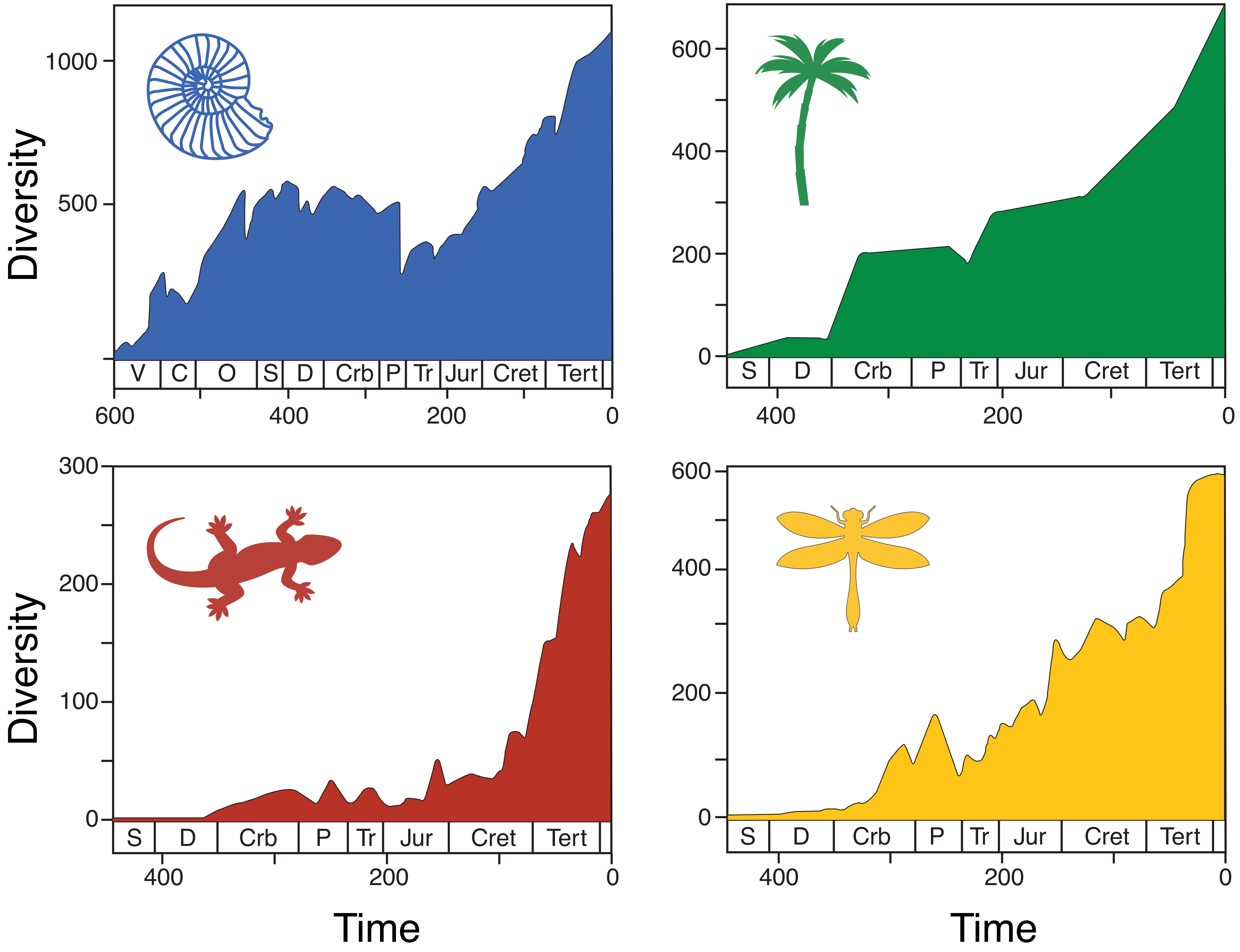
DeepDive – Estimation of diversity trajectories using deep learning

Estimating biodiversity through time



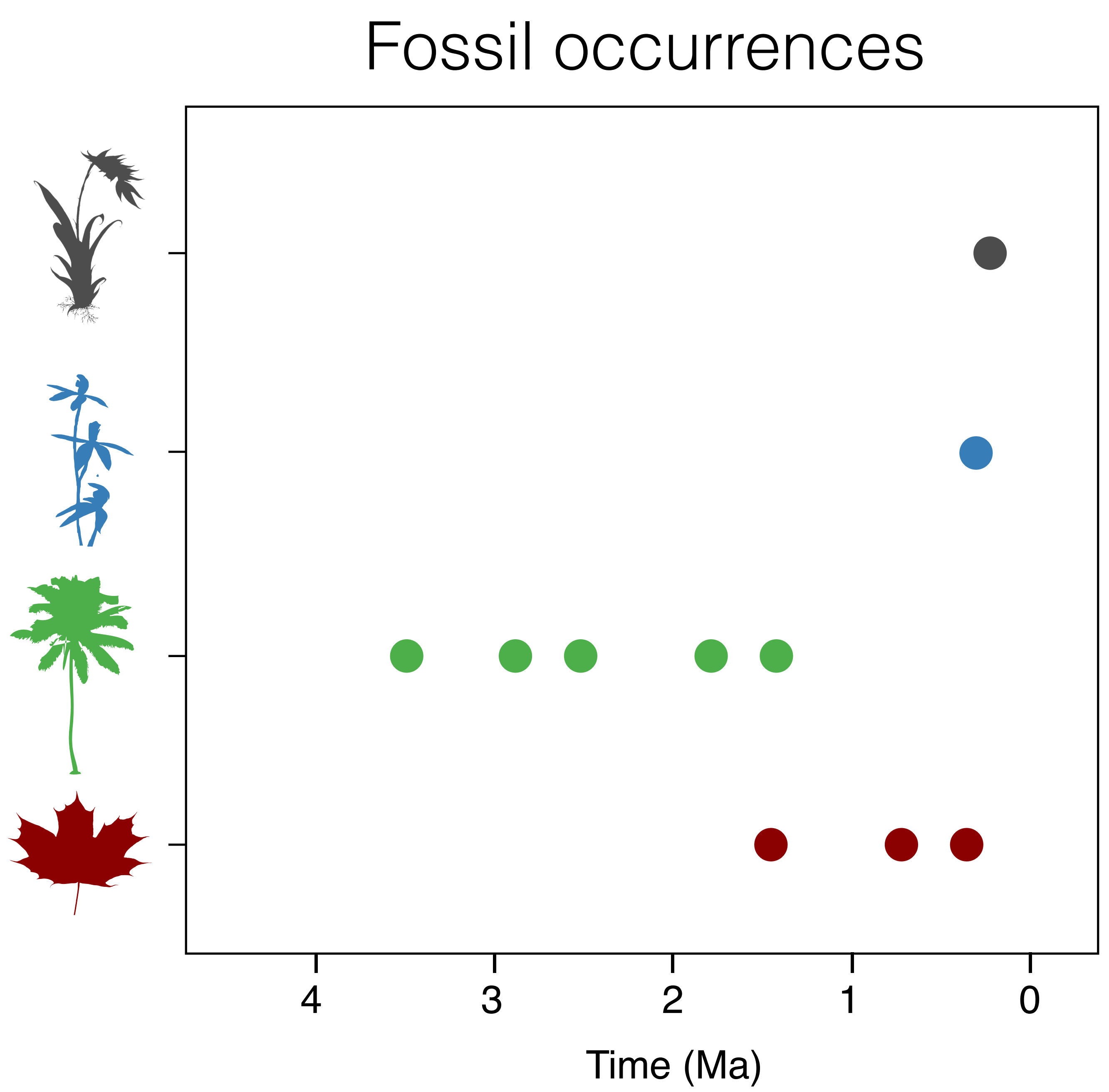
Is there a limit to biodiversity?

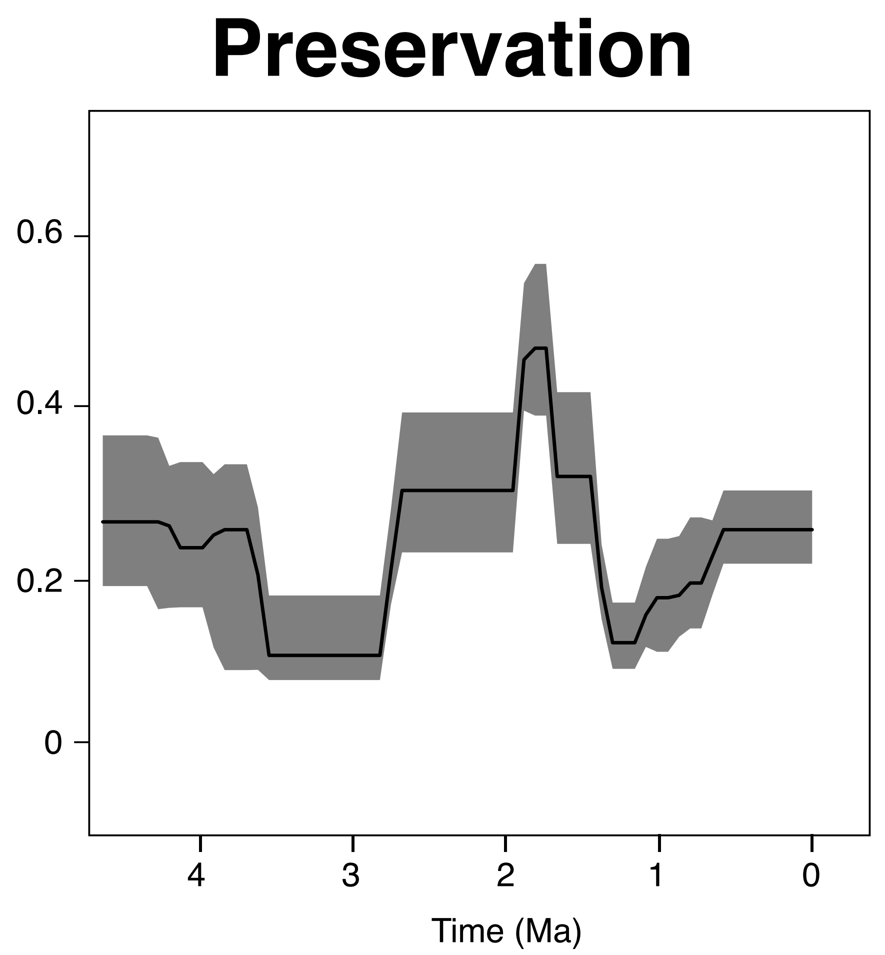
Does biodiversity increase over time?

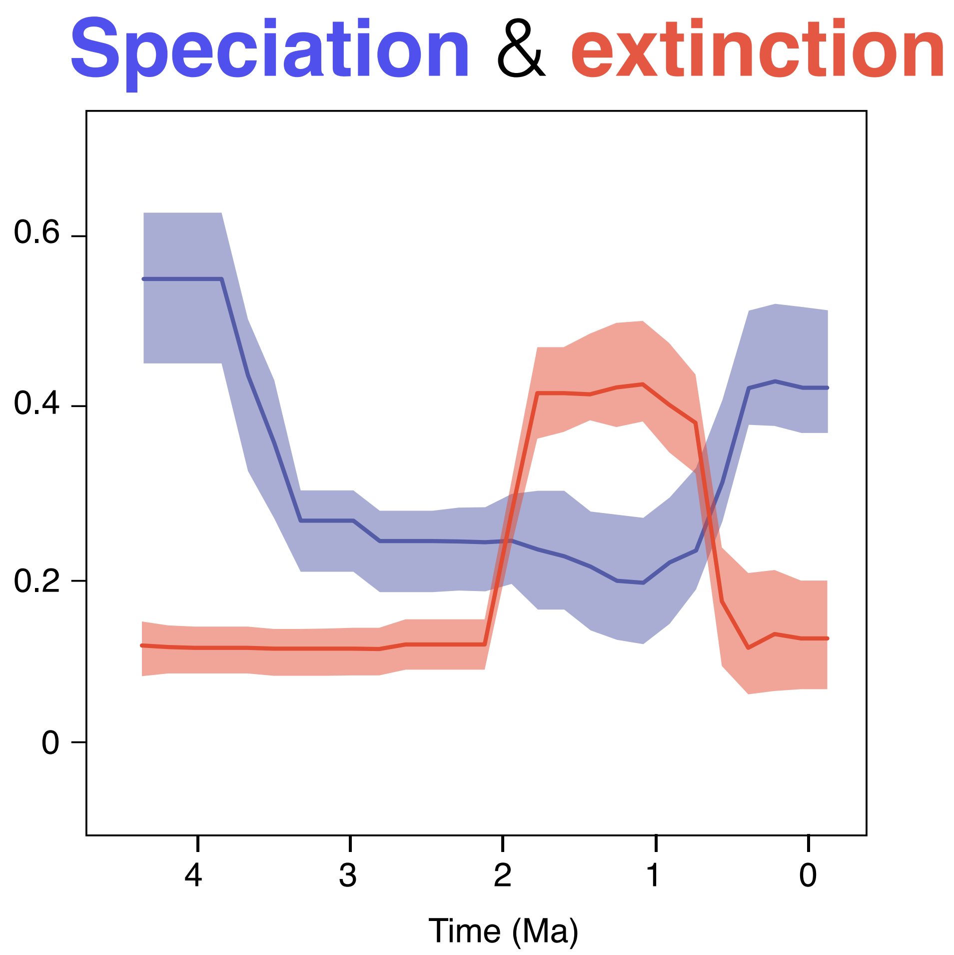
What are the mechanisms controlling biodiversity?

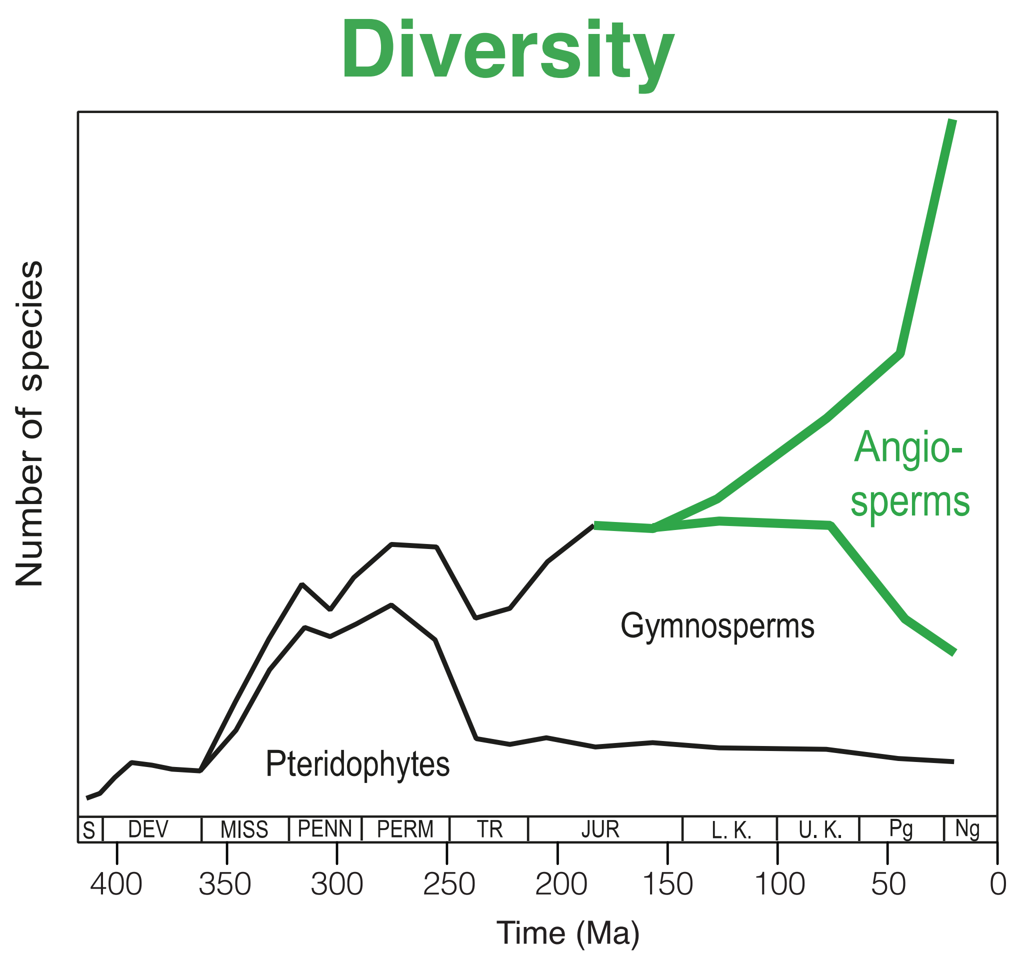
Benton 2007 Palaeontology

Bayesian estimation of diversity trajectories – mcmcDivE





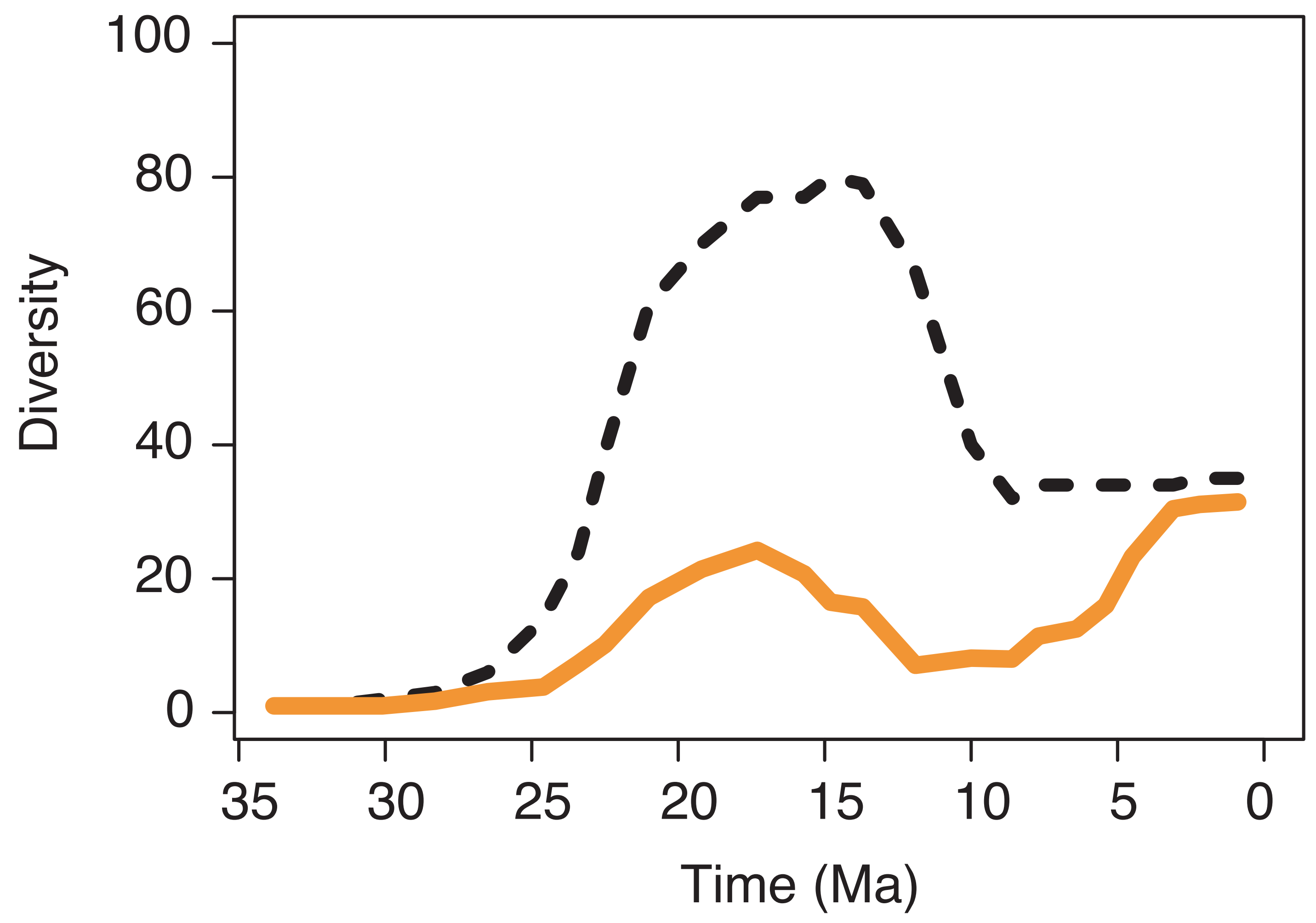




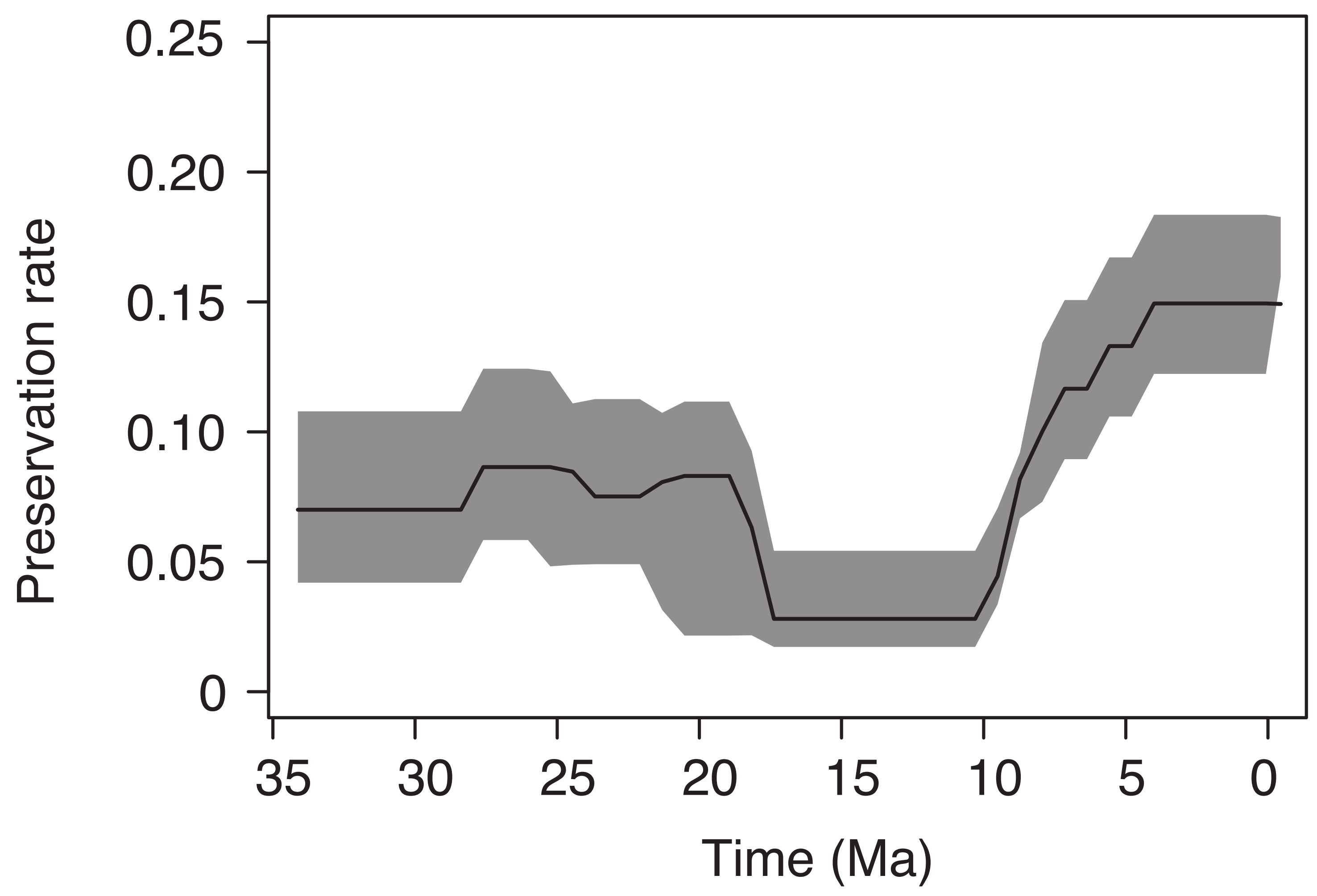
What true diversity could have generated the observed fossil record?

