OAK RIDGE National Laboratory

Mapping Biogeographical Regions with Latent Variable Models

Christopher Krapu

GeoAI Group

Geospatial Sciences & Human Security Division

ORNL is managed by UT-Battelle LLC for the US Department of Energy

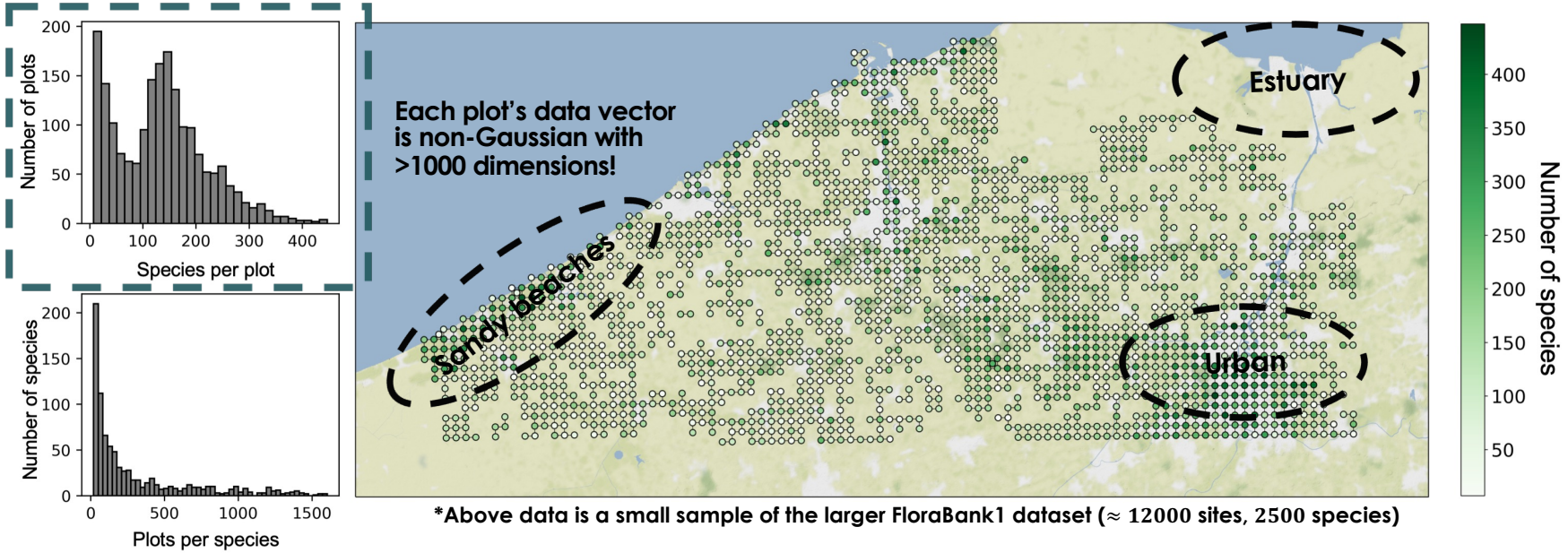
U.S. DEPARTMENT OF ENERGY

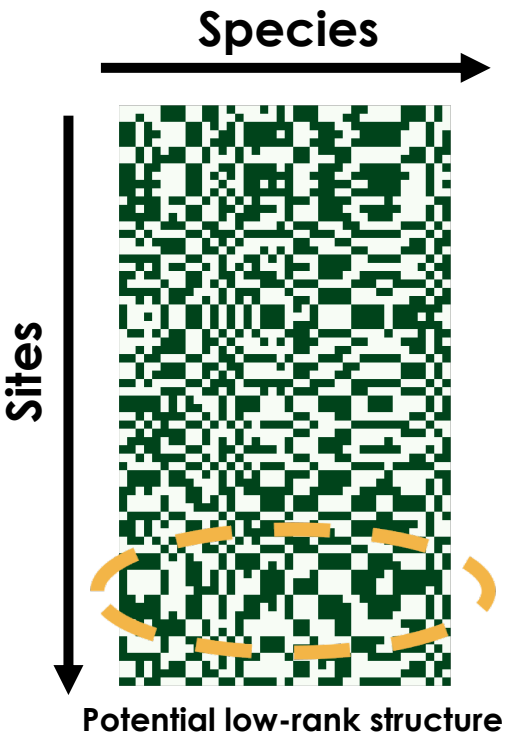
Rich Biodiversity Data

Species abundance or presence/absence data has grown dramatically richer with new datasets in recent decades!

