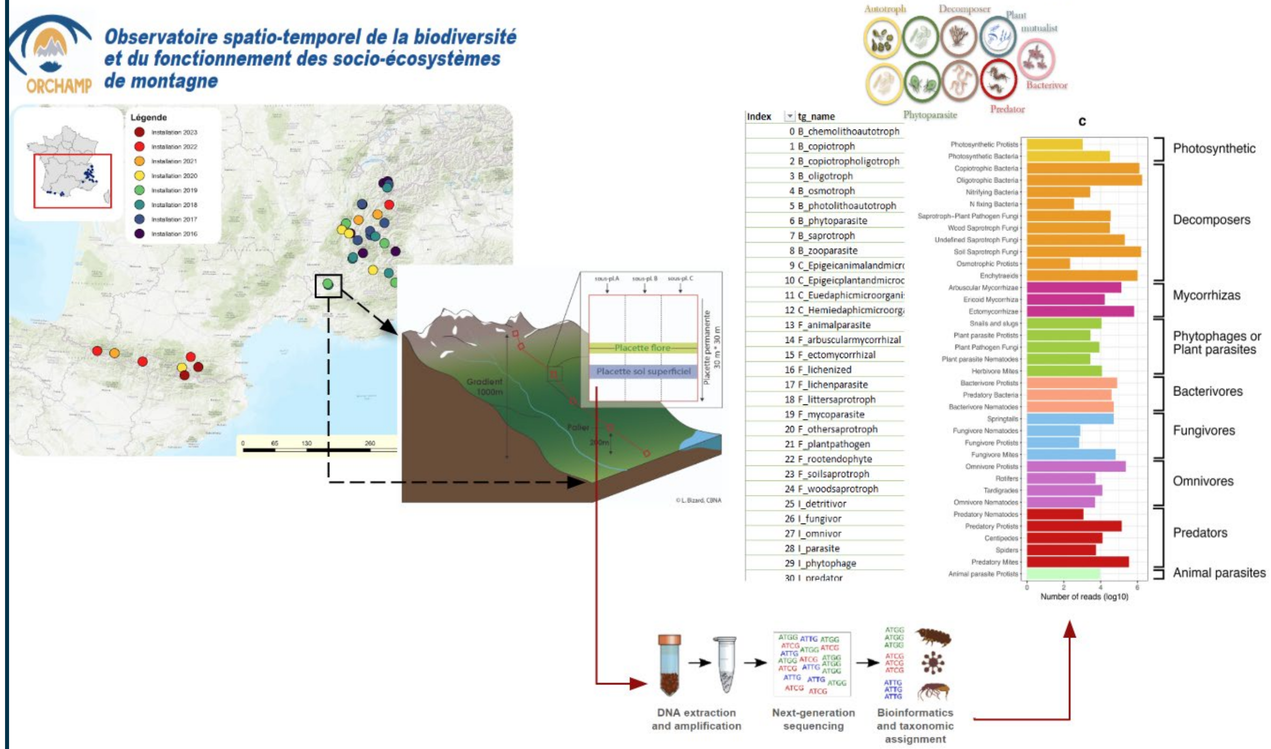
I. Comparative analysis of ML techniques and data configurations for soil trophic prediction.

II. Evaluate orthophoto features (embeddings).

III. Identify the environmental drivers.

Data Collection

Trophic group abundance (soil biodiversity)



Metabarcoding of Environmental DNA

Environmental Tabular Data

Climate

Interpolated rasters CHELSA

Mean annual temperature · Seasonal temperature · Total precipitation per year
Seasonal precipitation · Total solar radiation · Seasonal radiation

In-situ data SAFRAN-Crocus

Daily minimum temperature · Daily maximum temperature · Daily mean temperature · Daily sum of precipitation
Surface Incident Longwave Radiation · Surface Incident Diffuse Shortwave Radiation · Wind speed
Surface Pressure · Relative Humidity · Saturated water vapor pressure · Temperature of 1st cm of soil…

Soil

Interpolated rasters SOILGRIDS

Nitrogen · Clay content · Silt content · Sand content
Soil organic carbon content · Soil pH in H2O

In-situ data ORCHAMP

Acidity of the soil · Carbon to nitrogen ratio · Soil moisture · Soil nitrogen
Cation-exchange capacity · Clay content · Silt content · Sand content

Phenology (Copernicus - Vegetation)

Interpolated rasters

Seasonal productivity · Season length · Amplitude

Landscape (THEIA OSO Lancover Map)

Interpolated rasters

Hardwood forest · Meadow · Roads
Coniferous forest · Mineral surfaces · Dense builtup
Grassland · Sparse built up · Water bodies
Shrubs · Industrial zone · Rivers, streams

Orthophotos

RGB

IRC

IGN INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

Resulting dataset size: Aprox. 1000 samples

Methodology: Levarage EOF Models

Challenge: Lack of soil trophic data for building models from scratch (small dataset – aprox. 1000 samples)

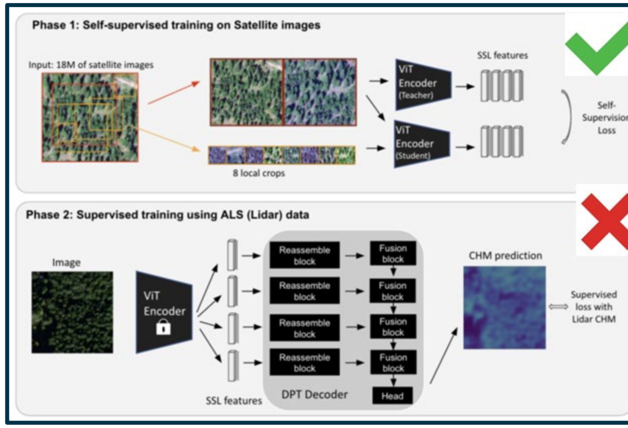
Proposed solution: Leverage pretrained EOF models to extract soil features (embeddings)

SatDINO Foundation Model by META AI Research:

Source: https://github.com/facebookresearch/HighResCanopyHeight
Paper: https://arxiv.org/abs/2304.07213 (Tolan et al.)

Why SatDINO?

- Model architecture: **ViT.**
- Training strategy: DINOv2 technique (**self-supervised learning**).
- Dataset: **forests**, **mountainous** terrains, and high degree of **tree biodiversity** (high-resolution RGB images).



