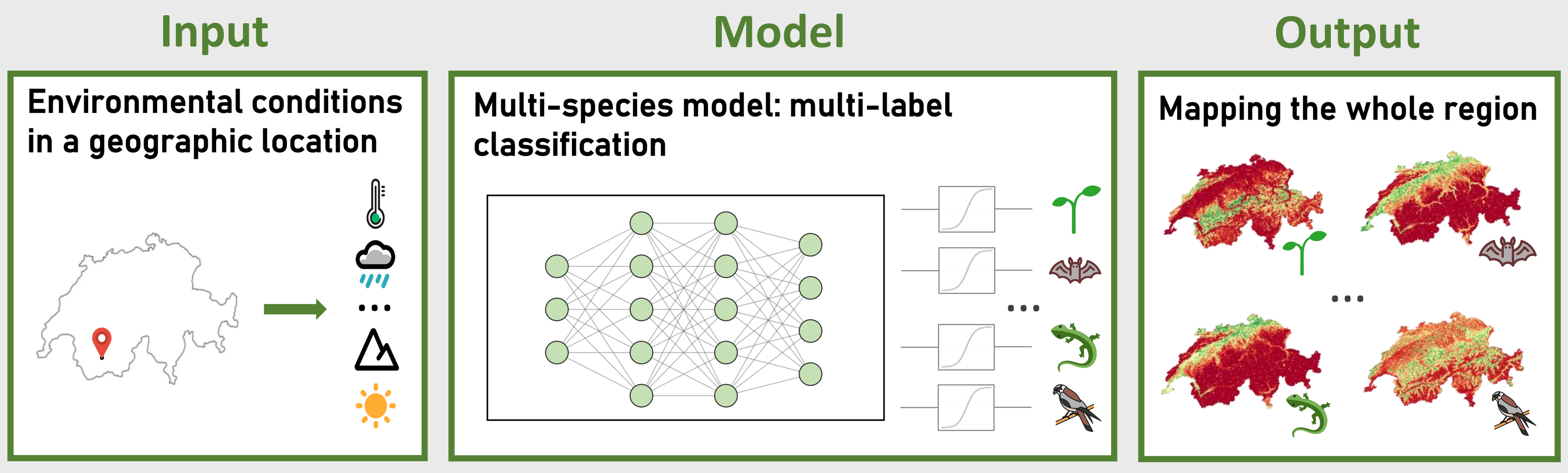
MaskSDM: Adaptive species distribution modeling through data masking

Robin Zbinden (EPFL)
Nina van Tiel (EPFL)
Gencer Sümbül (EPFL)
Benjamin Kellenberger (UCL)
Devis Tuia (EPFL)

1. Species Distribution Models (SDMs)

- **Relate species occurrence data with environmental variables**.
- **Numerous applications** to understand the: **geographic distribution** of a species, **ecological niche**, impact of **climate change on biodiversity**, and spread of **invasive species**.
- **Support decision-making for conservation** and restoration.

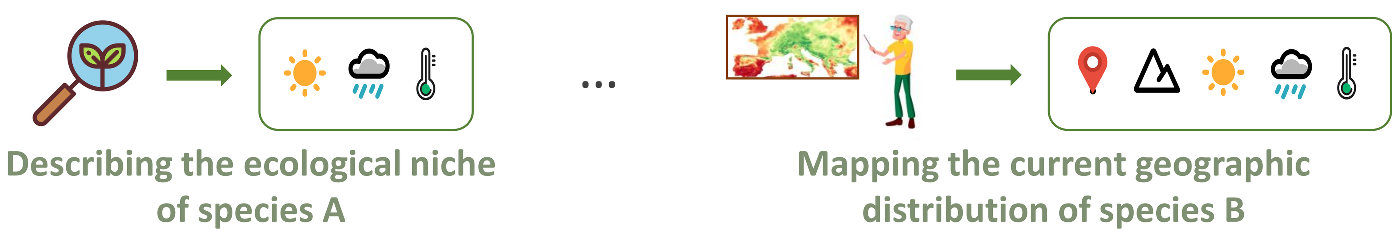


Critical aspect: the selection of appropriate environmental variables

2. Challenges with variable selection

Enabling flexibility for end-users

- Previous multi-species models use the same variables for all species, despite **differing needs**.
- **Different research questions** require different sets of input variables.



Analysis of variable contributions

- Identifying **which variables influence predictions and performance helps gain ecological insights**.
- Traditional ablation studies require retraining multiple times.

