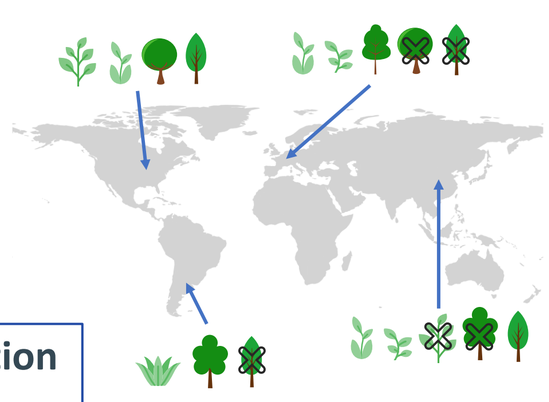
CISO: Species Distribution Modeling Conditioned on Incomplete Species Observations

Published in *Methods in Ecology and Evolution*

Hager Radi Abdelwahed* (MILA)
Mélisande Teng* (MILA, UdeM)
Robin Zbinden* (EPFL)
Laura Pollock (McGill)
Hugo Larochelle (MILA, UdeM)
Devis Tuia (EPFL)
David Rolnick (MILA, McGill)

Introduction: Conditioning SDMs on Incomplete Species Observations

- **Biotic interactions** play a key role in shaping species distributions
- Incorporating biotic variables into species distribution models (**SDMs**) is therefore essential
- However, biotic data are often **incomplete** or **inconsistently available** across species and locations



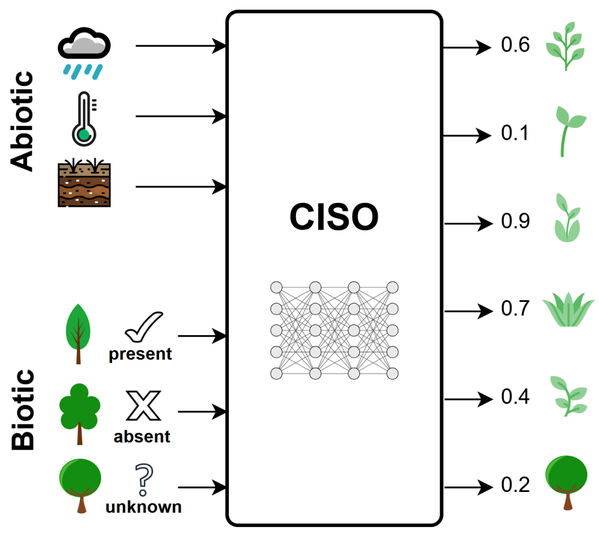
Our goal: leverage incomplete biotic information during both training and prediction

