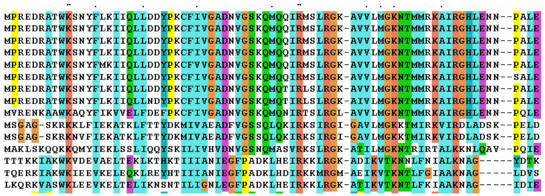
Inferring population sizes of bacterial populations
a deep learning approach

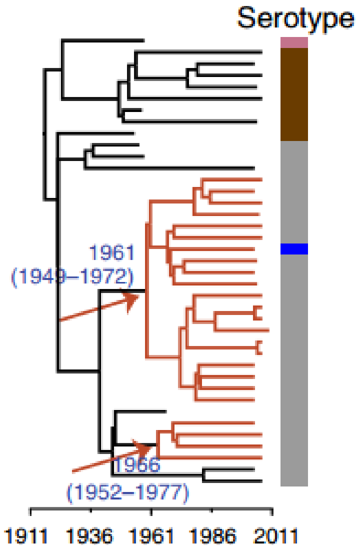
MLMicrobial Genomics - ECML -2022

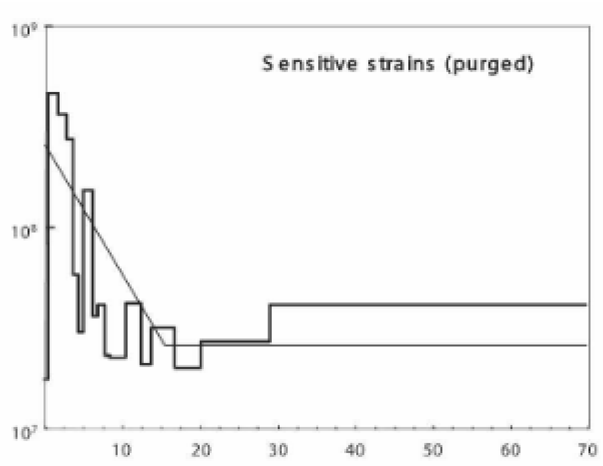
Jean Cury, Théophile Sanchez, Erik Bray, Jazeps Medina-Tretmanis, Maria Avila-Arcos, Emilia Huerta-Sanchez, Guillaume Charpiat, and Flora Jay

Bacterial population genetics



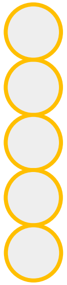
```
MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE
TTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK
TQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
LKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
```