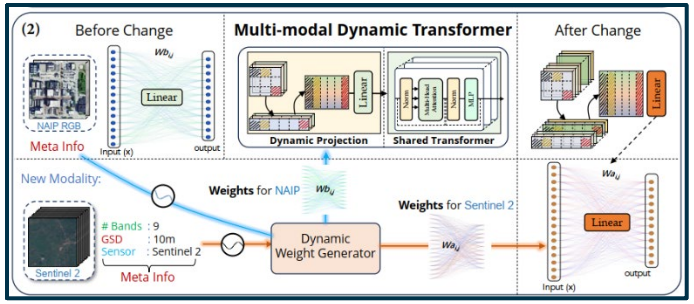
Overview of SatDINO approach for predicting canopy height. Author: Tolan et al.

Dynamic One-For-All (DOFA) Foundation Model:

Source: https://huggingface.co/XShadow/DOFA
Paper: https://arxiv.org/abs/2403.15356 (Xiong et al.)

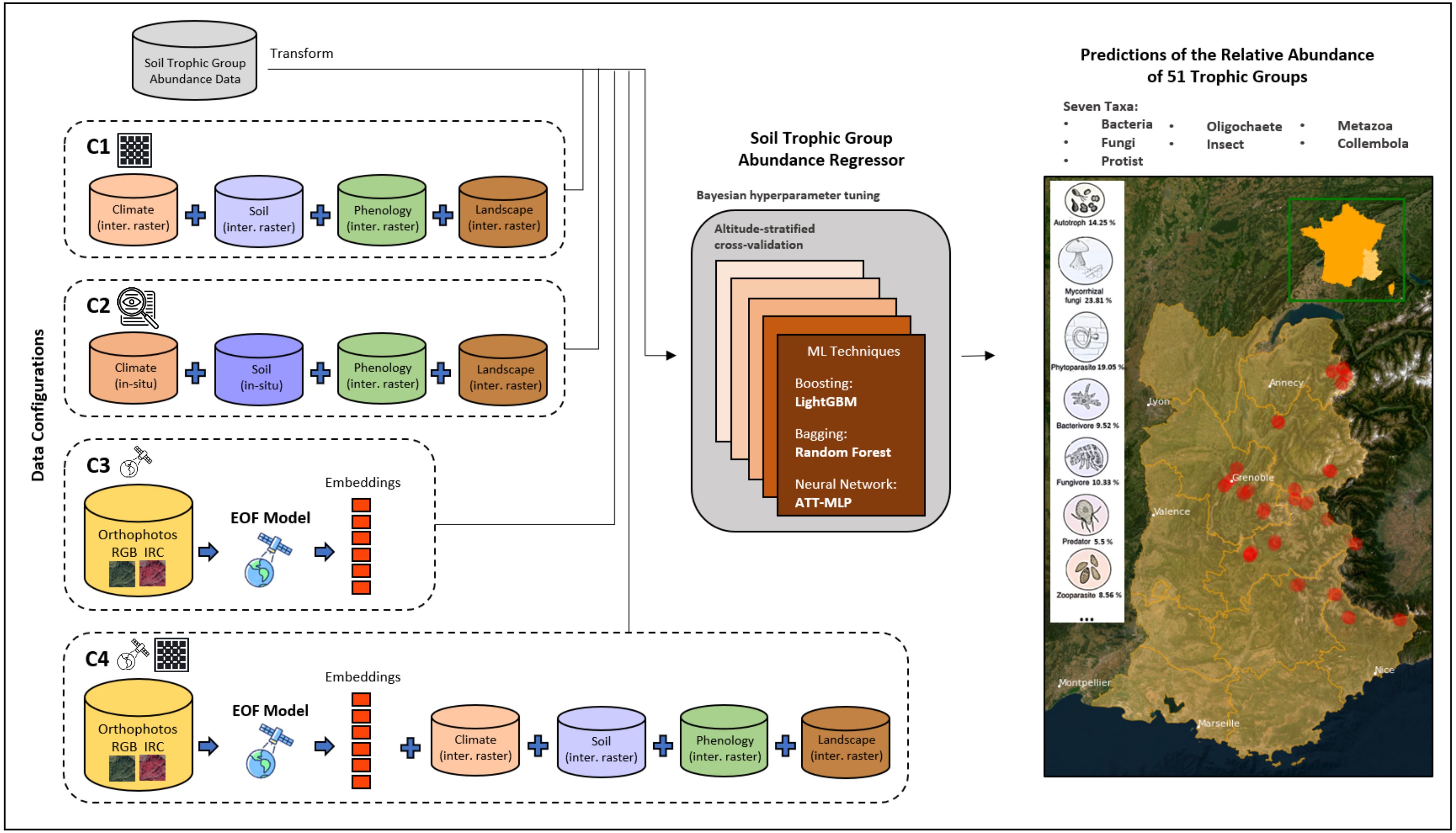
Why DOFA?

- Model architecture: **ViT.**
- Training strategy: **Masked** modeling, **wavelength-conditioned** dynamic patch embedding, and multimodal **distillation** pretraining.
- Dataset: Setnetinel-1 (**SAR**), Sentinel-2 (**multispectral**), NAIP (**RGB**), EnMAP (**hyperspectral**).



DOFA's architecture emulating the neuroplasticity mechanism for processing multimodal EO data. Author: Xiong et al.

