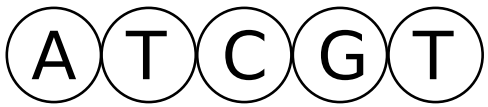
Simulation of ancient DNA sequences using transformer-based techniques.

Théo Boury [1,2], Jazeps Medina-Tretmanis [3], Maria Avila-Arcos [4], Emilia Huerta-Sanchez [3], Burak Yelmen [2,5], Flora Jay [2]

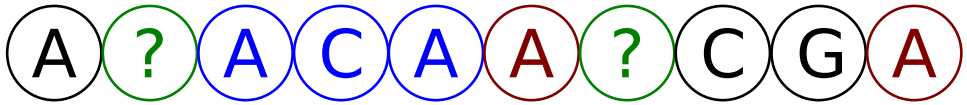
[1]Computer Science Department, Ecole Normale Supérieure de Lyon, France; [2]U Paris-Saclay, CNRS, INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, France; [3] Center for Computational Molecular Biology, Brown U, USA; [4] International Laboratory for Human Genome Research, U Nacional Autónoma de México, México; [5]Institute of Genomics, U of Tartu, Estonia.

Ancient DNA specificities

Undamamaged DNA



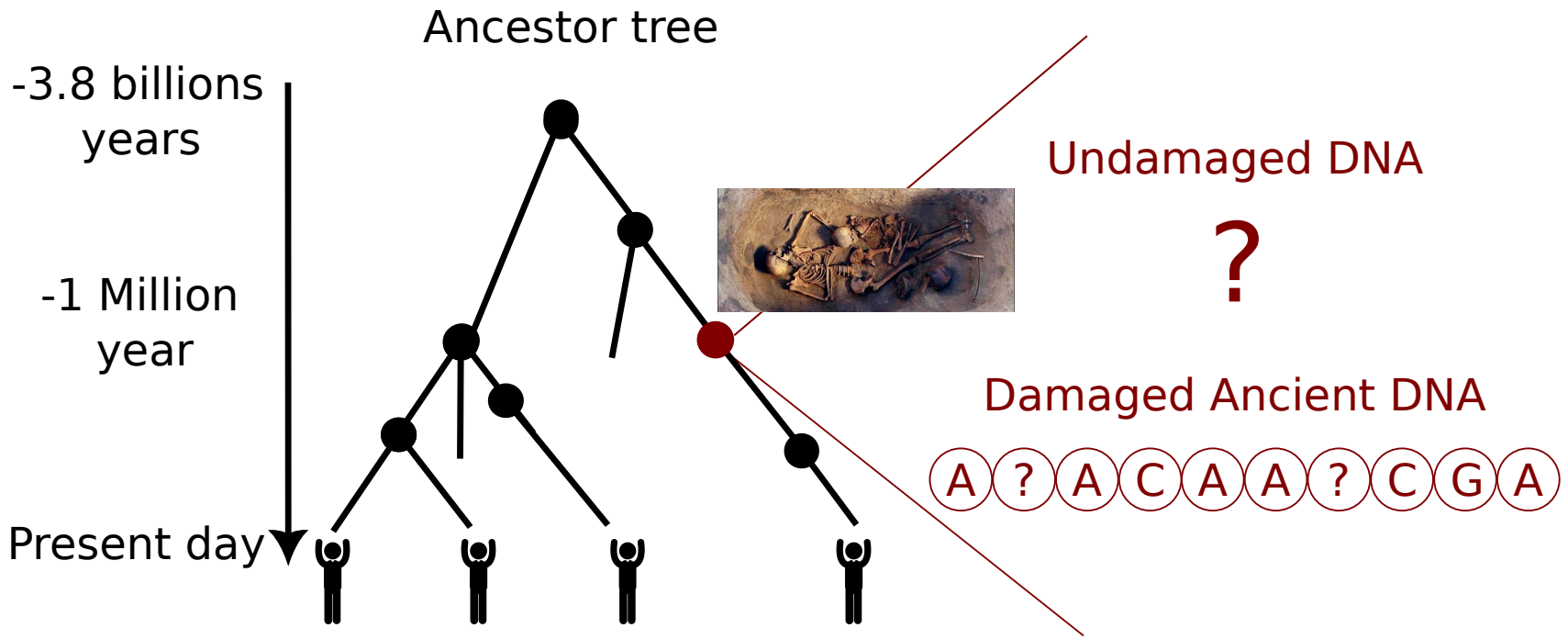
Damaged Ancient DNA



Difficulties with Ancient DNA (aDNA):

- Degrades over time
- Contaminated by external DNA
- More missing data and errors than modern DNA

Why study Damaged Ancient DNA?



ARTICLE

doi:10.1038/nature16544

Ancient gene flow from early modern humans into Eastern Neanderthals