Flannery-Sutherland et al. 2022 Nature Comm

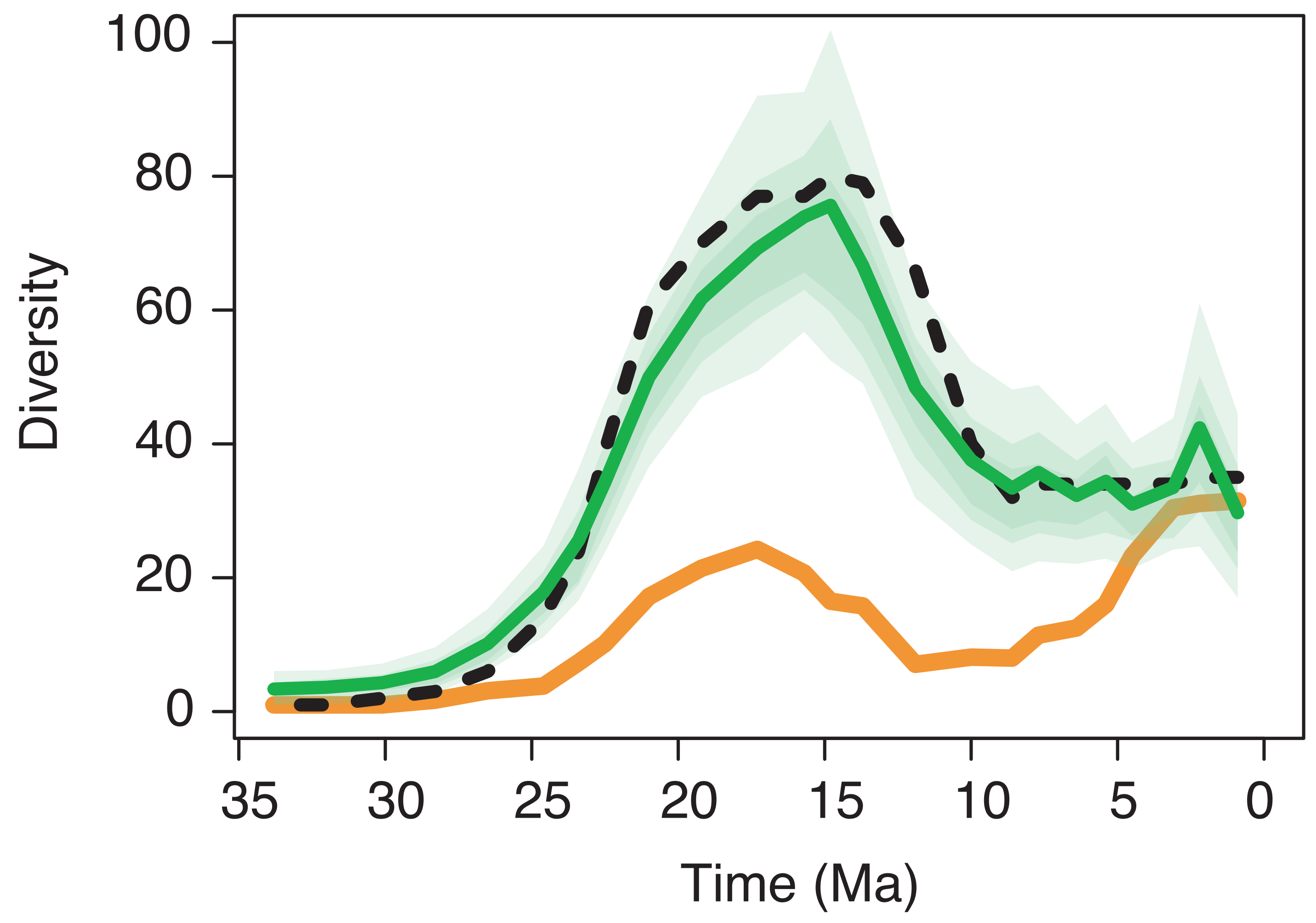
Observed diversity through time



Preservation rates



Inferred diversity through time

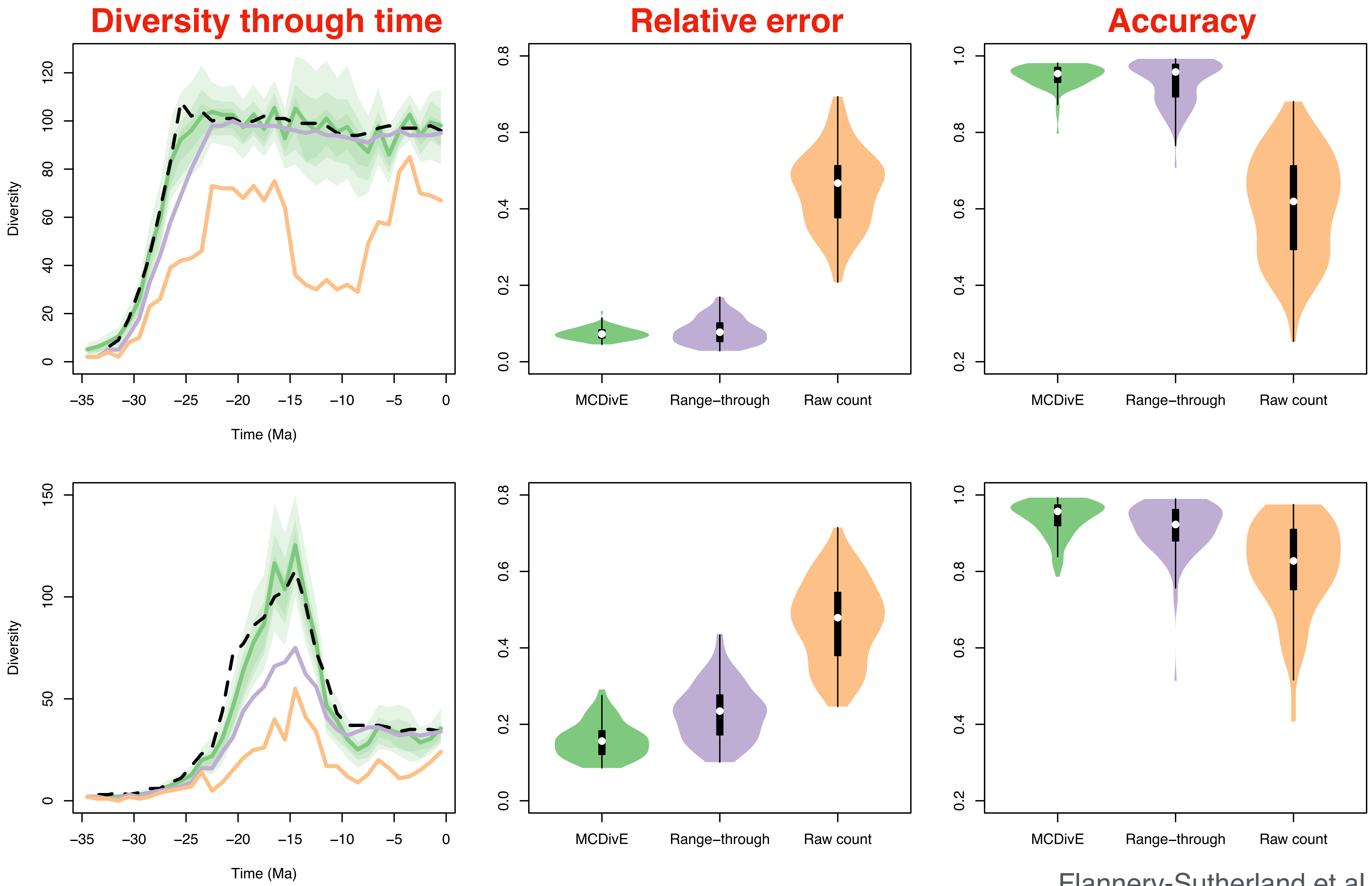


Likelihood function: Binomial distribution

Prior (smoothing) function: Brownian process

Flannery-Sutherland et al. 2022 Nature Comm

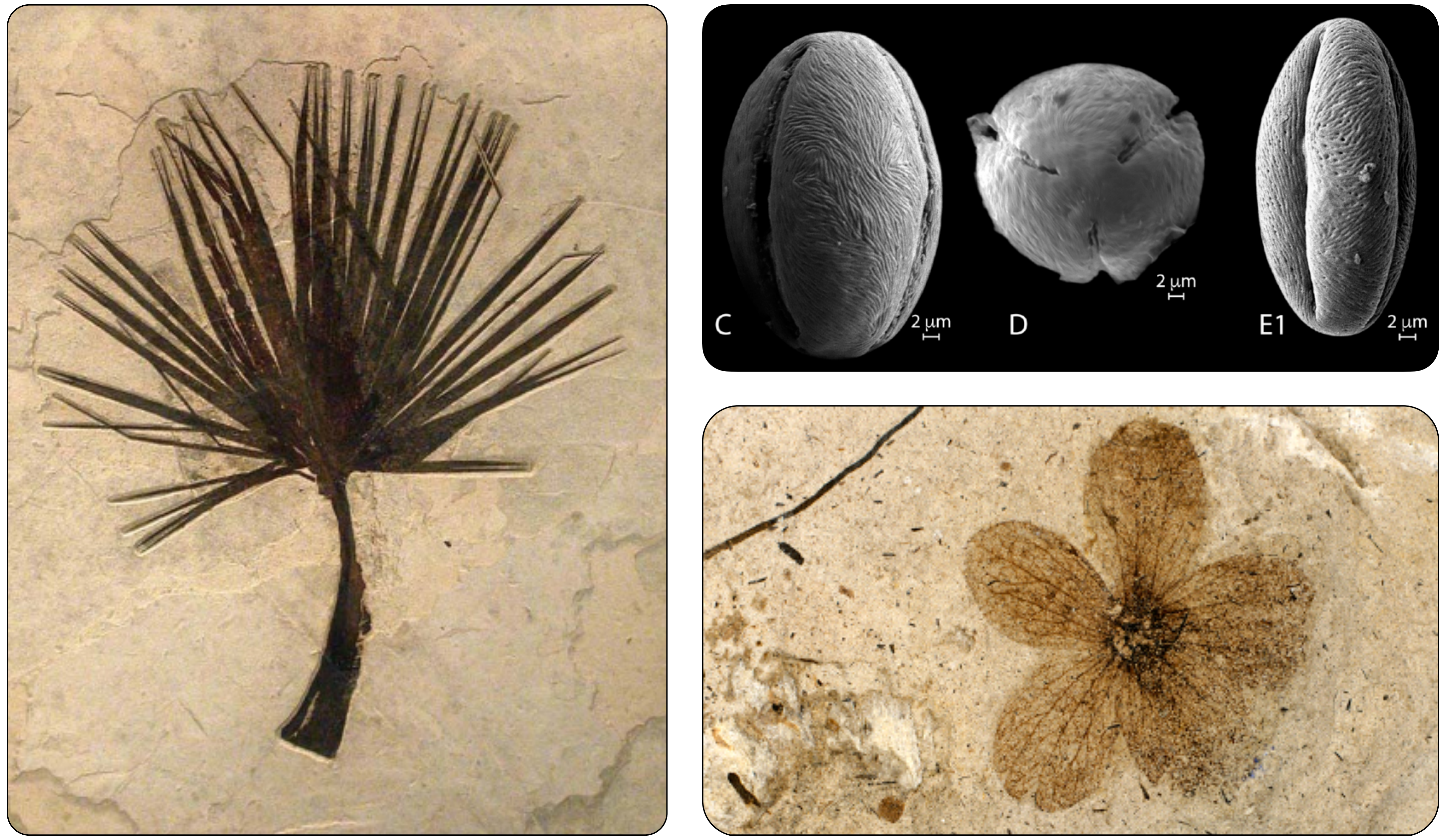
Bayesian estimation of diversity trajectories – mcmcDivE



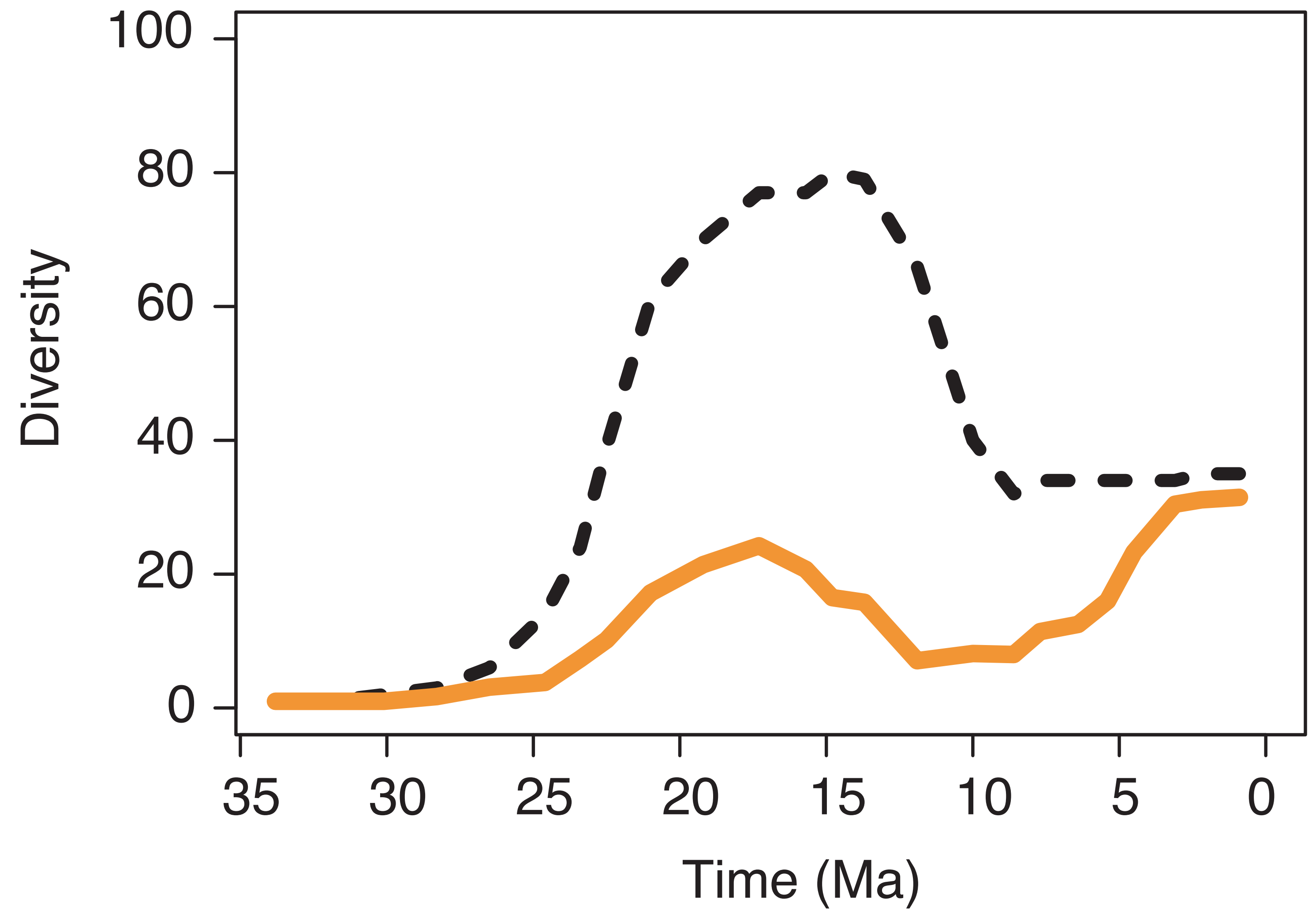
Flannery-Sutherland et al. 2022 Nature Comm

Estimating biodiversity through time from fossil data

Evidence of past biodiversity in the fossil record



Observed vs true diversity through time

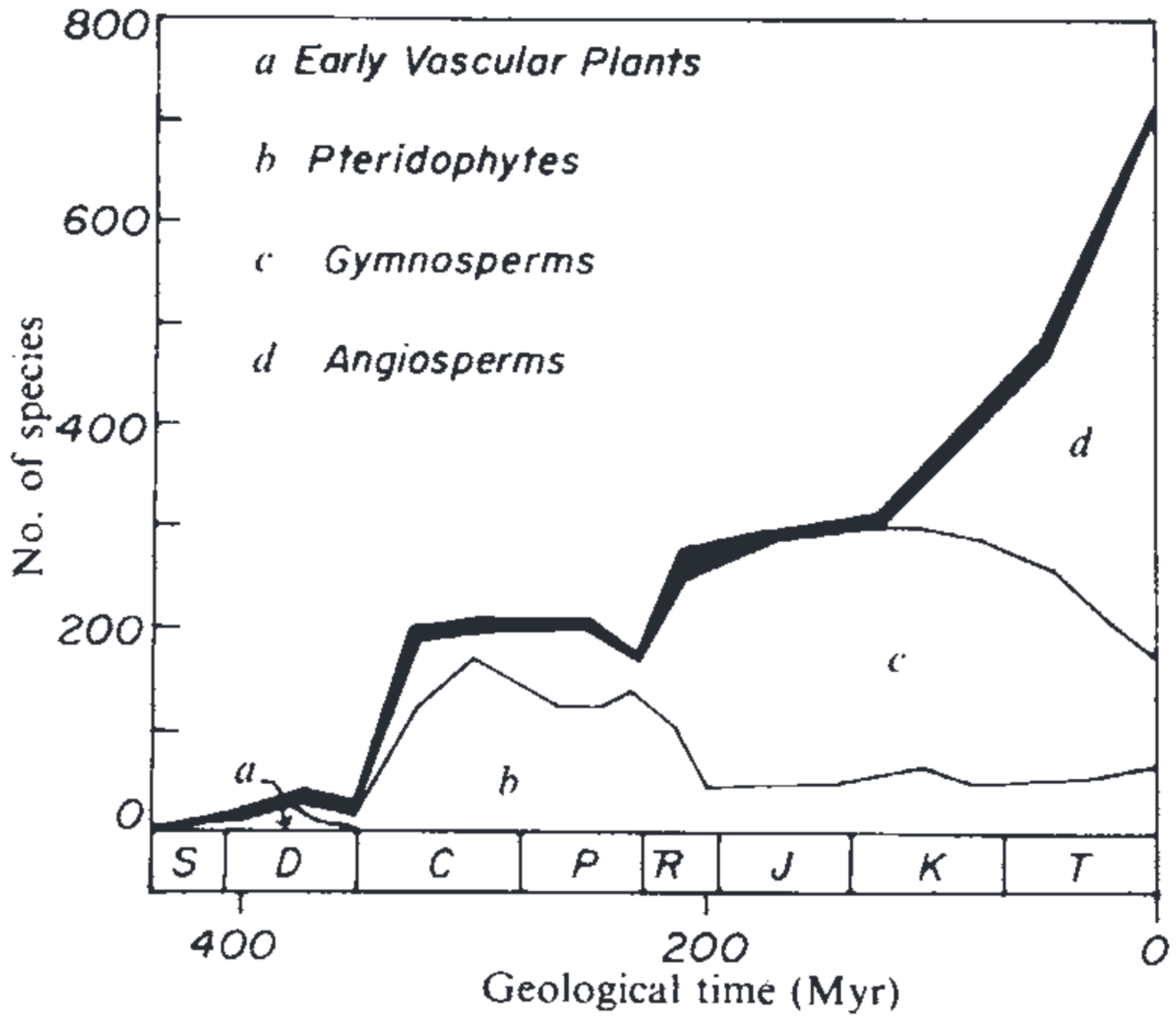


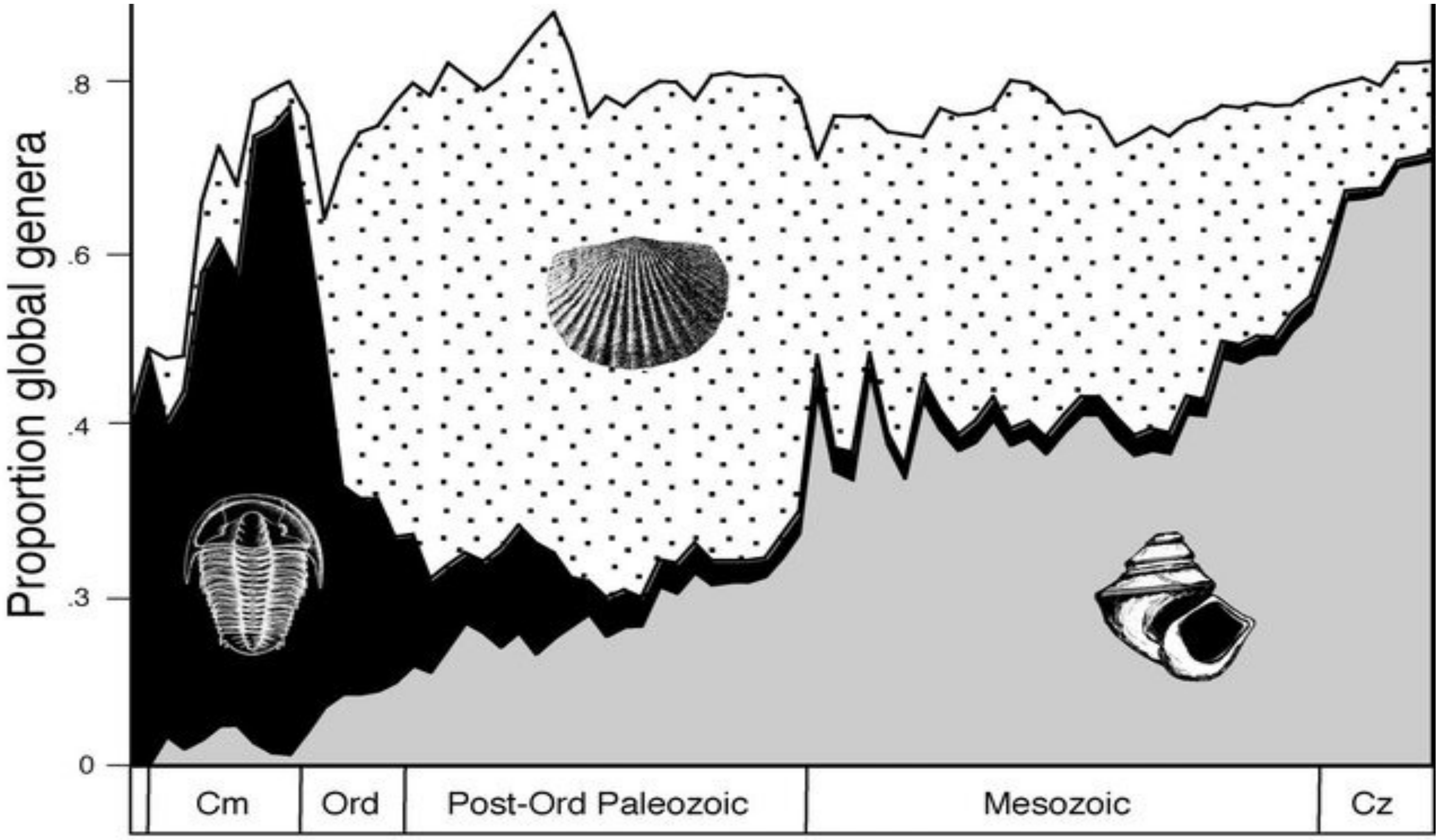
Non-random biases at taxonomic, temporal, and spatial scales 😱

