



# Data Integration Methods for Deep Species Distribution Models

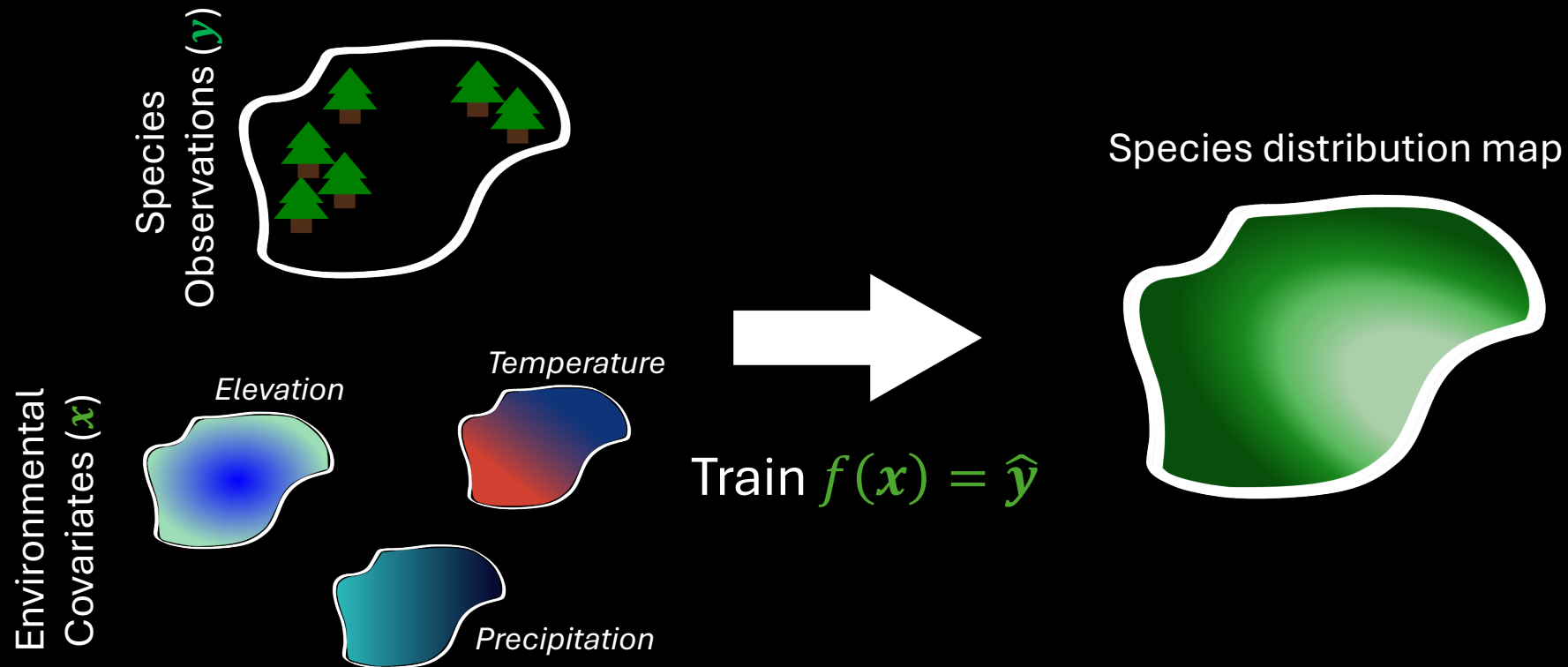
Pablo Ubilla Pavez, Diego Marcos, Christophe Botella



UNIVERSITÉ DE  
MONTPELLIER

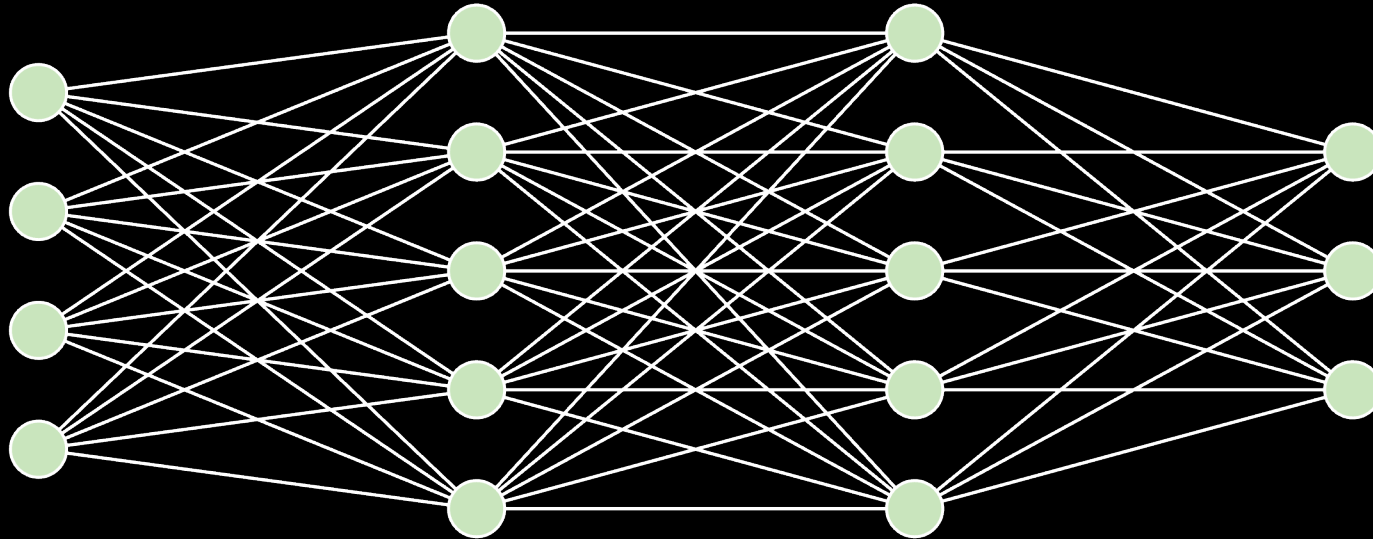
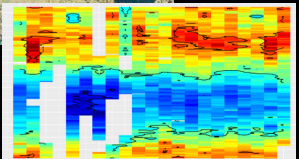
# Species Distribution Models

- We want to learn a mapping function  $f$  from environmental covariates  $x$  to a species response  $y$ .



# Deep Species Distribution Models

- Function  $f$  is a Neural Network
- Can learn non-linear relationships
- Can handle multiple modalities as input (e.g. environmental covariates, satellite images, time series)
- Can learn multiple species at the same time



Which kind of data feeds these models?



# Presence Absence (PA) Data

- PA is the ideal type of data for SDMs.
- Controlled surveys where experts track which species are present and which are not.
- Observed value for species  $j$  in location  $i$  would normally be a binary variable  $y_{ij} \in \{0,1\}$



# Presence Absence (PA) Data

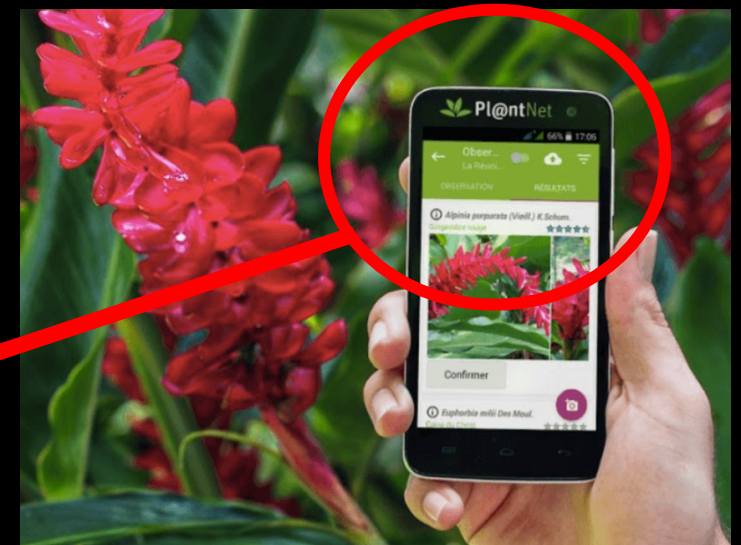


$$\mathbf{y}_i = (1, 0, 0, 1, \dots)$$

$$\mathbf{x}_i = \begin{matrix} Temp & Prec & Hum & \dots \\ 12 & 5 & 4 & \dots \end{matrix}$$