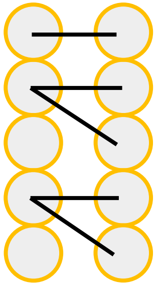


⇒ Focus on population size inference

Intuition



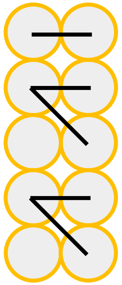
Intuition



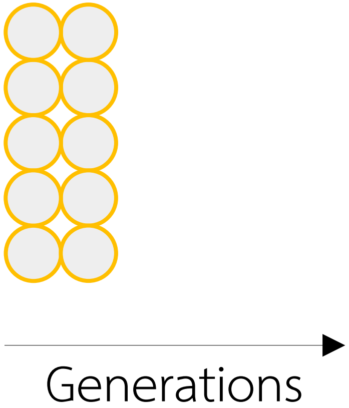
Parental

Daughter cells

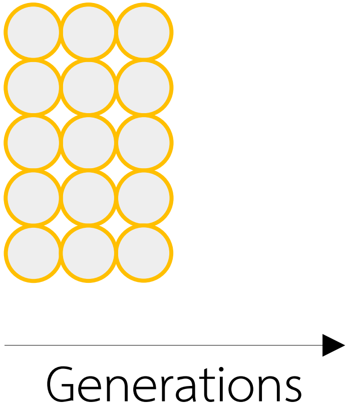
Intuition



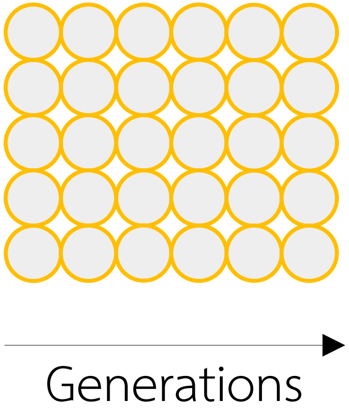
Intuition



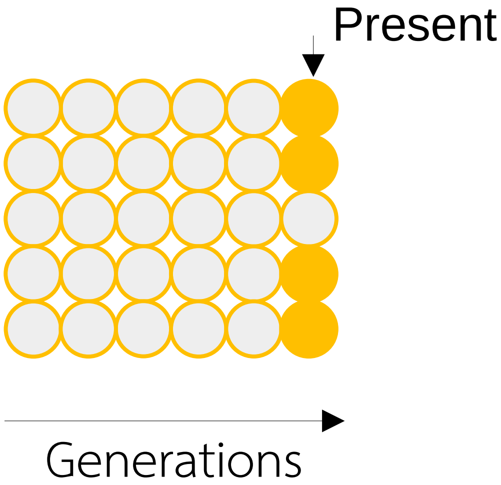
Intuition



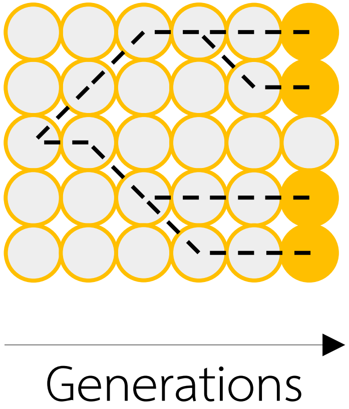
Intuition



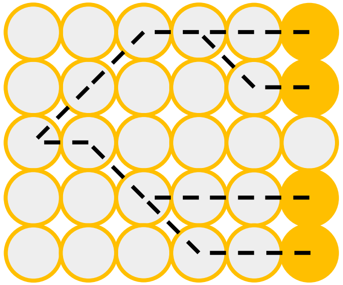
Intuition

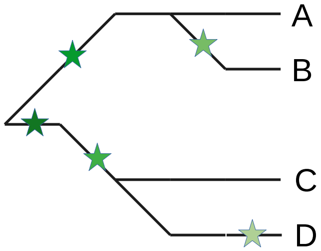


Intuition

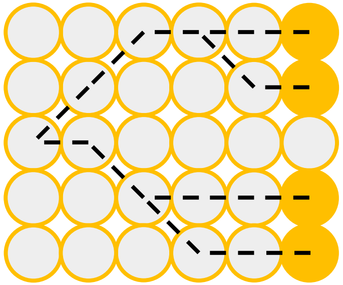


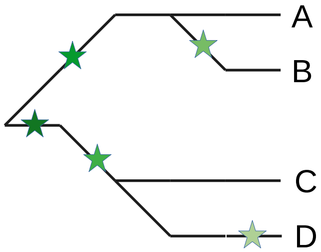
Intuition

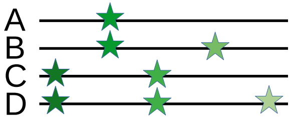




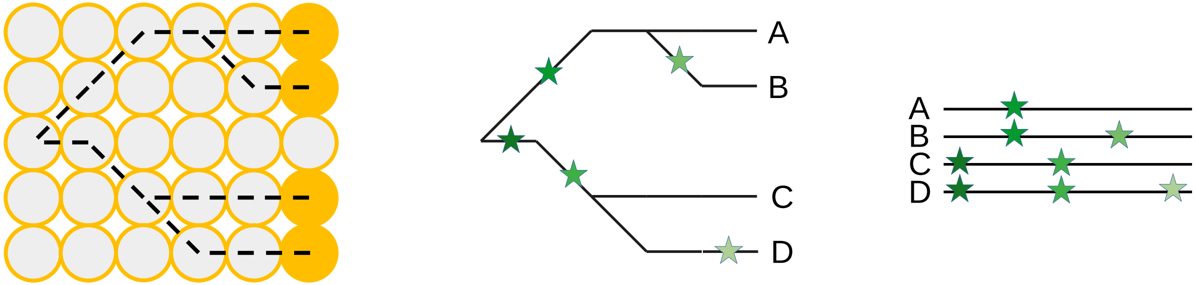
Intuition

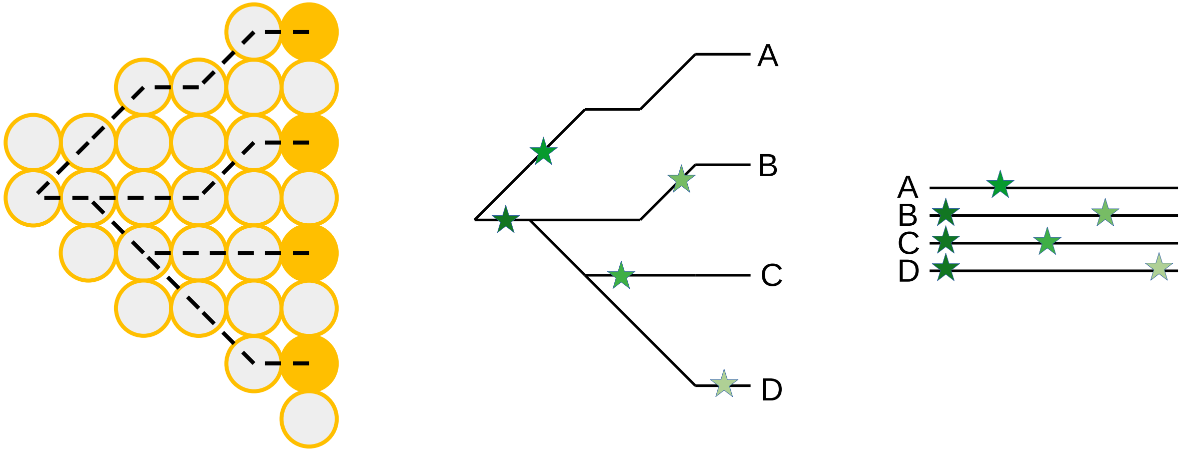




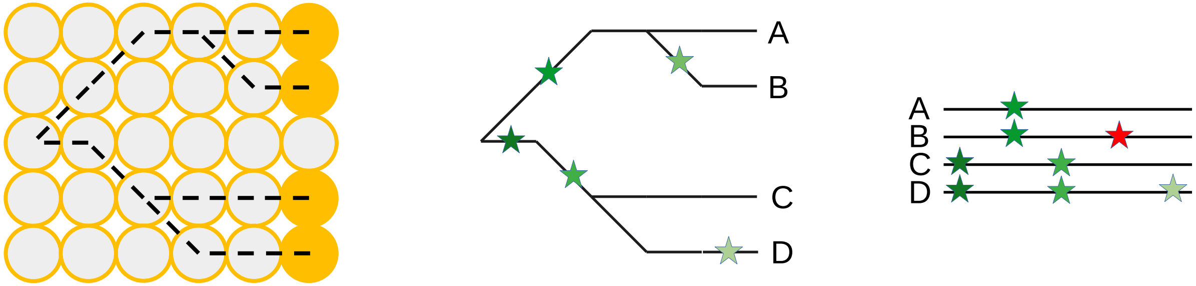


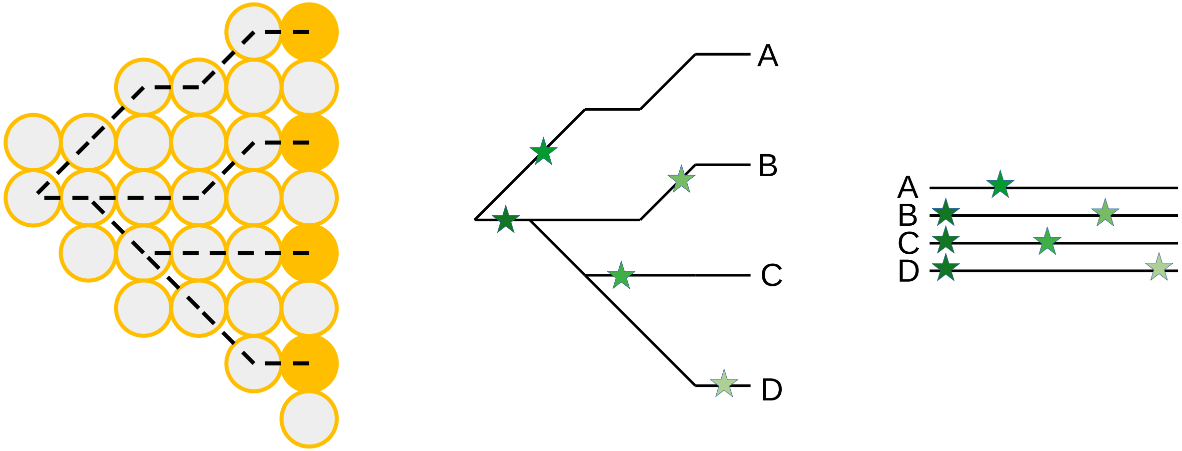
Intuition



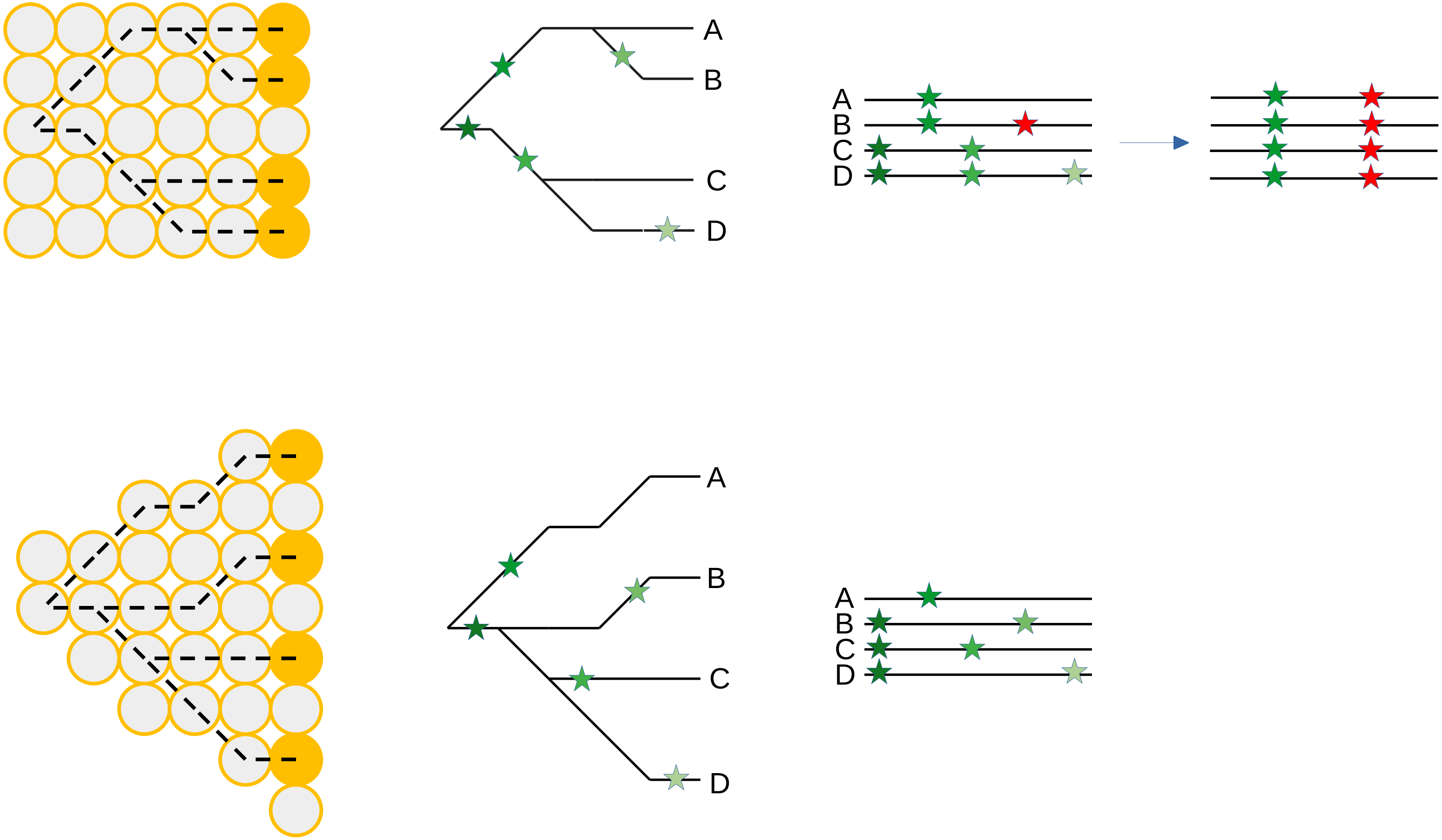


Intuition

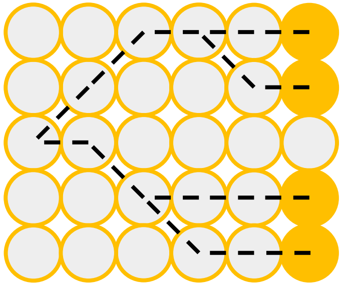