Methodology: Overview

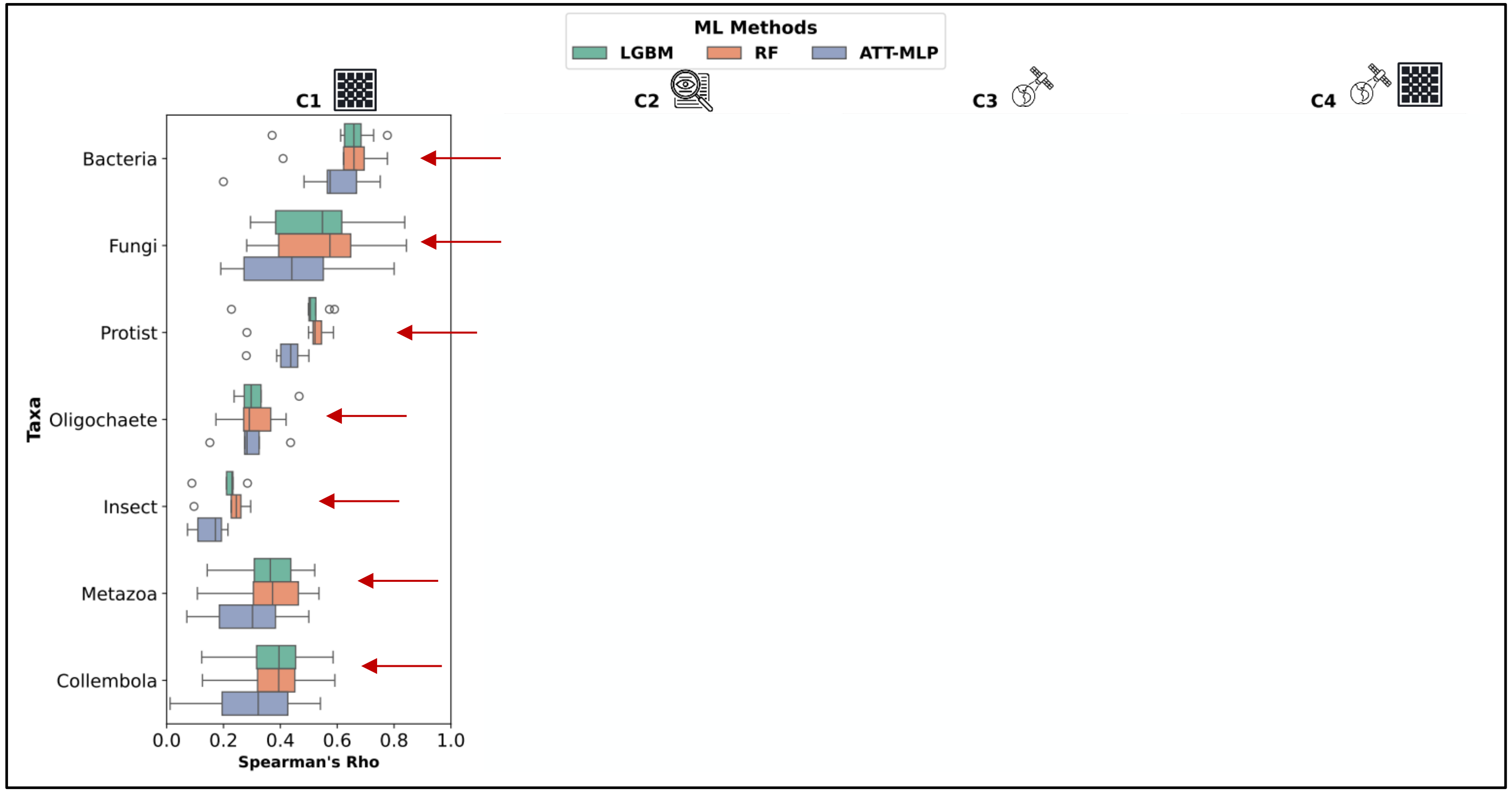


Can **embeddings** match or **improve** the performance of models based on **in-situ tabular data**?

Results: Performance of ML models

ML techniques:

- LGBM and Random Forest showed the best performances.
- LGBM is the fastest in all data configurations.



Average execution time in sec:
LGBM -> C1 and C2 = 0.71, C3 and C4 = 6.37
RF -> C1 and C2 = 2.43, C3 and C4 = 35.61
ATT-MLP -> C1 and C2 = 273.85, C3 and C4 = 354.83

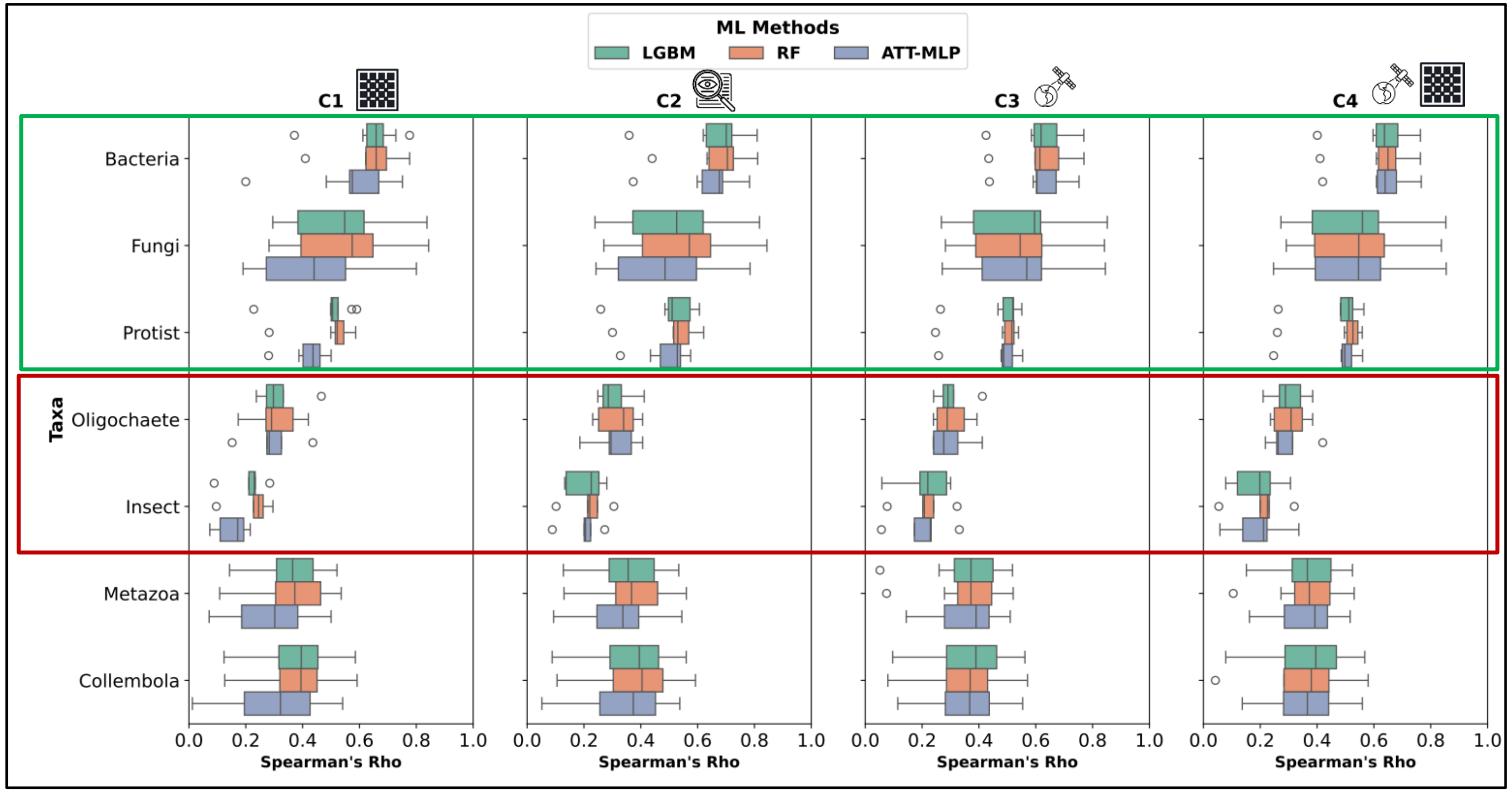
Results: Performance of ML models

ML techniques:

- LGBM and Random Forest showed the best performances.
- LGBM is the fastest in all data configurations.