Martin Kuhlwilm[1*], Ilan Gronau[2*], Melissa J. Hubisz[3], Cesare de Filippo[1], Javier Prado-Martinez[4], Martin Kircher[1,5], Qiaomei Fu[1,6,7], Hernán A. Burbano[1,8], Carles Lalueza-Fox[9], Marco de la Rasilla[10], Antonio Rosas[10], Pavao Rudan[11], Dejana Brajkovic[12], Željko Kucan[11], Ivan Gušic[11], Tomas Marques-Bonet[4,13,14], Aida M. Andrés[4], Bence Viola[15,16], Svante Pääbo[1], Matthias Meyer[1], Adam Siepel[3,17] & Sergi Castellano[1]

Article

Large-scale migration into Britain during the Middle to Late Bronze Age

https://doi.org/10.1038/s41586-021-04287-4
Received: 31 December 2020
Accepted: 29 November 2021
Published online: 22 December 2021

A list of authors and their affiliations appears at the end of the paper.

Present-day people from England and Wales harbour more ancestry derived from Early European Farmers (EEF) than people of the Early Bronze Age[1]. To understand this, we generated genome-wide data from 793 individuals, increasing data from the Middle to Late Bronze and Iron Age in Britain by 12-fold, and Western and Central

RESEARCH ARTICLE SUMMARY

HUMAN EVOLUTION

The formation of human populations in South and Central Asia

Vagheesh M. Narasimhan *et al.*

SCIENCE ADVANCES | RESEARCH ARTICLE

EVOLUTIONARY BIOLOGY

Genetic ancestry changes in Stone to Bronze Age transition in the East European plain