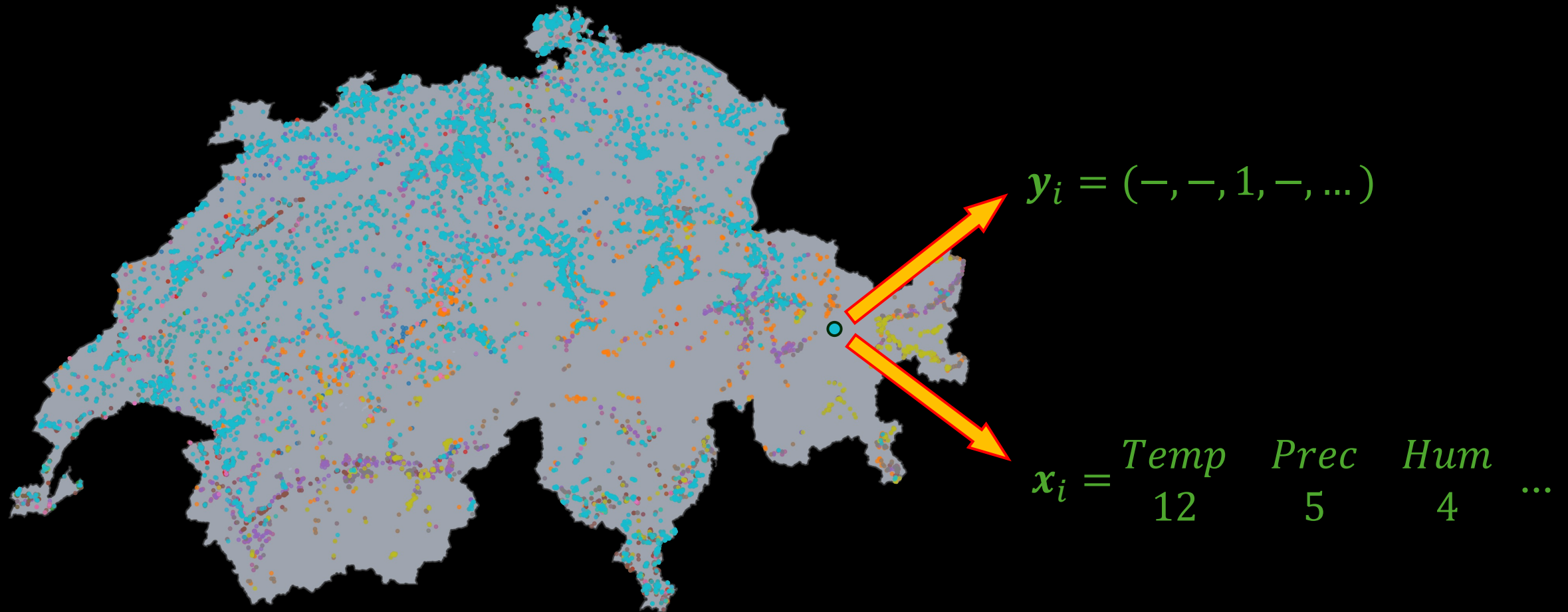
# Presence Only (PO) Data

- We are not always privileged with nice PA data
- PO data consists in opportunistic observations, normally from phone applications
- As the name says, it only gives us the presence, hence  $y_{ij} \in \{-, 1\}$



**Download Pl@ntNet!**

# Presence Only (PO) Data





# Presence Only (PO) Data

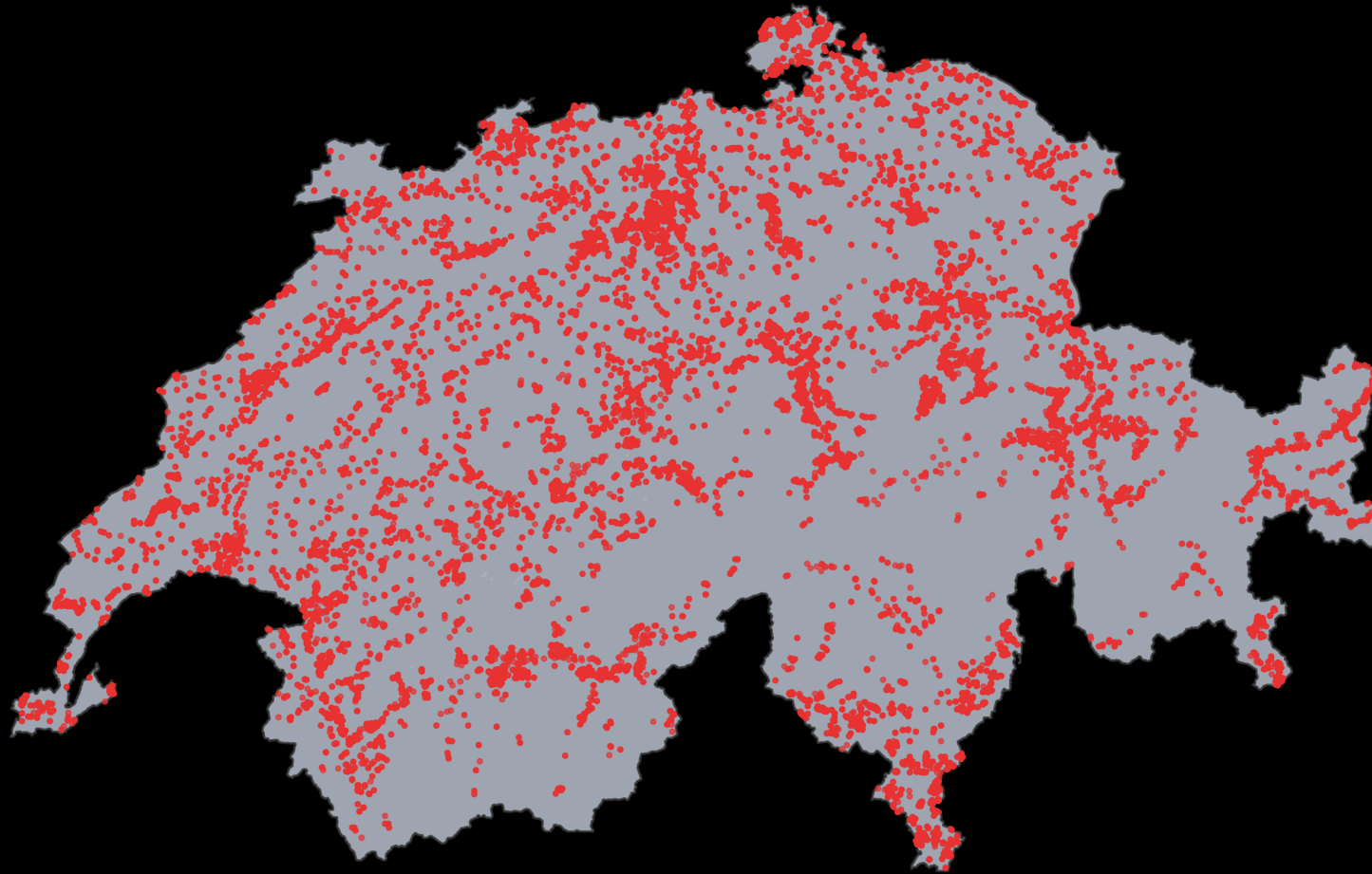


## Target Group Background:

Consider other observation for one species as background for the others.

$$y_{ij} = 0 \text{ if } \exists j' \neq j \mid y_{ij'} = 1$$

# Presence Only (PO) Data



## Target Group Background:

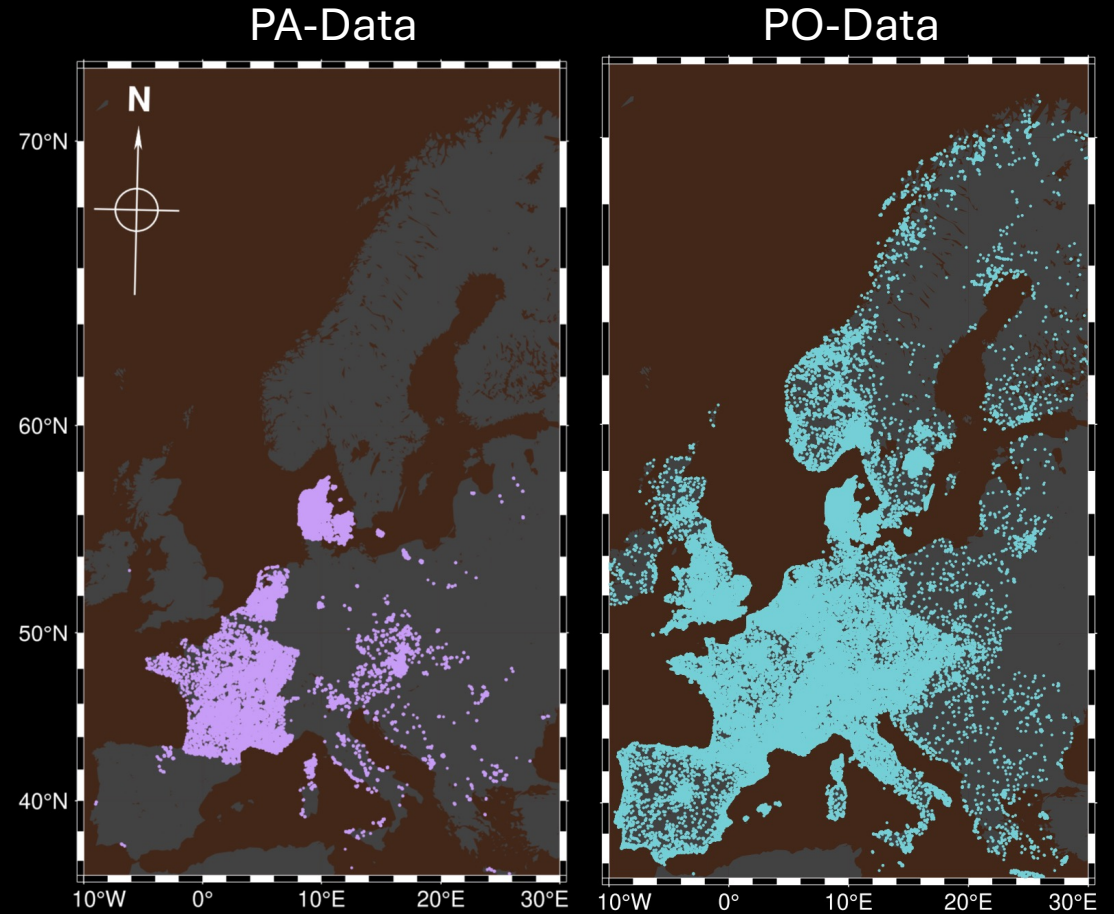
Consider other observation for one species as background for the others.

$$y_{ij} = 0 \text{ if } \exists j' \neq j \mid y_{ij'} = 1$$

We get a similar input as PA but with background instead of absences.