Handling missing or noisy variables

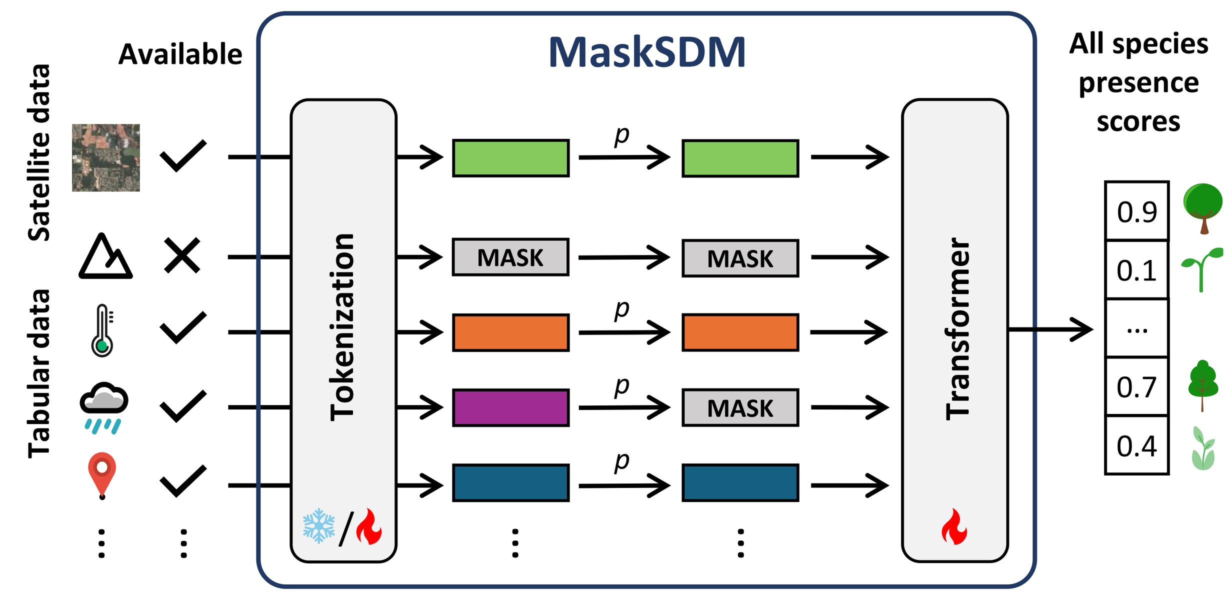
- Geospatial data usually contains many samples with **missing variables**.
- **Geographic biases** can lead to **noisy, unreliable data in certain areas**.
- **Meta-data**, though highly predictive, is **inconsistently available**.

3. Our approach

- **MaskSDM:**
  - Enables the **selection of relevant variables during inference**
  - Offers **insights into variable contributions to predictions and performance**
  - Effectively **handles missing data** during both training and inference.
- It uses **supervised masked data modeling**.
- Each modality/variable is **independently tokenized** and then **input into a Transformer encoder**.

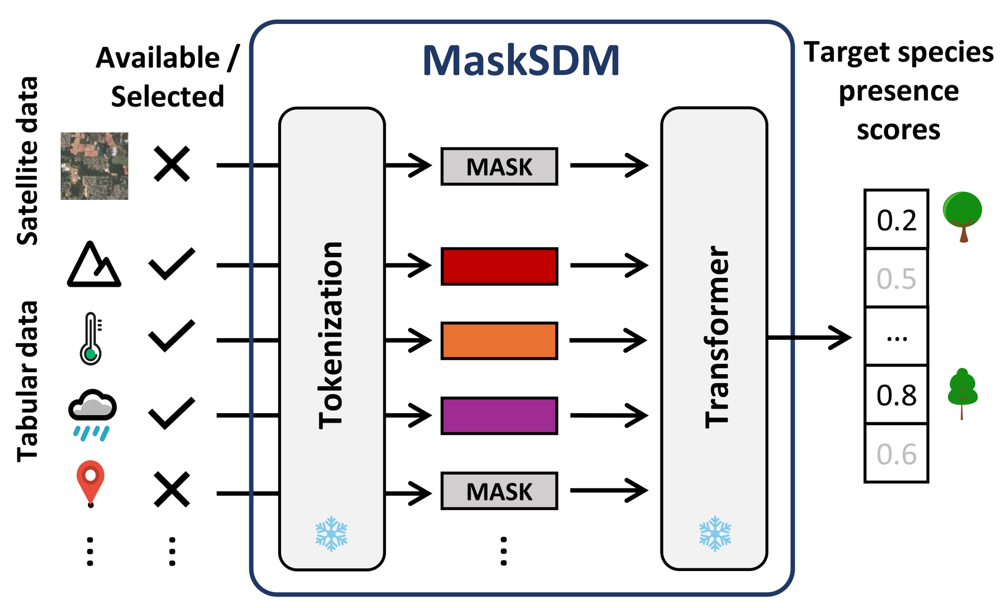
Training

- We use a **mask token** to indicate missing input variables to the Transformer.
- Additionally, this mask token is used to **randomly mask** each input variable with a **varying probability *p***, enhancing robustness to any subset of variables.