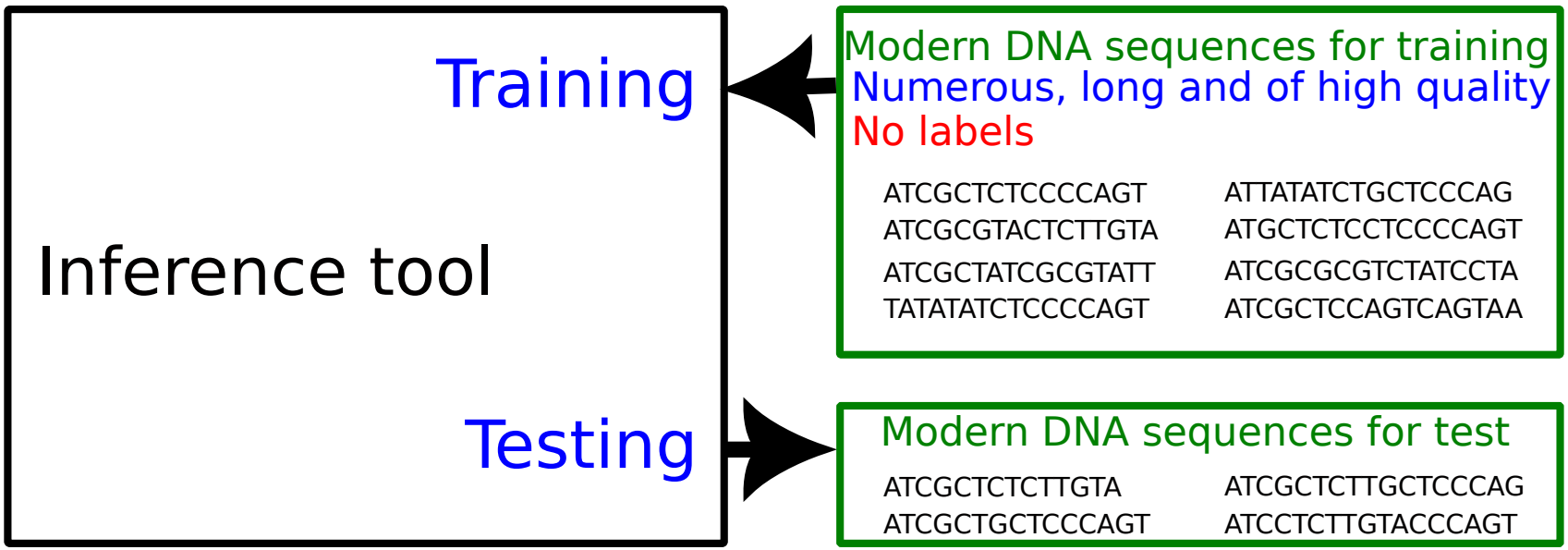
Lehti Saag[1]*, Sergey V. Vasilyev[2], Liivi Varul[3], Natalia V. Kosorukova[4], Dmitri V. Gerasimov[5], Svetlana V. Oshibkina[6], Samuel J. Griffith[1], Anu Solnik[1], Lauri Saag[1], Eugenia D'Atanasio[7], Ene Metspalu[1], Maere Reidla[1], Siiri Rootsi[1], Toomas Kivisild[1,8], Christiana Lyn Scheib[1,9], Kristiina Tambets[1], Aivar Kriiska[10]*, Mait Metspalu[1]*