Taxa:

- Bacteria, Fungi, and Protist show consistently higher Spearman's Rho correlation.
- Oligochaete and Insect perform poorly.



Average execution time in sec:
LGBM -> C1 and C2 = 0.71, C3 and C4 = 6.37
RF -> C1 and C2 = 2.43, C3 and C4 = 35.61
ATT-MLP -> C1 and C2 = 273.85, C3 and C4 = 354.83

Results: Performance of ML models

ML techniques:

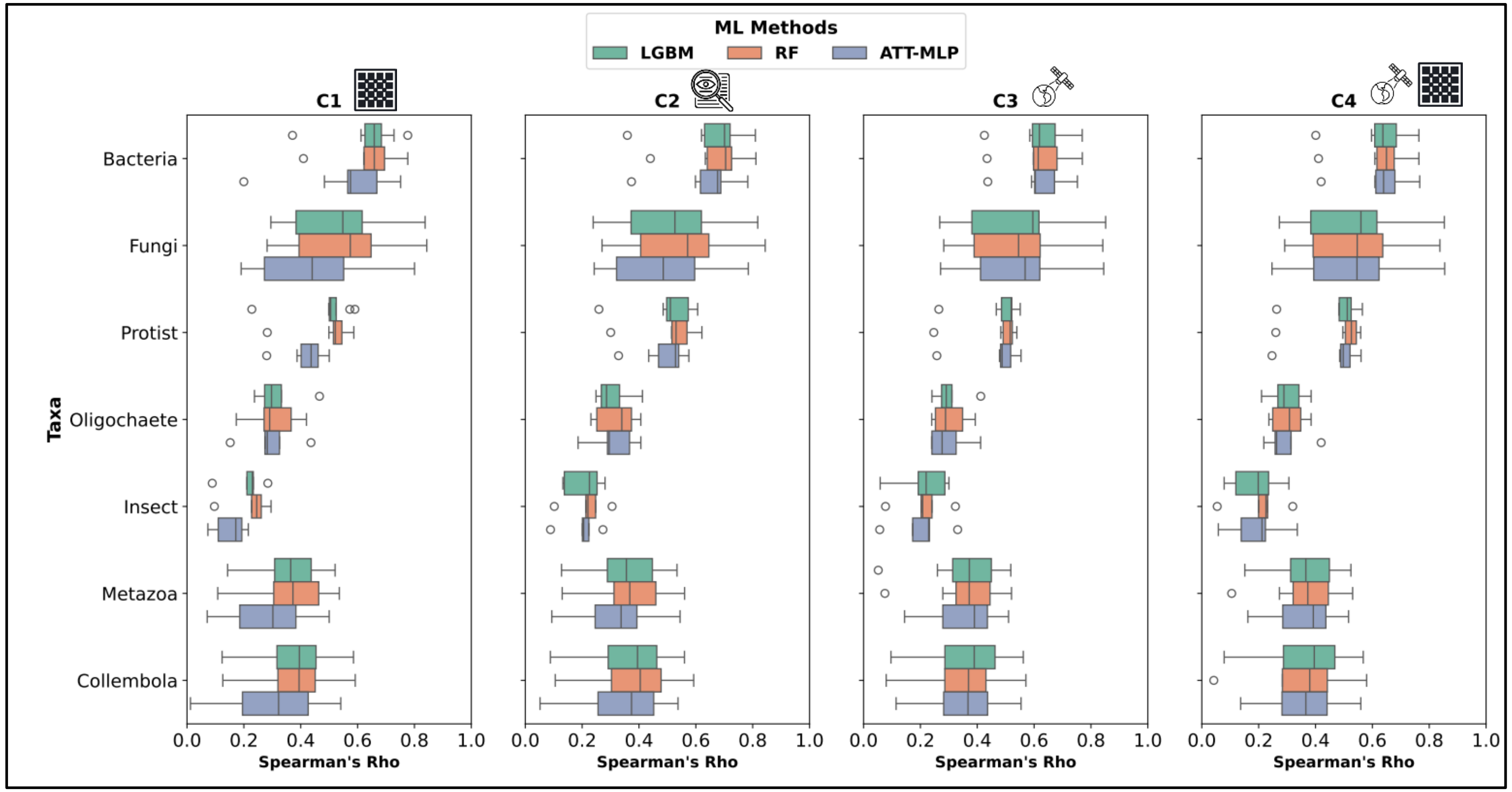
- LGBM and Random Forest showed the best performances.
- LGBM is the fastest in all data configurations.

Taxa:

- Bacteria, Fungi, and Protist show consistently higher Spearman's Rho correlation.
- Oligochaete and Insect perform poorly.

Data configurations:

- C2 generates the best performance.
- C3 (embeddings) captures relevant information but not enough to overcome C1 or C2 or C4… **WHY**?