- *Each species is an **indicator** of the type of environment it is in, but also **competes** with or aids other species nearby*
- *Identifying patterns of frequently co-occurring species can be very helpful for determining how **geography influences biodiversity***

Example: FloraBank1 dataset from Northern Belgium

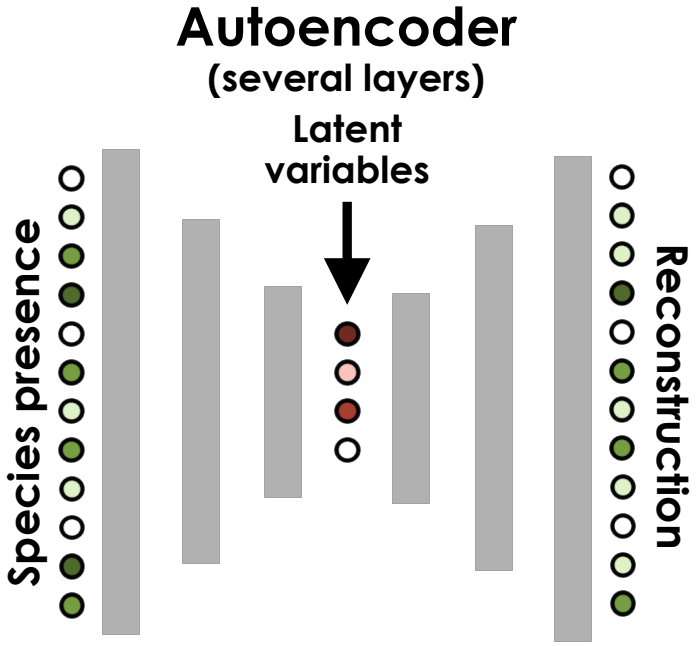


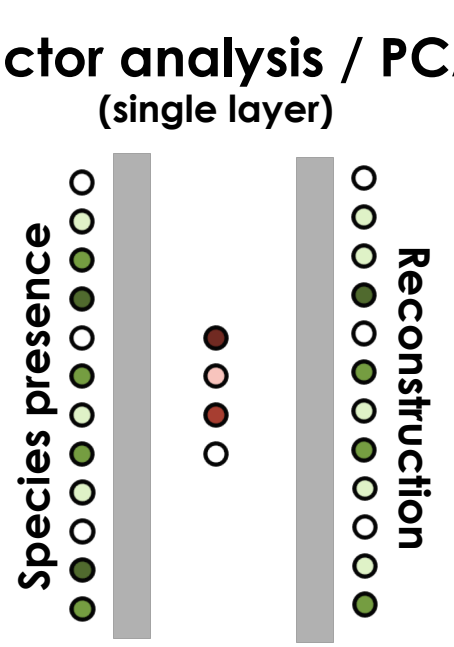


OAK RIDGE National Laboratory

Latent Variable Models

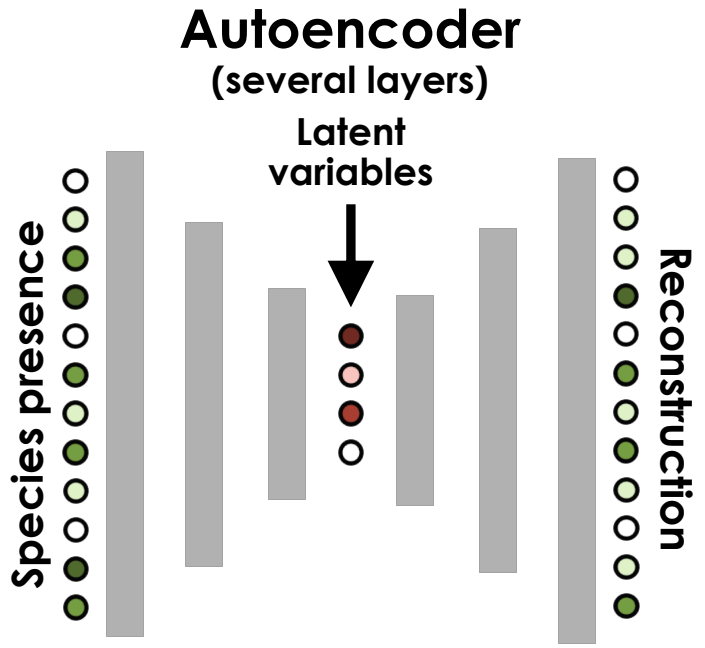
- Identifying simplified structure in the data is largely synonymous with
  - Optimal compression (information theory)
  - Unsupervised learning (AI / machine learning)
  - Dimensionality reduction (statistics)
- **Ecological hypothesis #1:** minimizing reconstruction error leads to latent variables significantly correlated with biogeographical features





Variational autoencoder

- x is a d-dimensional vector of binary observations
- $\hat{p}$ is the reconstruction of x with each element indicating probability of occurrence for a single species
- β, λ are hyperparameters for regularization and imbalance correction, respectively
- $\mu_\phi(x), v_\phi(x)$ are neural networks mapping data vectors into mean/variance of latent state distribution



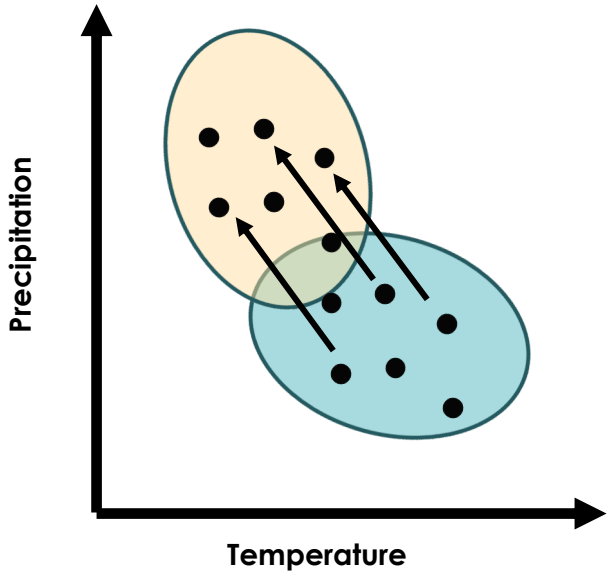
Single-point loss function:

$$L(x,\hat{p}) = \underbrace{\beta \cdot D_{KL}\left[\mathcal{N}\left(\mu_\phi(x), v_\phi(x)\right) || \mathcal{N}(0, I_L)\right]}_{\text{Regularization pushing latent codes to zero}} - \underbrace{\sum_{k=1}^{K} x_k \log \hat{p}_k + \lambda(1-x_k)\log(1-\hat{p}_k)}_{\text{Reconstruction loss summed over species}}$$