Available methods are based on rarefaction or simple models assuming random sampling

Alroy 2010 Science; Starrfelt & Liow 2016 Phil Trans B; Flannery-Sutherland et al. 2022 Nature Comm

Inferring biodiversity dynamics from the fossil record





Niklas et al. 1983 Nature; Sepkoski 1981 Paleobiology

Problems in estimating biodiversity through time from fossil data

Spatial biases make global biodiversity patterns unidentifiable using the current methods

Science AAAS

PALEONTOLOGY

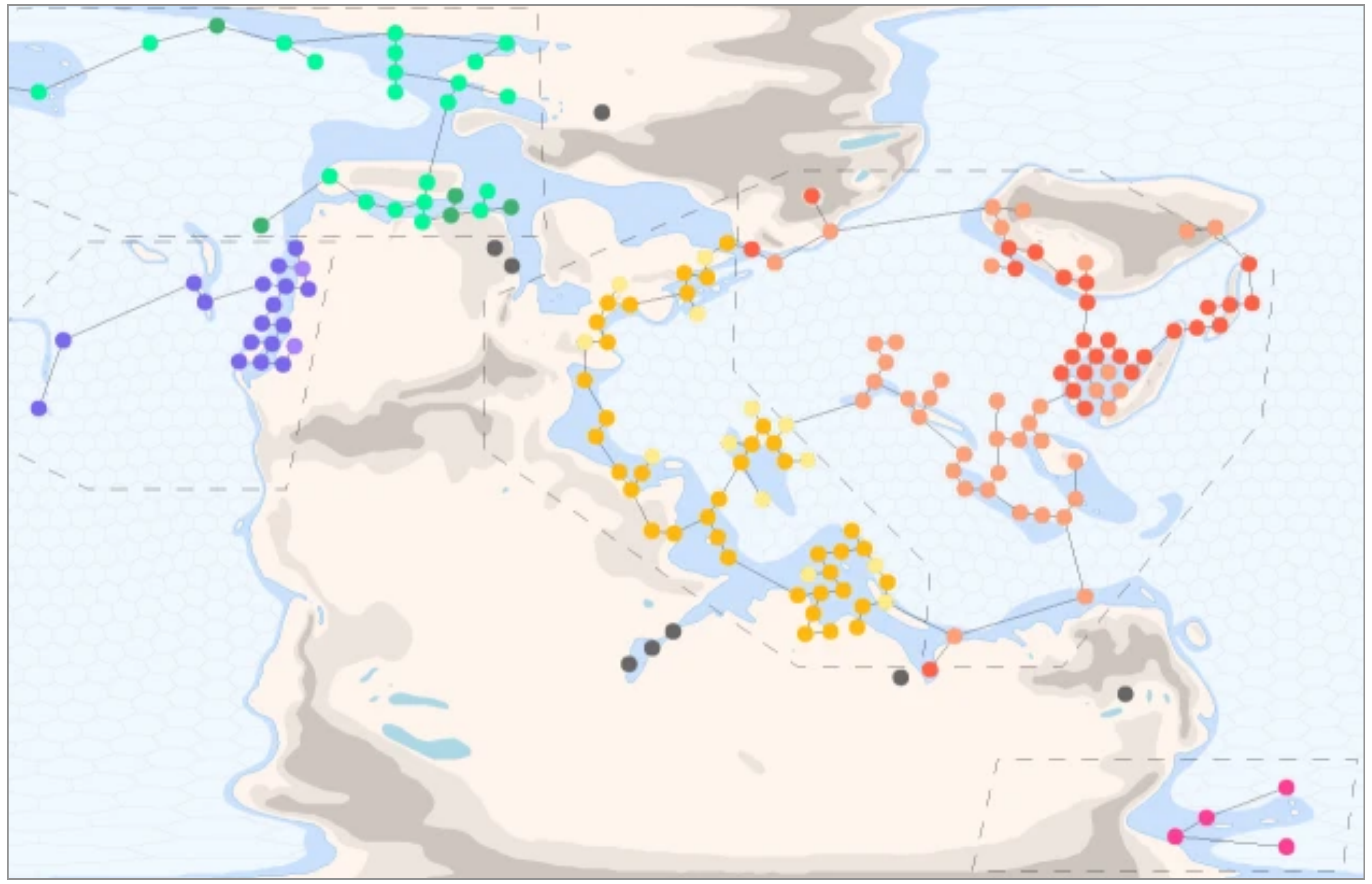
The spatial structure of Phanerozoic marine animal diversity

R. A. Close[1]*, R. B. J. Benson[2], E. E. Saupe[2], M. E. Clapham[3], R. J. Butler[1]

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

The apparent exponential radiation of Phanerozoic land vertebrates is an artefact of spatial sampling biases

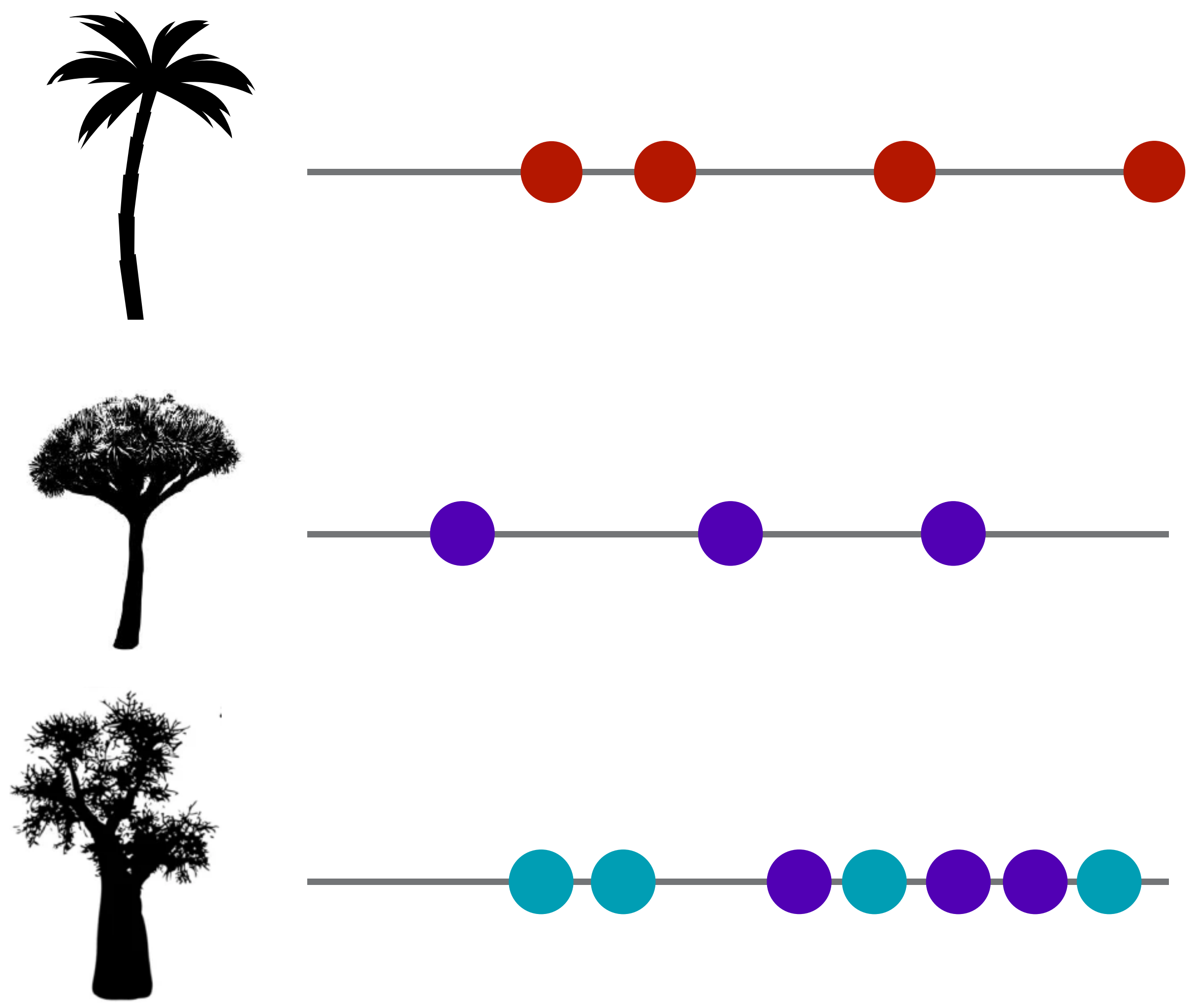


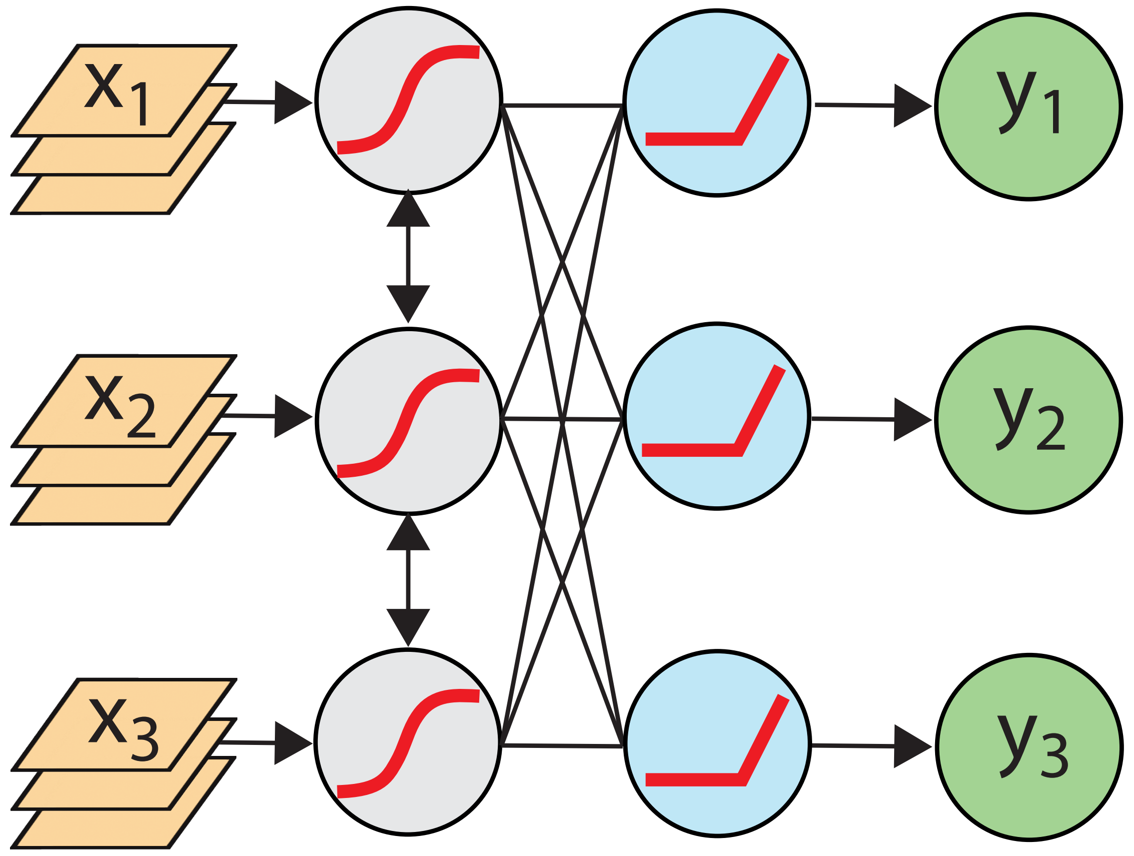
Alroy 2010 Science; Starrfelt & Liow 2016 Phil Trans B; Flannery-Sutherland et al. 2022 Nature Comm

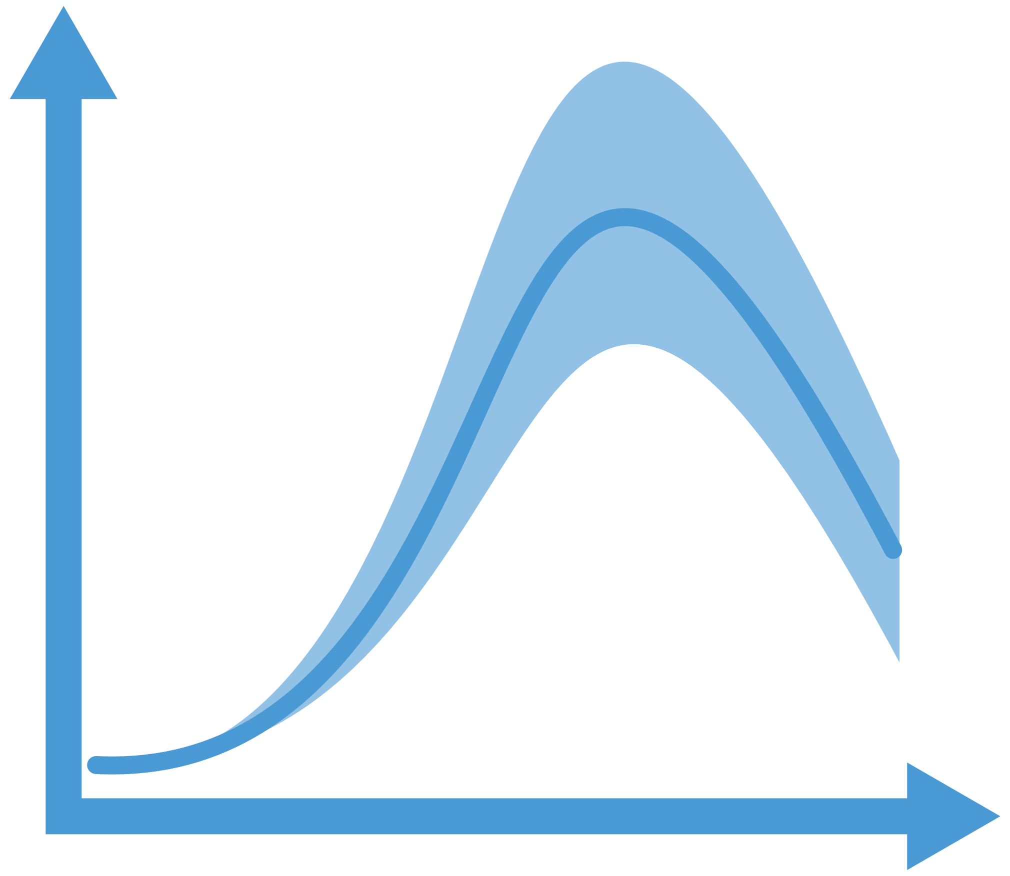
DeepDive: deep learning estimation of biodiversity through time from fossil data