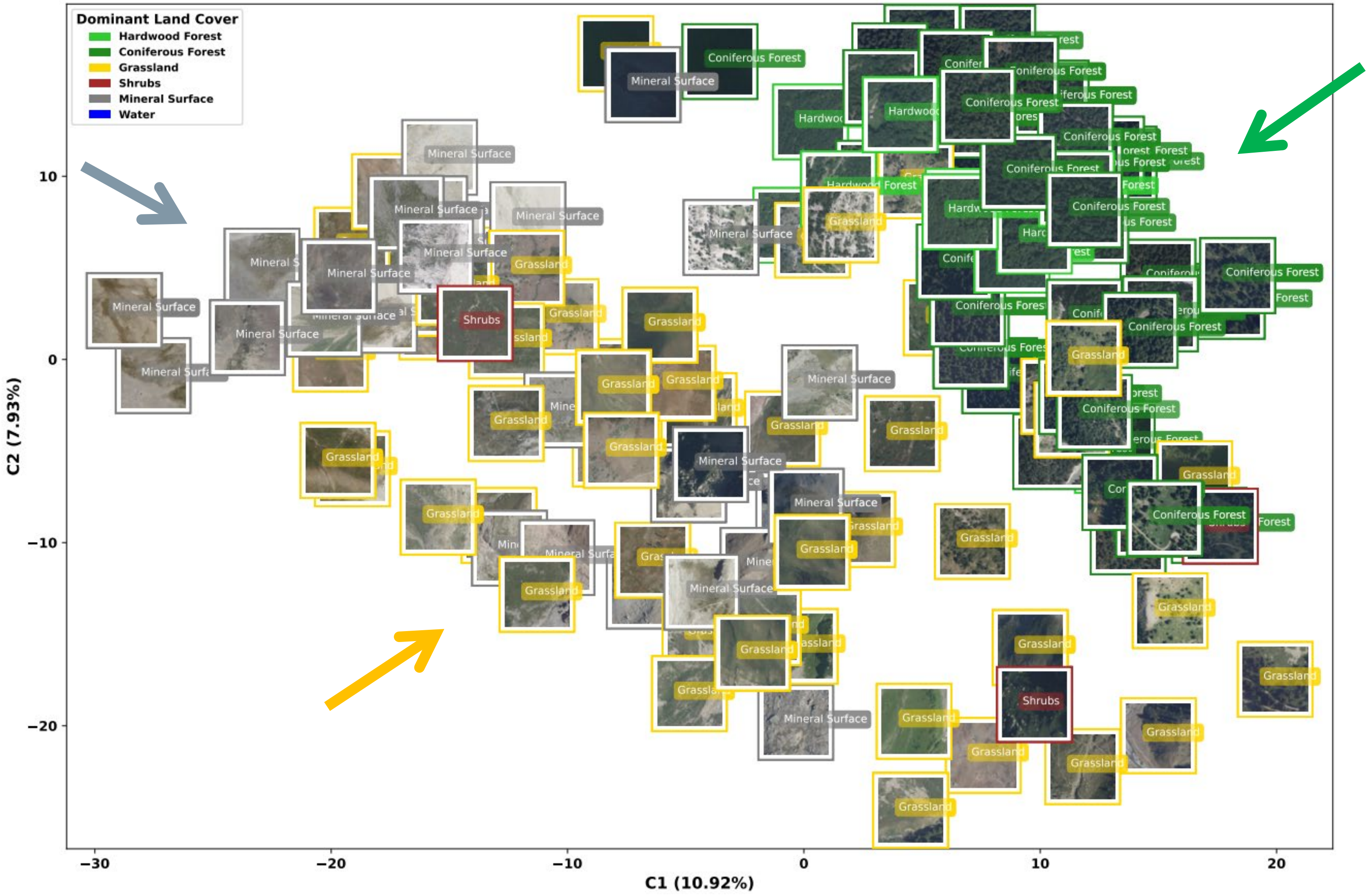


	SatDINO		DOFA			
	C1	C2	C3	C4	C3	C4
B_Chemolithoautotroph	0.716	0.745	0.714	0.714	0.716	0.719
B_Copiotroph	0.623	0.633	0.615	0.630	0.607	0.626
…	…	…	…	…	…	…
Overall Mean Value	0.463	0.474	0.449	0.455	0.442	0.462

Average execution time in sec:
LGBM -> C1 and C2 = 0.71, C3 and C4 = 6.37
RF -> C1 and C2 = 2.43, C3 and C4 = 35.61
ATT-MLP -> C1 and C2 = 273.85, C3 and C4 = 354.83

Results: Evaluating Embeddings

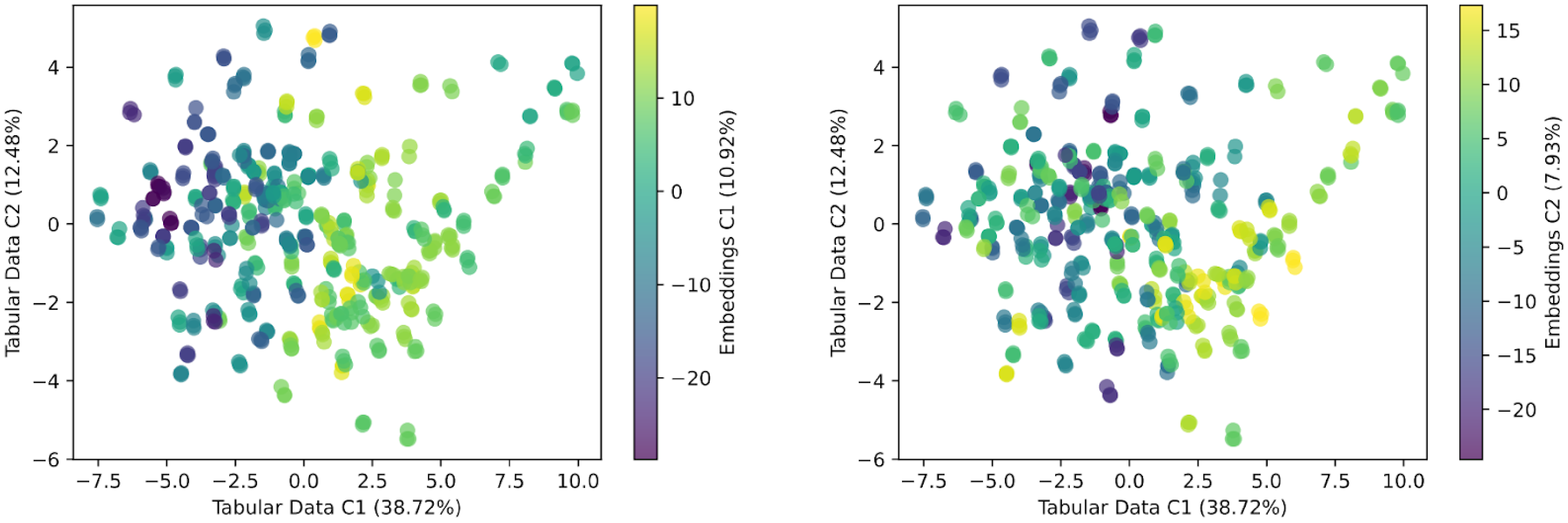
PCA projection and visualization of land cover clustering based on the SatDINO embeddings.