# Data Availability

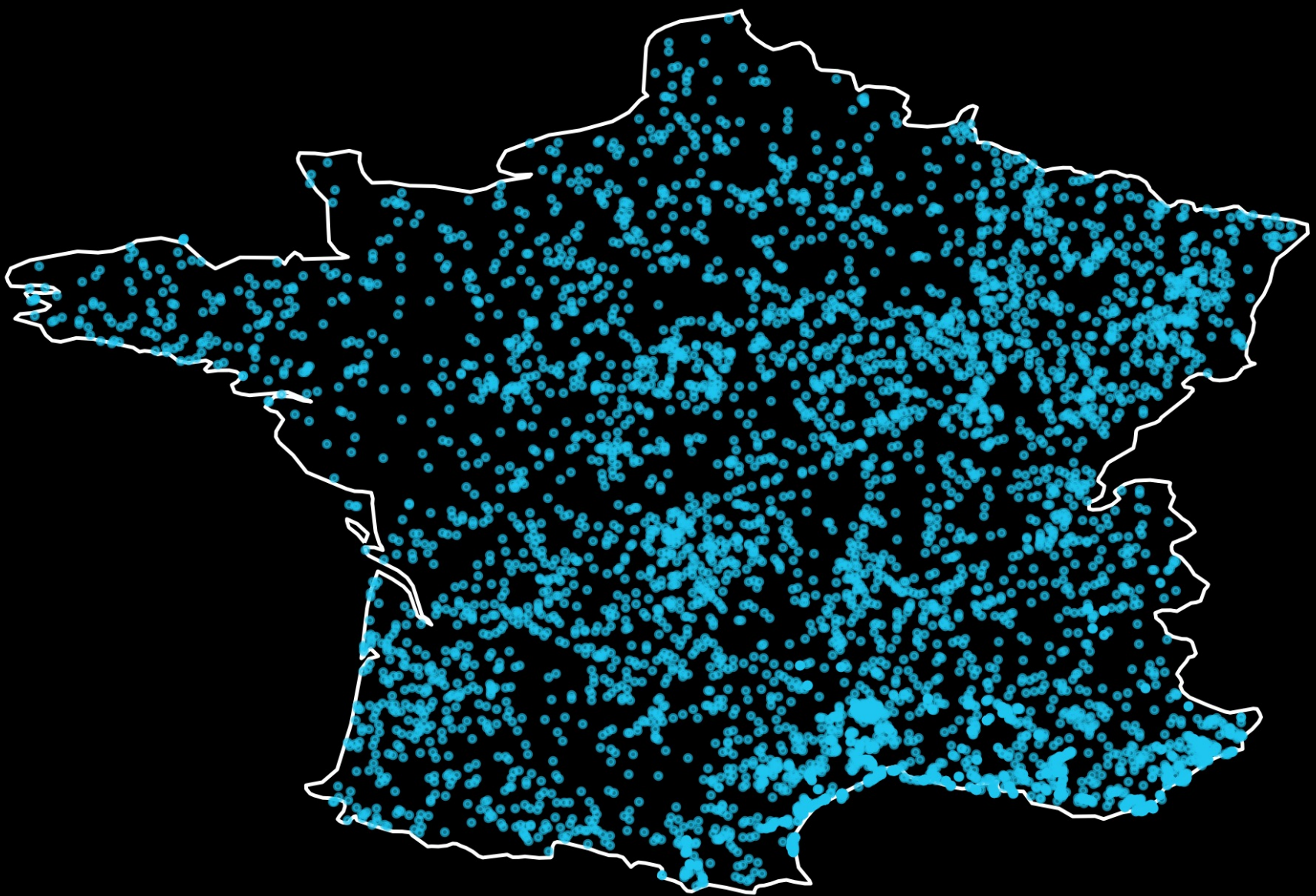
- PA data is often only available for certain regions/countries
- Even though PO data is biased, it has a much wider geographical coverage

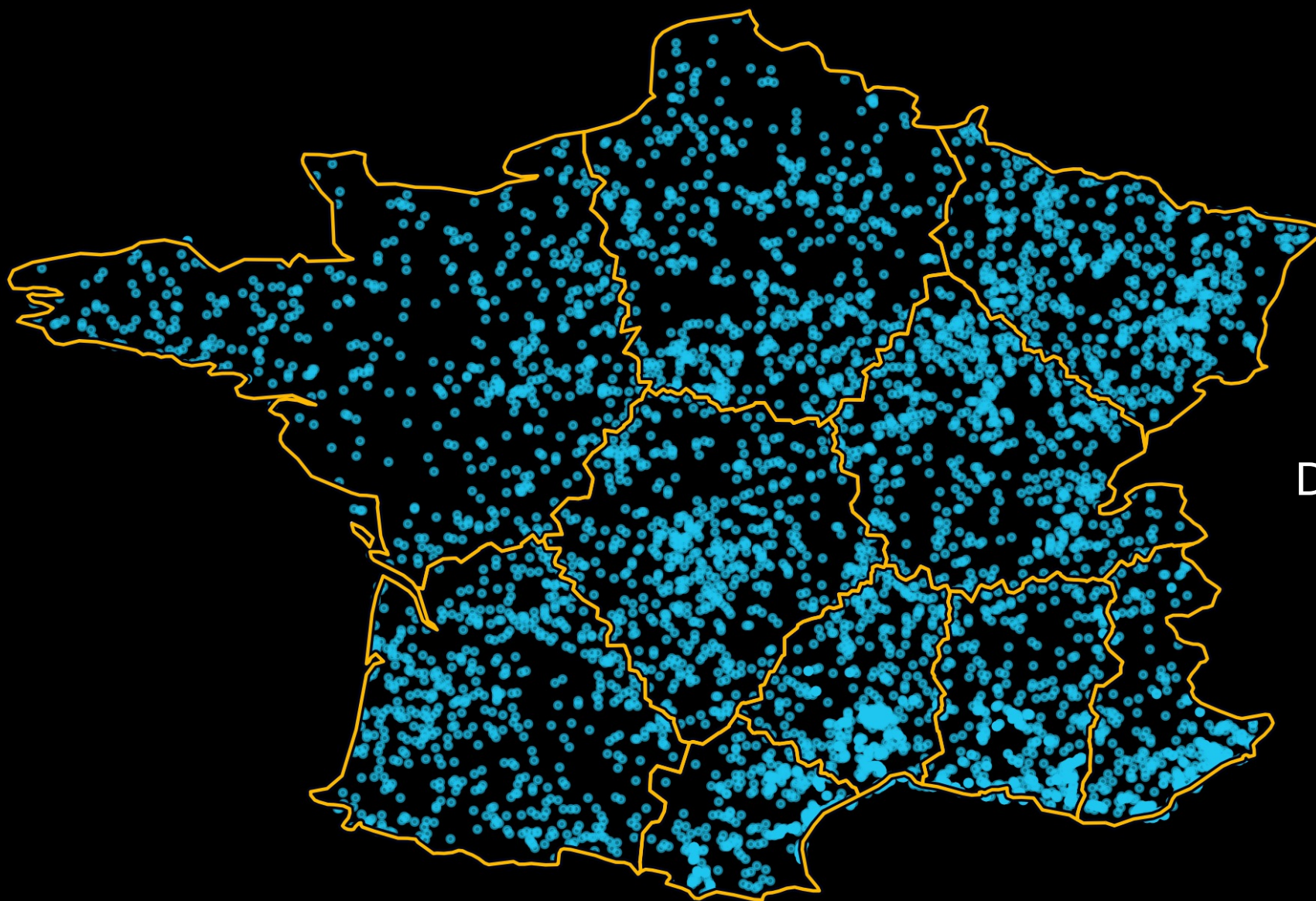


*GeoLife Clef 2023 (Botella et al.)*

How can we test if Data Integration helps  
when PA data is not available?