Input: fossil distribution in space and time

Output: diversity estimates through time







R Cooper

B Allen

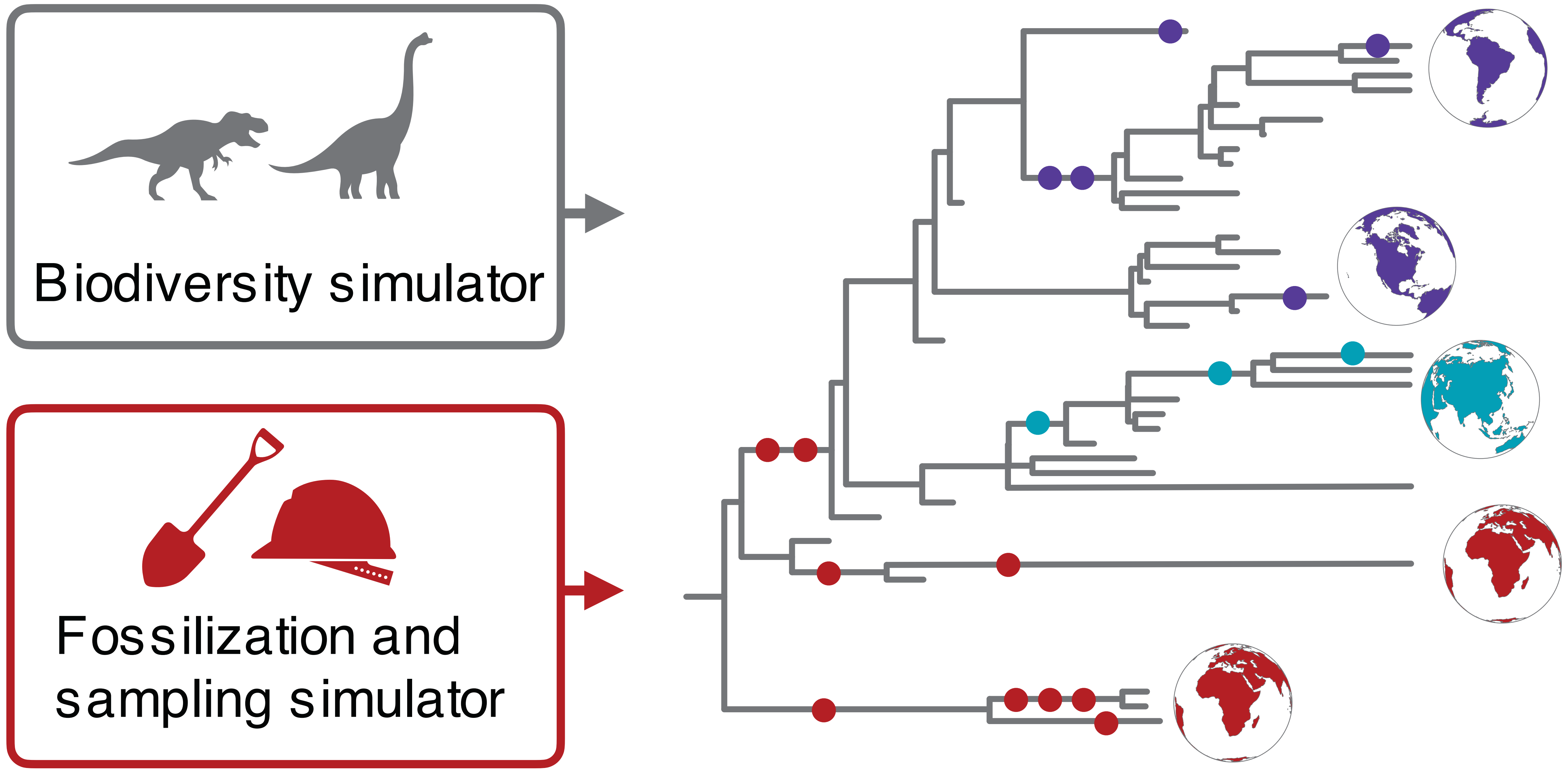
github.com/DeepDive-project

Cooper et al. 2024 Nature Comms

Application note: Cooper et al. 2024 bioRxiv

DeepDive – Deep learning models to estimate Diversity trajectories

Mechanistic model of species diversification and extinction in time and space

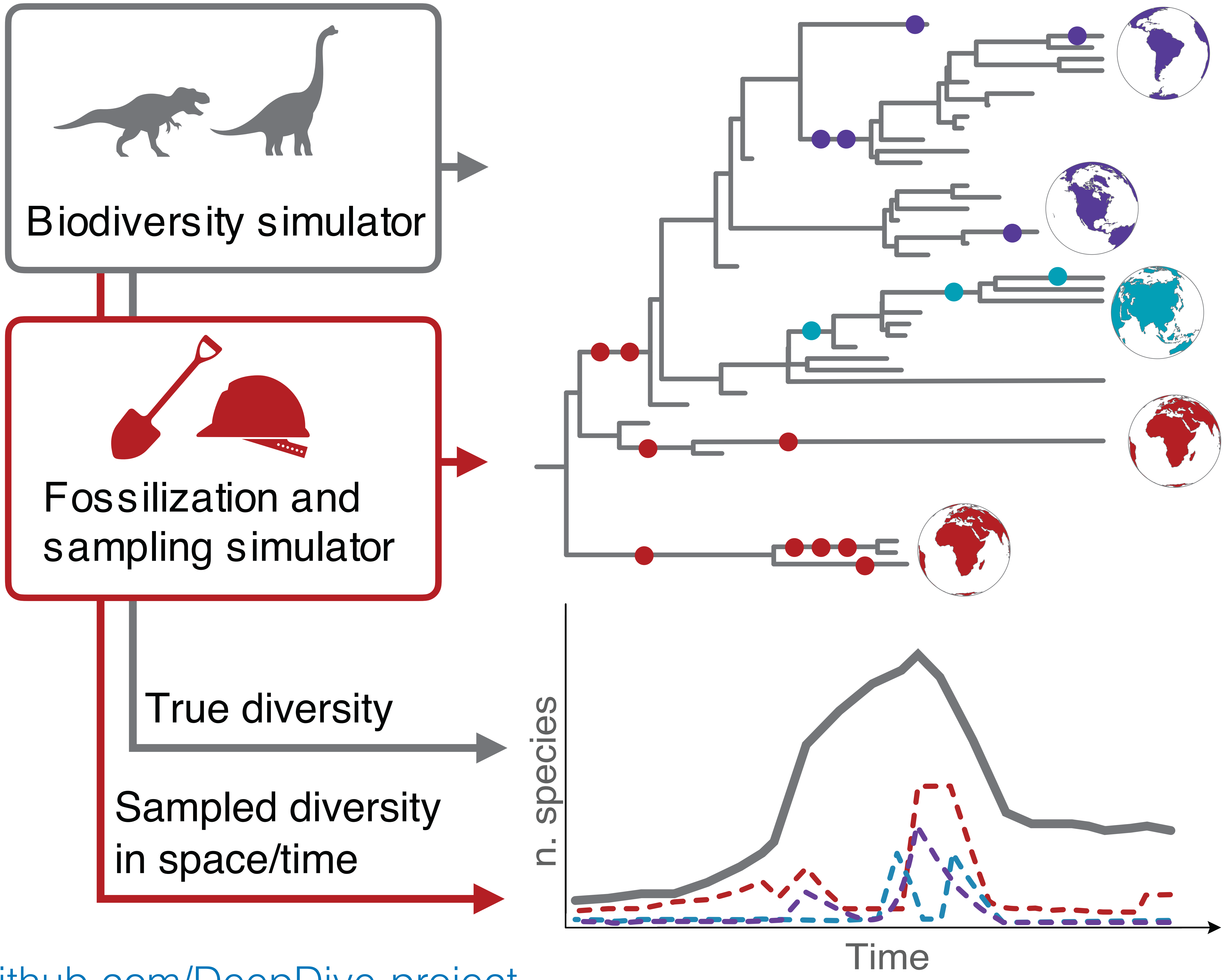


Clade evolution based on spatially-explicit birth-death processes

Simulated fossil data with spatial, temporal, and taxonomic biases

DeepDive – Deep learning models to estimate Diversity trajectories

Mechanistic model of species diversification and extinction in time and space



Clade evolution based on spatially-explicit birth-death processes

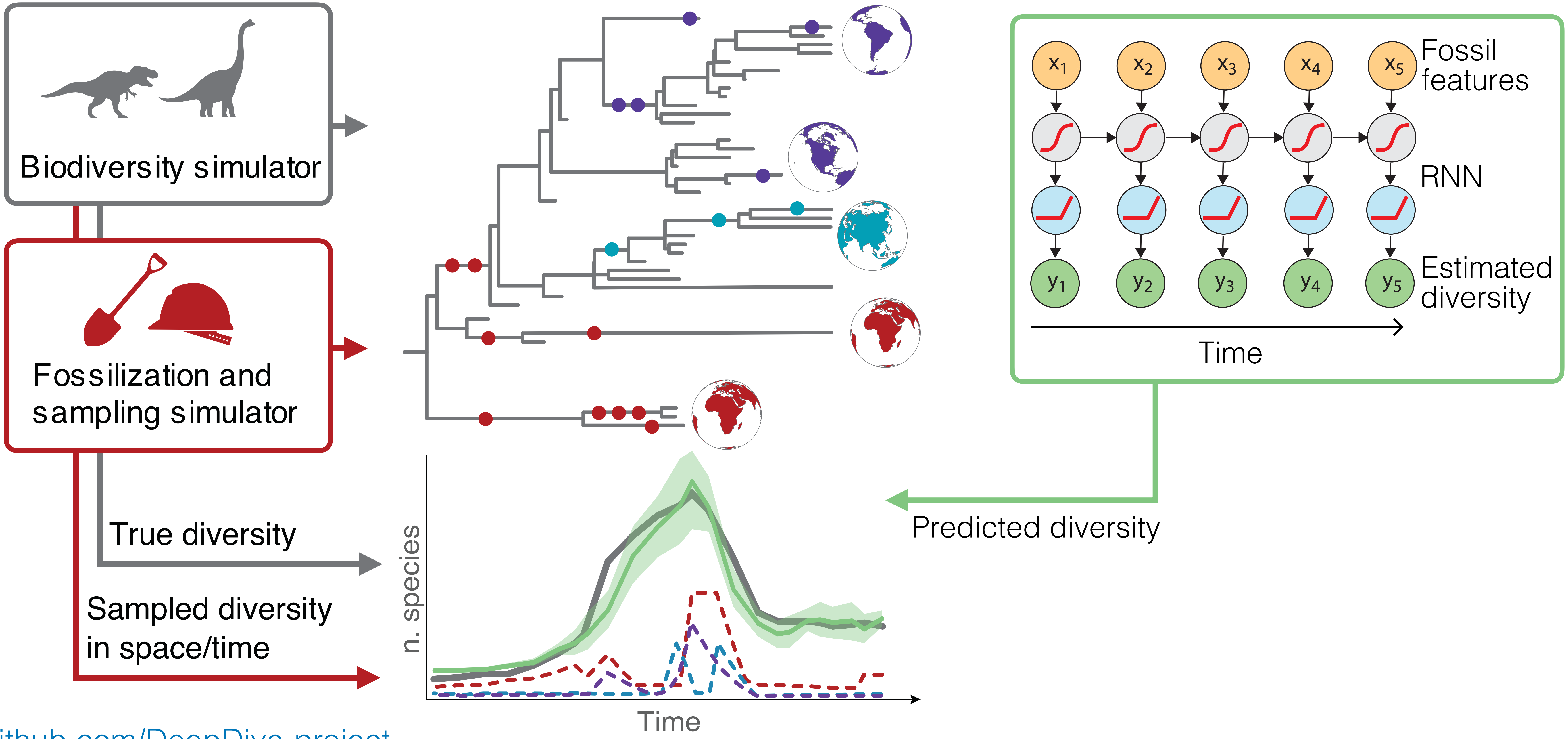
Simulated fossil data with spatial, temporal, and taxonomic biases

github.com/DeepDive-project

Supervised deep learning models to estimate diversity trajectories

Mechanistic models of speciation, extinction, and fossilization in time and space

Deep learning model predicting diversity time series from fossil features

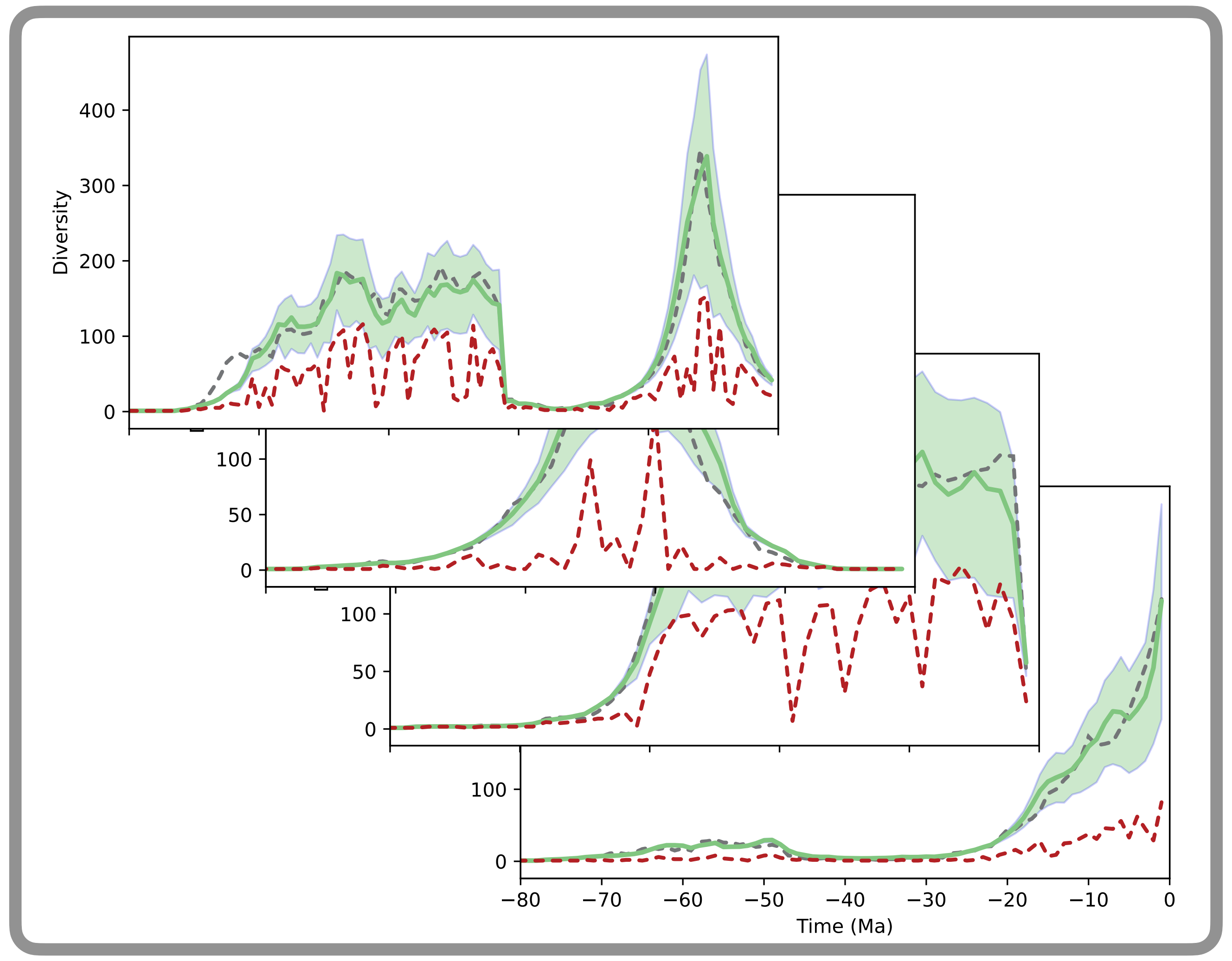


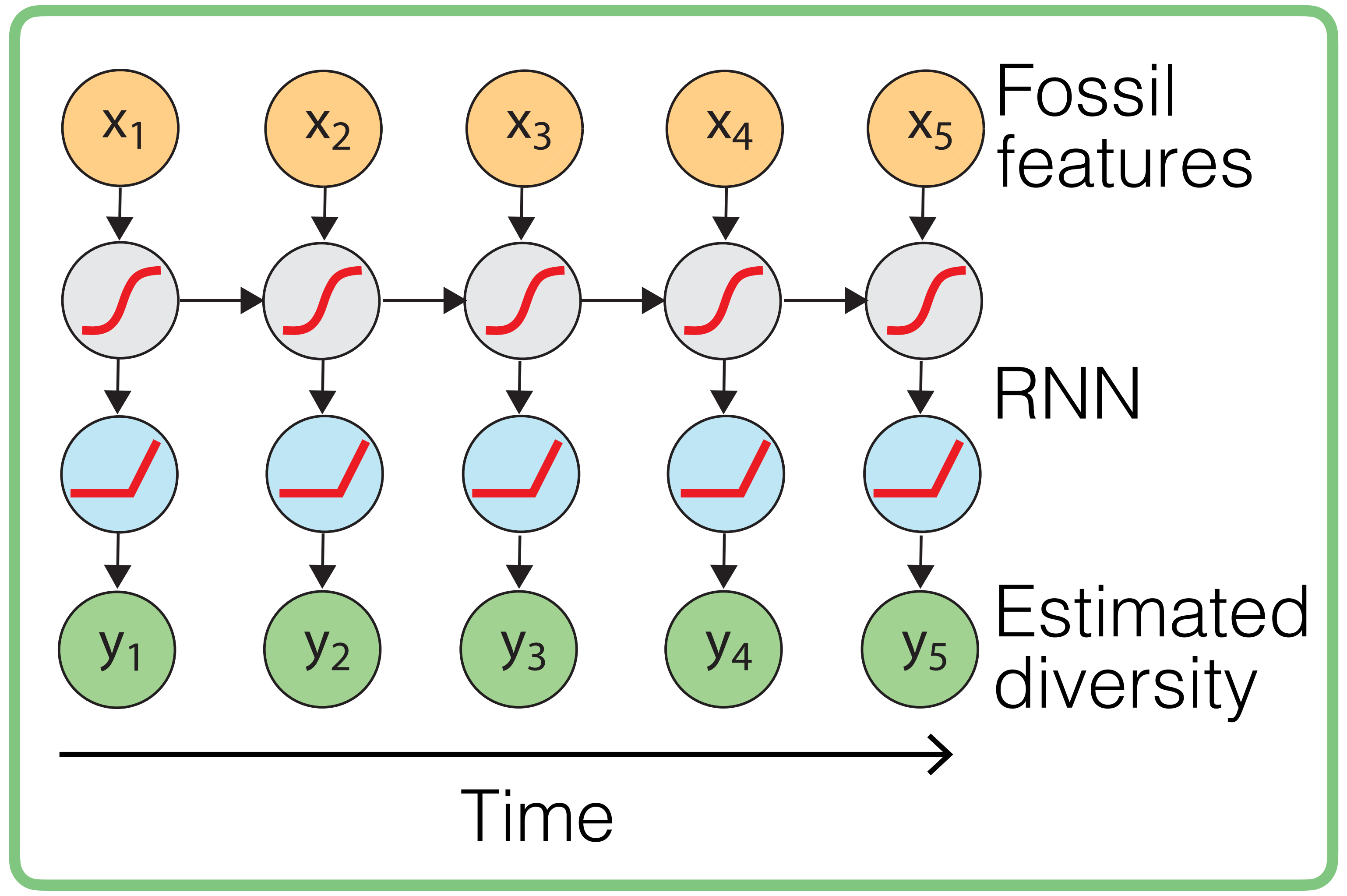
github.com/DeepDive-project

DeepDive – Deep learning models to estimate Diversity trajectories

Model optimization (training) and validation

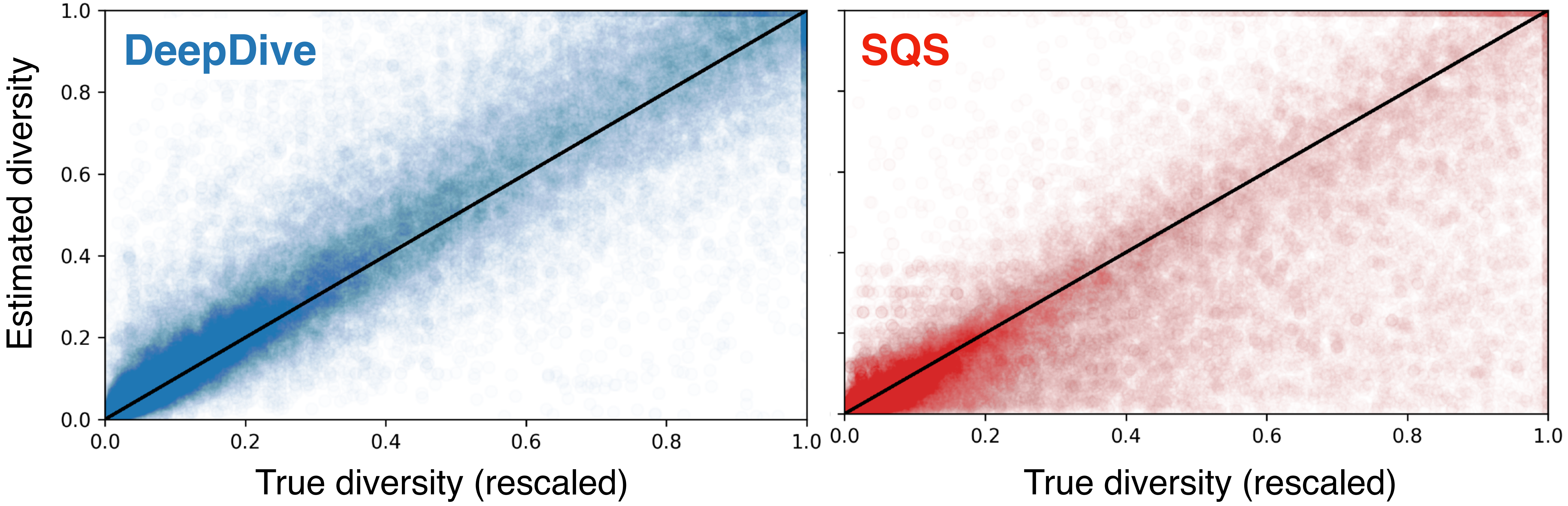
Samples from 'prior' distributions of speciation, extinction, migration, preservation rates



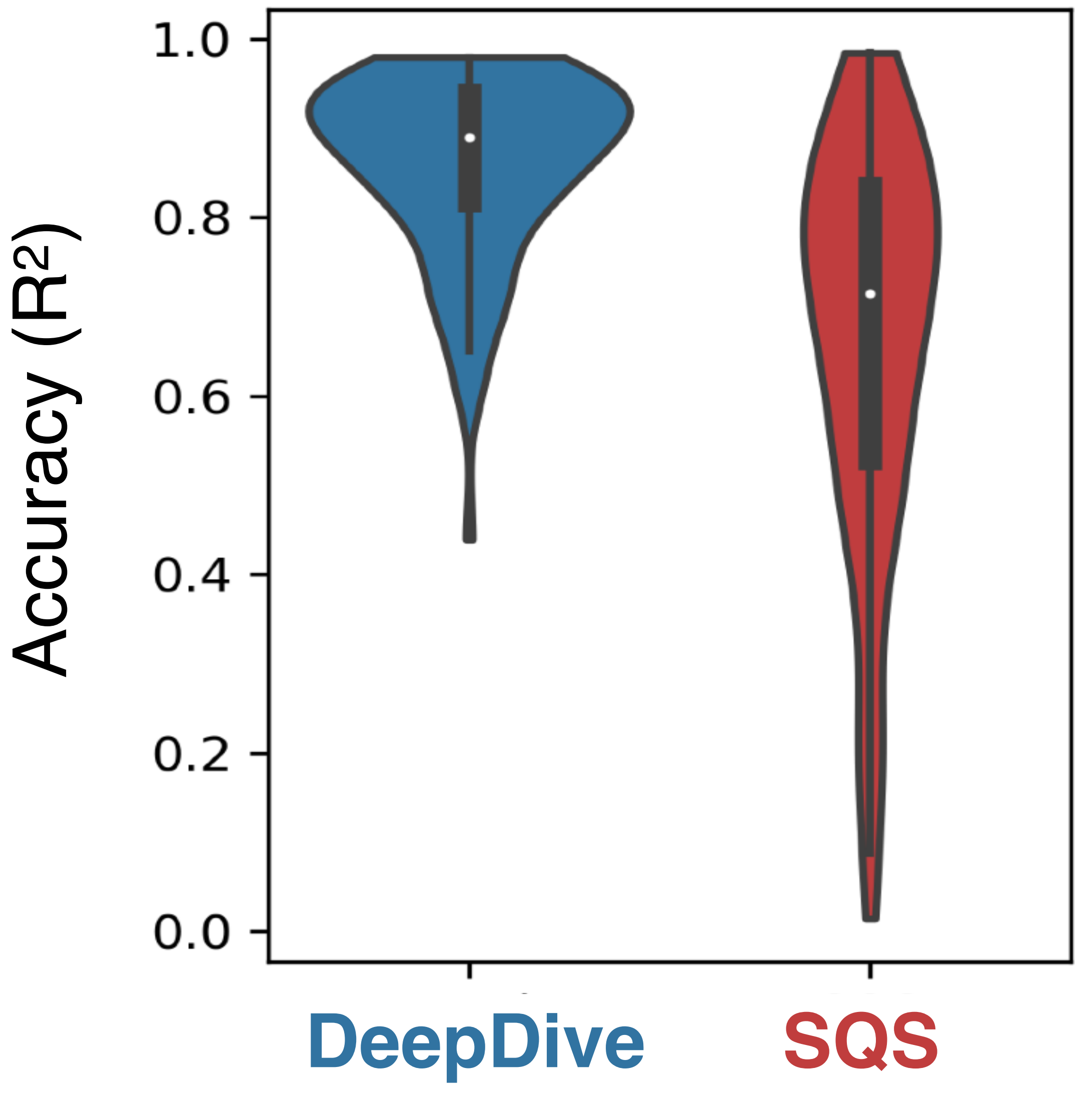


github.com/DeepDive-project

DeepDive performance

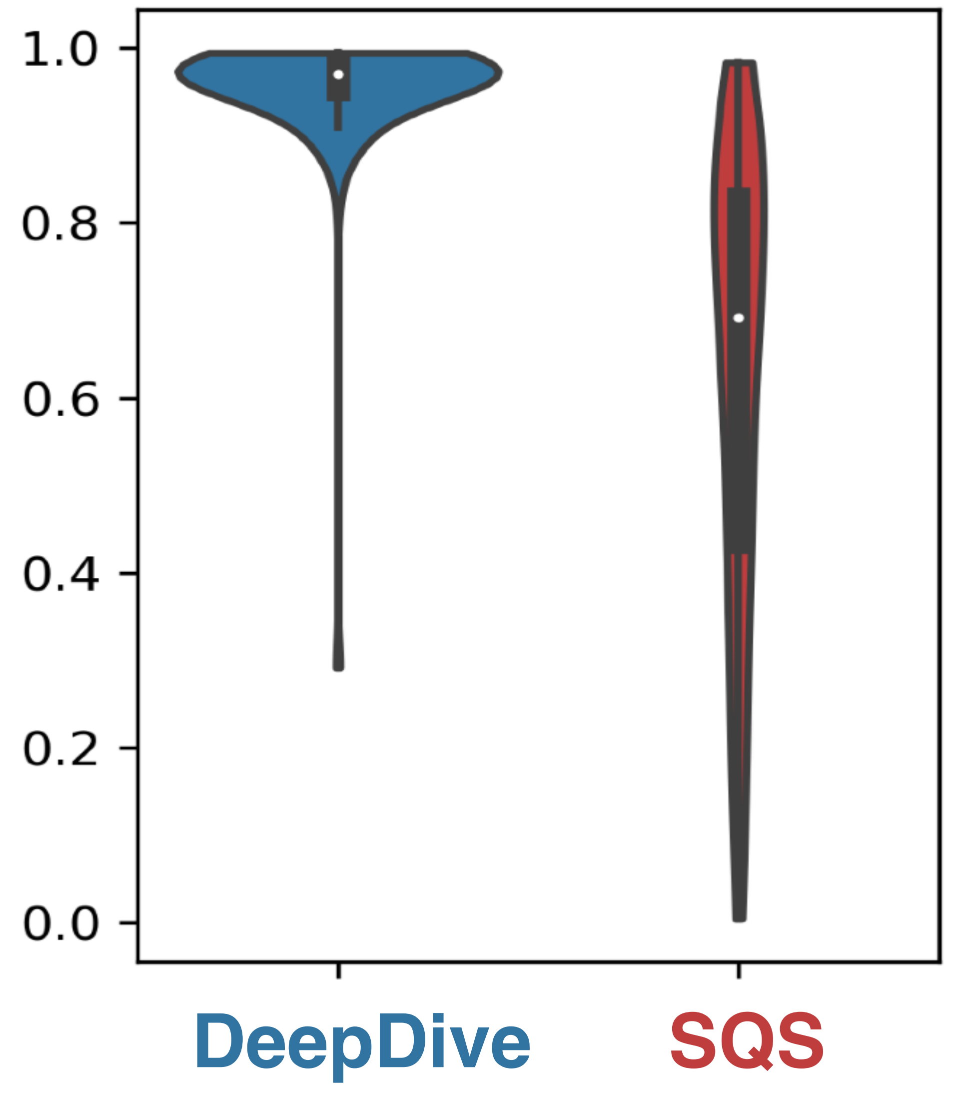


Strong temporal bias



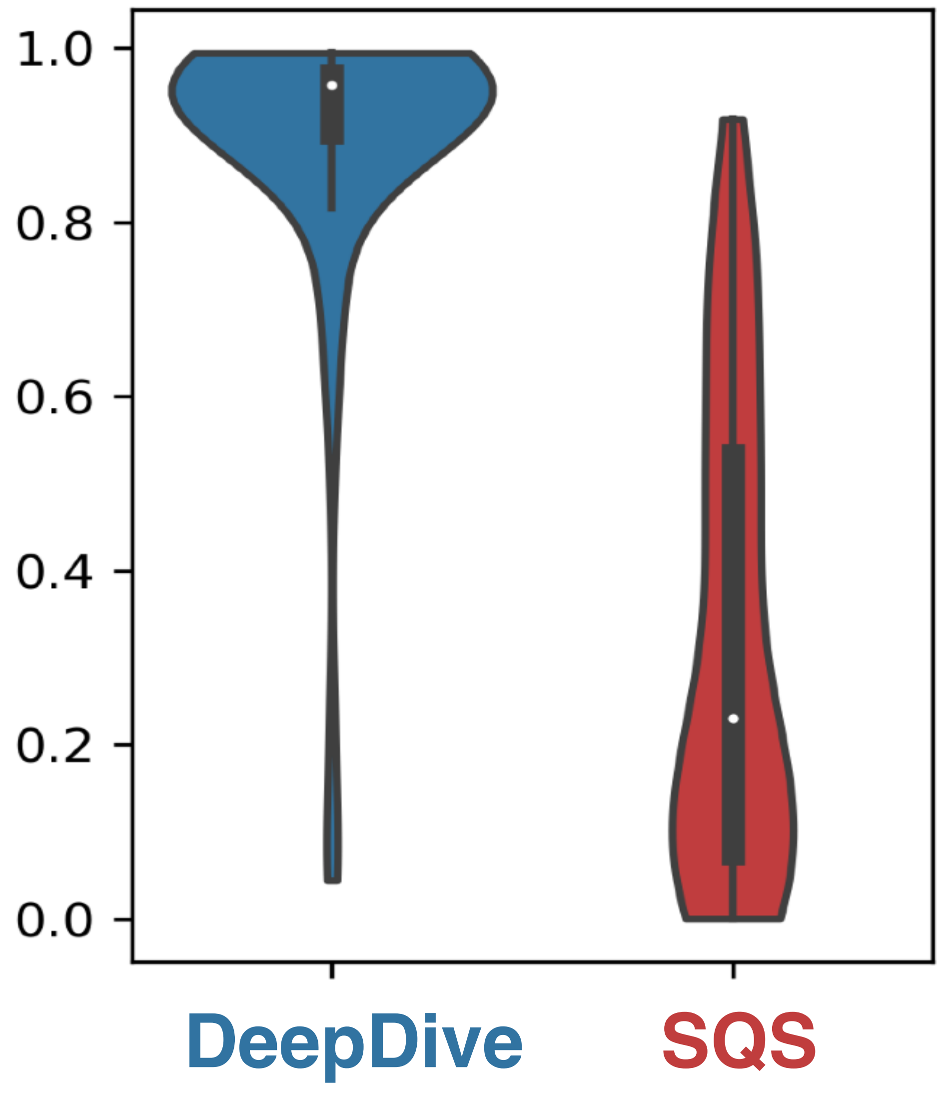
100fold sampling rate variation through time