- SatDINO captures some level of differentiation based on land cover types.
- Grassland and Shrubs might have inherently similar visual features, making them harder to separate in the embedding space.

Results: Evaluating Embeddings

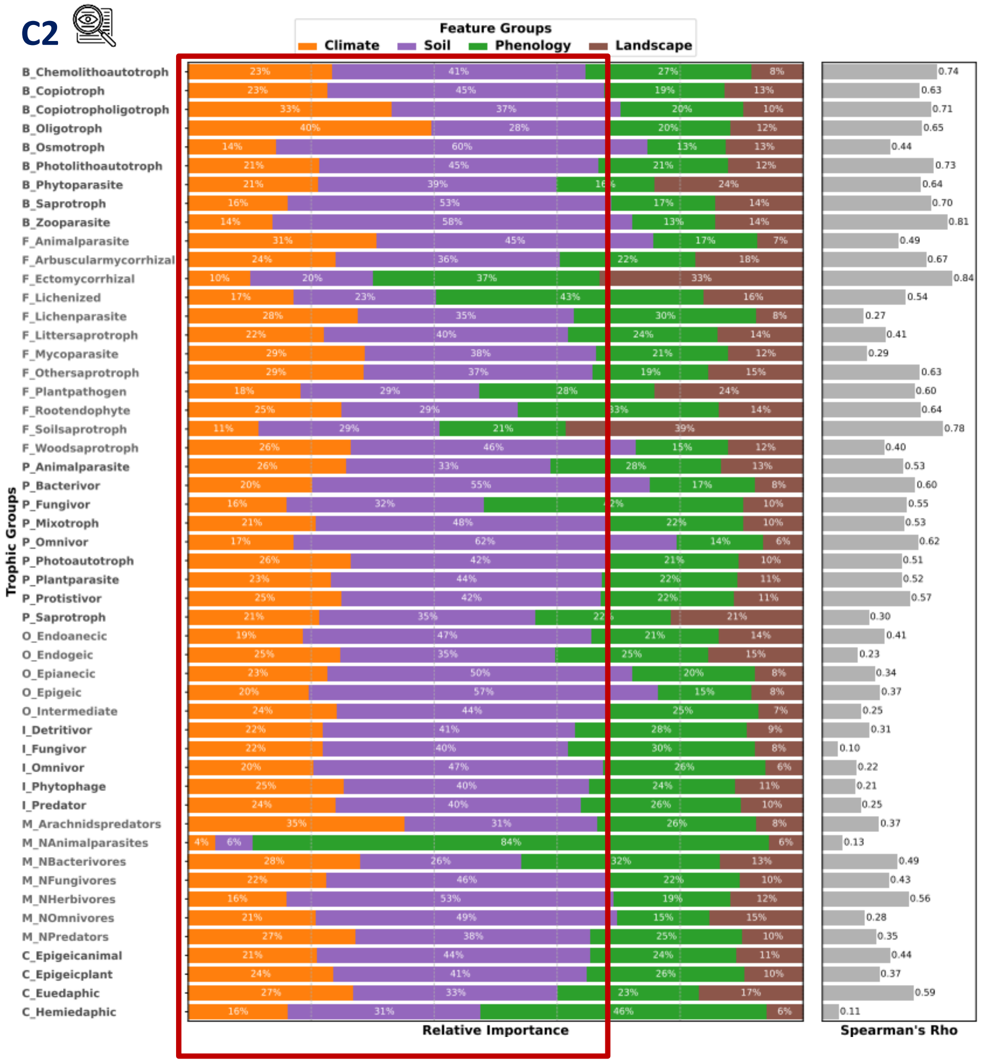
Relationship between in-situ tabular data and PCA components of SatDINO embeddings.



The color gradients in the embeddings align with patterns in the tabular data components (redundancy!).
This explains why adding embeddings (configuration 4) doesn't improve performance.