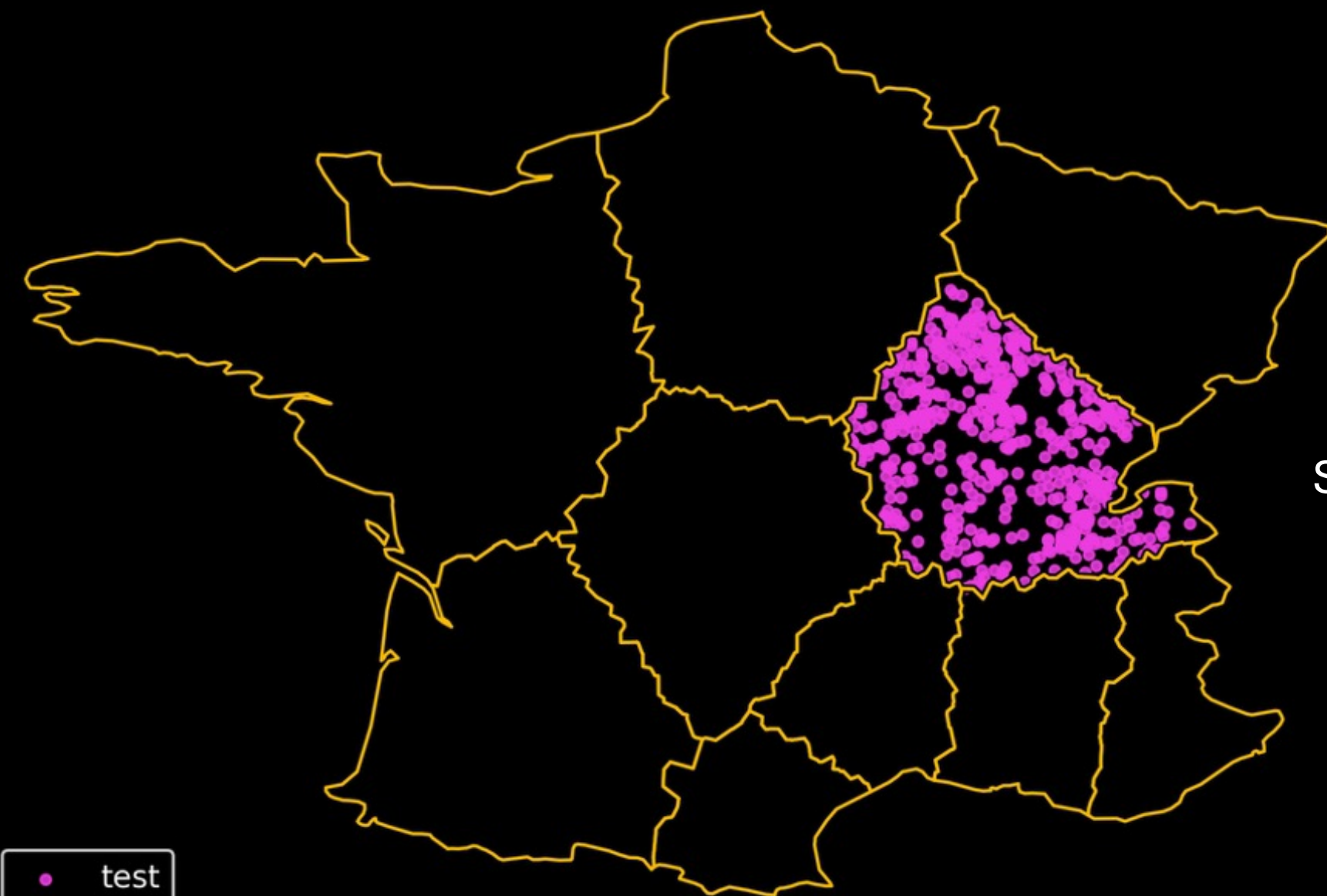






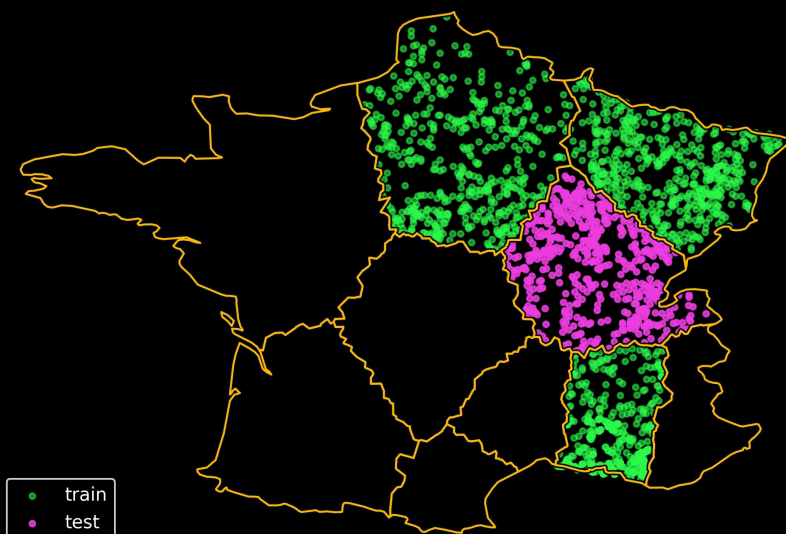
Divide space into equally  
sized clusters  
(*constrained K-means*)



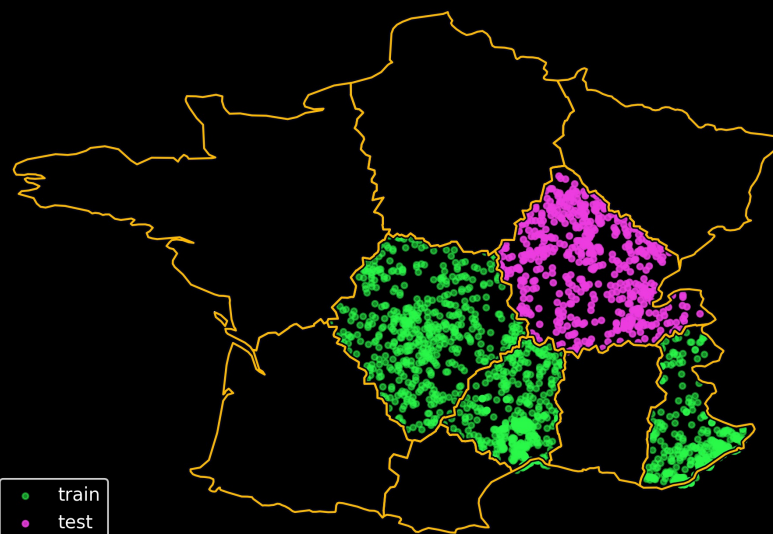


Select one cluster for test

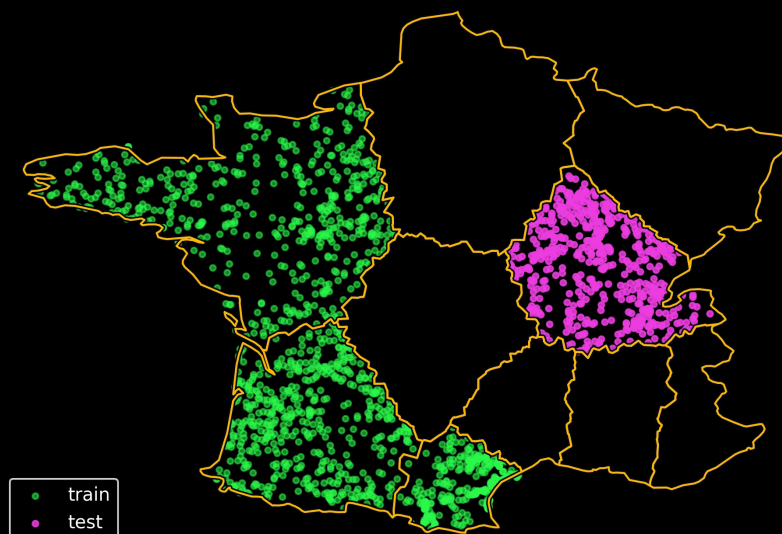
Closest percentile range  
[0,0.33]



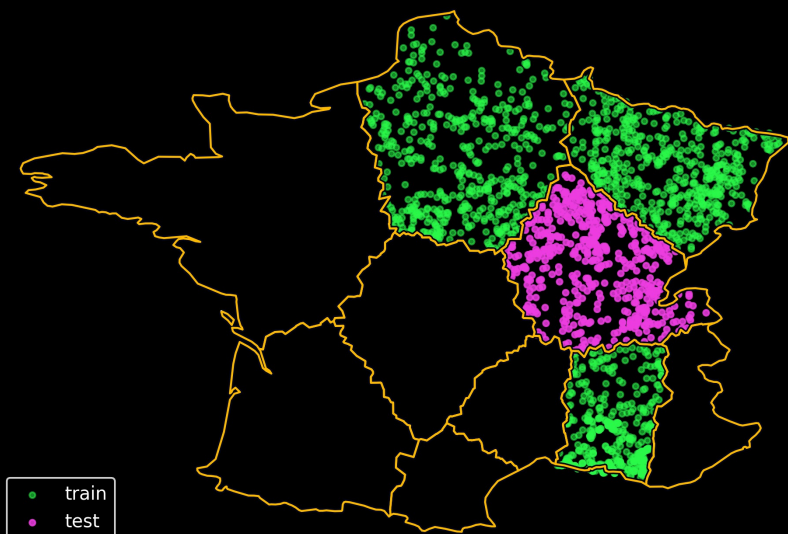
Middle percentile range  
(0.33,0.66]