Cell

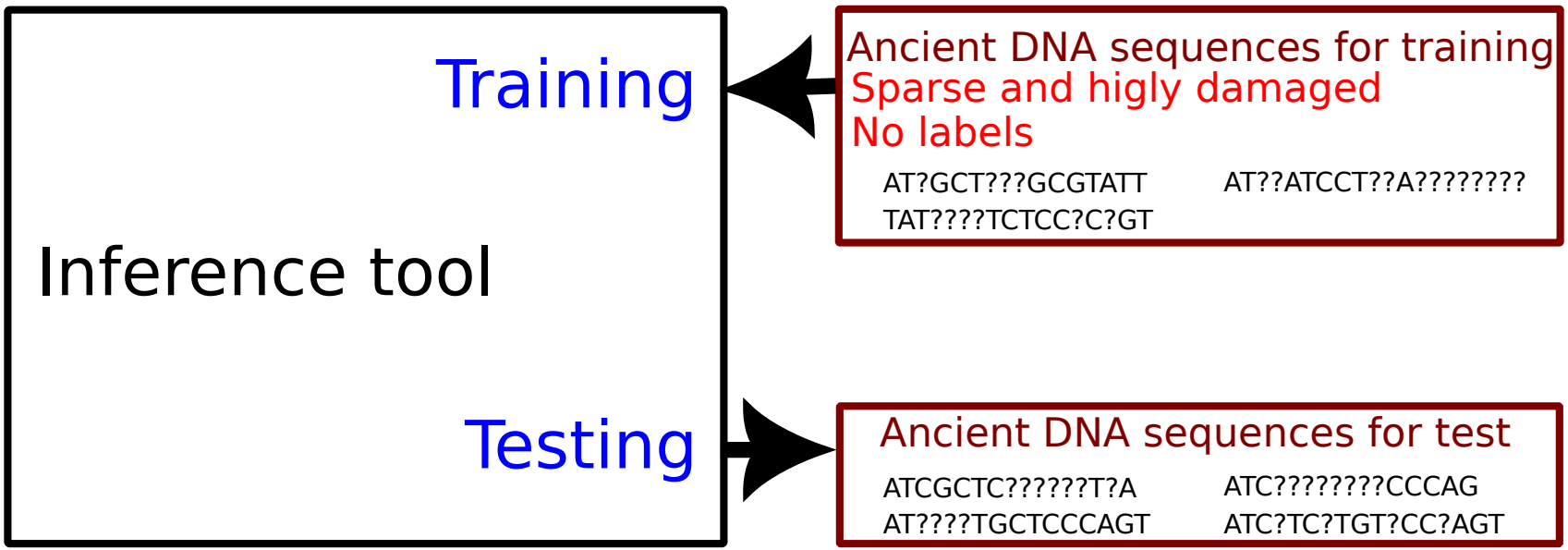
Article

The genomic origins of the world's first farmers

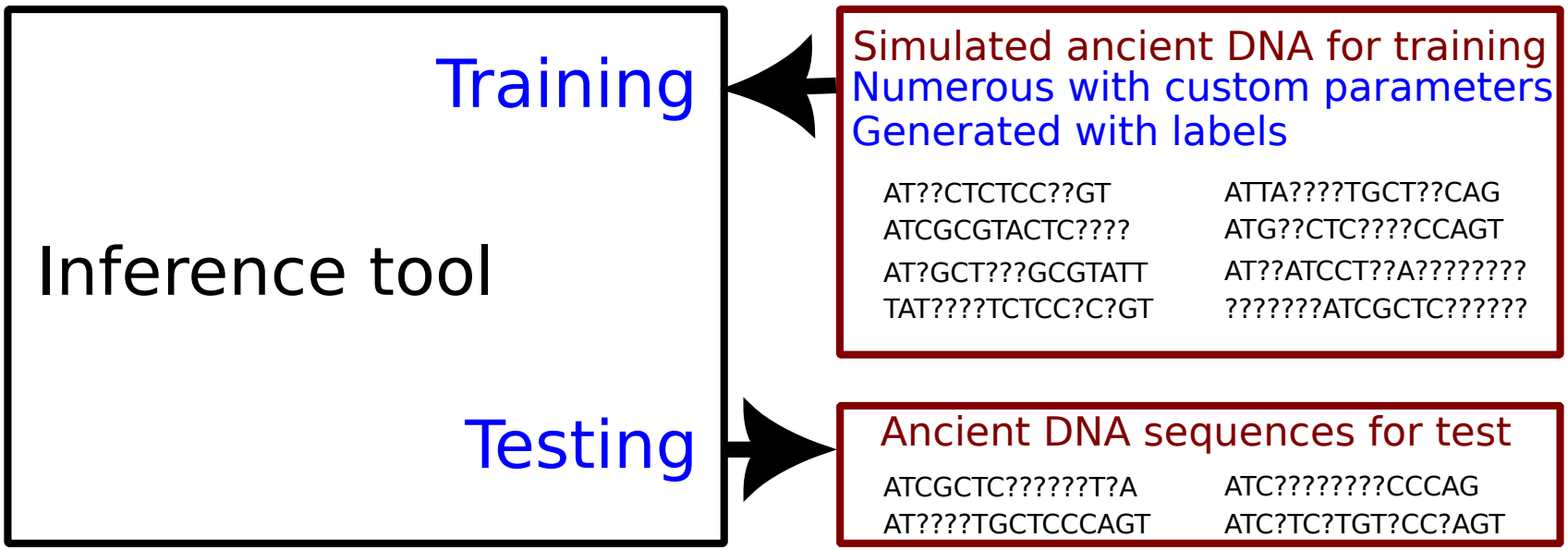
DNA sequences for inference purposes....