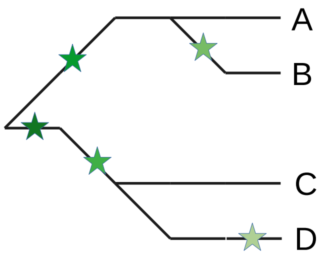


Intuition

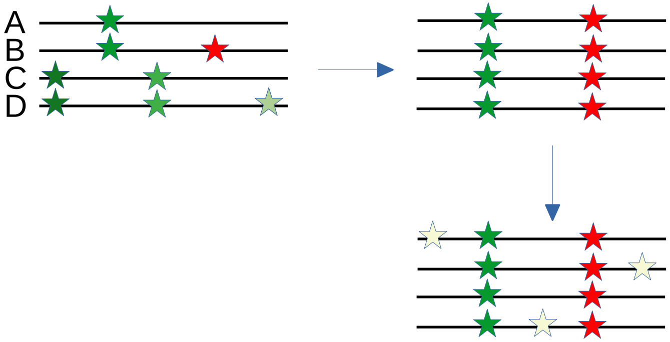


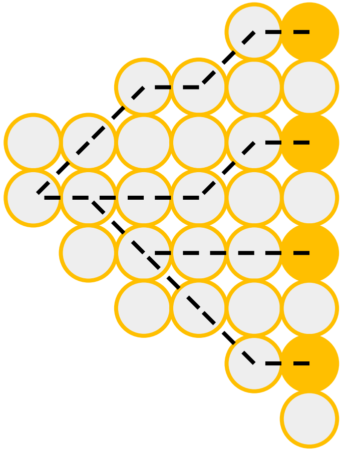
Intuition

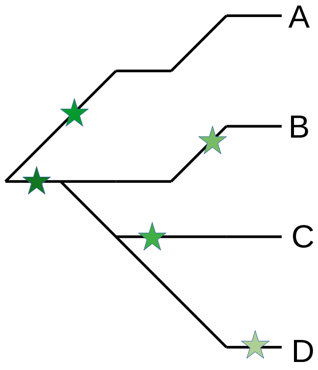


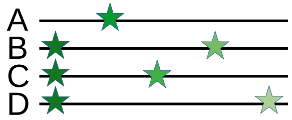


Add selection sign

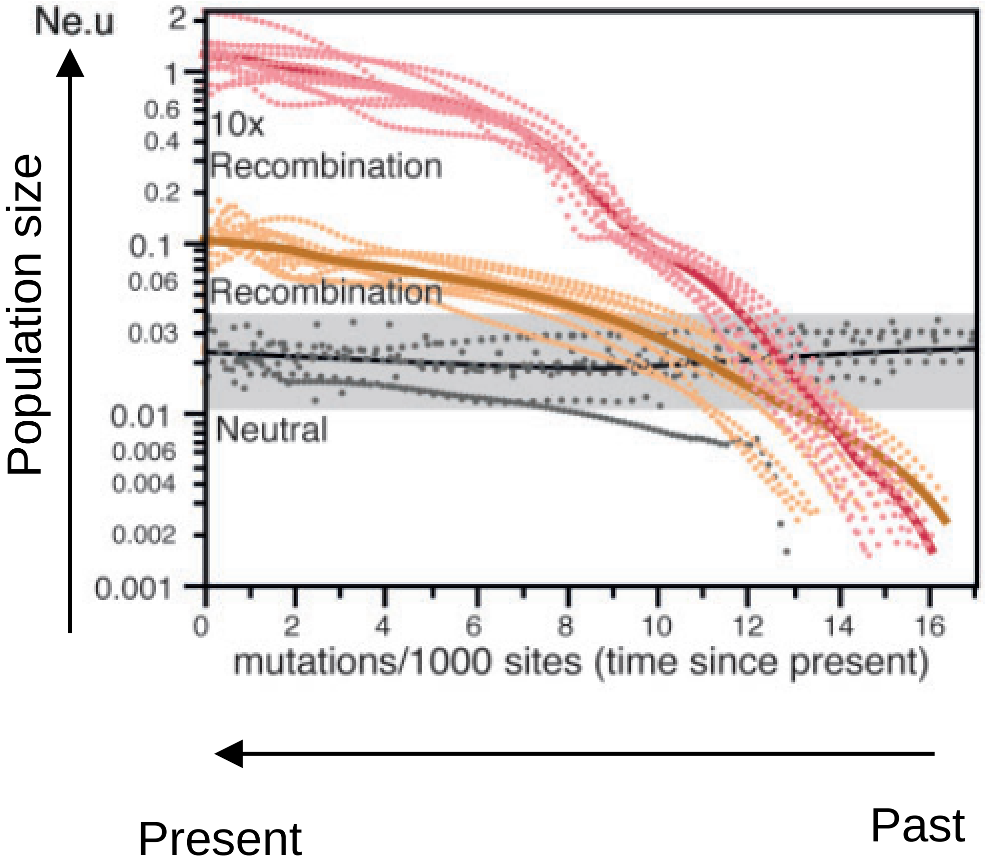








Inference : Skyline Plot

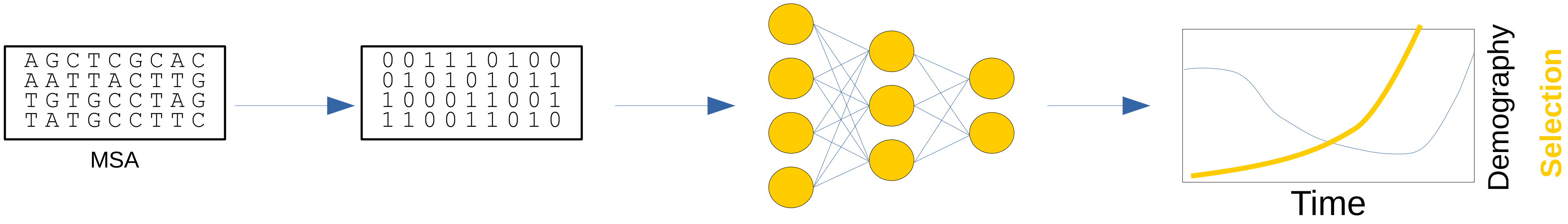


- Using estimated coalescent time, it infers population size
- Non-parametric
- Does not require simulation
- Problem : does not work in bacteria

Lapierre et al. 2016

Project

- End to end deep learning approach for bacterial popgen



```
A G C T C G C A C
A A T T A C T T G
T G T G C C T A G
T A T G C C T T C
```

```
0 0 1 1 1 0 1 0 0
0 1 0 1 0 1 0 1 1
1 0 0 0 1 1 0 0 1
1 1 0 0 1 1 0 1 0
```

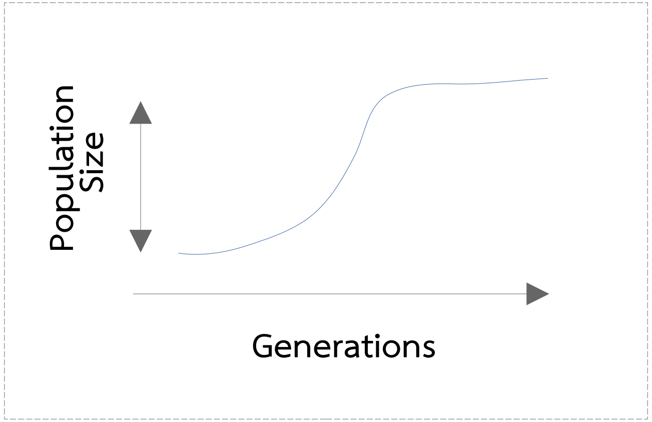
- Problem: No ground truth data
- We need a population genetic simulator that is :
  - fast
  - Implement bacterial recombination (homologous HGT)
  - Demography, selection, etc..