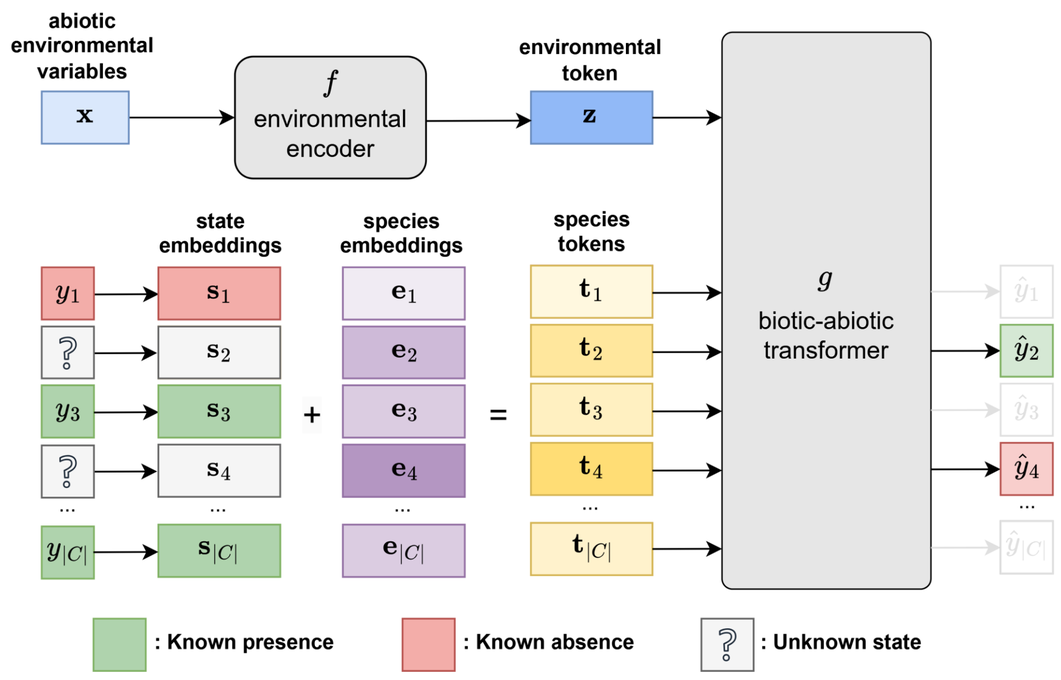
Limitations of JSDMs

- Species-species associations represented with a residual covariance matrix:
  → Assumes **symmetric interactions**
  → Capture only **pairwise relationships**
- Usually don't handle incomplete observations during training
- Struggle to scale to thousands of species and samples, especially when conditioning on other species

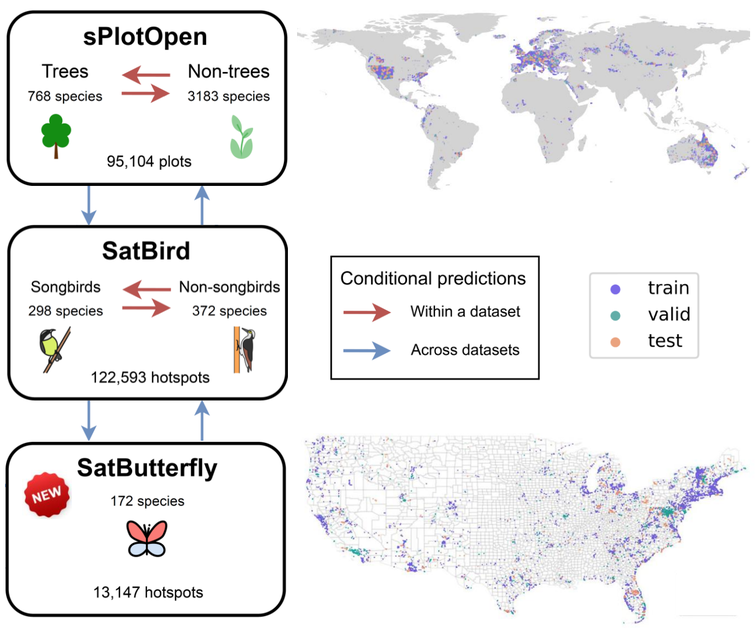
CISO: A Multi-Species Deep Learning Approach

- We propose **CISO**, an SDM that can be **conditioned on any available species presence or absence** during both training and prediction
- **Biotic and abiotic information are non-linearly combined** to predict all species
- Species presence, absence, or unknown states are explicitly encoded as **state embeddings**
- State embeddings are combined with learned **species embeddings**, which can integrate species-specific information
- **During training, we randomly mask species observations** to simulate incomplete-observation scenarios