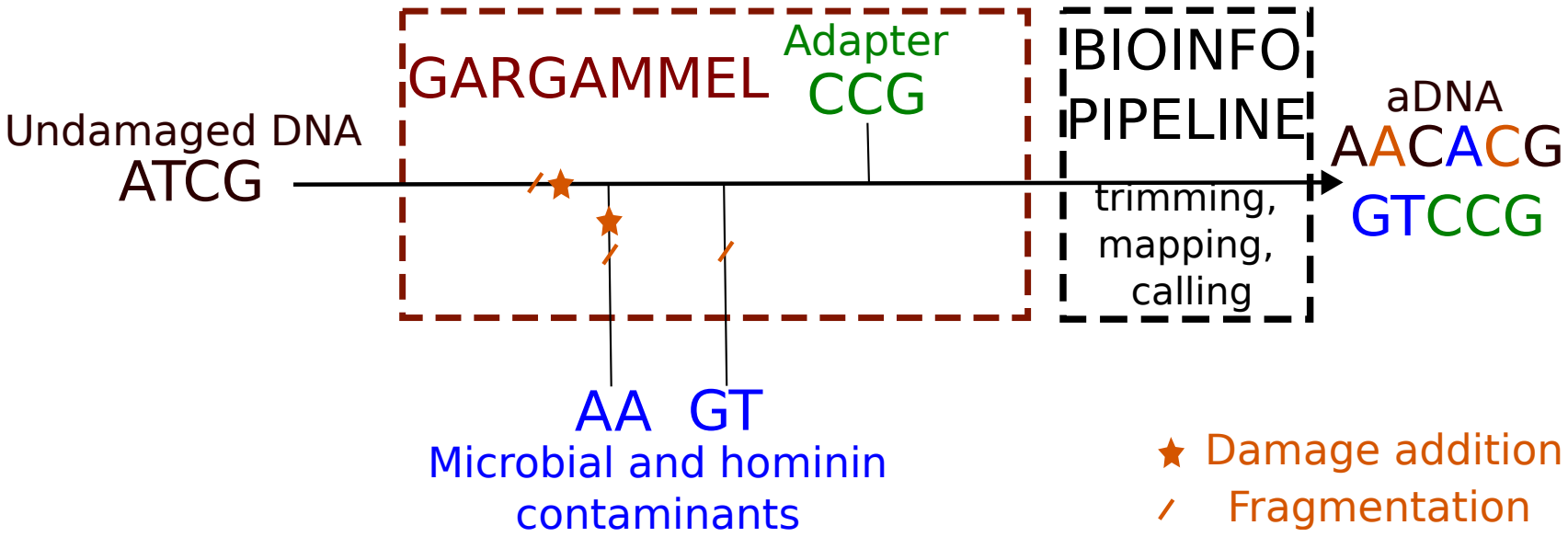
aDNA sequences for inference purposes....