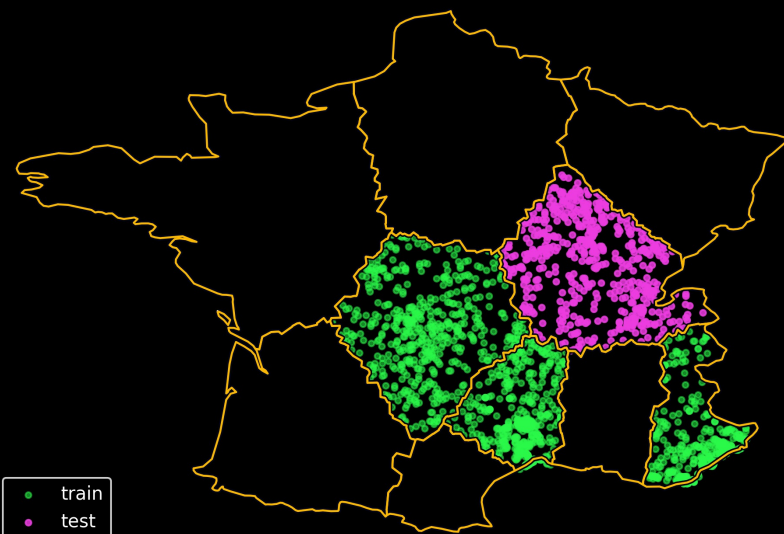
Farthest percentile range  
(0.66,1]



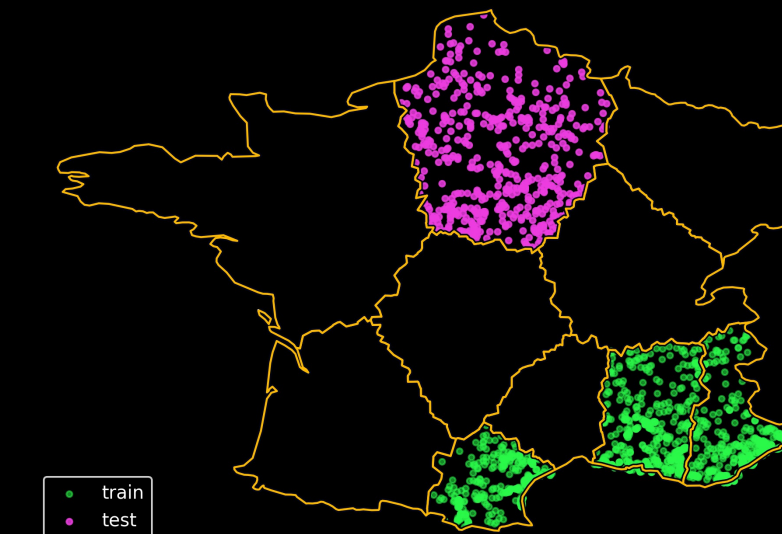
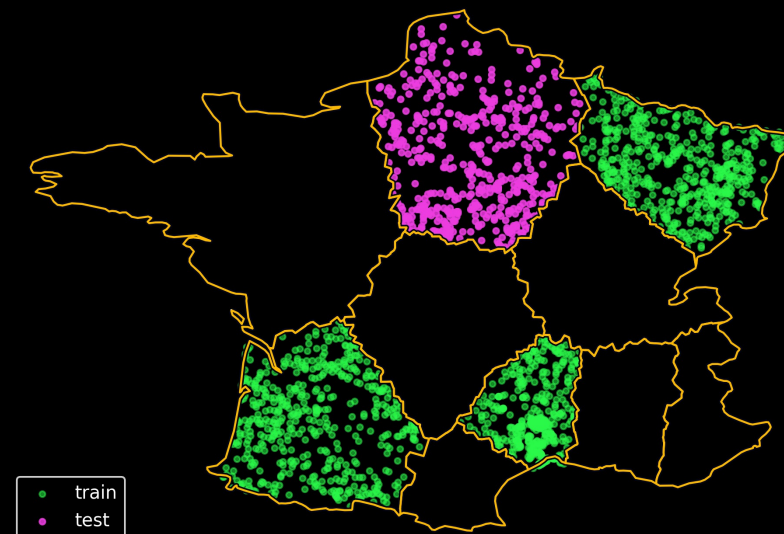
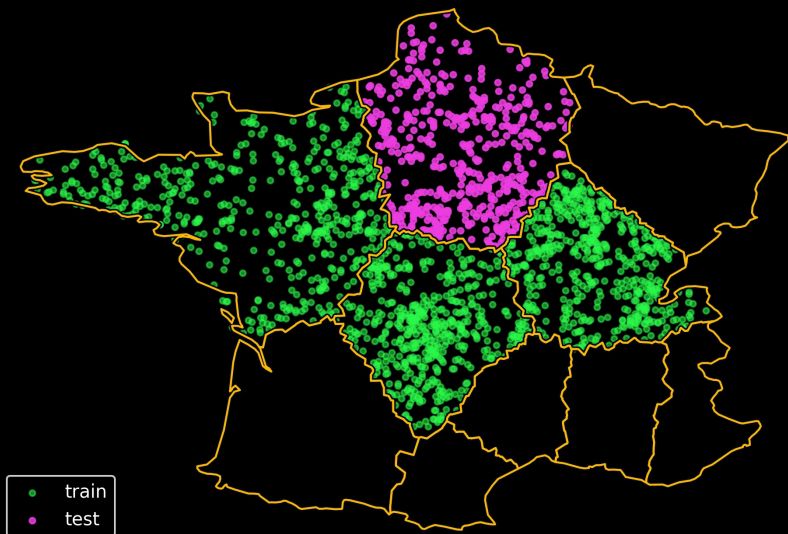
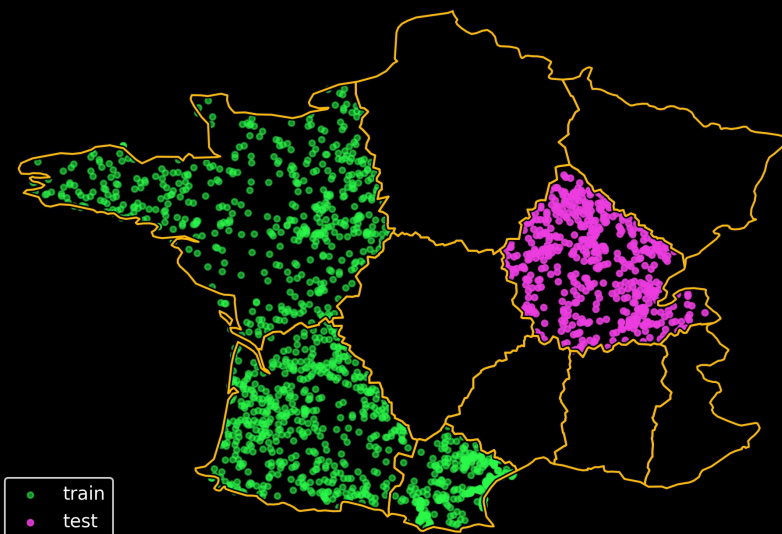
Closest percentile range  
[0,0.33]



Middle percentile range  
(0.33,0.66]