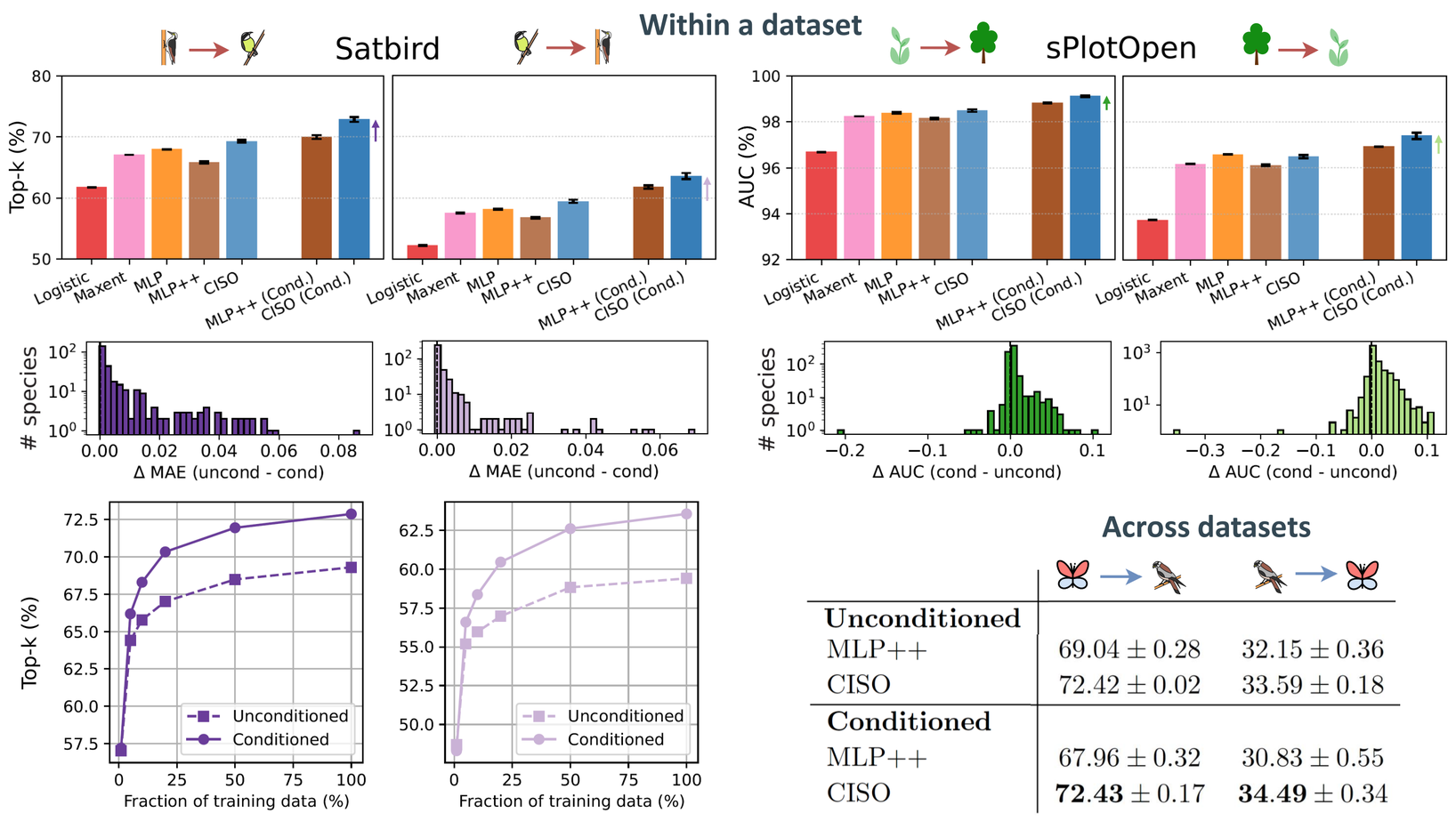


Datasets



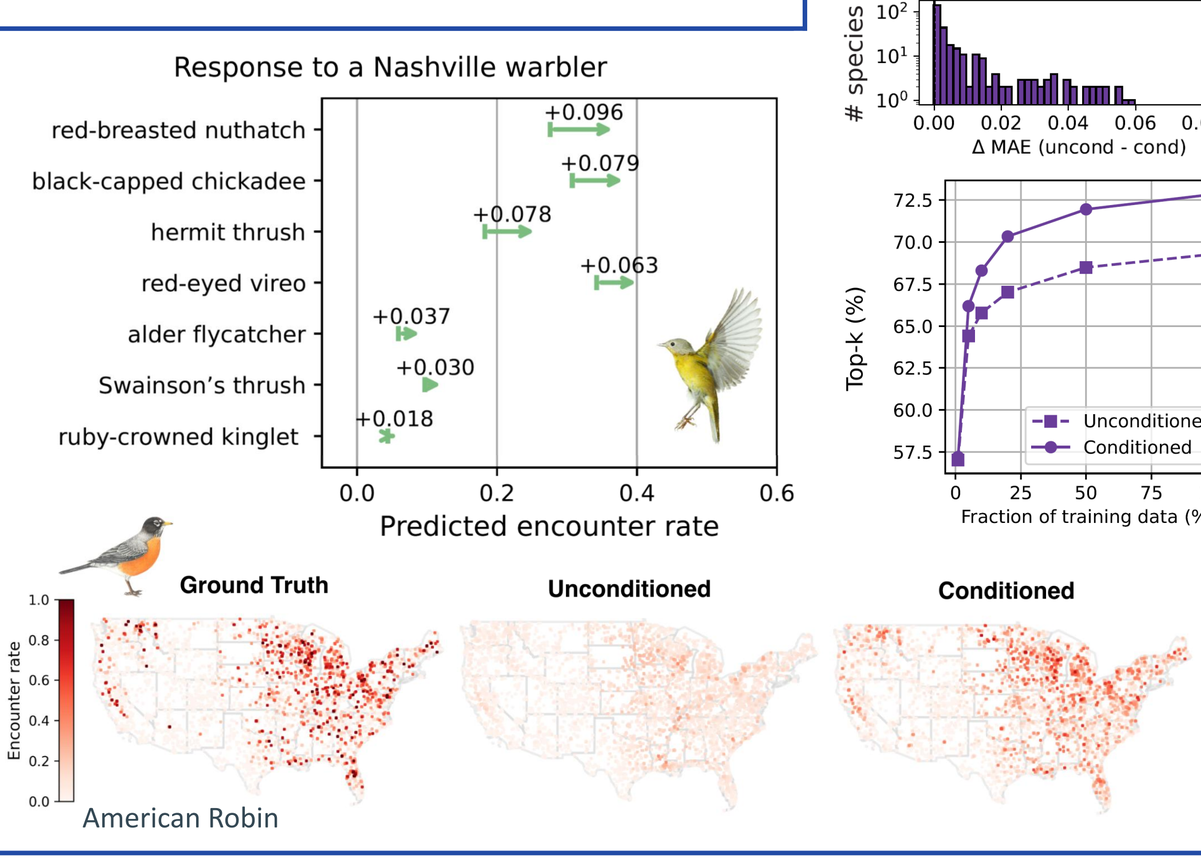
Results

- **CISO outperforms alternative methods** in predicting the remaining species group
- **Biotic information improves predictive performance,** consistently across most species
- Enable **analysis of potential interactions** (or shared habitats)
- **Conditioning gains increase with more training data**
- Across datasets conditioning requires a **large amount of co-located data** to be effective



Across datasets

Unconditioned		
MLP++	69.04 ± 0.28	32.15 ± 0.36
CISO	72.42 ± 0.02	33.59 ± 0.18
Conditioned		
MLP++	67.96 ± 0.32	30.83 ± 0.55
CISO	$\mathbf{72.43} \pm 0.17$	$\mathbf{34.49} \pm 0.34$



Conclusion

- **CISO**: a promising tool to condition SDMs on **any arbitrary, incomplete set of presence and absence data**
- Moves SDMs toward unified, **multi-dataset integration**