Input Data

MSA

ATGCGACAG
CTGCG**T**C**G**G
ATG**A**G**T**CAG
CTGCG**T**CAG
123456789

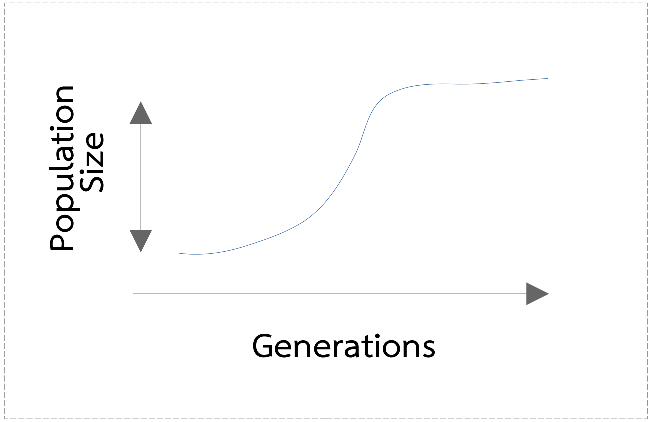
Observation



Input Data

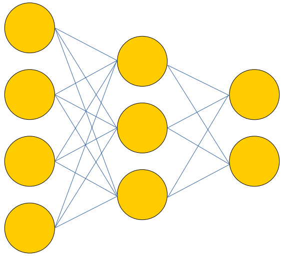
MSA		SNP matrix
ATGCGACAG	000000000	0000
CTGCG**T**C**G**G	**1**0000**1**0**1**0	**1**0**11**
ATG**A**G**T**CAG	000**1**0**1**000	0**11**0
CTGCG**T**CAG	**1**0000**1**000	**1**0**1**0
123456789	123456789	1468

Observation

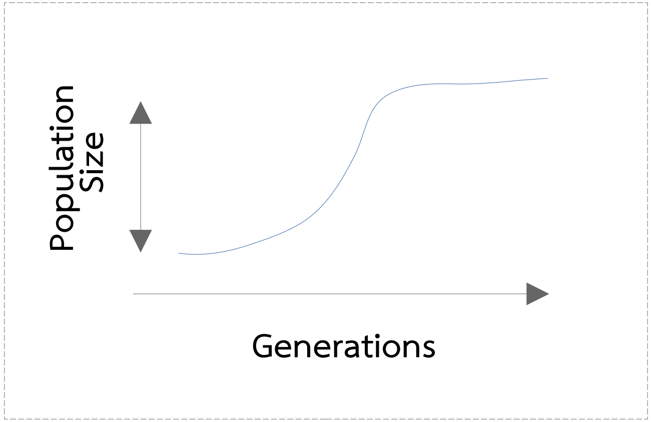


Input Data

MSA		SNP matrix
ATGCGACAG	000000000	0000
CTGCG**T**C**G**G	**1**0000**1**0**1**0	**1**0**11**
ATG**A**G**T**CAG	000**1**0**1**000	0**11**0
CTGCG**T**CAG	**1**0000**1**000	**1**0**1**0
123456789	123456789	1468

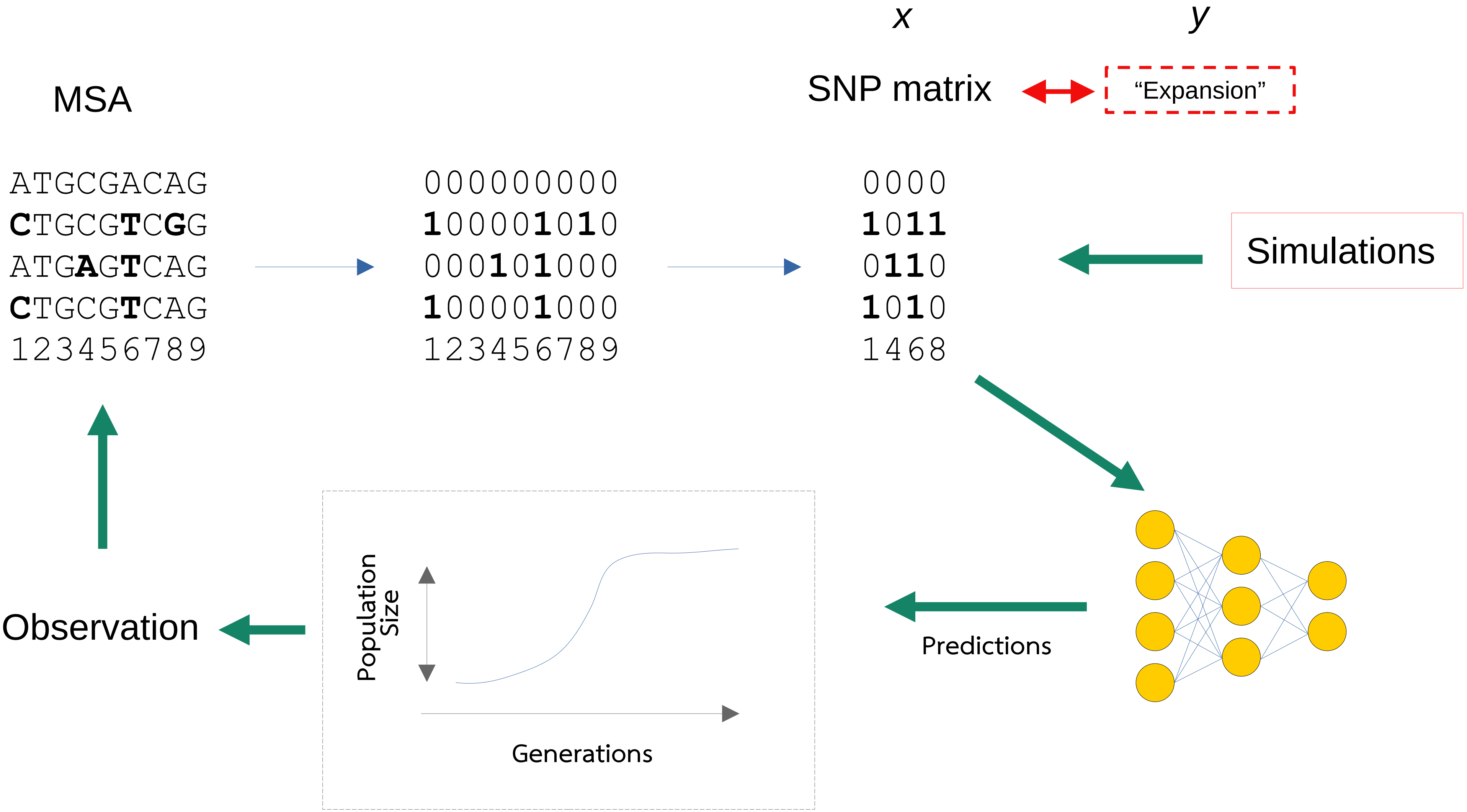


Predictions



Observation

Input Data



x

y

MSA

SNP matrix

“Expansion”