Strong taxonomic bias



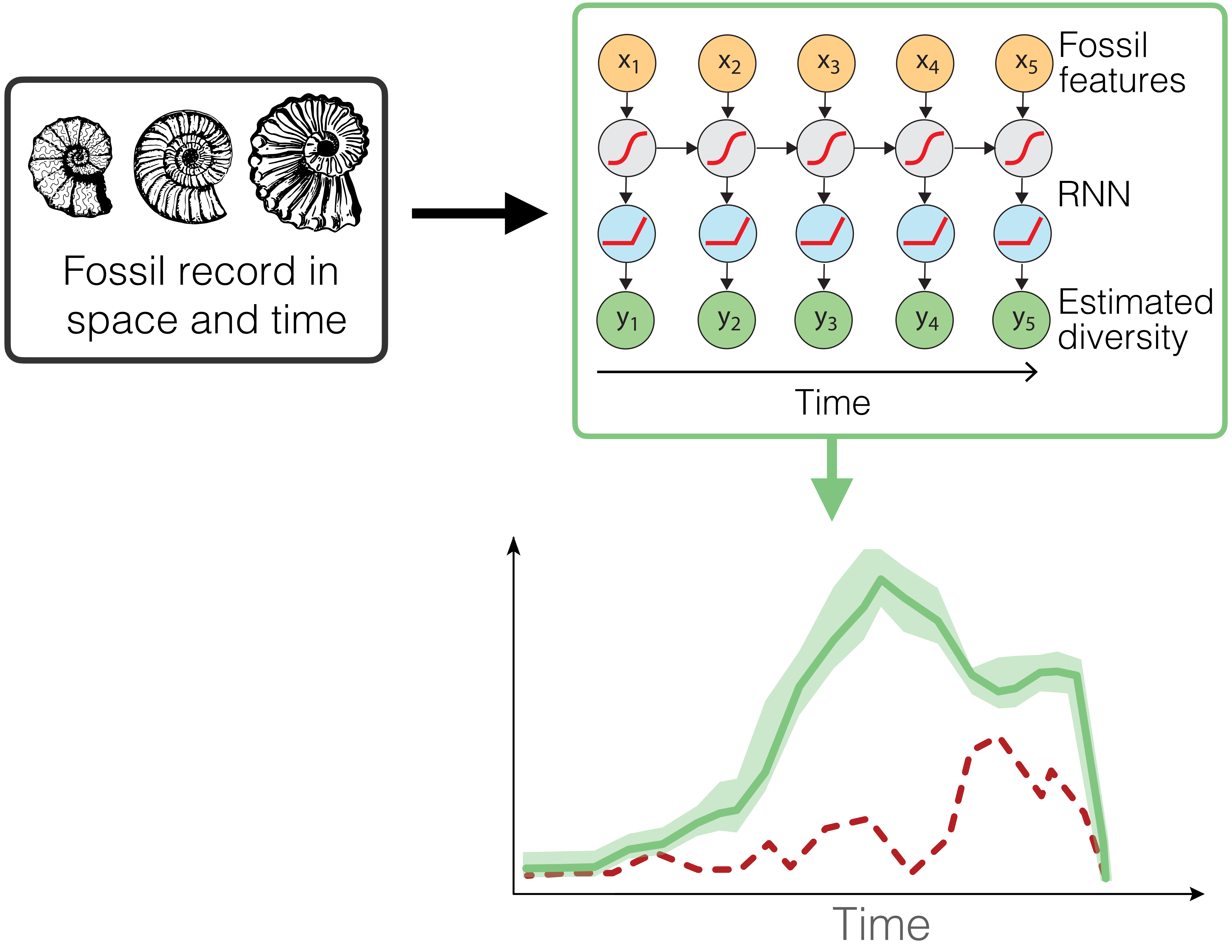
1–3 orders of magnitudes in sampling rate variation across species

Strong spatial bias

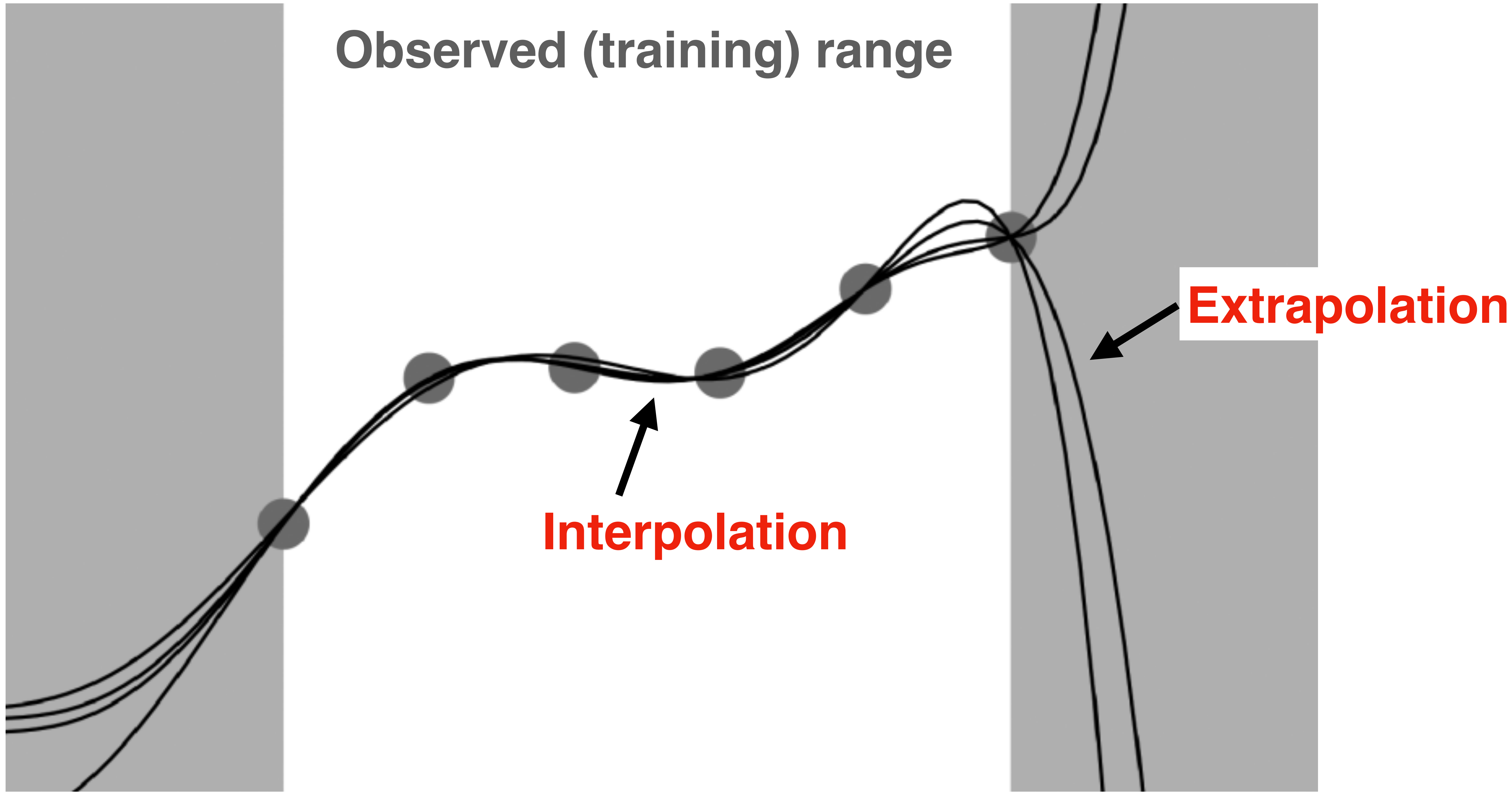


100fold sampling rate variation across areas

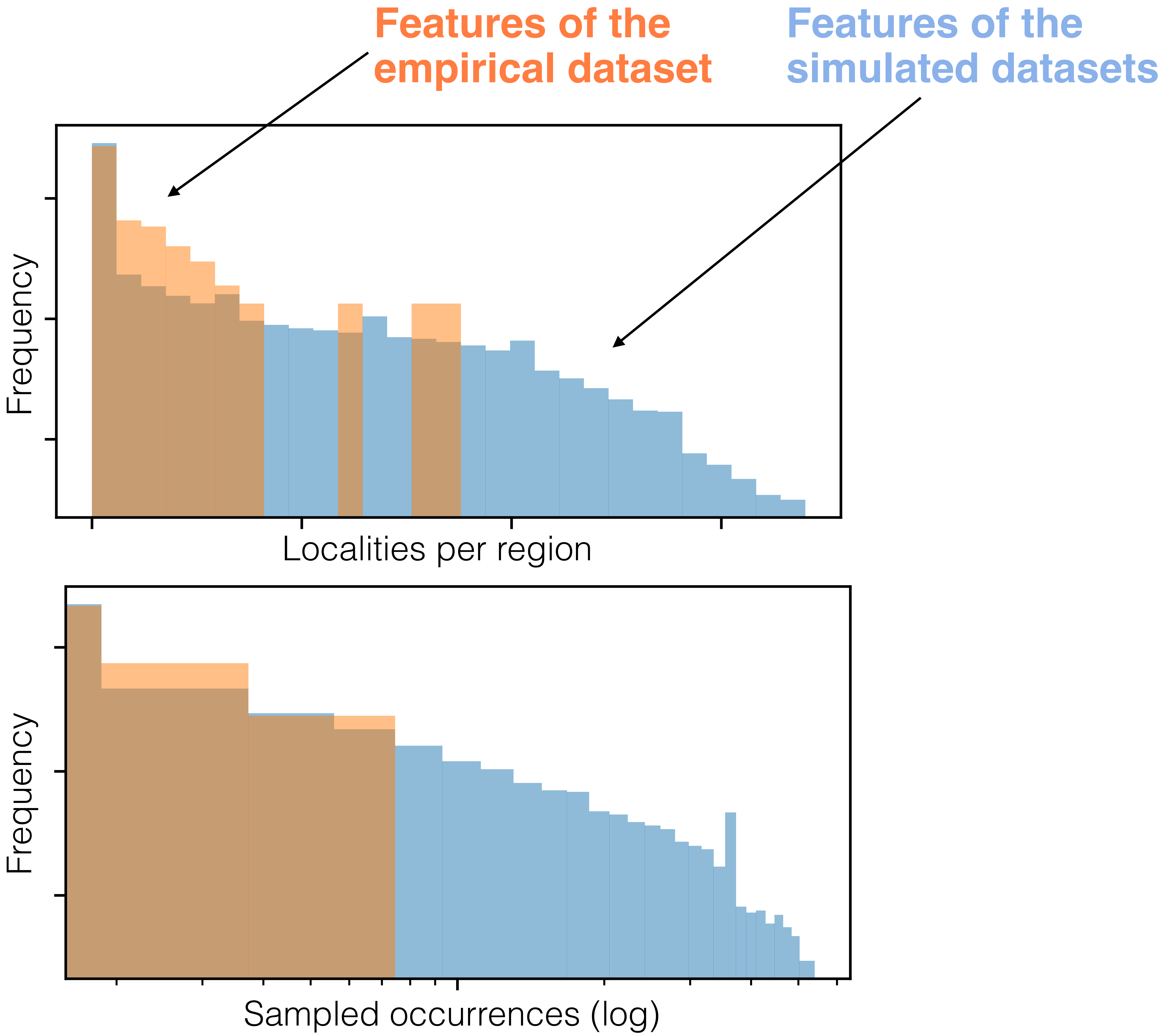
DeepDive – Predictions of empirical data