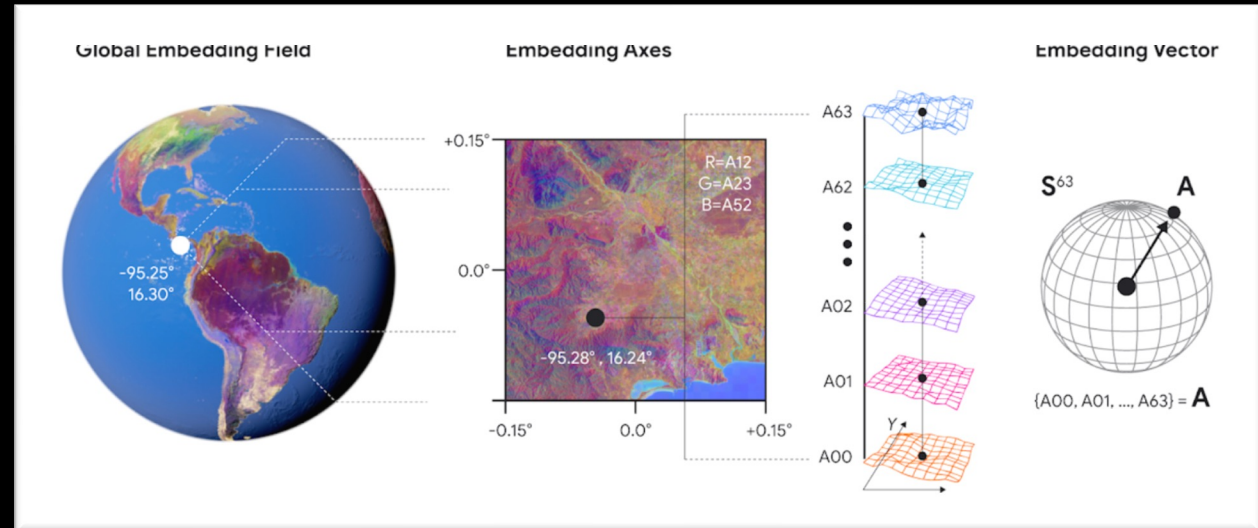
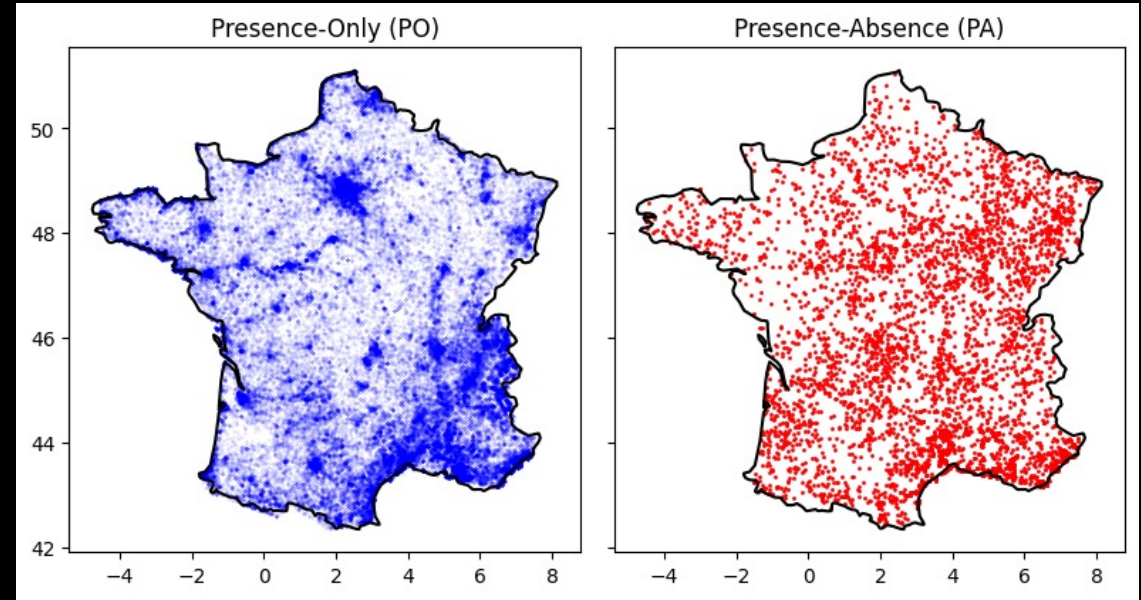


Farthest percentile range  
(0.66,1]



# Case Study

- Subset of the GeoPlant dataset in France
  - ~5000 PA plots
  - ~80000 PO observations
  - ~1600 species
- For covariates: Alpha Earth Embeddings (64 bands)





# Methods

- Multi-Species Distribution Models can be treated as a **Multi-Label Classification Problem**, where the most natural Loss is **Binary-Cross-Entropy**

$$L(y_i, f(x_i)) = - \sum_j [y_{ij} \cdot \log(f(x_i)_j) + (1 - y_{ij}) \cdot \log(1 - f(x_i)_j)]$$

- This Loss is tested with
  1. Only PO data with TGB
  2. Only PA data
  3. Concatenated PO (with TGB) and PA data

# Integration Strategy: 4. Separate Loss

- Both PO and PA come from the same process modeled by  $\lambda$ 
  - PA is considered as observing at least one species

$$Y_{ij} \sim \text{Ber}(1 - e^{-a \cdot \lambda_{ij}})$$

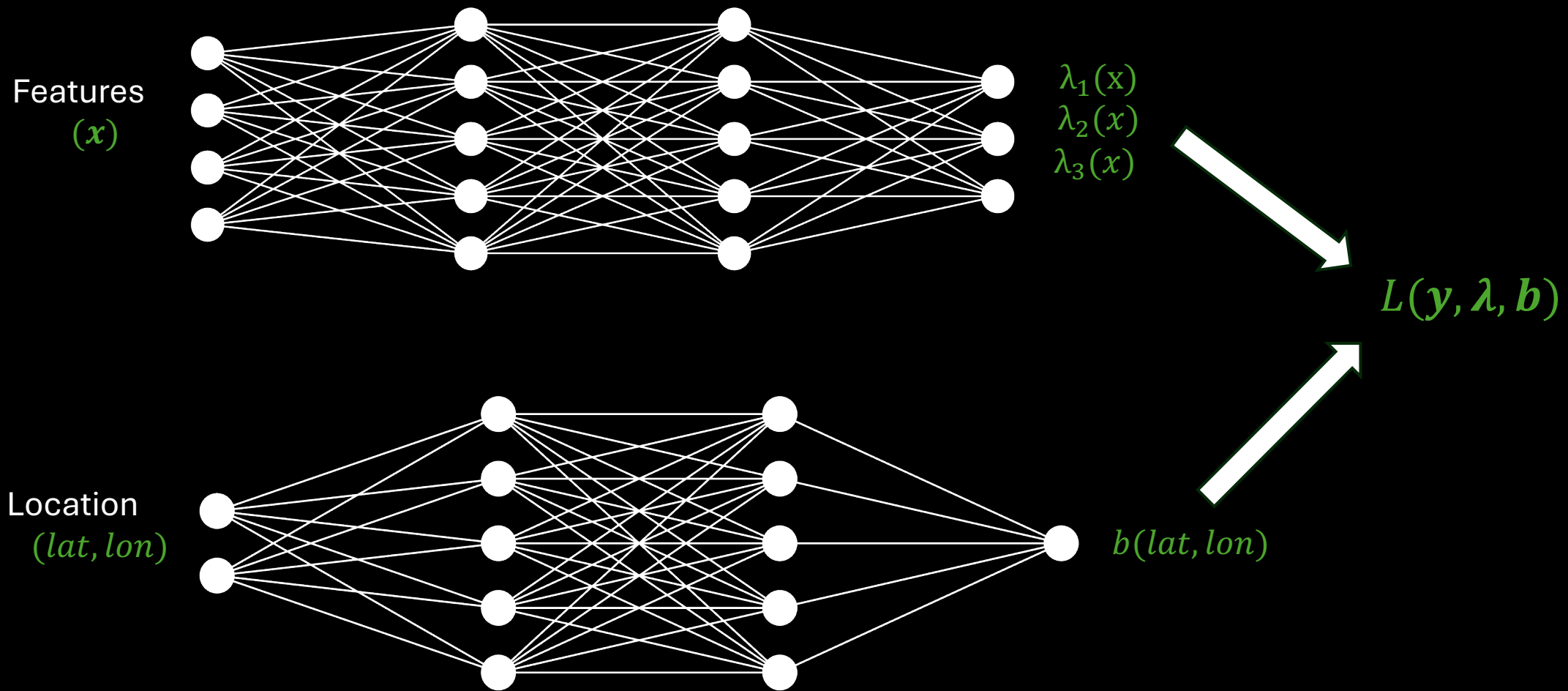
- PO comes with a bias term per site:

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij} \cdot b_i)$$

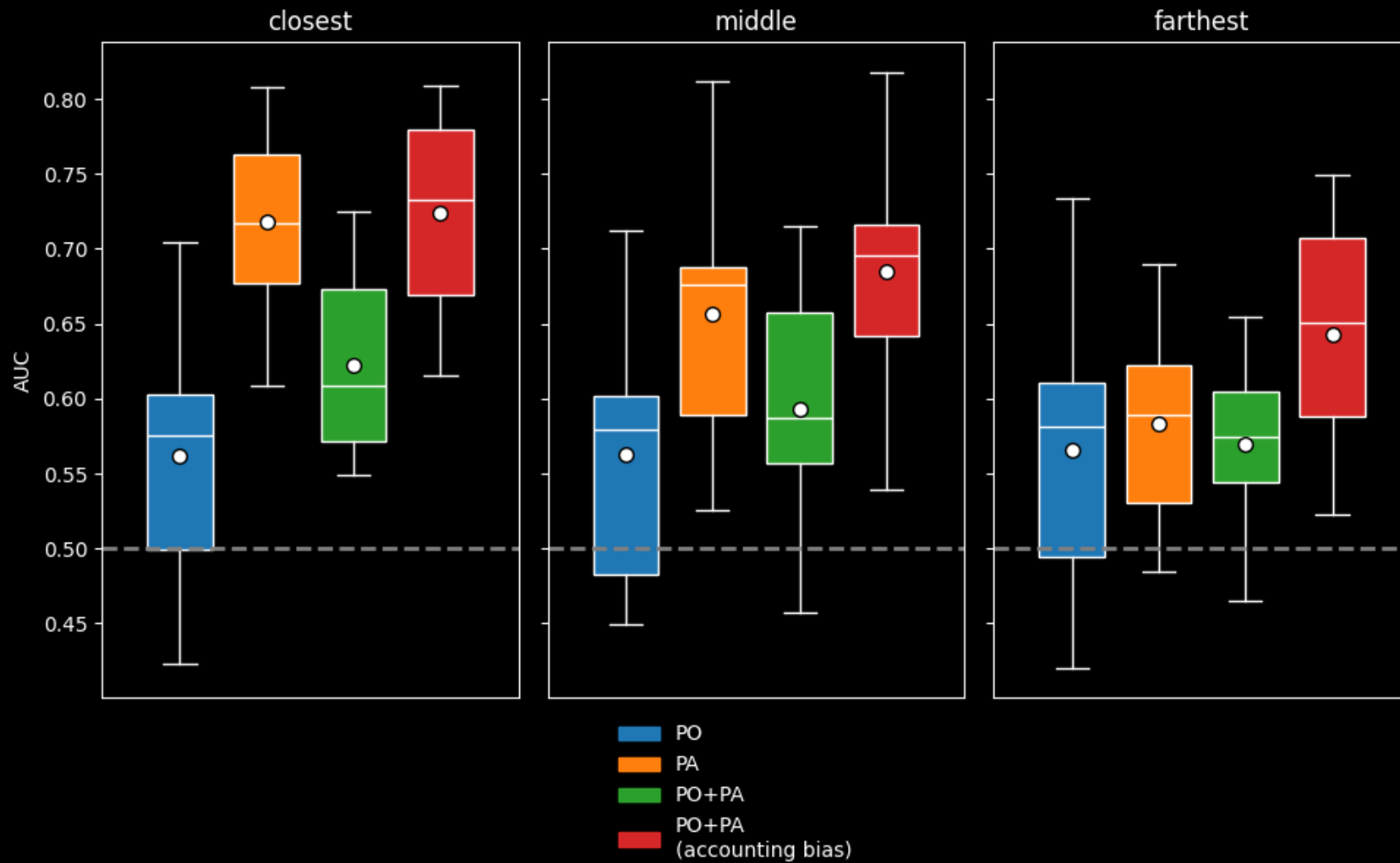
- The final loss is shared and weighted

$$L(\mathbf{y}_i, \lambda_i, b_i) = L_{PA}(\mathbf{y}_i, \lambda_i) + w_{PO} \cdot L_{PO}(\mathbf{y}_i, \lambda_i, b_i)$$