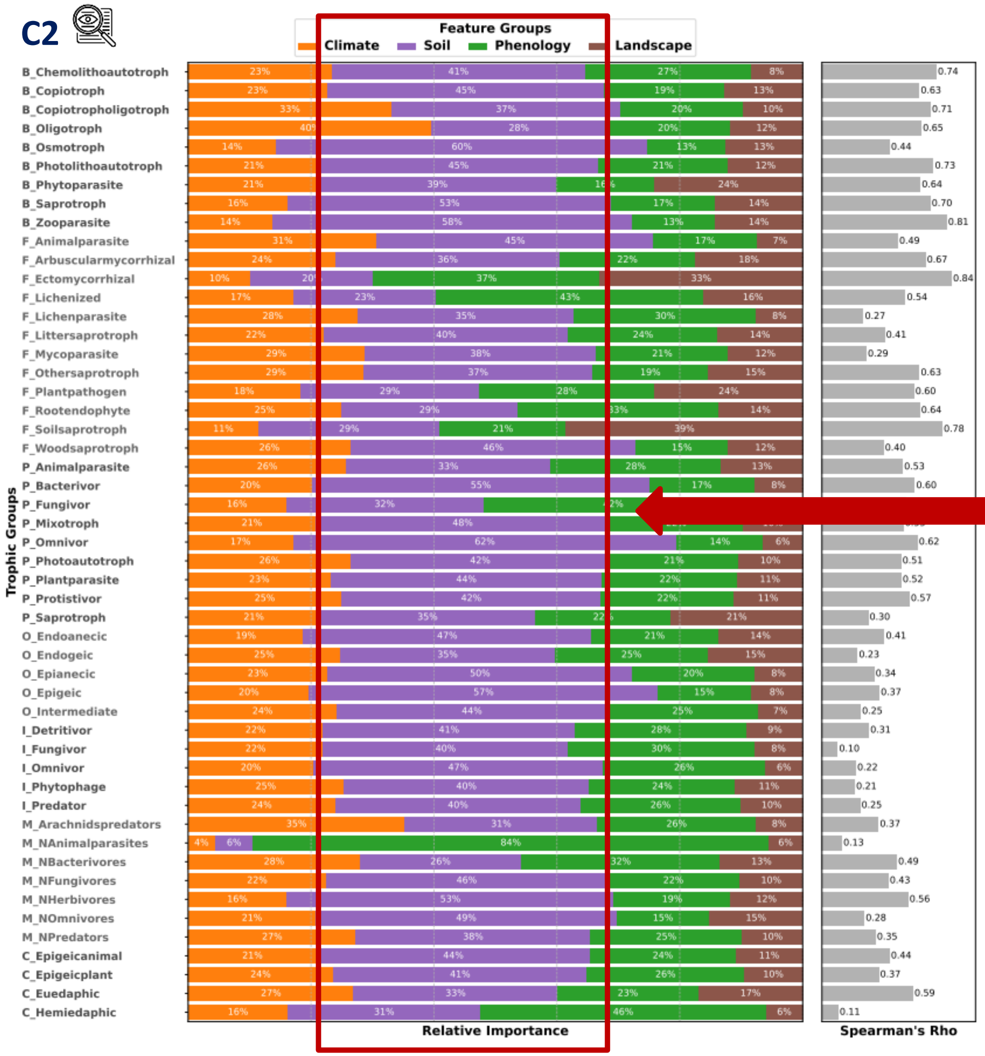
Results: Environmental Drivers

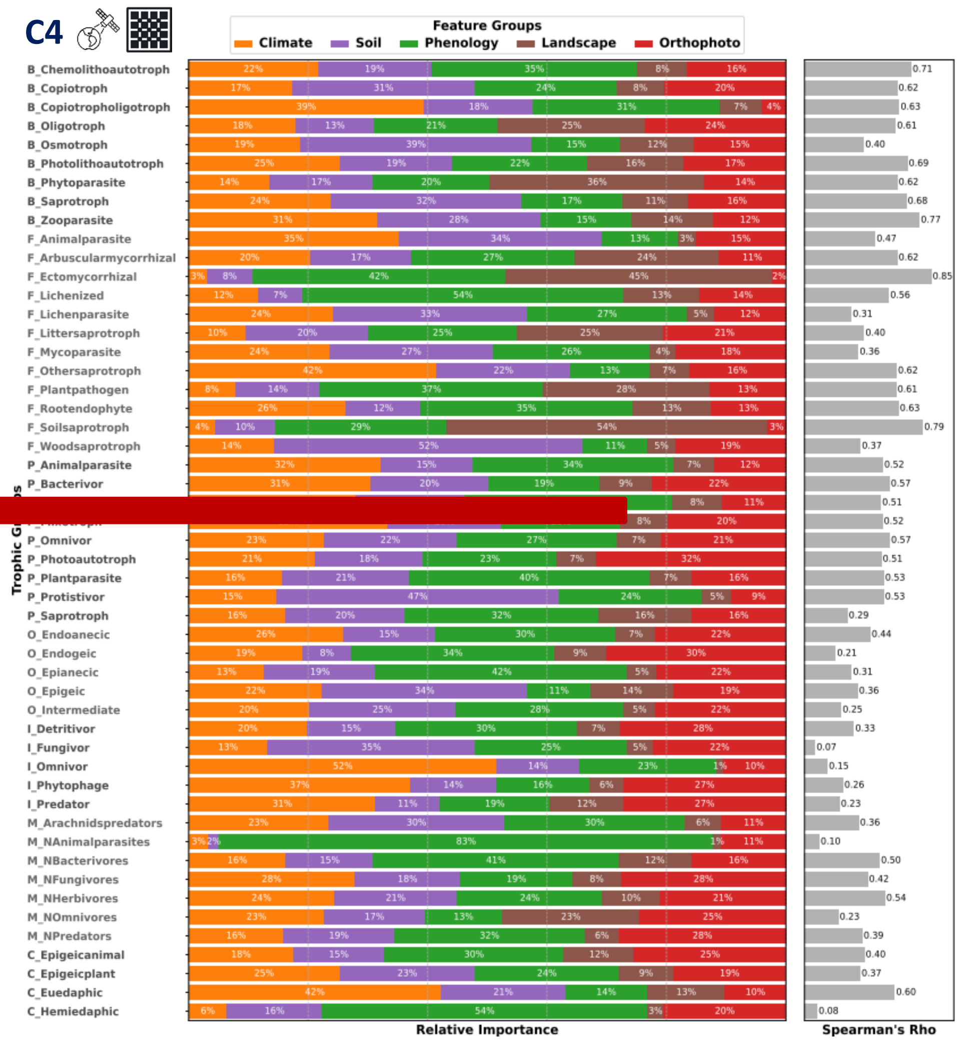
- Climate and soil are main contributors in most cases.
- Soil variables compensate the absence of embeddings in C2.
- Landscape contributions remain modest in most cases and phenology's relevance is highly group-specific.