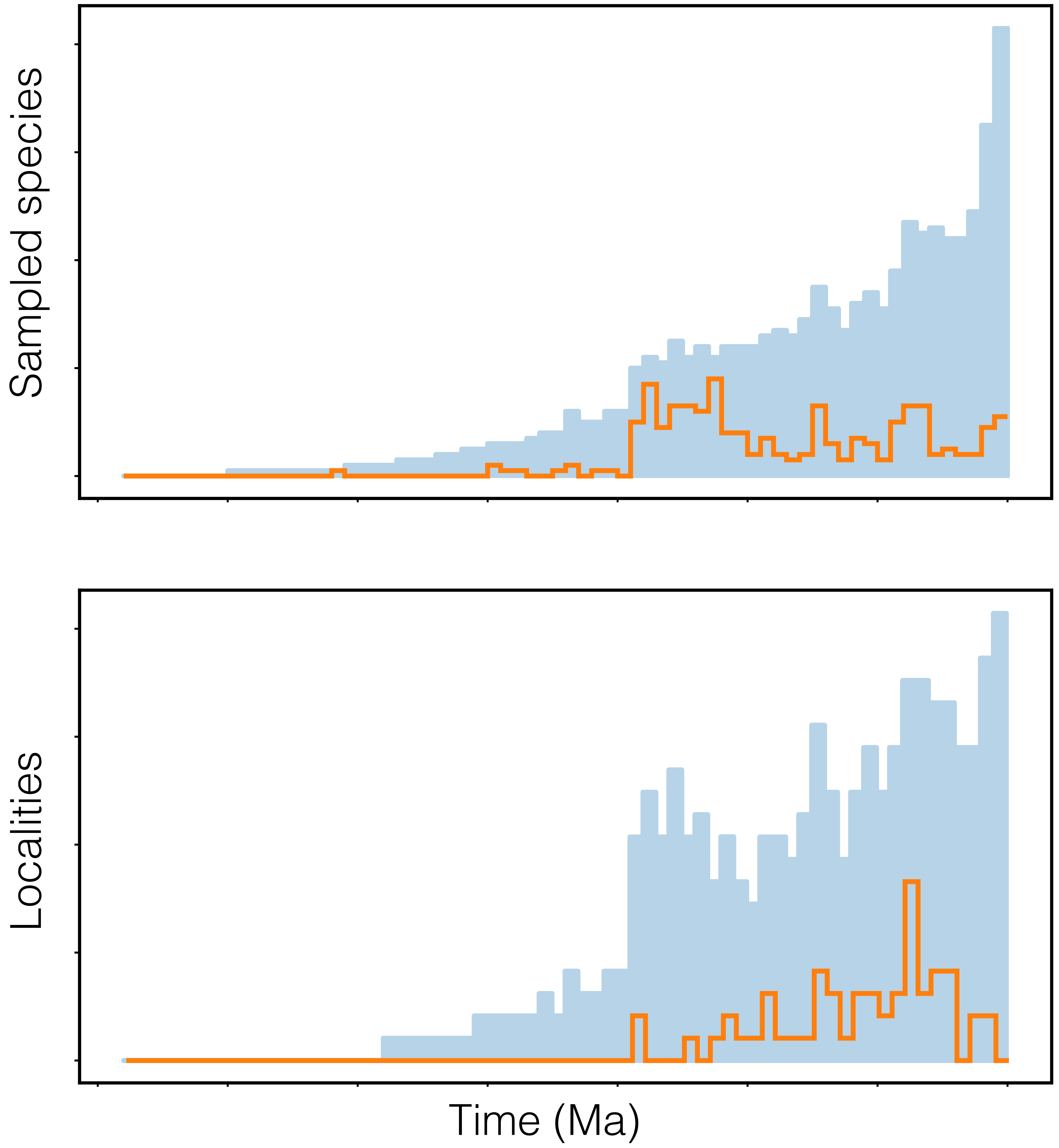


Caveat: the training simulations must resemble the true data

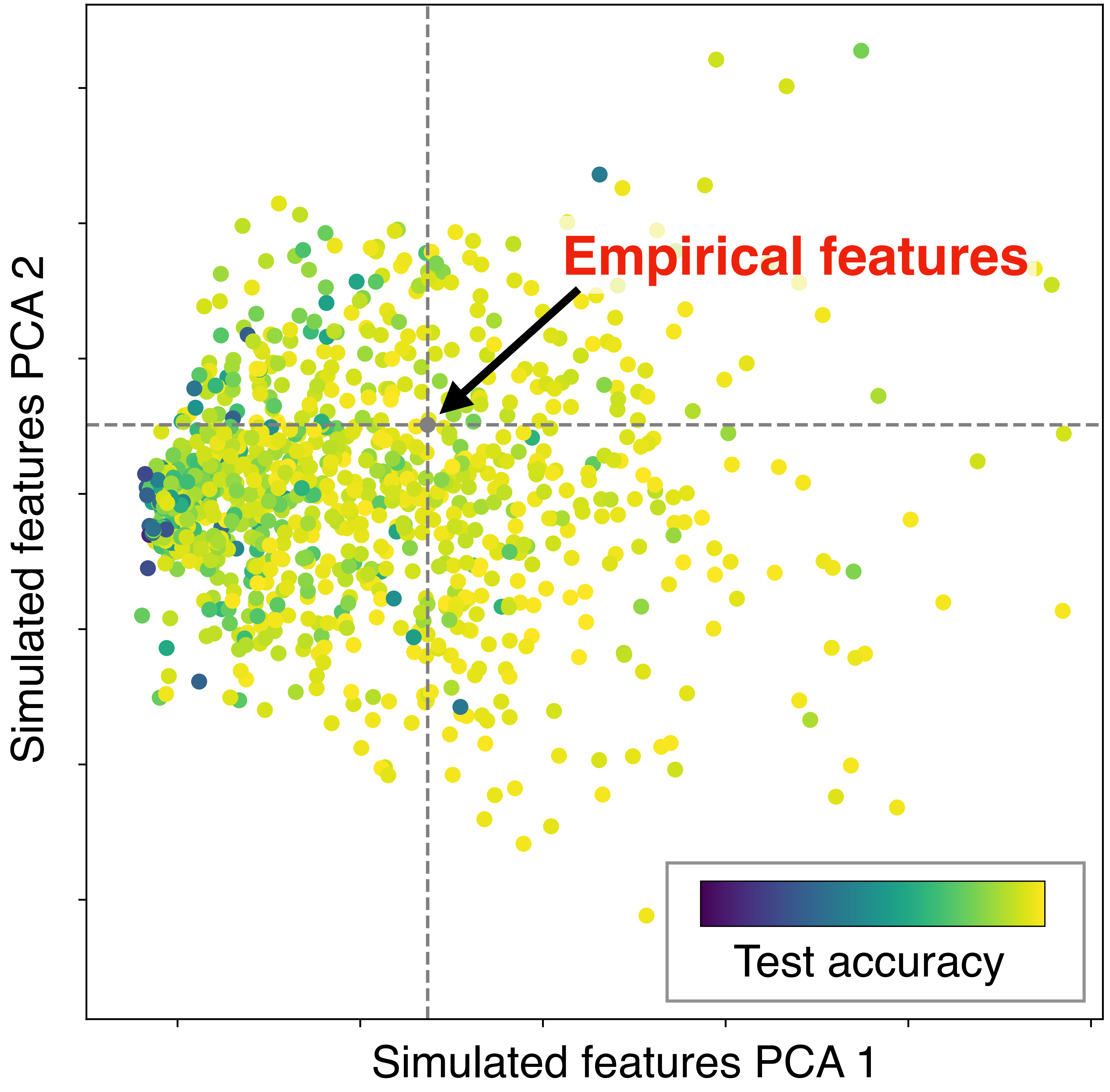


Checking that the simulations are in the right ballpark

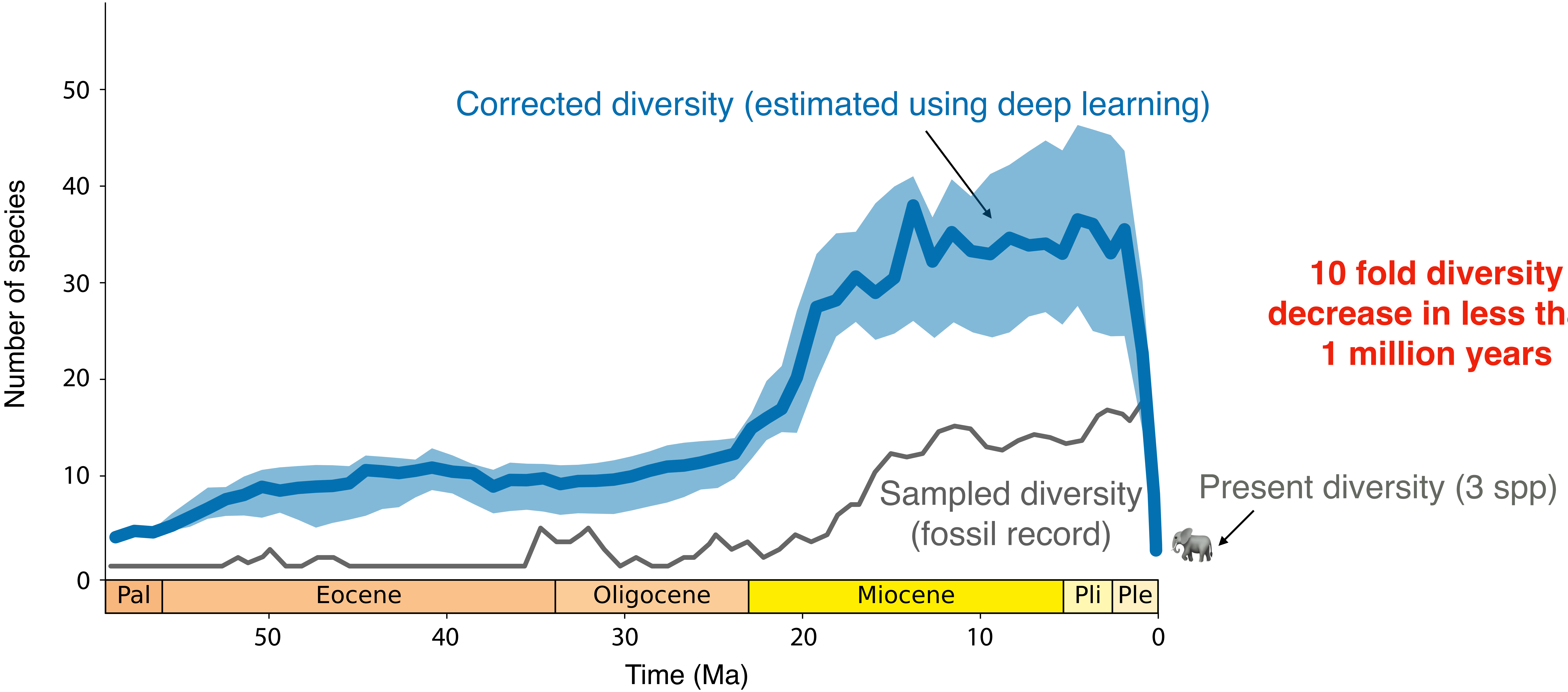




Checking that the simulations are in the right ballpark

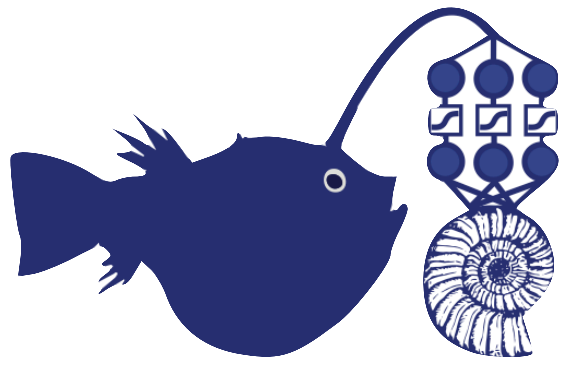


The rise and fall of elephants



10 fold diversity decrease in less than 1 million years

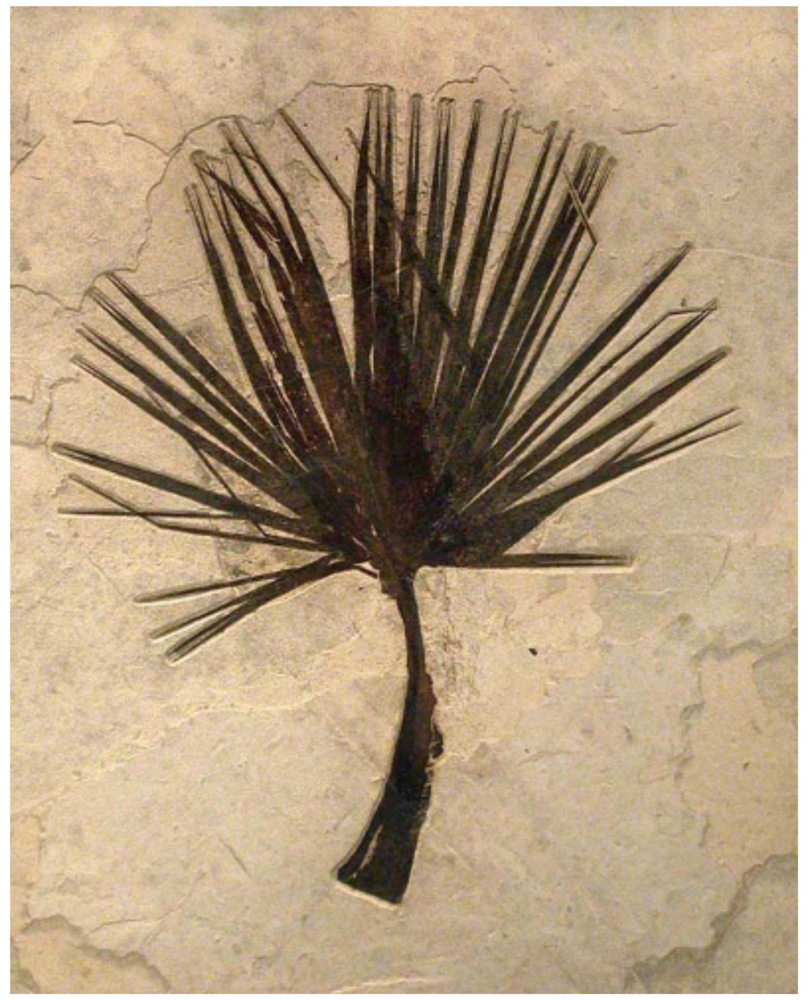
R Cooper



github.com/DeepDive-project

Cooper et al. 2024 Nature Comms

How did palms diversify through time?



A new palm fossil database

>2000 records from 1300 references

>400 verified micro- and macro-fossil occurrences

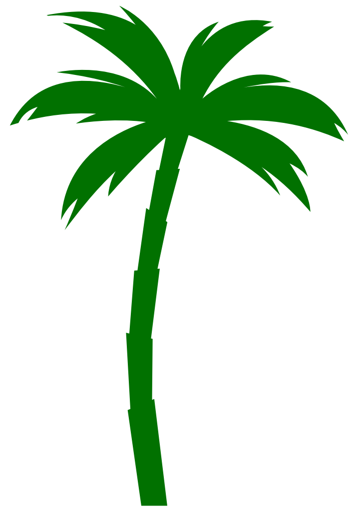
Cretaceous and Cenozoic localities from all continents

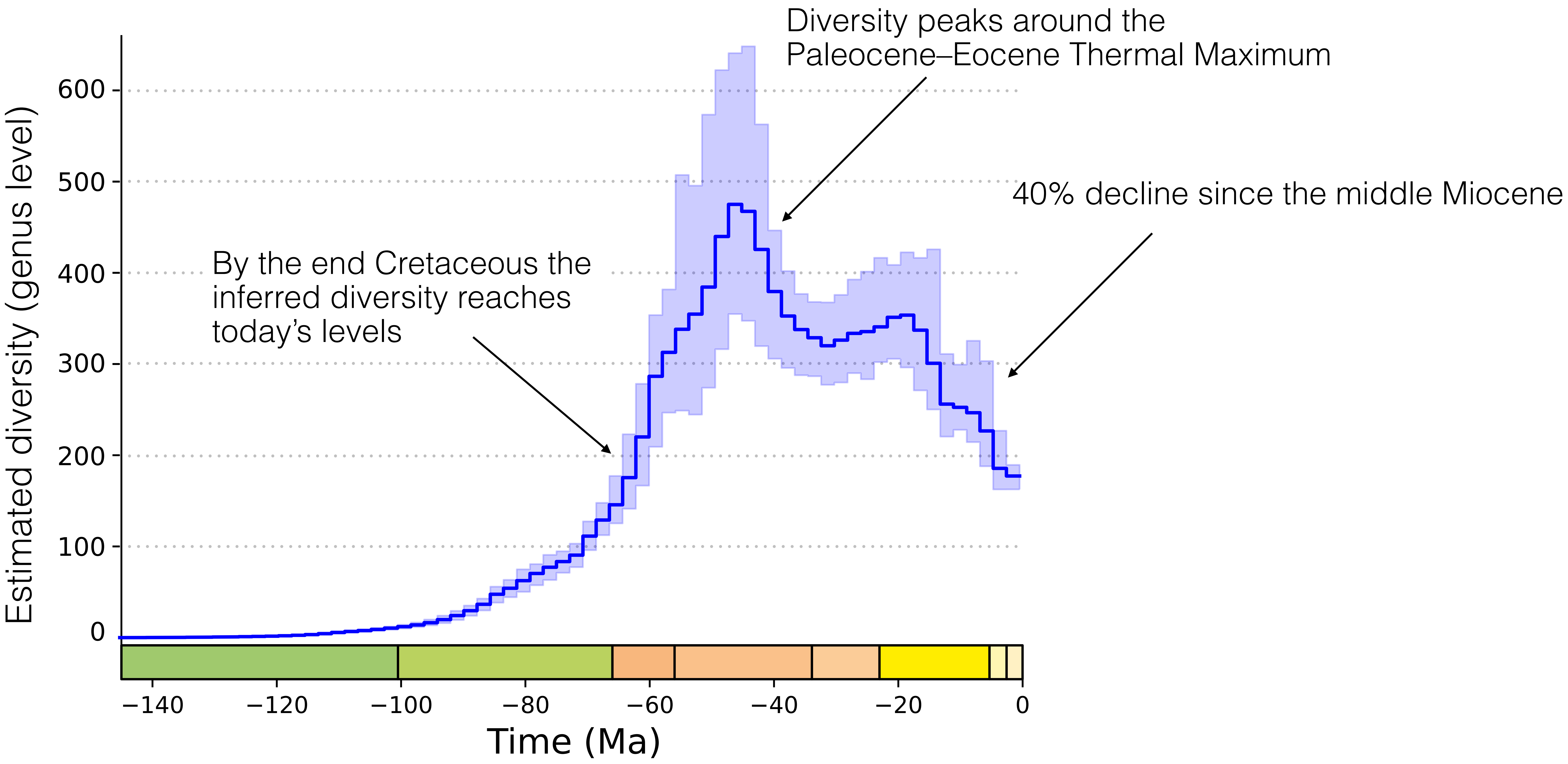

C. D. Bacon

A team (ongoing) effort

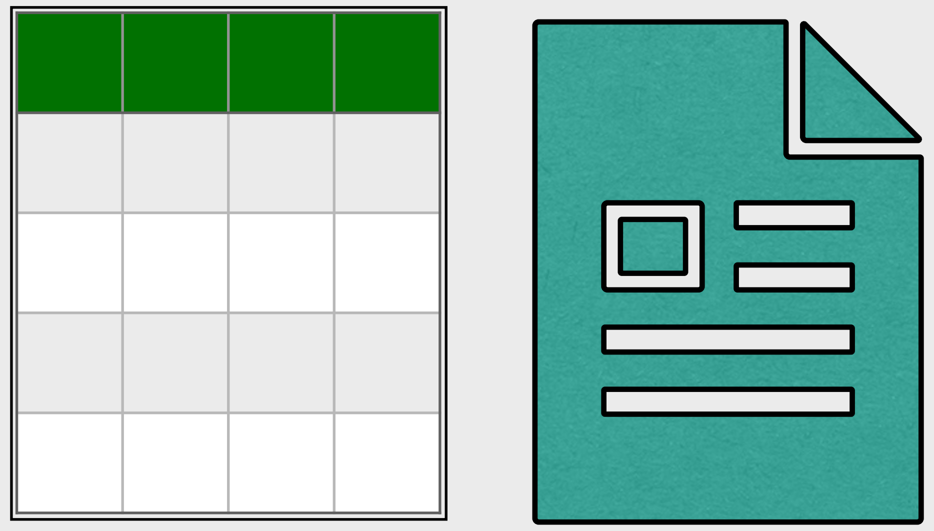
Rosane Collevatti
Carina Hoorn
Huasheng Huang
Viktoria Keller
Luis Palazzesi
Shalani Parmar
Vandana Prasad
David Sunderlin
Kelly Matsunaga
Robert Morley
Yaowu Xing

The inferred palm evolutionary trajectory



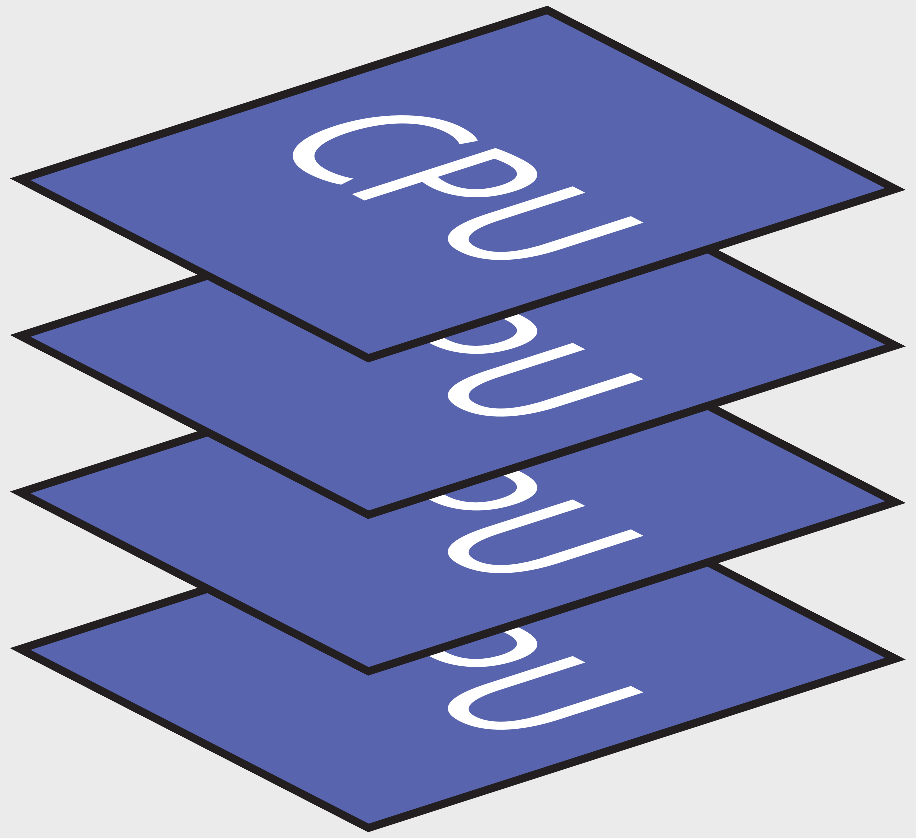


How DeepDive works



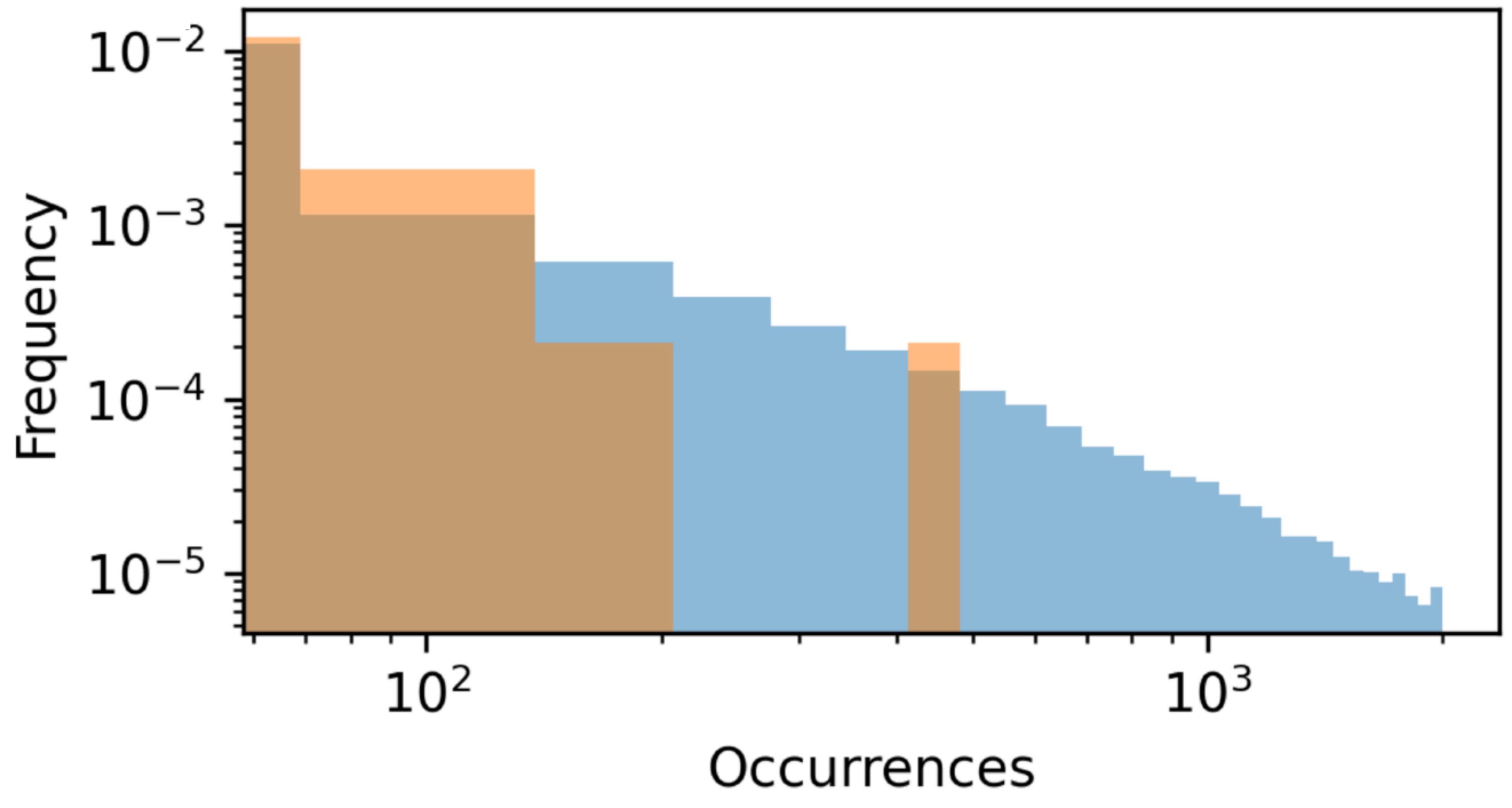
Fossil table and config preparation

DeepDiveR



Parallel simulations

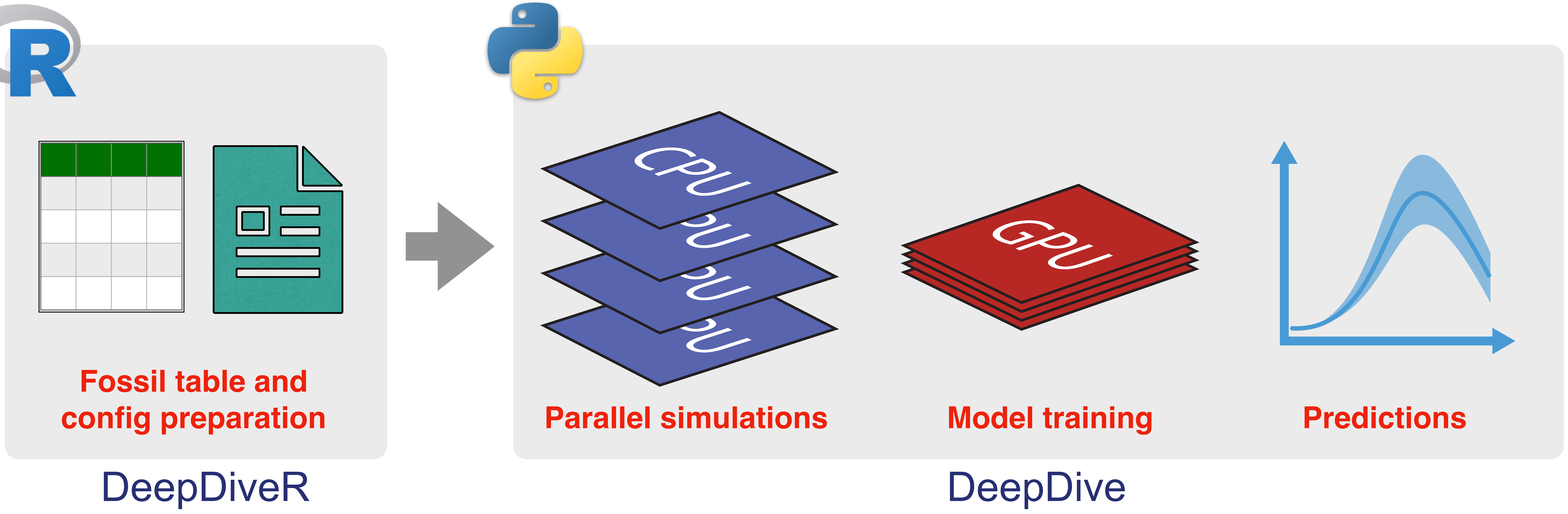
DeepDive



Simulations settings can be adjusted manually but are otherwise automatically calibrated to match the empirical data

github.com/DeepDive-project

How DeepDive works



github.com/DeepDive-project

Setting up an analysis using DeepDiveR

Taxon	Area	MinAge	MaxAge	Locality
Ailurus	Asia	0.6	1.3	loc_1
Ailurus	Asia	9.5	10.35	loc_2
Alopecocyon	Asia	9.5	10.35	loc_2
Alopecocyon	Europe	9.7	11.11	loc_3
…	…	…	…	…

Fossil occurrence data with taxonomic, temporal and spatial info

Parse fossil table and prepare config file

```
library(DeepDiveR)

fossil_tbl <- read.csv("Fossil_table.csv")

# define time bins (here of 1 myr)
bins <- seq(23, 0, by=-1)

# prep DeepDive-formatted fossil dataset
prep_dd_input(fossil_tbl, bins=bins, output_file="fossil_input.csv", r=100)
```

Define timelines and number of age randomizations

Setting up an analysis using DeepDiveR

Taxon	Area	MinAge	MaxAge	Locality
Ailurus	Asia	0.6	1.3	loc_1
Ailurus	Asia	9.5	10.35	loc_2
Alopecocyon	Asia	9.5	10.35	loc_2
Alopecocyon	Europe	9.7	11.11	loc_3
…	…	…	…	…