ATGCGACAG
CTGCG**T**C**G**G
ATG**A**G**T**CAG
CTGCG**T**CAG
123456789

000000000
10000**1**0**1**0
000**1**0**1**000
10000**1**000
123456789

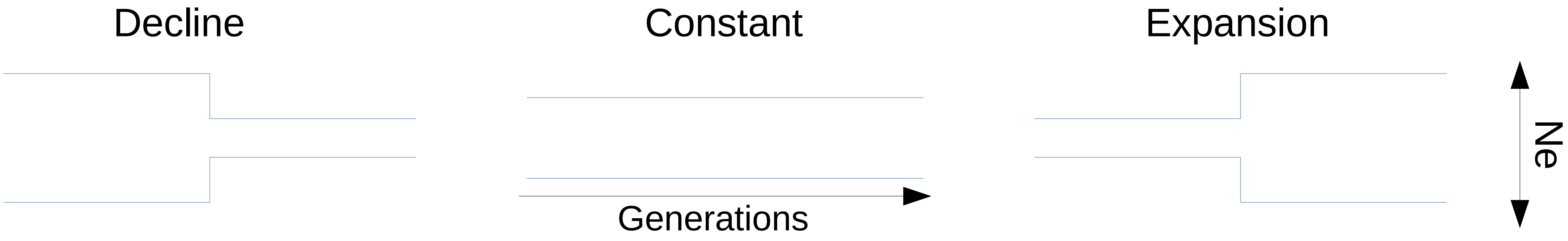
0000
10**11**
0**11**0
10**1**0
1468

Simulations

Observation

Predictions

What we simulate



5000x100 simulations

1000x100 simulations

5000x100 simulations

→ 50% with and 50% without selection

→ Variable parameters:

- **initial population size** (~Ne)
- mutation rate
- recombination rate (ratio r/m)
- coefficient of selection
- time of selection
- **time of demographical change**
- **strength of bottleneck/expansion**

→ Generated with a generalized Halton sequence

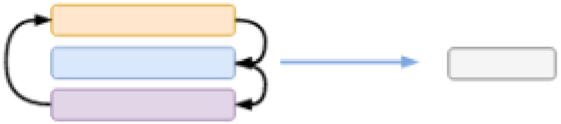
→ Fixed parameters:

- chromosome size
- mean size of gene conversion tracts
- Number of generations

Using SLiM, adapted for
Bacterial population
(Cury et al., 2021)

Approach

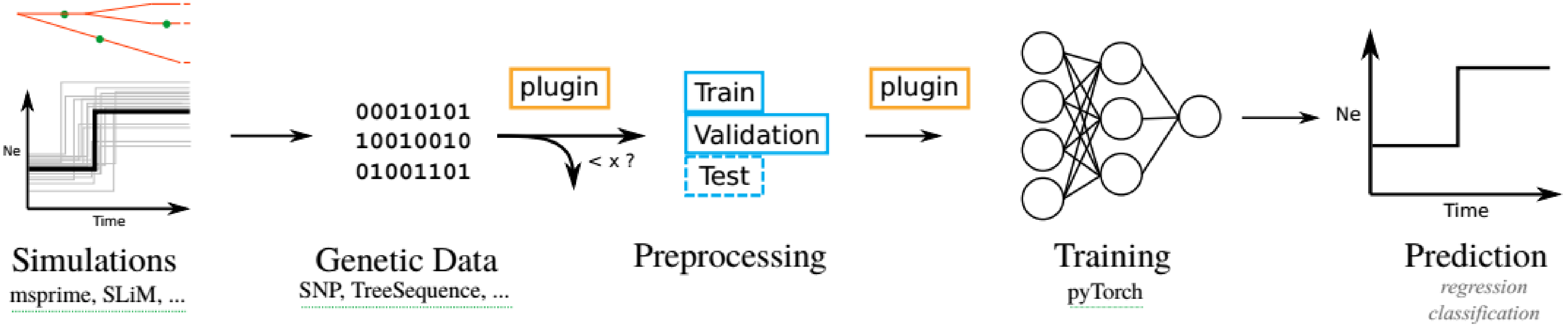
- Use of `dnadna`, a package that help to reproduce, share and develop DL methods for population genetics
- Use of SPIDNA architecture
  - Invariant to permutation of individuals
  - Adaptive to input dimension
  - Good performance on human populations



- Add uncertainty estimation

Sanchez et al. 2022
Sanchez et al. 2020

dnadna : Package for DL in population genetics



Simulations
msprime, SLiM, ...

Genetic Data
SNP, TreeSequence, ...

plugin

Preprocessing
Train
Validation
Test

plugin

Training
pyTorch

Prediction
regression
classification

`dnadna simulation`

`dnadna preprocess`

`dnadna train`

`dnadna predict`

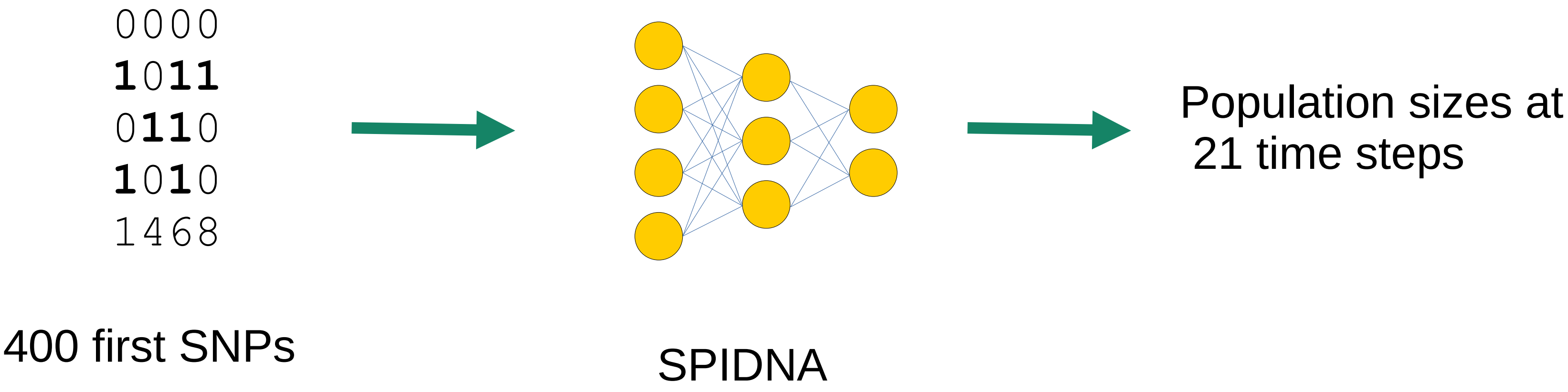
Package that allow:
- Development of network
- Reuse of someone else’s network
- Reproduce training/prediction

→ Without coding skill (YAML)