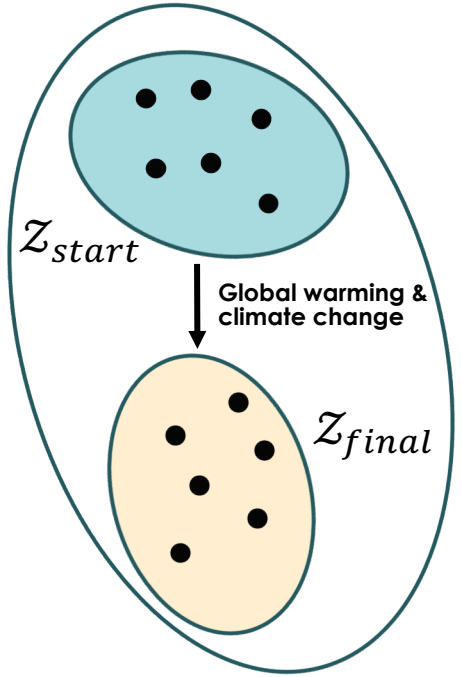
Mapping Expected Biogeographical shifts

Ecological hypothesis #2:
Clusters identified in latent space $\mathcal{Z}$ can be interpreted as qualitatively and semantically distinct groupings of spatial sites into ecoregions

Ecological hypothesis #3:
Latent representations are sufficiently correlated with geographical predictor variables $\mathbf{u} \in \mathcal{U} \subseteq \mathbb{R}^p$ such as rainfall and temperature that future shifts in species' distributions can be modeled using shifts in geographic variables via regression of latent representations upon predictors