→ THE EUROPEAN SPACE AGENCY

Results: Environmental Drivers

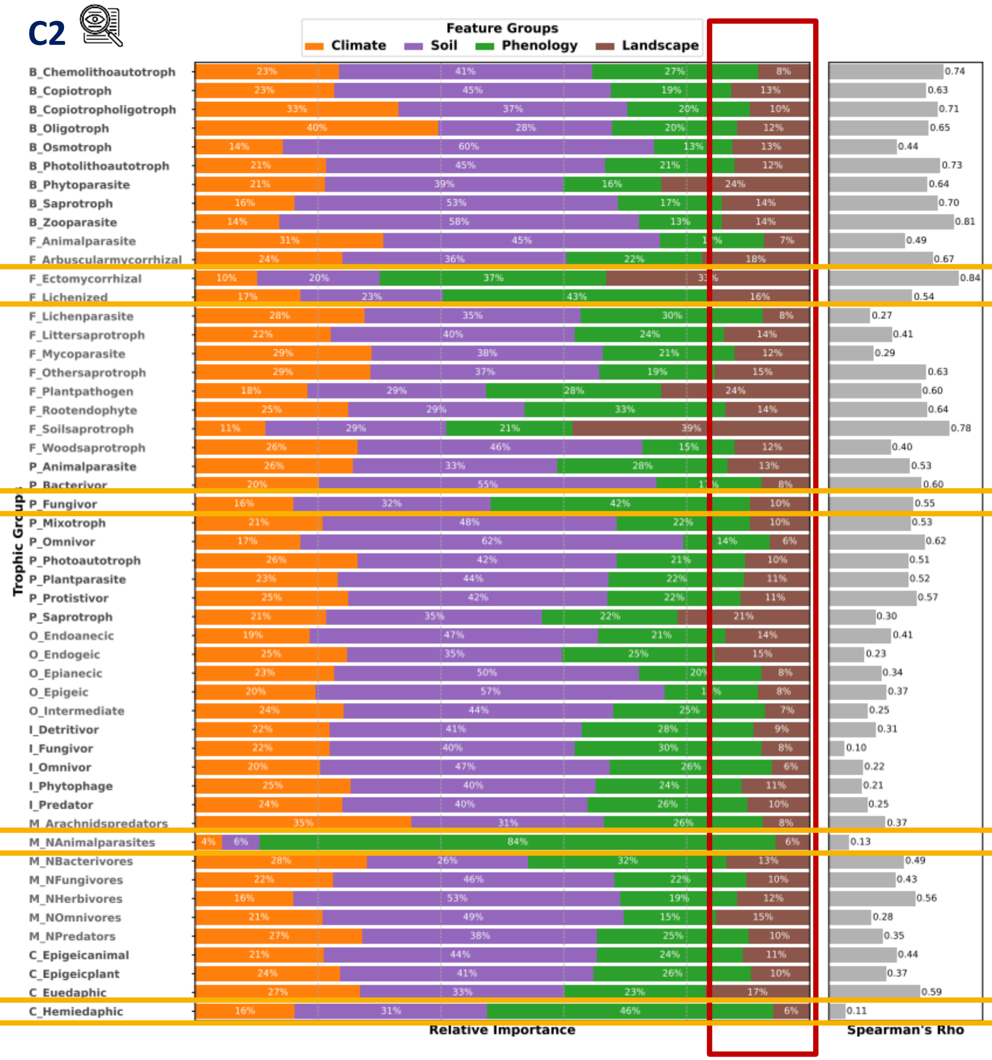


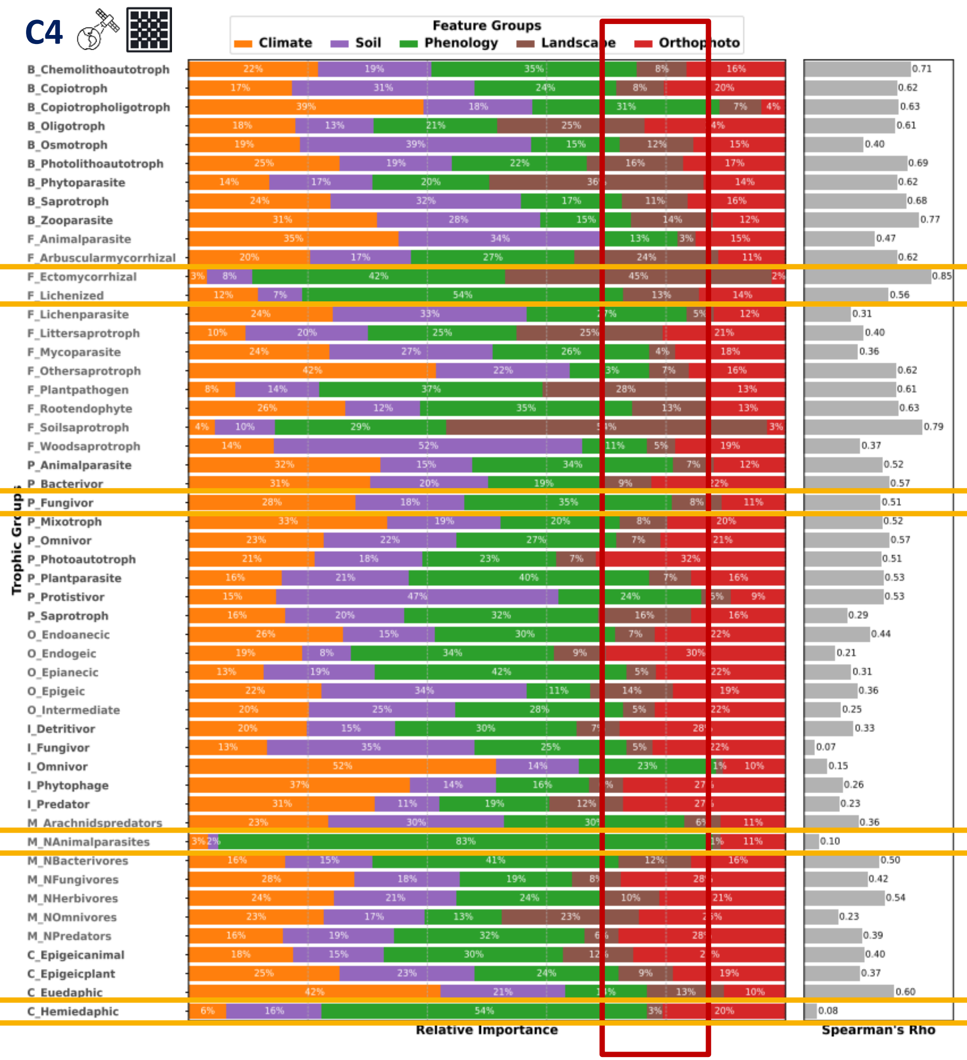


- Climate and soil are main contributors in most cases.
- Soil variables compensate the absence of embeddings in C2.
- Landscape contributions remain modest in most cases and phenology's relevance is highly group-specific.

Results: Environmental Drivers

- Climate and soil are main contributors in most cases.
- Soil variables compensate the absence of embeddings in C2.
- Landscape contributions remain modest in most cases and phenology's relevance is highly group-specific.





Conclusions

Conclusions:

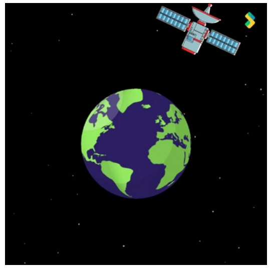
- There are **group-specific** prediction **challenges** across trophic groups.
- **Tabular** data **outperforms** embeddings.
- Embeddings capture partial information.
- **Embeddings offer an alternative** where in-situ data is scarce.

Future Work:

- **Explore multimodal** EOF models to integrate environmental data for soil biodiversity modeling.
- **Model** trophic networks to uncover species **interdependencies**.
- Investigate the feasibility of **incorporating high-resolution remote sensing** data.

Three Recommendations for the conference organizers:

- ❑ Traditional tabular environmental data remains essential for robust biodiversity modeling.
- ❑ Integrating higher-resolution hyperspectral/multispectral remote sensing data could refine environmental characterizations for biodiversity assessment.
- ❑ A coordinated evaluation of different EOF models across multiple research teams would facilitate standardized comparisons.



Thank you for your attention!

selene.cerna@univ-grenoble-alpes.fr

Université Grenoble Alpes

LECA Laboratoire d'Ecologie Alpine

MIAI Grenoble Alpes Multidisciplinary Institute In Artificial Intelligence

cnrs