https://gitlab.com/mlgenetics/dnadna

Sanchez et al. 2022

Inference of demography



0000
10**11**
0**11**0
10**1**0
1468

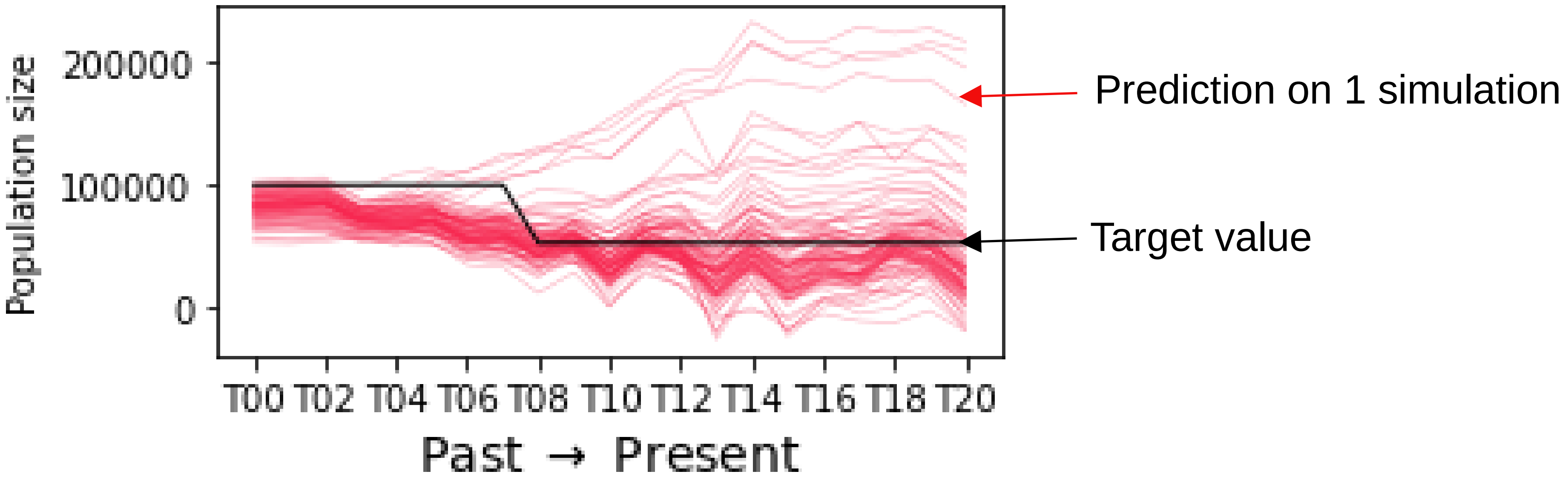
400 first SNPs

SPIDNA

Population sizes at
21 time steps

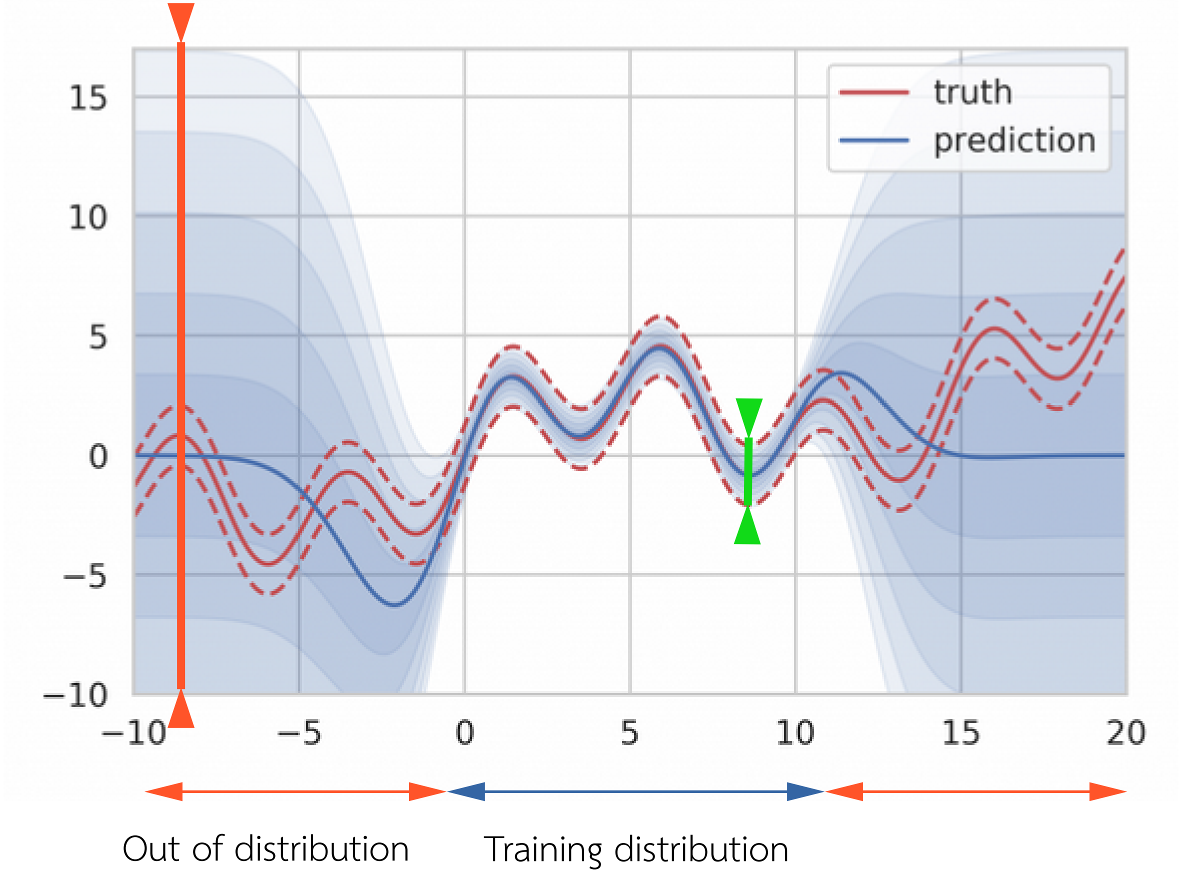
Inference of demography

Example with 100 simulations with the same set of parameters



What about uncertainty ?

- DNN output a single value without notion of uncertainty :
  - Aleatoric : due to the underlying process that is intrinsically stochastic
  - Epistemic : Your sample is out of the distribution of the simulations

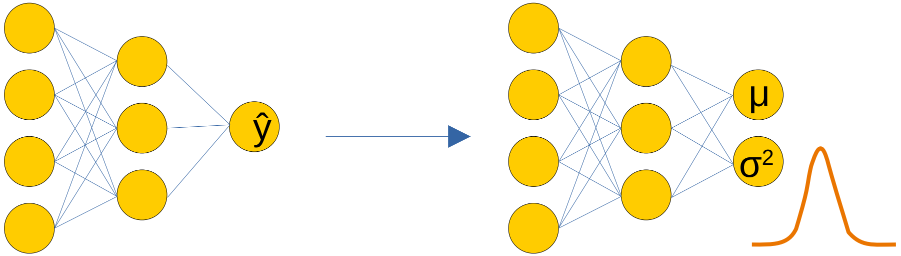


https://www.inovex.de/de/blog/uncertainty-quantification-deep-learning/

What about uncertainty ?

- DNN output a single value without notion of uncertainty :
  - Aleatoric : due to the underlying process that is intrinsically stochastic
  - Epistemic : Your sample is out of the distribution of the simulations

Use of Gaussian Negative Log Likelihood Loss to learn a gaussian with parameters μ and σ^2

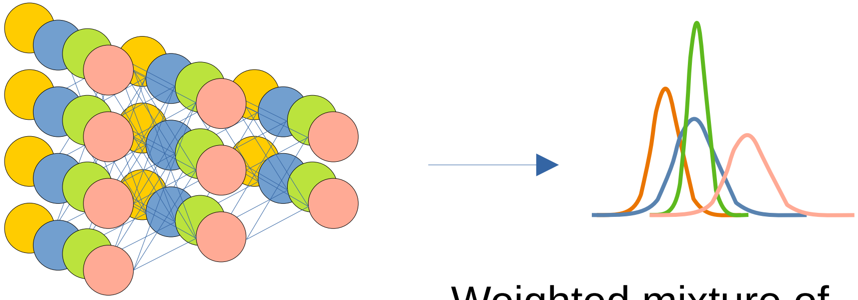


Nix and Weigend, 1994
Lakshminarayanan et al, 2017

What about uncertainty ?

- DNN output a single value without notion of uncertainty :
  - Aleatoric : due to the underlying process that is intrinsically stochastic
  - Epistemic : Your sample is out of the distribution of the simulations