... required simulation of aDNA sequences



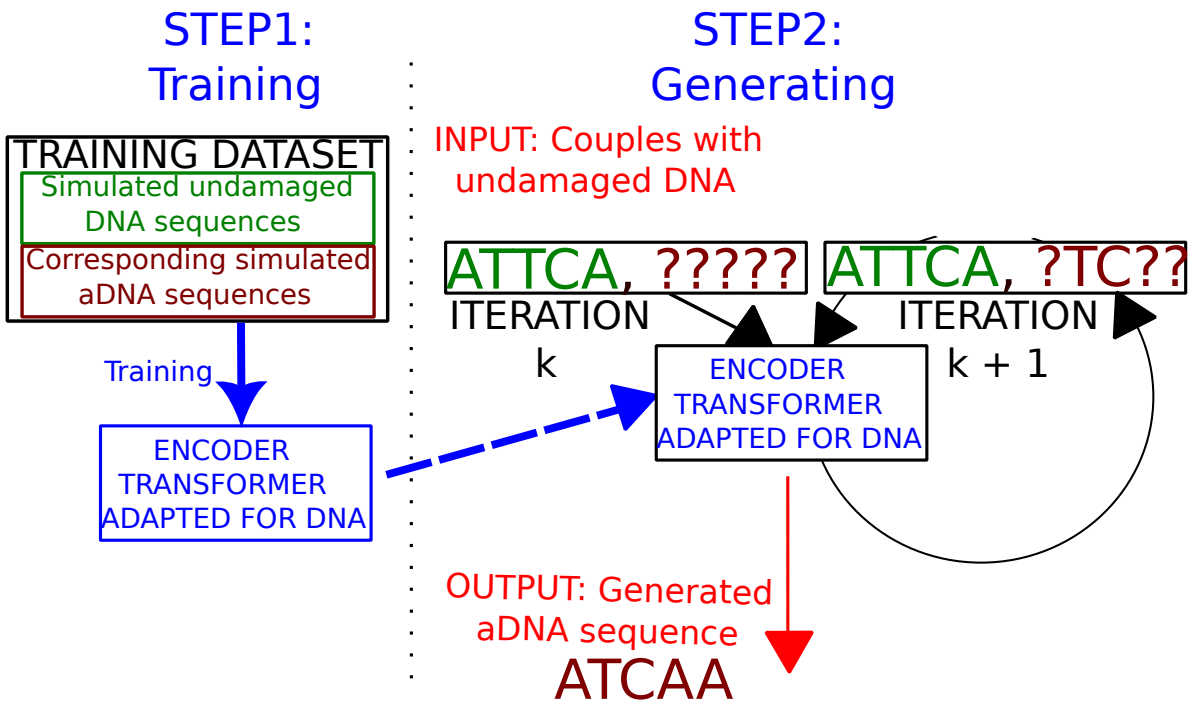
State-of-the-art aDNA simulator: Gargammel[1]



- Gargammel complexity: $O(n \times c \times f)$
- With c, the desired coverage and f, the number of fragments sampled by Gargammel
- f can lead to a large overhead in practice

[1] Renaud et al, 2016, Bioinformatics

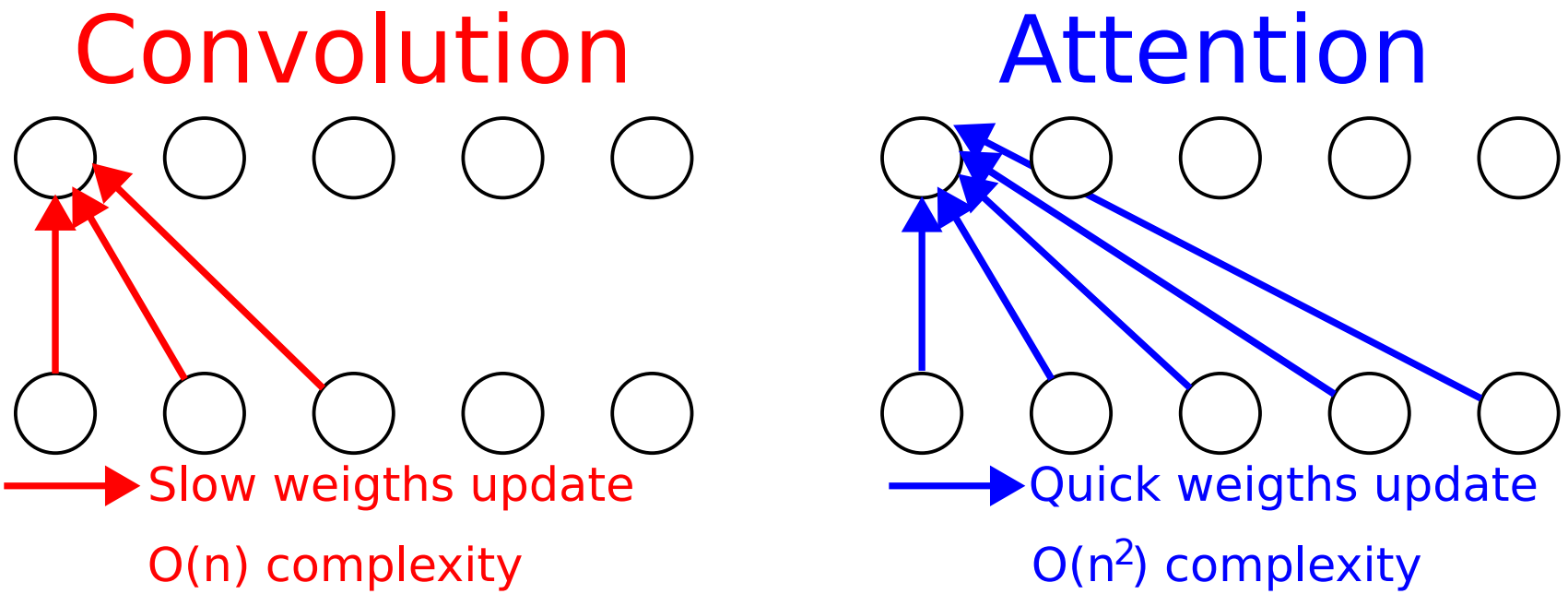
Achieved result: our new seq-to-seq aDNA simulator