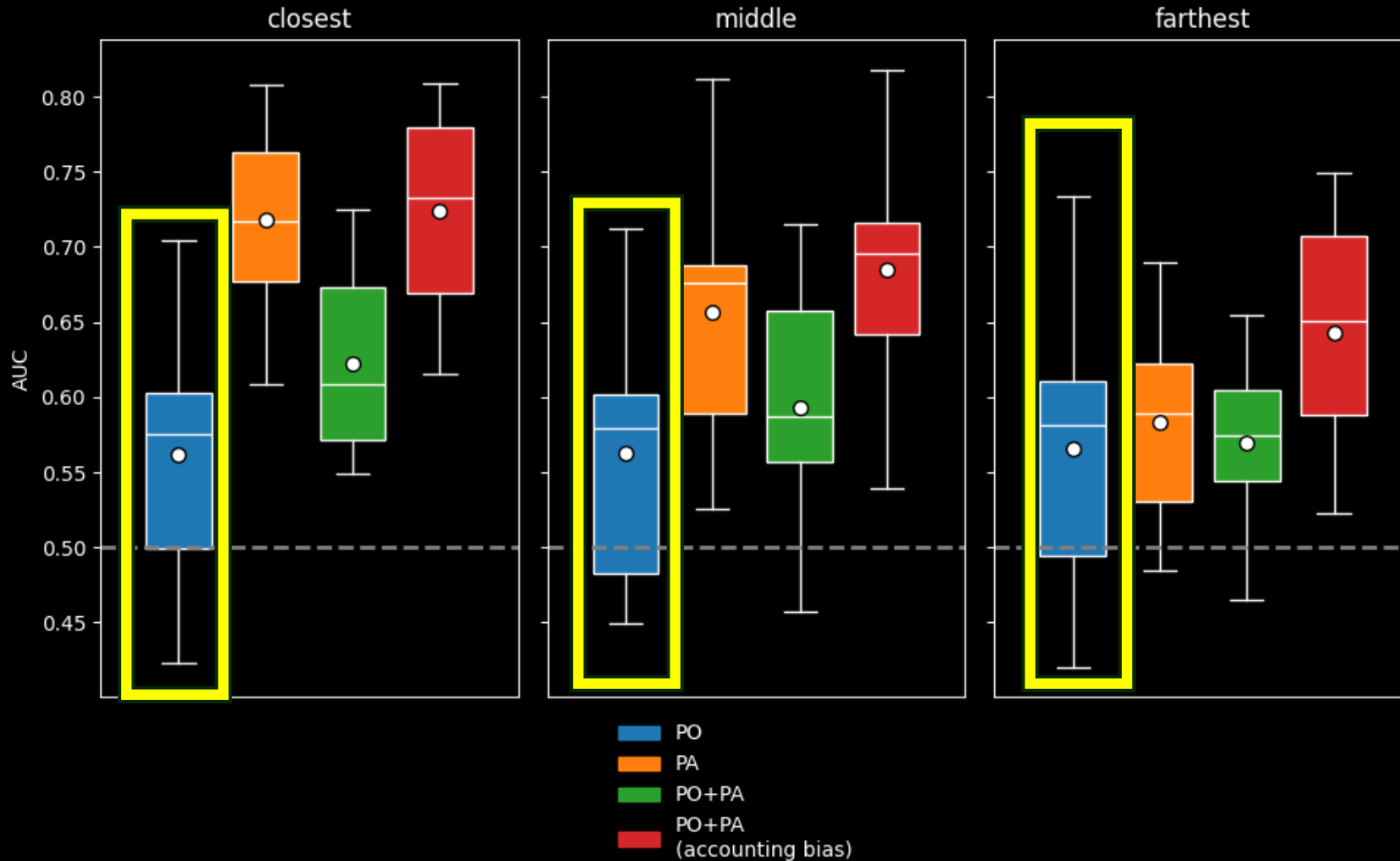


# Results

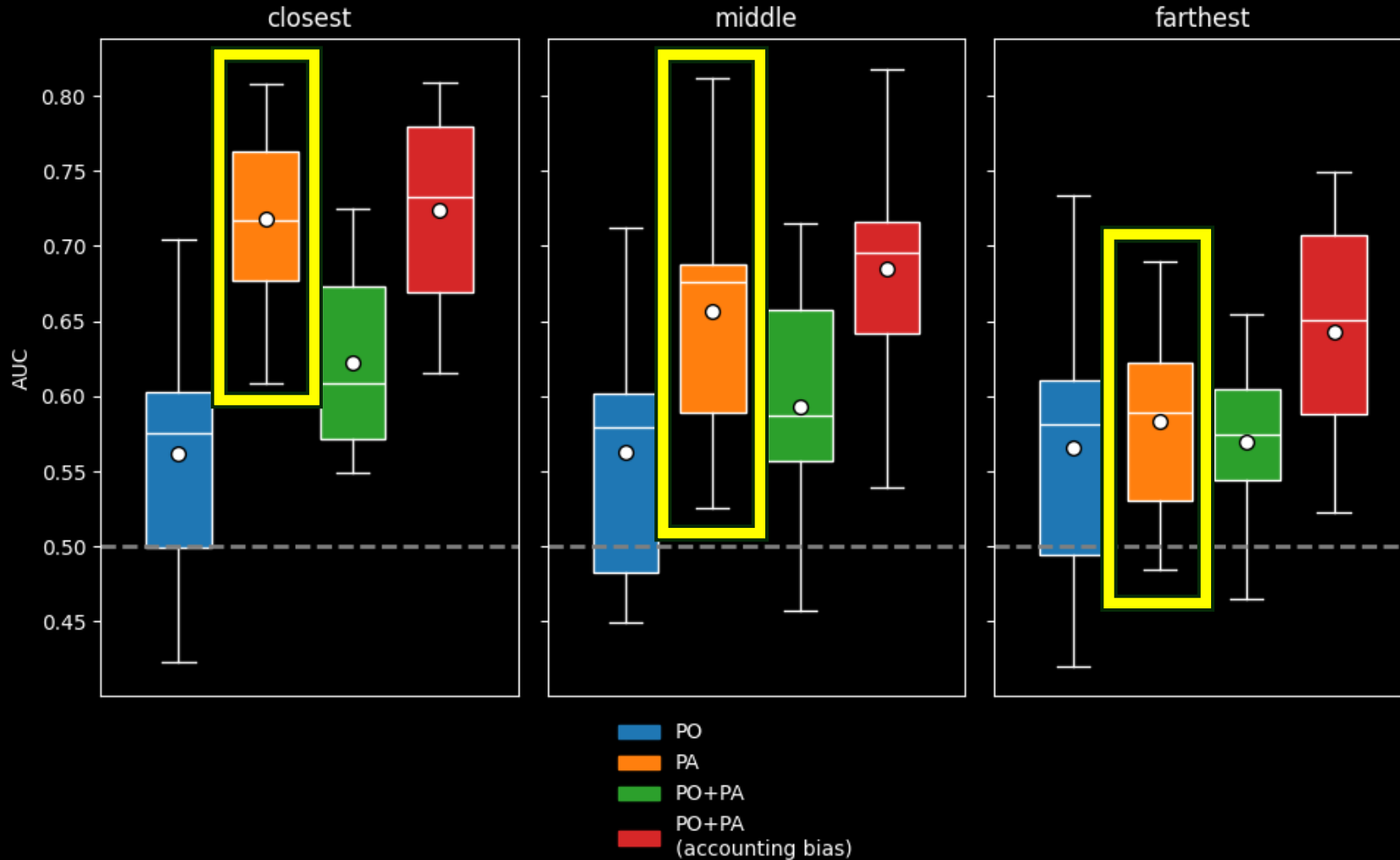


# Results

- Using only PO gives **constant** AUC as distance increases

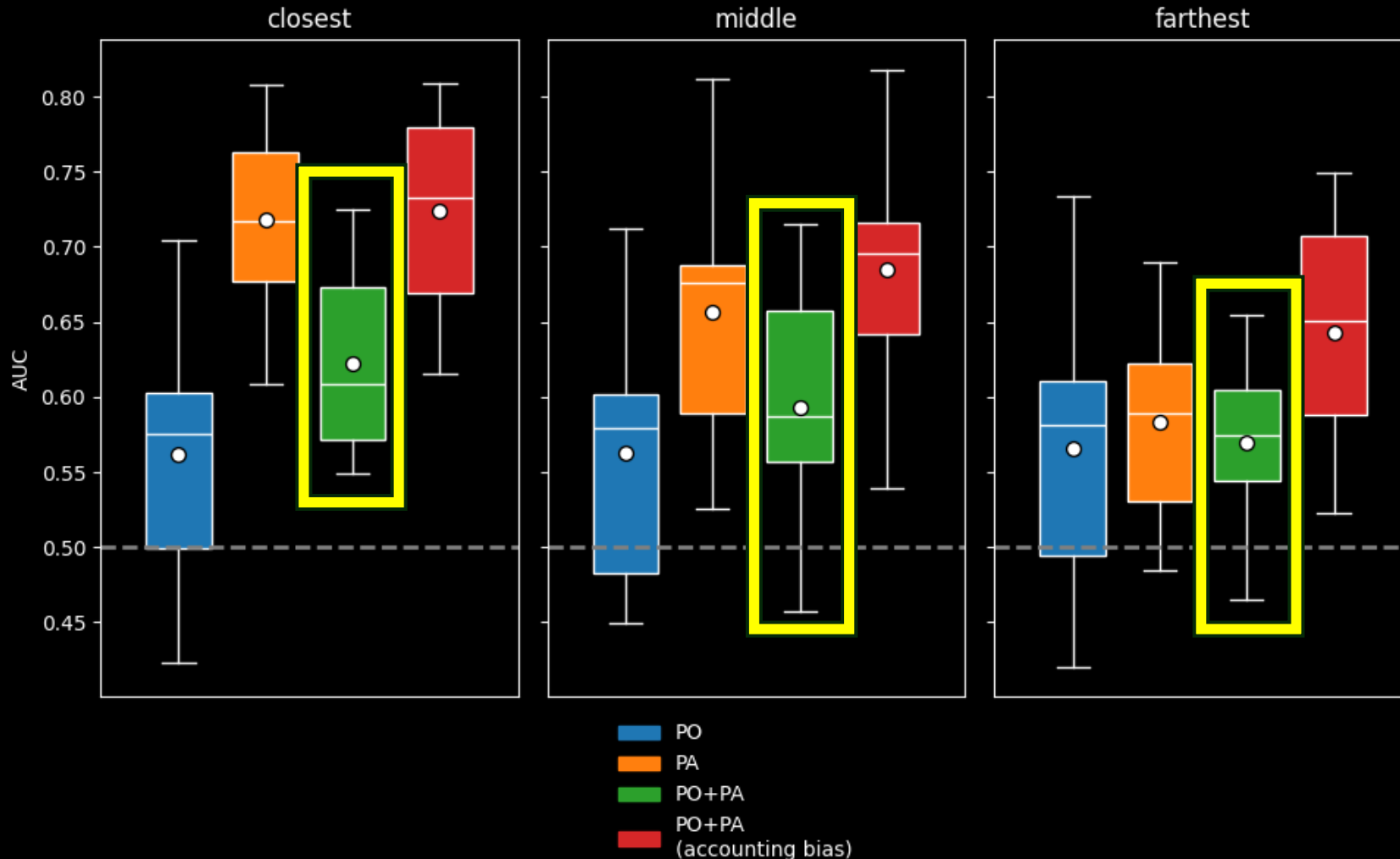


# Results



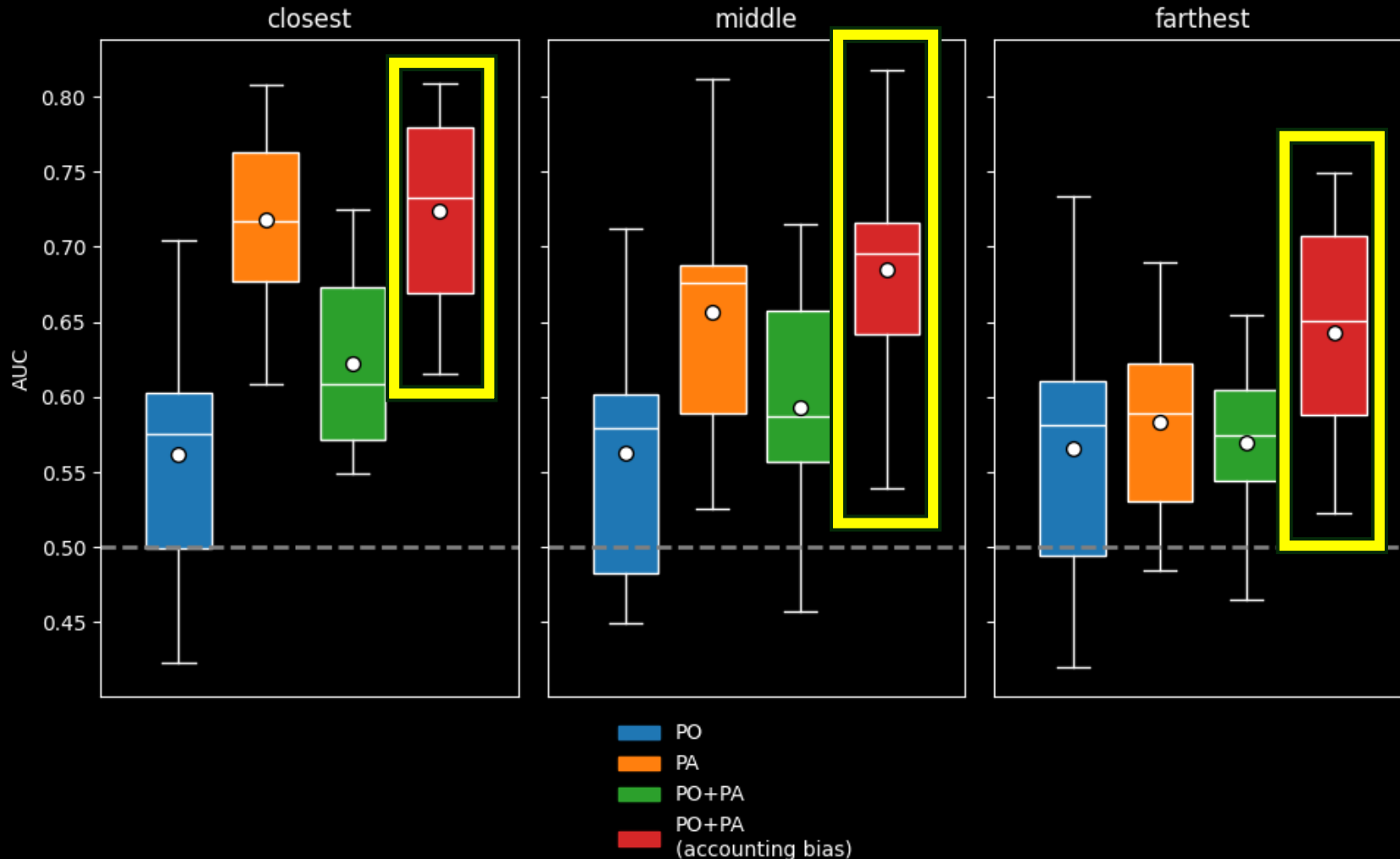
- Using only PO gives **constant** AUC as distance increases
- Using only PA **decreases** AUC as distance increases

# Results



- Using only PO gives **constant** AUC as distance increases
- Using only PA **decreases** AUC as distance increases
- Naive integration is **worst** than only PA

# Results



- Using only PO gives **constant** AUC as distance increases
- Using only PA **decreases** AUC as distance increases
- Naive integration is **worst** than only PA
- Accounting for bias is the **most robust**

Merci Beaucoup