Ensemble of Networks

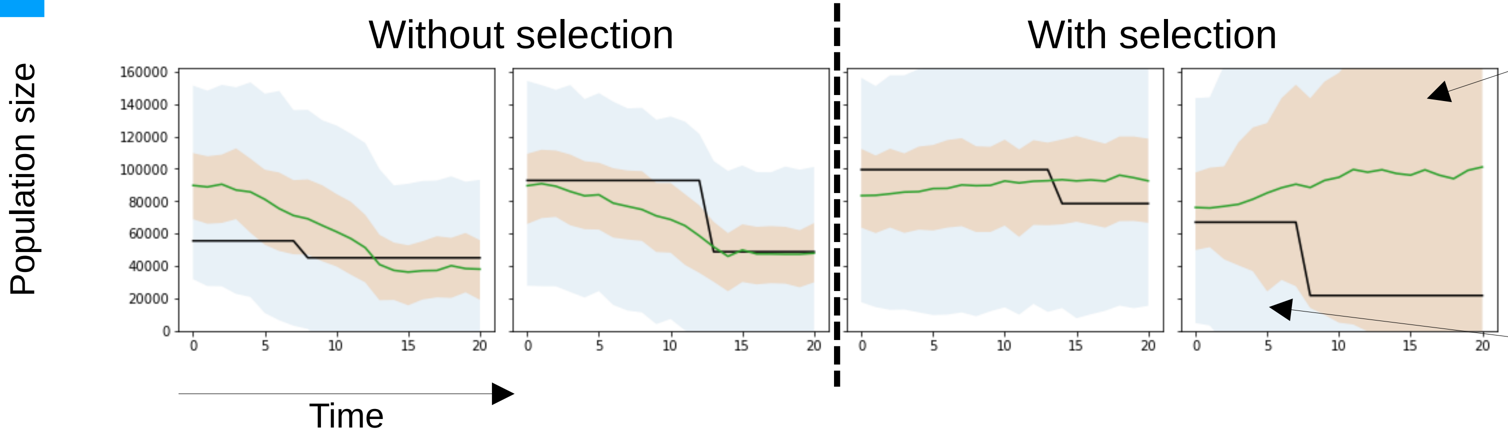


Weighted mixture of Gaussian distribution

weights ∝ $1/\sigma^2$

Lakshminarayanan et al, 2017

Uncertainty estimation



Without selection

With selection

50% High density Interval

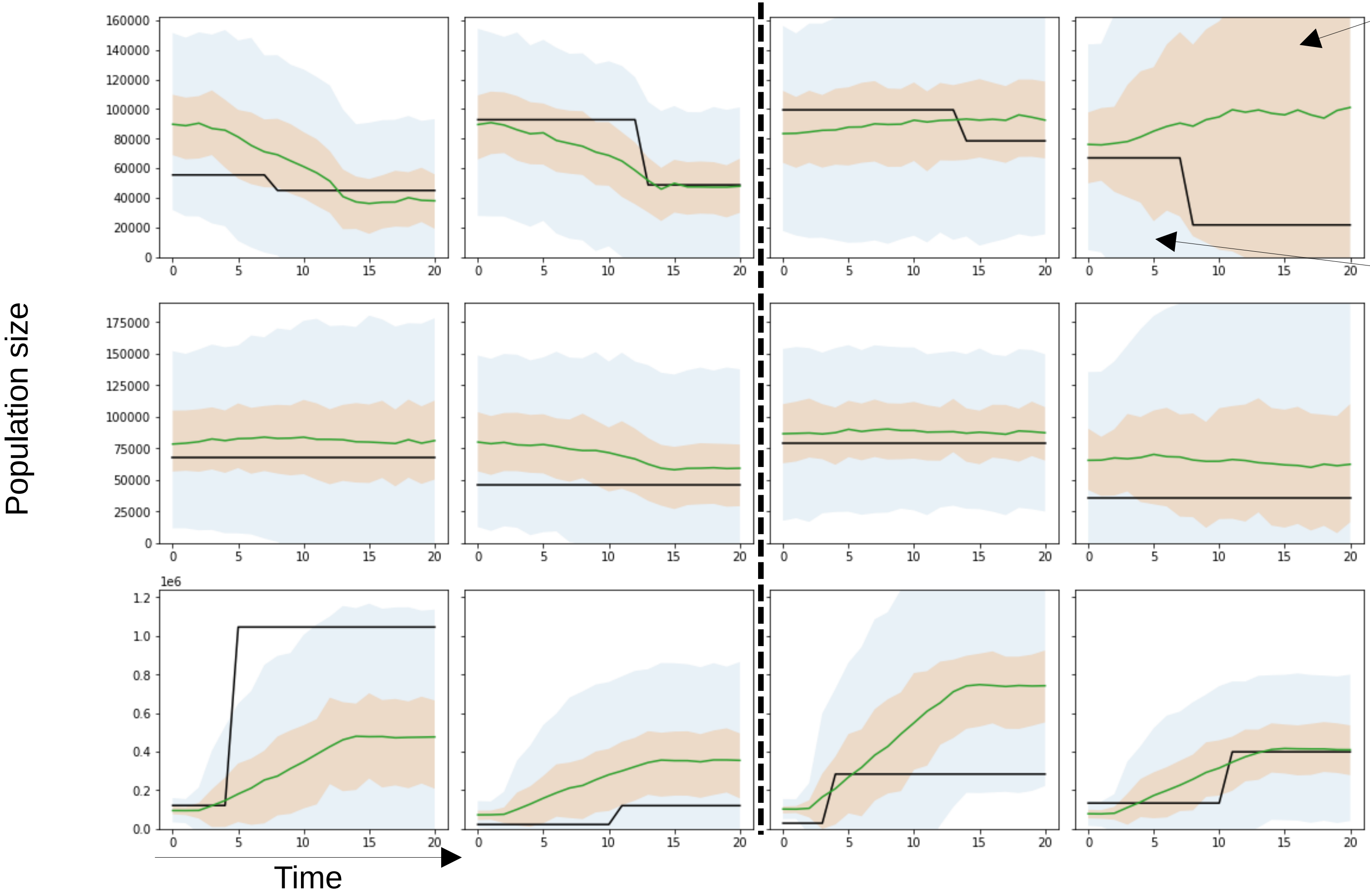
Decline

90% High density Interval

Uncertainty estimation

Without selection

With selection



50% High density Interval

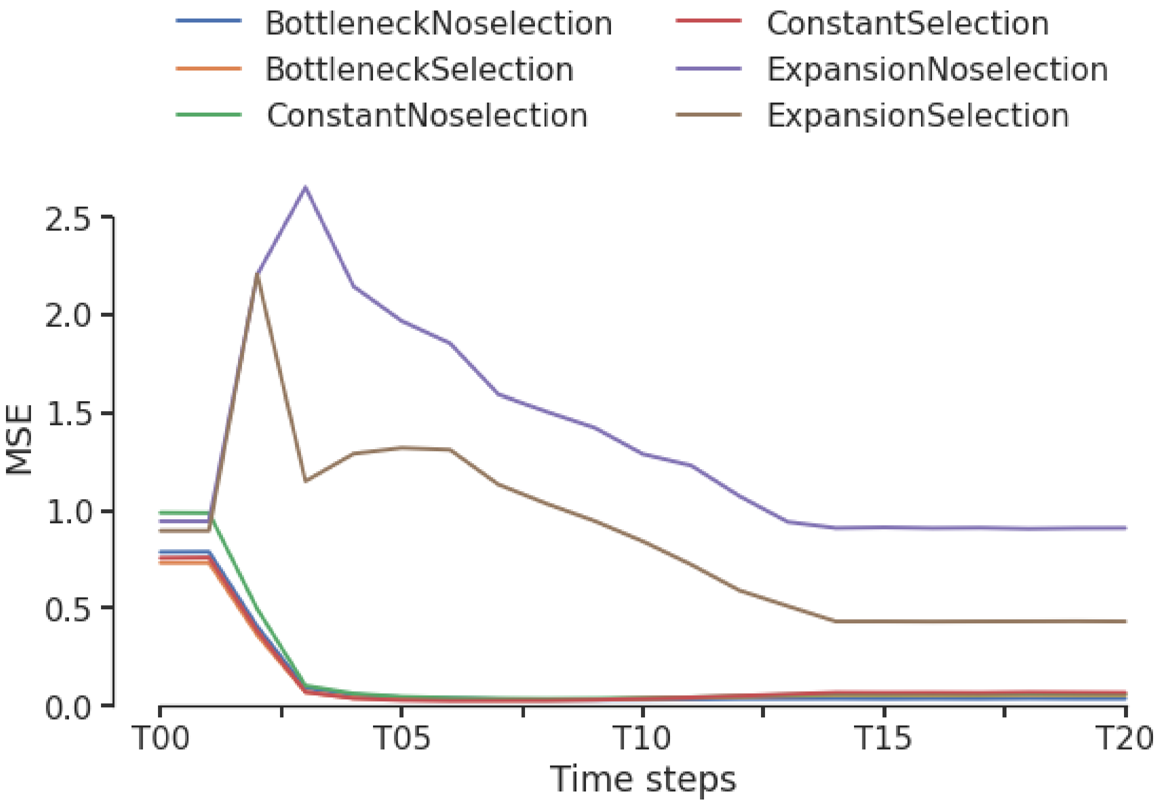
90% High density Interval

Decline

Constant

Expansion

Error on Test set (the lower the better)