- Generate aDNA sequences from undamaged ones
- Method: Iterative process over a specific encoder-only transformer
- Data simulation: Undamaged sequences: Msprime[2].
  Damaged sequences: pipeline around Gargammel[3]
- Generation complexity: $O(n^3)$, of interests compared to Gargammel

[2] Baumdicker et al, 2021, Genetics
[3] Jazeps et al, 2023, ?

Attention interest[4] versus convolution



[4] Vaswani et al, 2017

Our use of the pretrained DNABERT model[5]

- Max seq. length: 512 nucleotides. Comp.: 12 Transformer layers

[5] Ji et al, 2021, Bioinformatics

Our data and our ne-tunings

- A mask prediction task to predict the aDNA part
- A binary classi cation task to measure if the aDNA part is a plausible \translation" of the undamaged DNA

Generation algorithm using Mask Prediction only[6]

- Complexity with Mask Prediction alone $\mathcal{O}(n^3)$
- With n, size of the input

[6] Inspired by bert-gen, Wang et al, 2019

Use of classi cation as a complement for mask predictio in the generation

- Classi cation: Is the aDNA sequence a \plausible" translation for the undamaged sequence?
- Complexity with theK -Top table: $O(Kn^3)$

Results

- We do alignments[7] with 30 chunks of aDNA
- Aligned sequences are identical at 74 percent in average
- We count similar positions to the exclusion of gaps and missing data

[7] Wheeler et al, 2000, Nucleic Acids Research

Perspective: push further complexity and performances

- First leverage: Sparse[8] or linear[9] attention instead of full attention
  Reduce attention to $O(n)$ instead of $O(n^2)$.
- Second leverage: Use of SNPs
  Counteracts the 512 nucleotide limitation and "diminishes"

[8] Zaheer et al, 2021
[9] Nesterenko et al, 2022

Conclusion and future work

- A new seq-to-seq simulation technique for ancient DNA sequences
- Complexity in $O(n^3)$ in simpler case, complemented to the use of batches in practice

Future work:

- De ne new criteria to assess the quality of sequences
- Do our own pre-training instead of using DNABERT's one.
- De ne an "in-between" classi cation task when using k-Top table
- General advance in transformers will be at our advantage: Integration of sparse attention

Acknowledgements

Contributors:

- Jazeps Medina-Tretmanis
- Maria Avila-Arcos
- Emilia Huerta-Sanchez
- Antoine Szatkownik (\Generative model for genomics" poster!)
- Burak Yelmen
- Flora Jay
- Guillaume Charpiat

Laboratories and funding agencies:

Our ancient DNA simulation using encoder-transformer