Inference

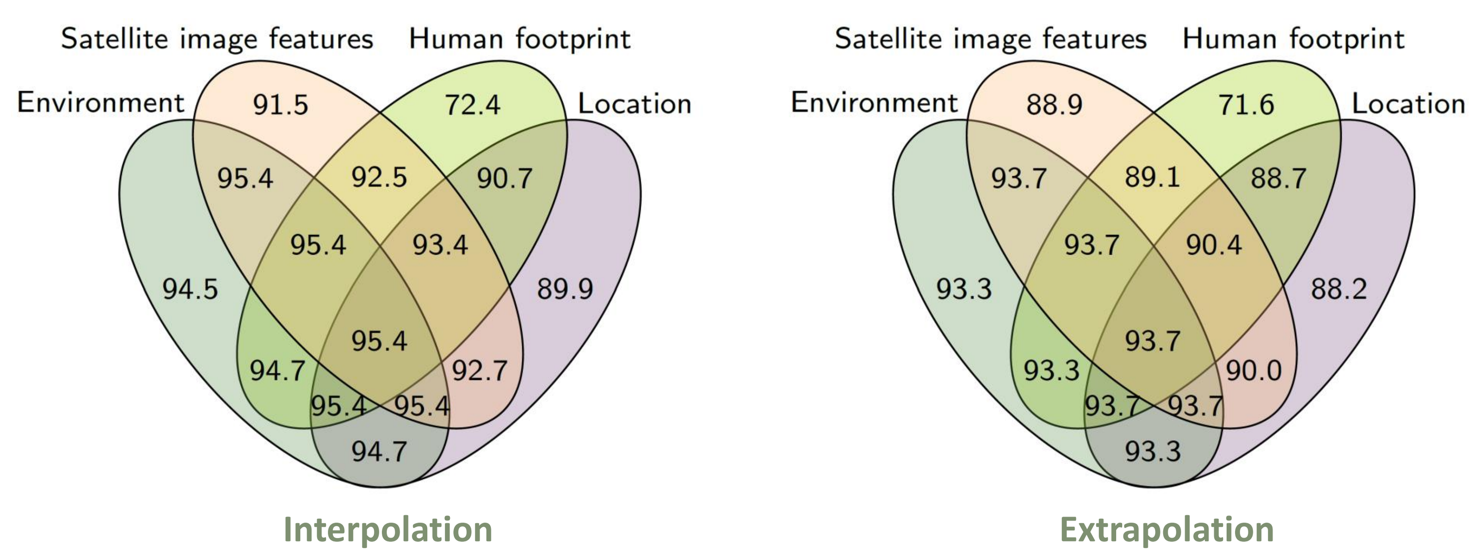
- **MaskSDM can take any subset of variables as input** to predict the presence of target species.
- Missing or undesired variables are replaced by the mask token.