Fossil occurrence data with taxonomic, temporal and spatial info

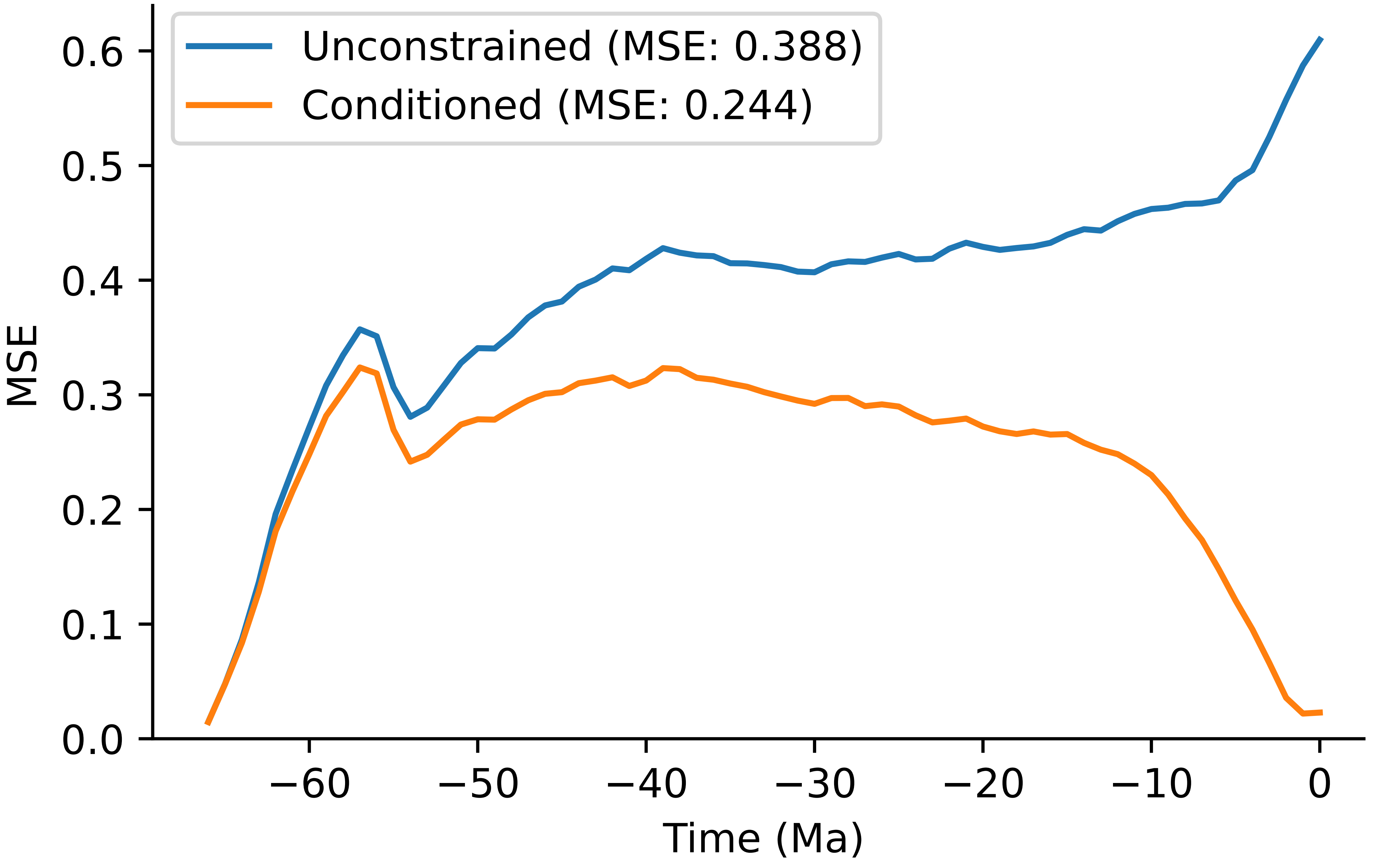
Parse fossil table and prepare config file

```
library(DeepDiveR)

fossil_tbl <- read.csv("Fossil_table.csv")

# define time bins (here of 1 myr)
bins <- seq(23, 0, by=-1)

# prep DeepDive-formatted fossil dataset
prep_dd_input(fossil_tbl, bins=bins, output_file="fossil_input.csv", r=100)

# create config file
config <- create_config(wd = "my_working_dir",
                        bins = bins,
                        present_diversity=137,
                        include_present_diversity = TRUE)

config$write("config_file.ini")
```

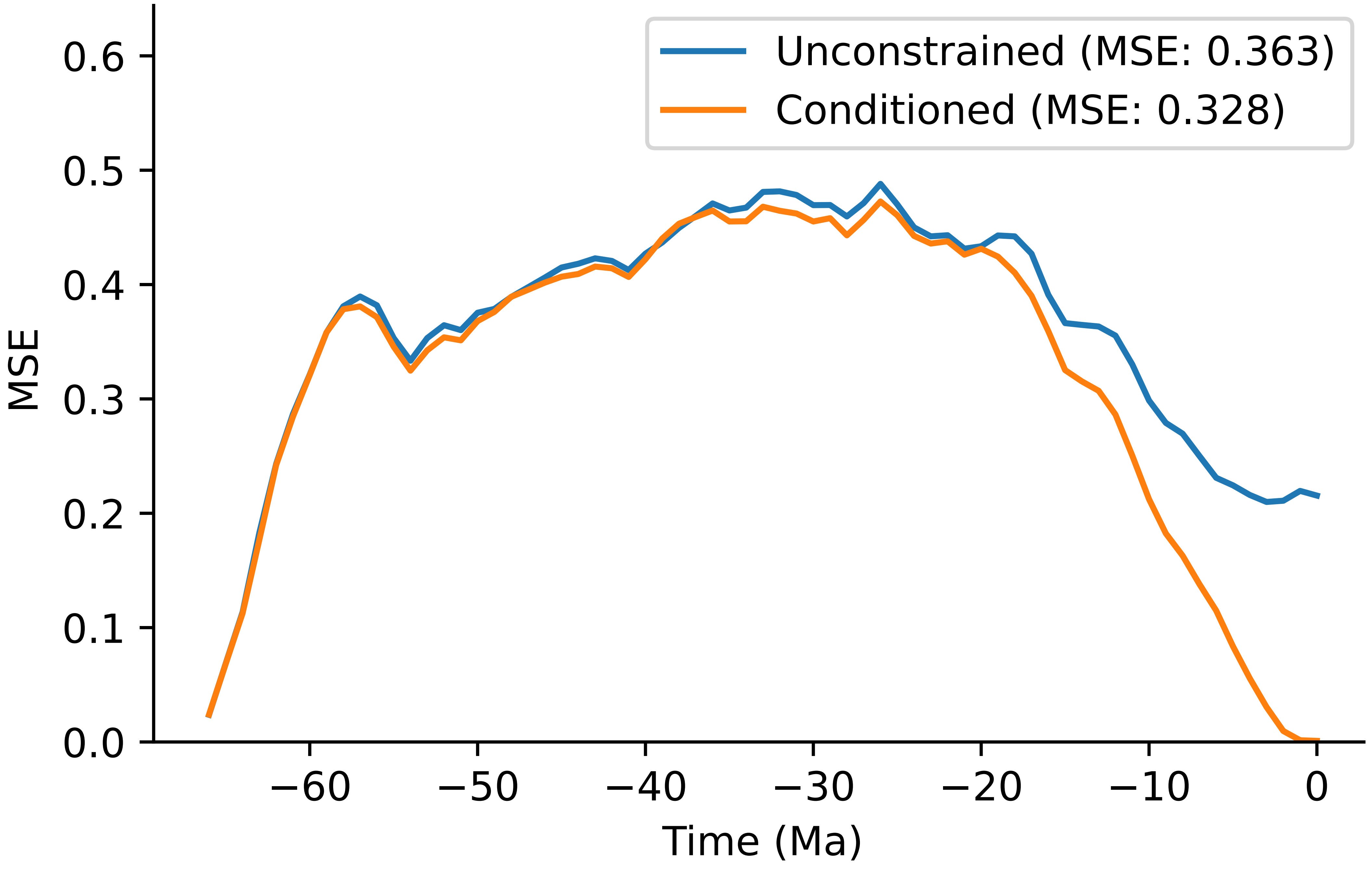
Include present diversity (if applicable)

Conditioning on present diversity

Extant clades



Extinct clades



Setting up an analysis using DeepDiveR

Adding biogeographic information (timing and availability of areas)

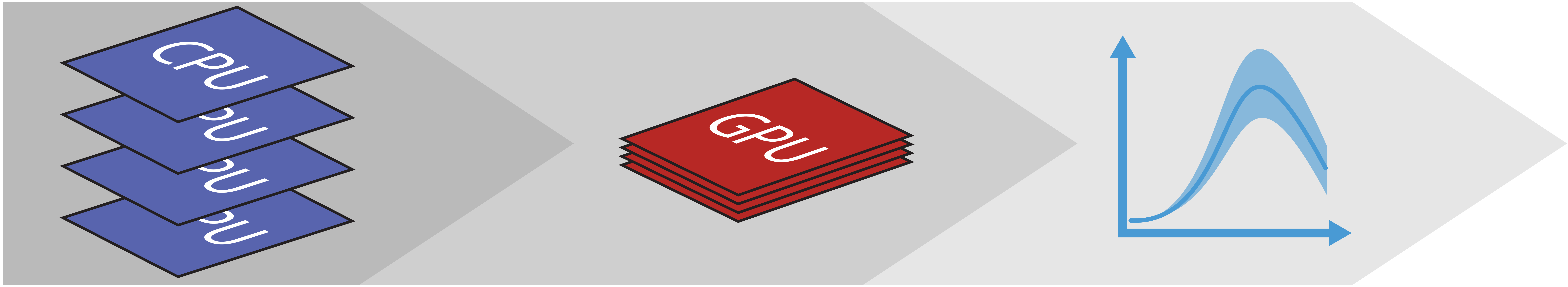
```r
area_ages <- rbind(
     c(0, 0),    # Africa
     c(30, 40), # Antartica
     c(0, 0),    # Asia
     c(0, 0),    # Europe
     c(0, 0),    # N America
     c(0, 0),    # Oceania
     c(0, 0)     # S America
     )

areas_matrix(area_ages, n_areas = n_areas, label = "end", config)
```

e.g. make Antartica inhabitable after ice sheet forms

Running a DeepDive analysis from config

Run in an interactive Python console

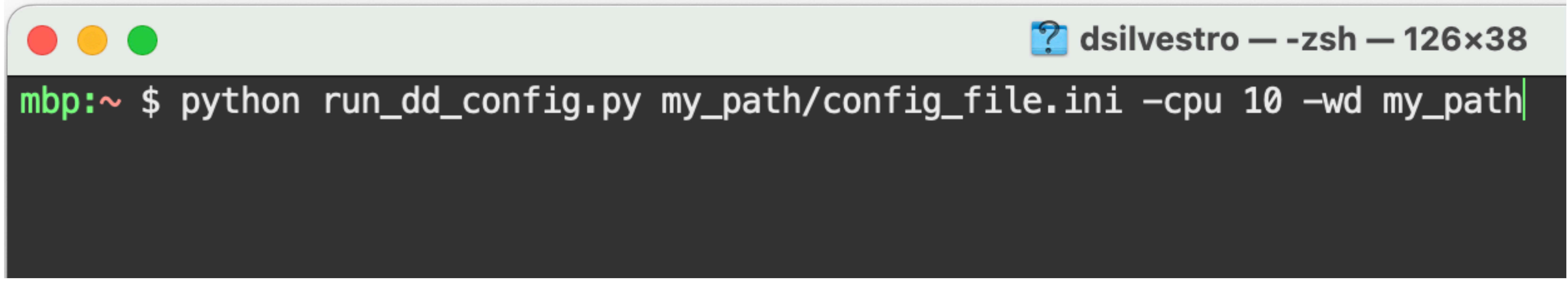
```python
# import the DeepDive python library
from deepdive import config_runner

wd = "my_working_dir"

# launch the analyis (simulations, training, predictions)
config_runner.run_config("config_file.ini", wd=wd, CPU=64)

```

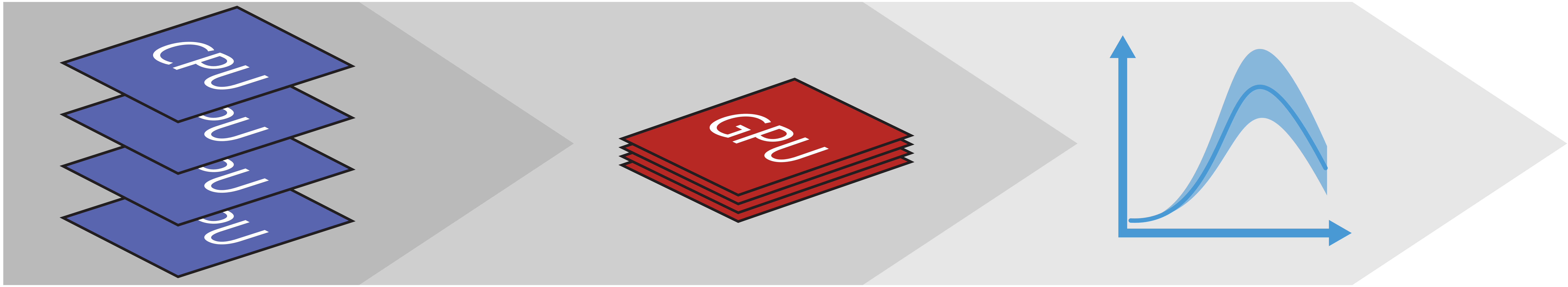


Parallel simulations **Model training** **Predictions**

Running a DeepDive analysis from config

Or execute as a command-line program from a Terminal window





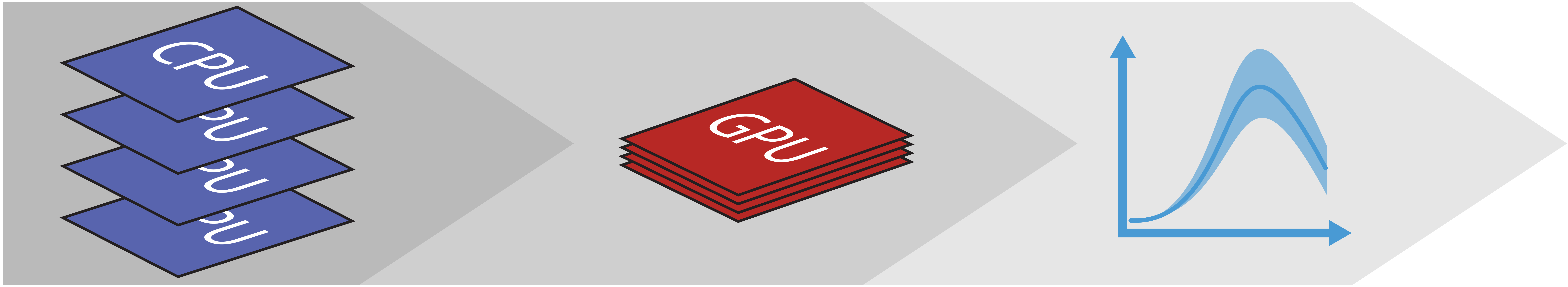
Parallel simulations

Model training

Predictions

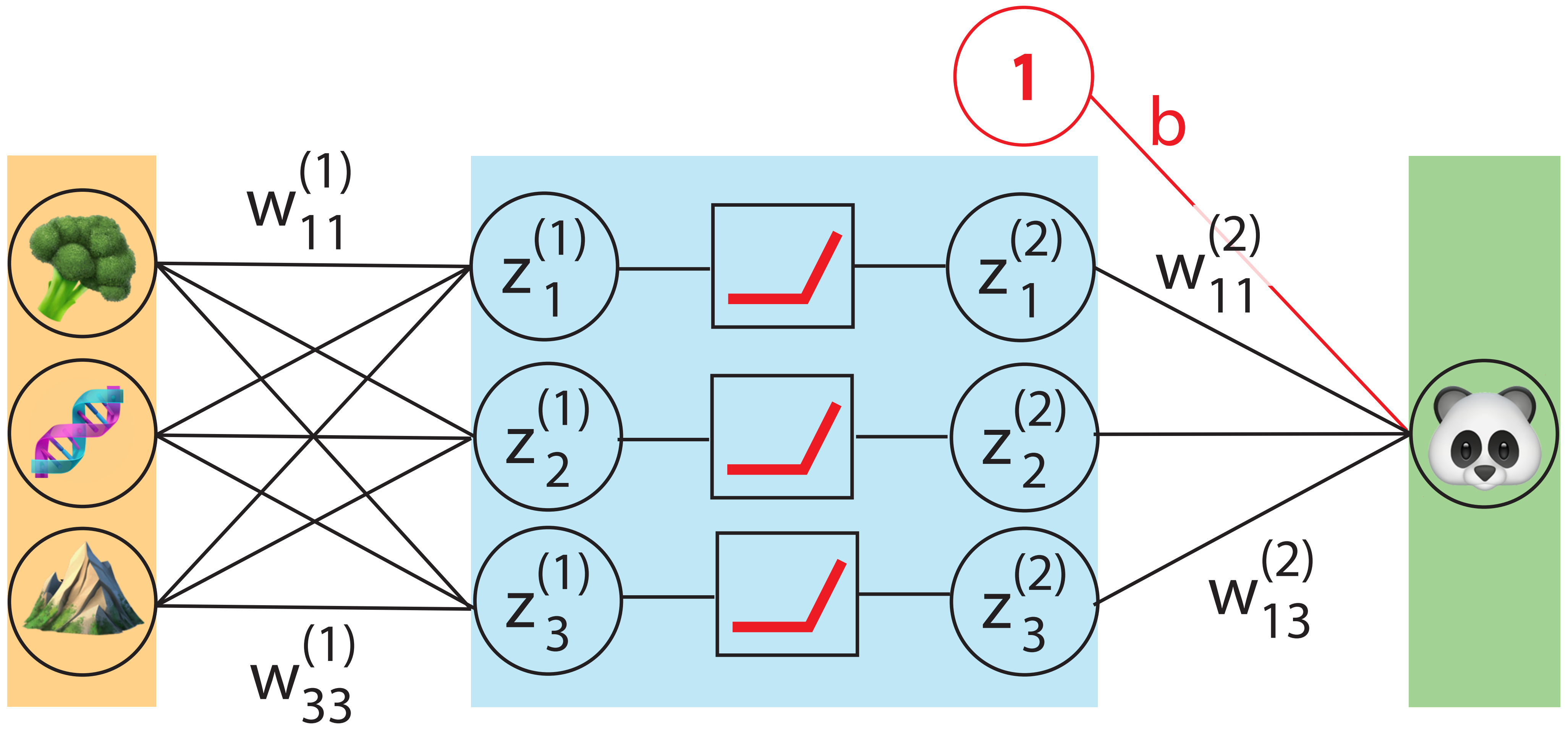
Check out our tutorials on Github

https://github.com/decoding-the-past/decoding_the_past/



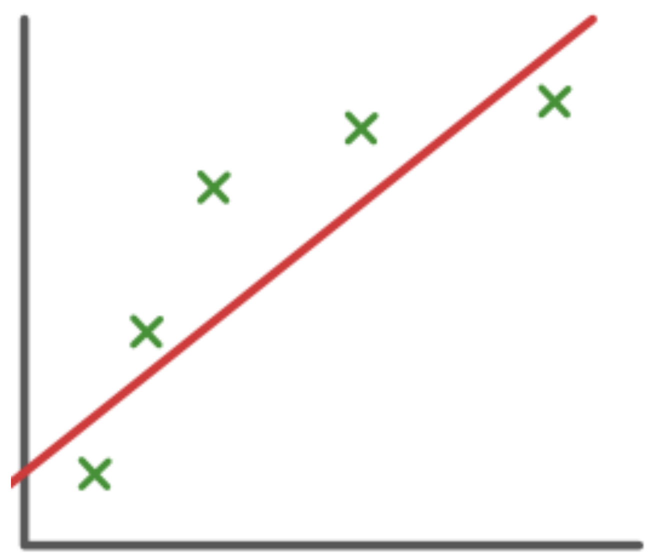
Parallel simulations **Model training** **Predictions**

Optimization of a NN model

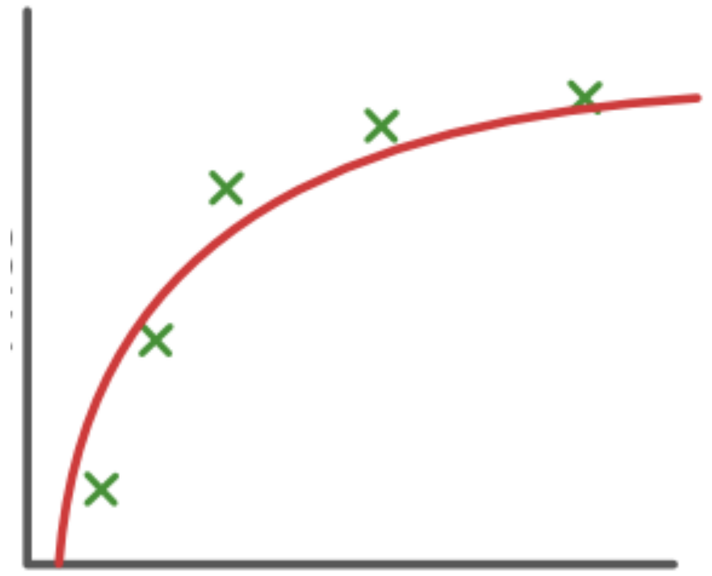


A maximum likelihood optimization of an NN would inevitably lead to over-fitting because NNs are over-parameterized