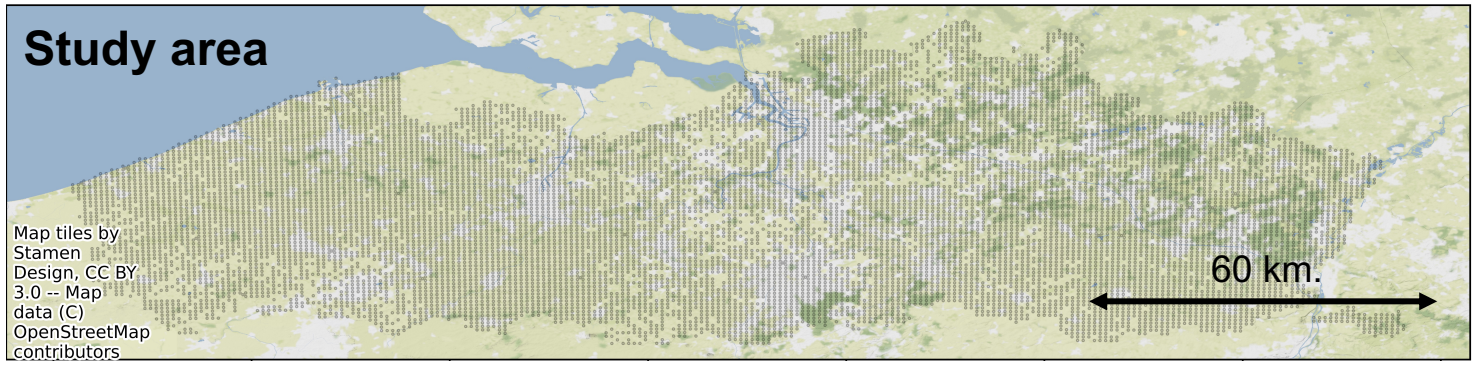


Predictor space $\mathcal{U}$



Latent space $\mathcal{Z} \subseteq \mathbb{R}^l$

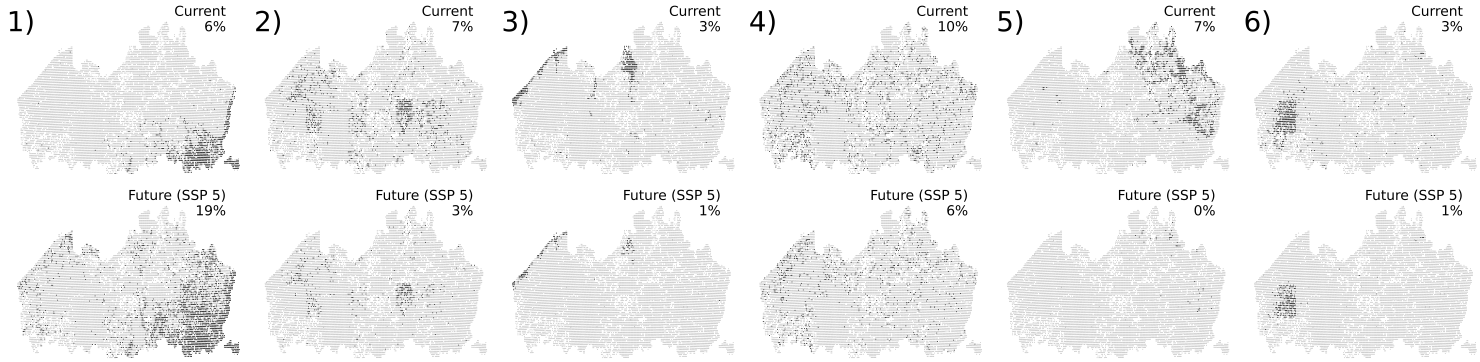
Mapping Expected Biogeographical shifts



Region 3

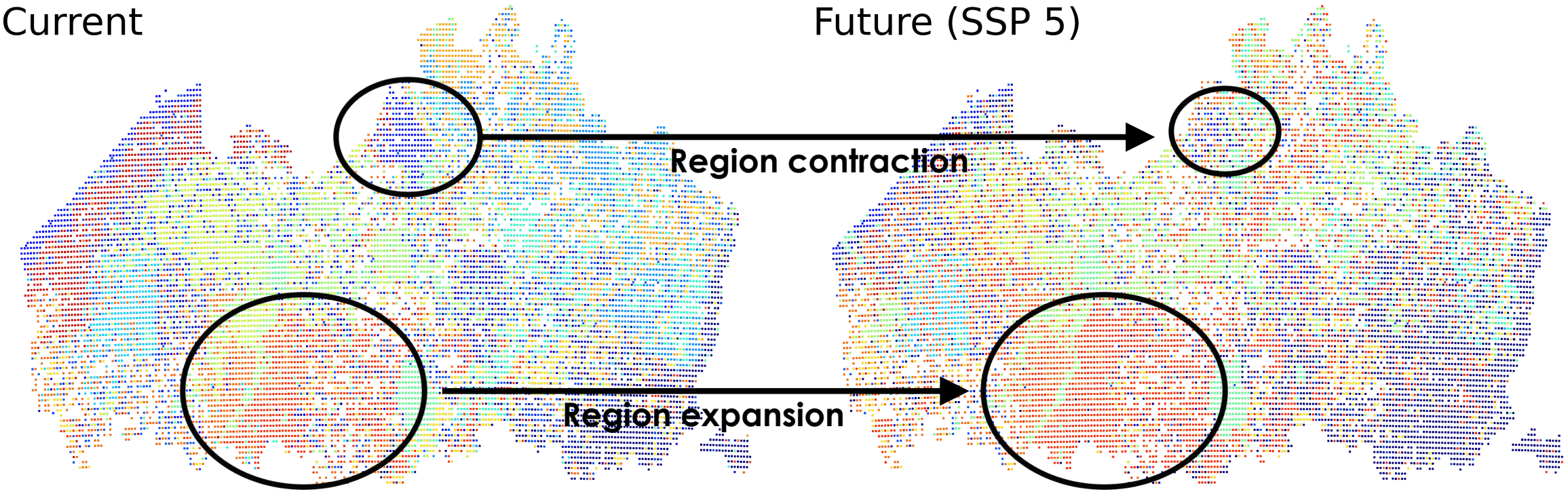
High probability	Low probability
Equisetum sylvaticum	Atriplex laciniata
Myosotis nemorosa	Diphasiastrum tristachyum
Geranium versicolor	Malva pusilla
Carex strigosa	Ammophila arenaria
Polystichum aculeatum	Armeria maritima
Dactylis polygama	Euphorbia paralias
Helleborus viridis	Vicia faba
Gagea spathacea	Suaeda maritima
Anemone ranunculoides	Wahlenbergia hederacea
Luzula forsteri	Thlaspi caerulescens

Cluster regions (6/15)



OAK RIDGE National Laboratory

Clustering of locations before and after climate shifts



Thanks for your time!