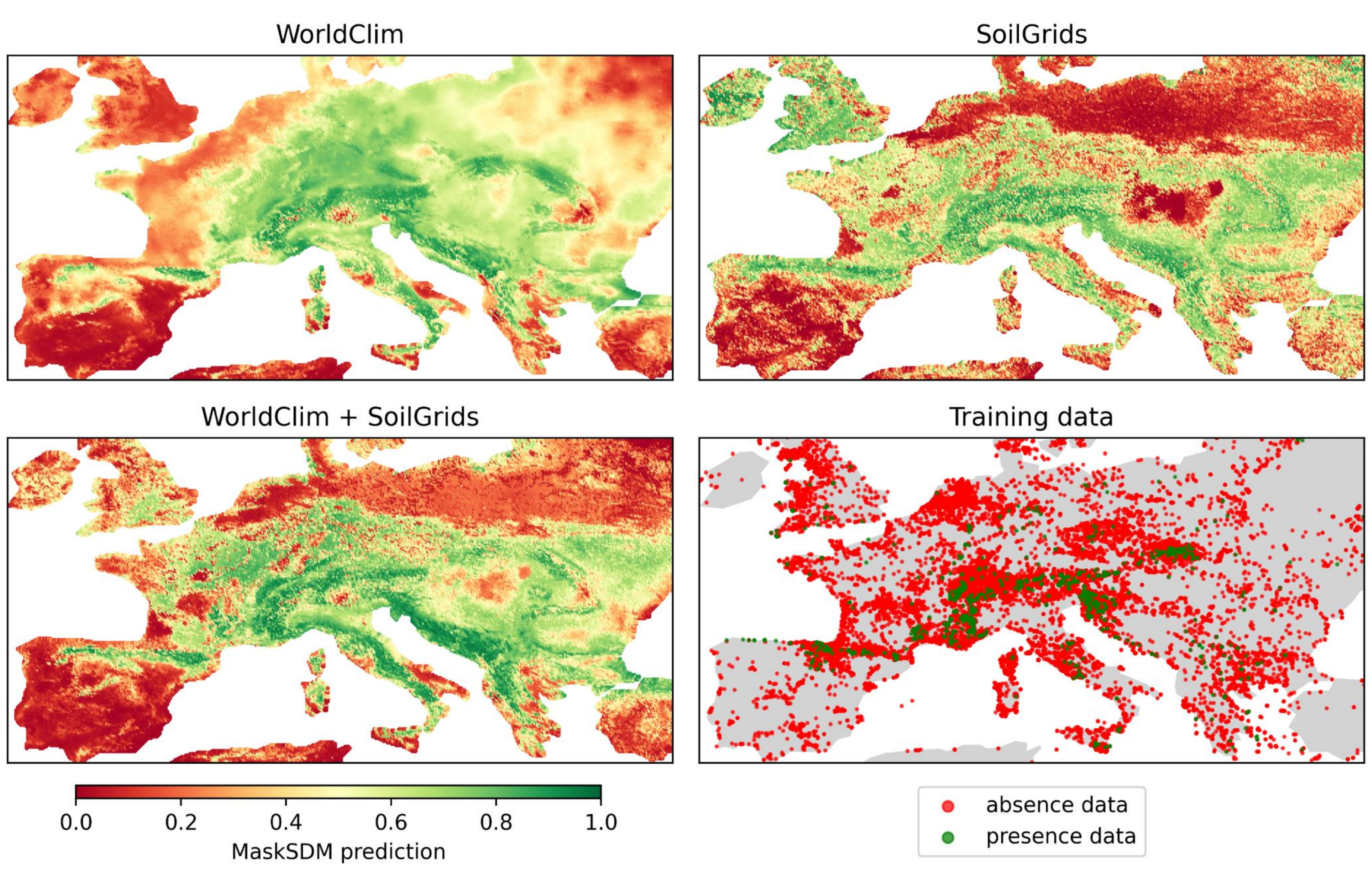
4. Experiments and Results

- We train and evaluate our approach on the global **sPlotOpen dataset** which includes presence-absence observations of plants species.
- We split the data using **spatial block cross-validation**.
- MaskSDM is assessed with **various groups of input variables**.
- Baseline models handle missing data using **mean imputation**.
- Evaluation metric: **Mean AUC across all species**.

Input Variable (#)	Avg. Temperature (1)	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓
	WorldClim (19)	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓
	SoilGrids (8)	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
	Topographic (3)	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
	Location (2)	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓
	Human footprint (9)	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
	Plot metadata (20)	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
	Satellite image features	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓
Method	MLP	69.9	75.5	N/A	88.1	89.0	89.7	91.1	91.2	91.5	N/A
	ResNet	72.5	80.7	N/A	87.3	90.7	91.5	**93.4**	**93.4**	**94.7**	N/A
	FTTransformer	72.2	75.3	70.2	82.1	86.0	87.3	91.8	91.9	93.7	94.3
	MaskSDM (ours)	**80.3**	**88.2**	**88.9**	**91.6**	**92.6**	**93.3**	93.3	**93.4**	**94.7**	**94.8**



MaskSDM map of *Anthyllis Vulneraria* with different input variables



Conclusions

- **MaskSDM consistently outperforms the baselines**, with the performance gap widening as fewer variables are available.
- **Environmental variables alone provide strong performance.** Adding **human footprint and location data offers little improvement** when combined with other variables.
- **MaskSDM can take any subset of variables as input.**

This work was funded by the SNSF under grant number 204057 (deepHSM)