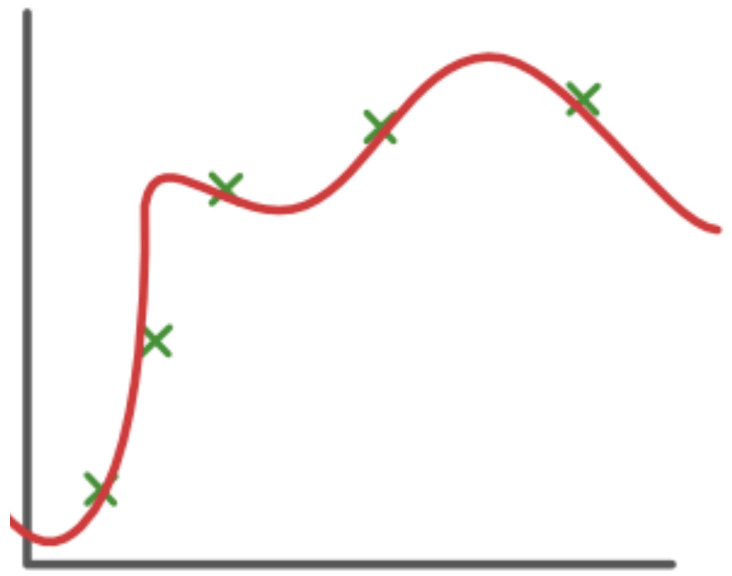
Under-fitting

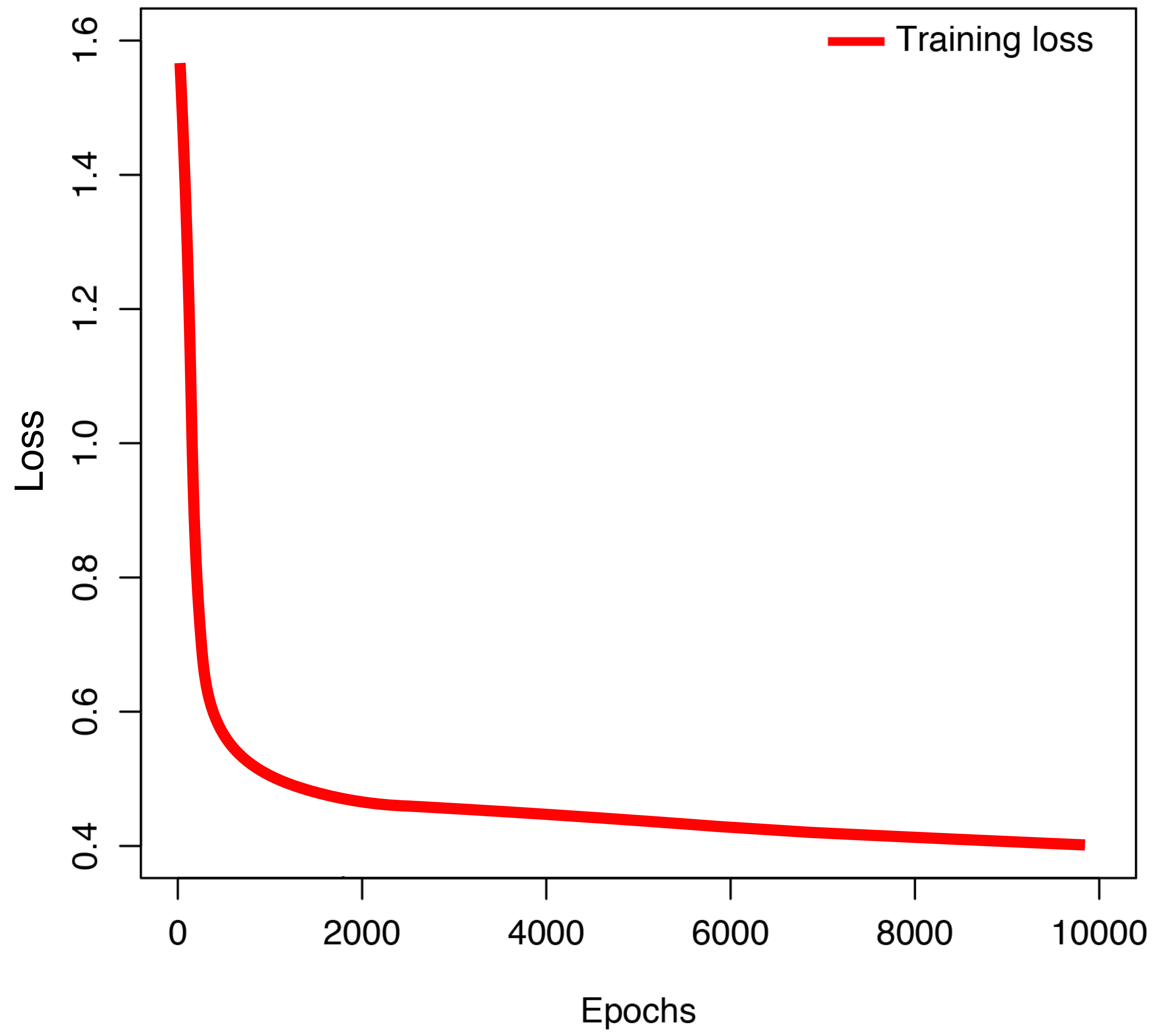


Just right

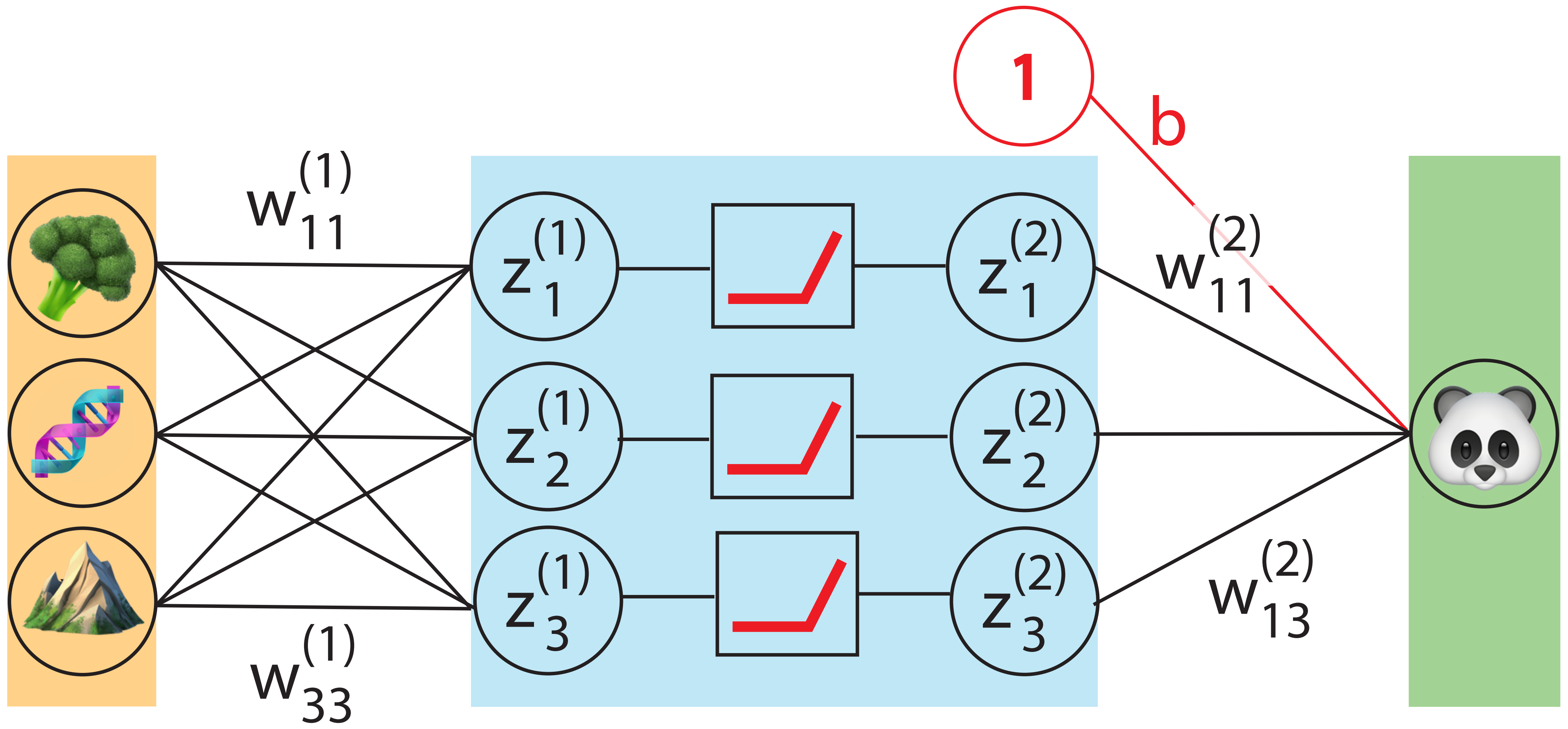


Over-fitting

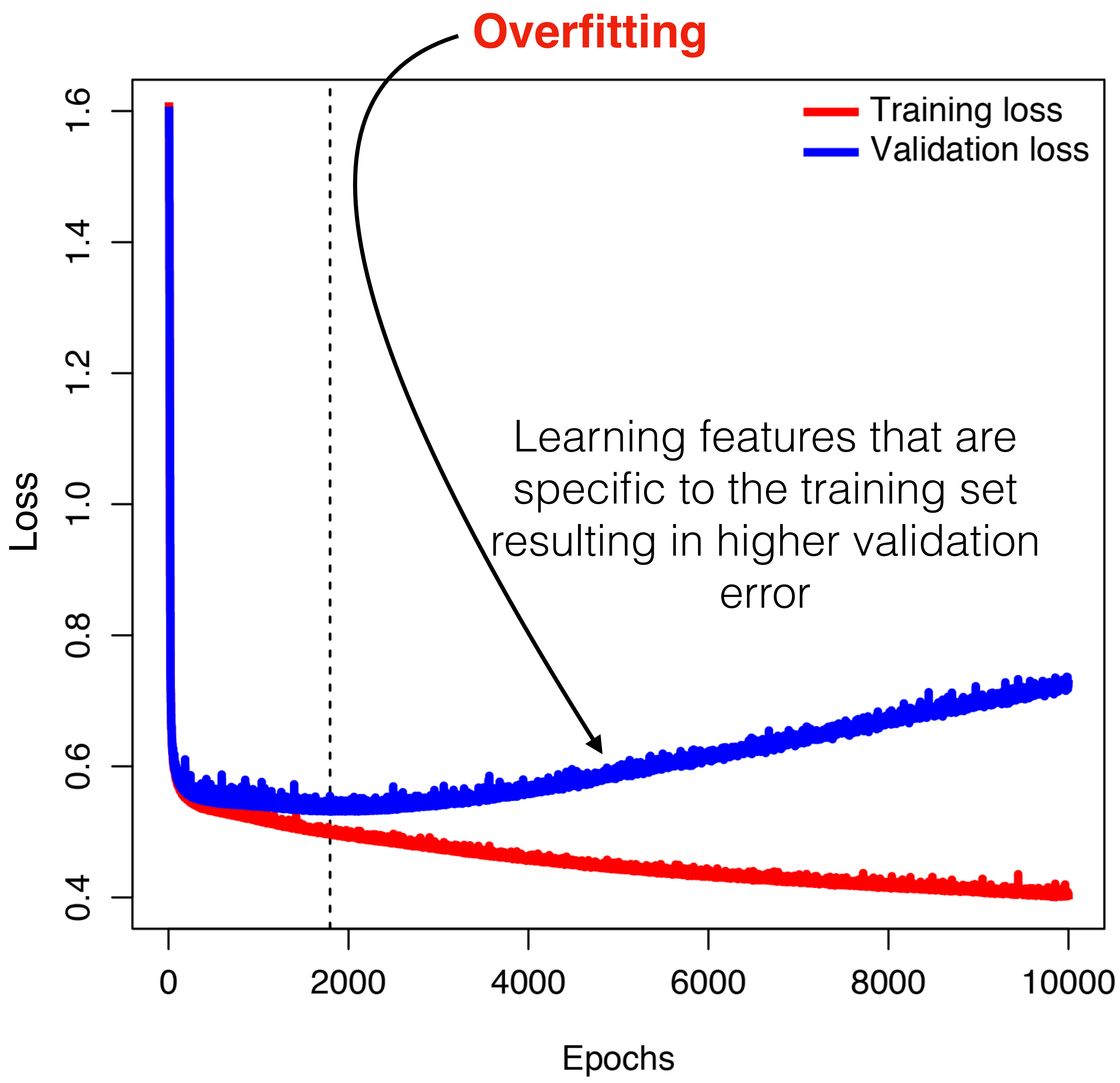




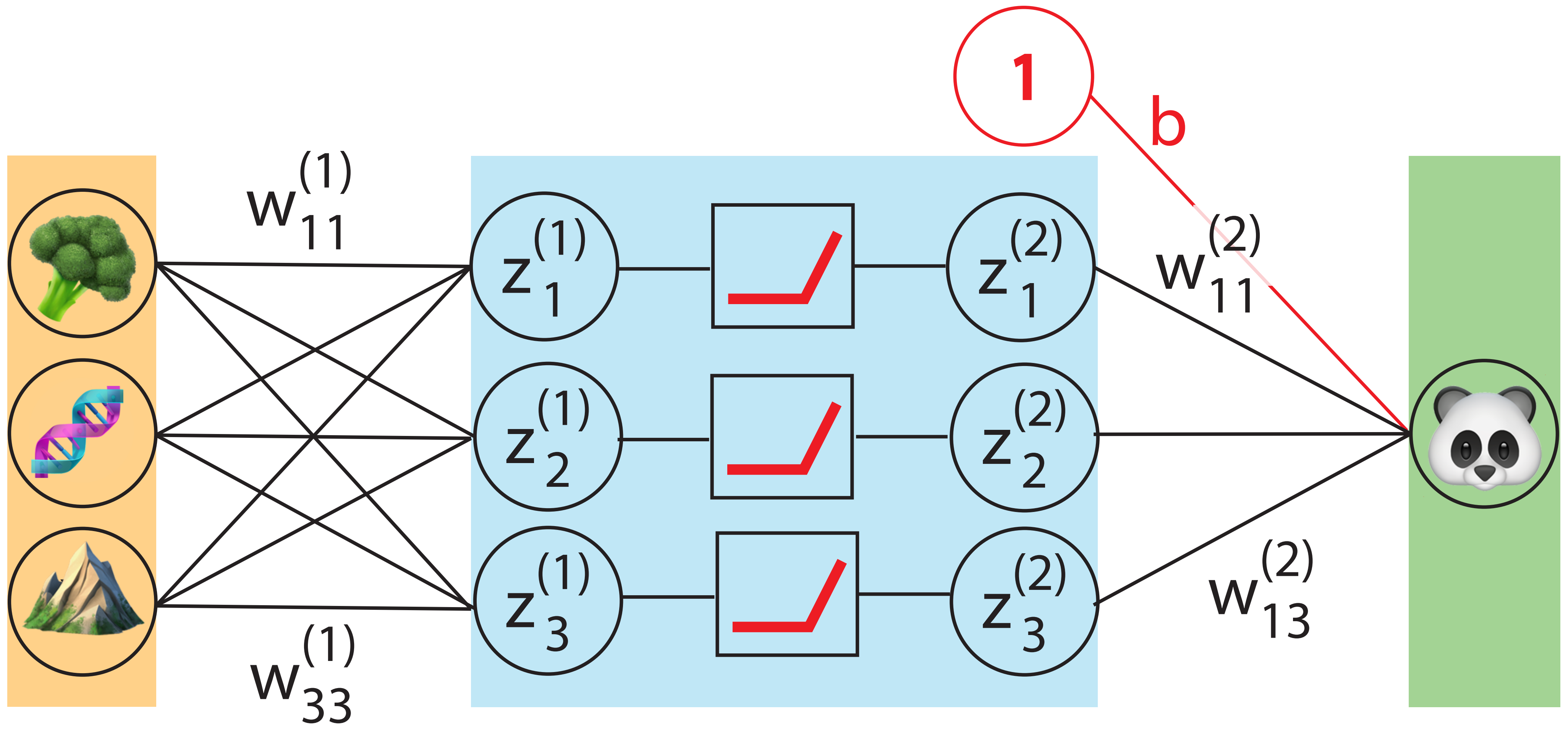
Optimization of a NN model



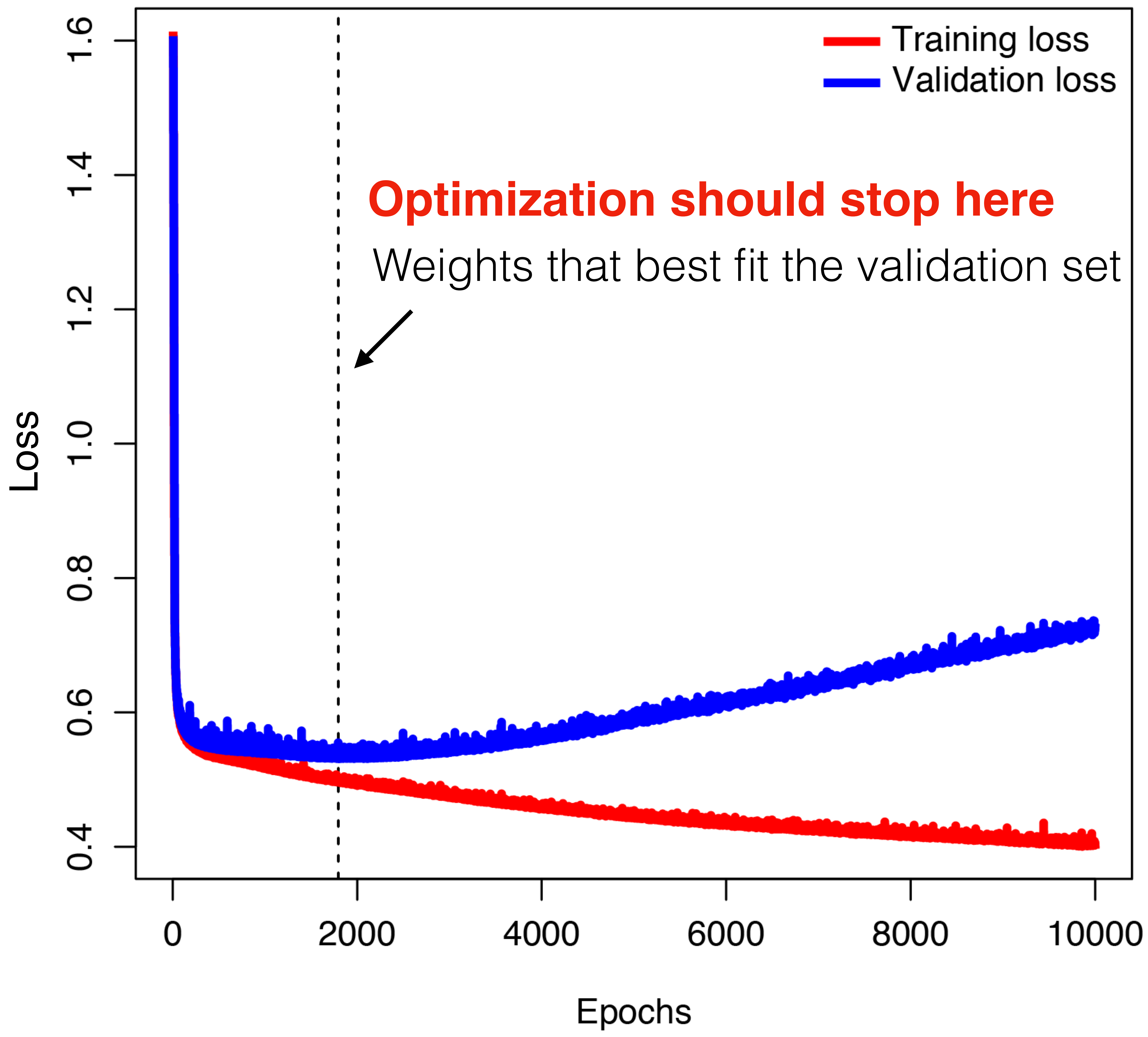
In machine learning a separate validation set is used to stop the optimization process before it starts overfitting



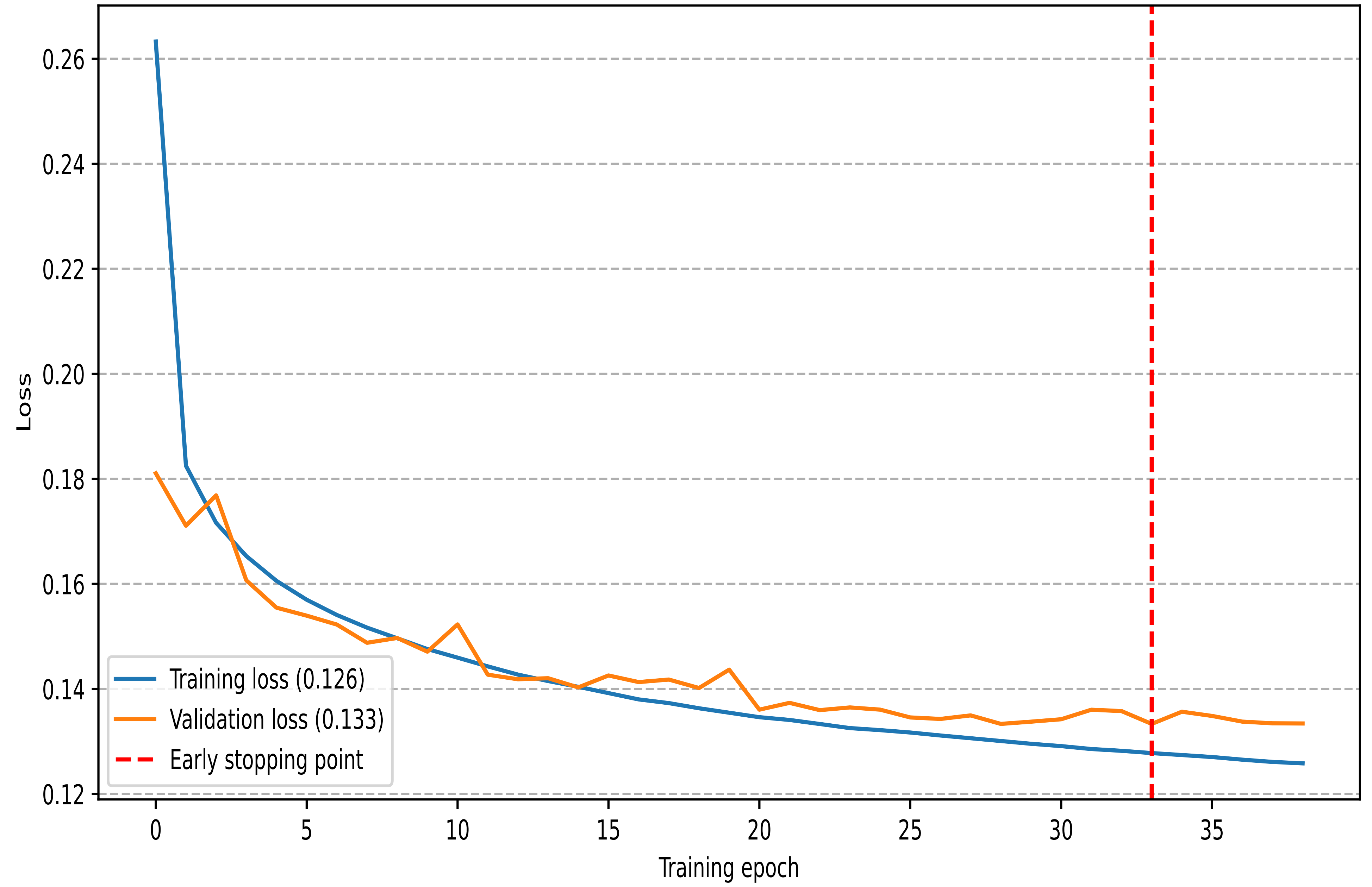
Optimization of a NN model



In machine learning a separate validation set is used to stop the optimization process before it starts overfitting



DeepDive results



Output files:

- A trained model (Keras model)
- Model accuracy (test set)
- Empirical predictions (CSV and PDF)
- Simulated vs empirical features (PDF plots)