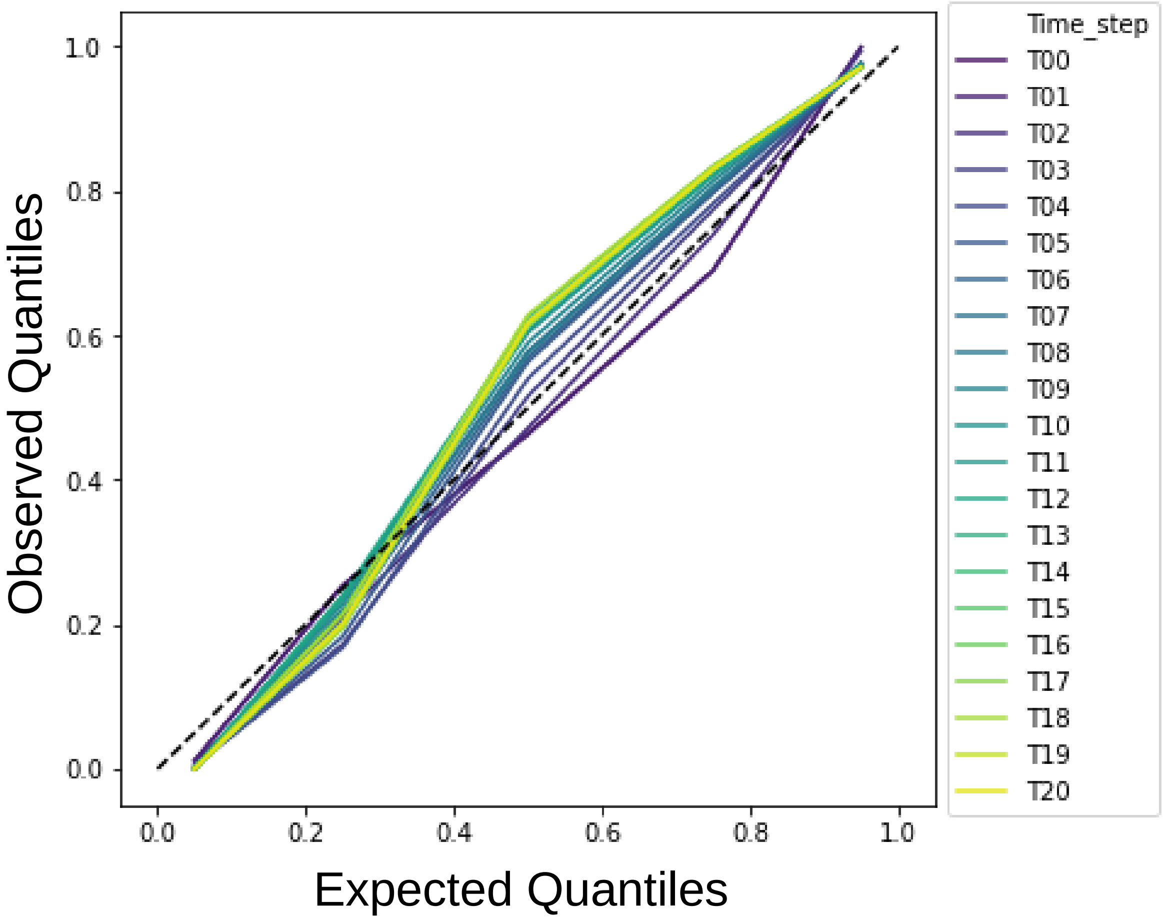


Bad predictions for Expansion

Good prediction otherwise

Except for ancient times where predictions follows the prior of the training set.

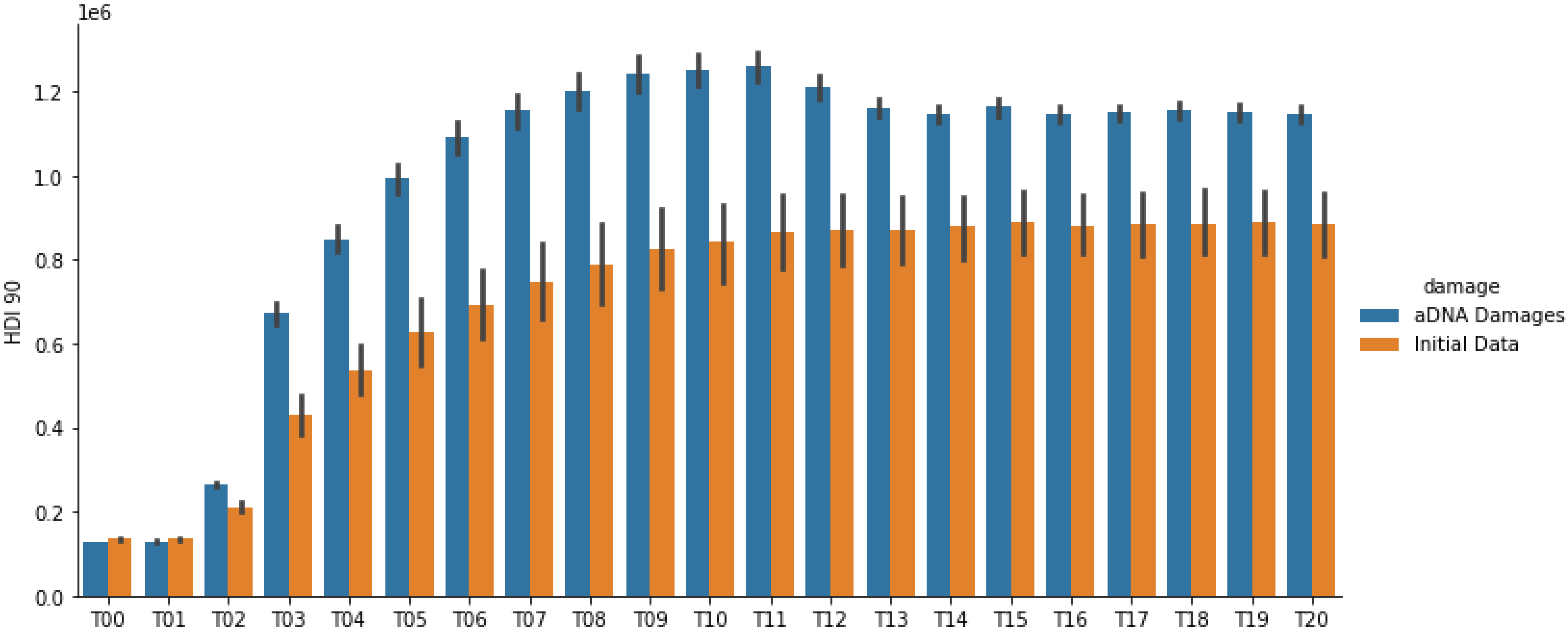
Calibration of the Gaussian mixture



ancient DNA

- Increasing amount of aDNA sequenced as technology improves
- Can help palaeontologist / historian understand distant past
- Problem : low quality of sequences
    - Due to degradation of DNA
    - Higher rate of sequencing error
    - Poor coverage (small amount of DNA)

Uncertainty on ancient DNA



Summary

- Prediction of bacterial population size through time
    - Irrespective of the underlying selection regime and other parameters
- Using **`dnadna`** package → easy to reuse / reproduce
- Estimation of the aleatoric and epistemic uncertainties

- Transfer learning with aDNA
- Assess interest of transfer learning from other net trained on similar task
- Improve training procedure with SPIDNA (something else than 400 SNP)
- Test on real data

Thanks

- Flora Jay
- Theophile Sanchez
- Guillaume Charpiat
- Erik M. Bray
- Ben Haller

- Jazeps Medina-Tretmanis
- Maria Avila-Arcos
- Emilia Huerta-Sanchez
- Mathieu Michel

LISN

DIM 1health
Un monde, une seule santé

îledeFrance

HFSP

Inria