Generative modelling in genomics and a perspective on uncertainty quantification

Burak Yelmen

University of Tartu, Institute of Genomics, Tartu, Estonia

burak.yelmen@ut.ee

UNIVERSITY OF TARTU Institute of Genomics

REALMENT

CoMorMent

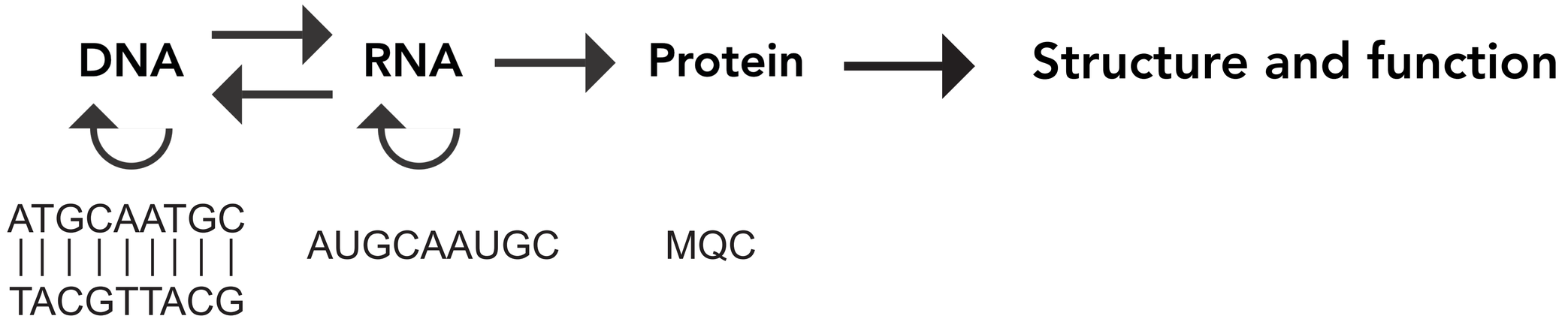
LISN

université PARIS-SACLAY

Genetics, genomics, transcriptomics, proteomics

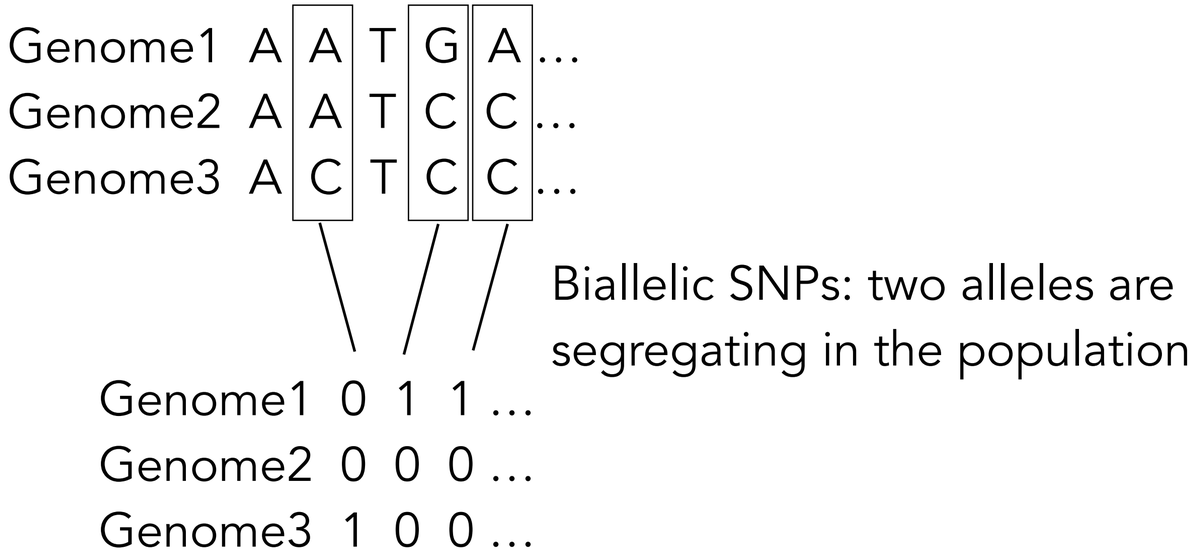
DNA: Deoxyribonucleic acid - **genetic information**
RNA: Ribonucleic acid - **transcribed** genetic information
Protein: Amino acid chains with 3D structure - **translated** genetic information
Gene: A sequence of DNA transcribed into a functional RNA - could be protein **coding** or **non-coding**
Genome: Entirety of DNA in an organism - 3 billion base pairs in human genome

Flow of genetic information (Central dogma)

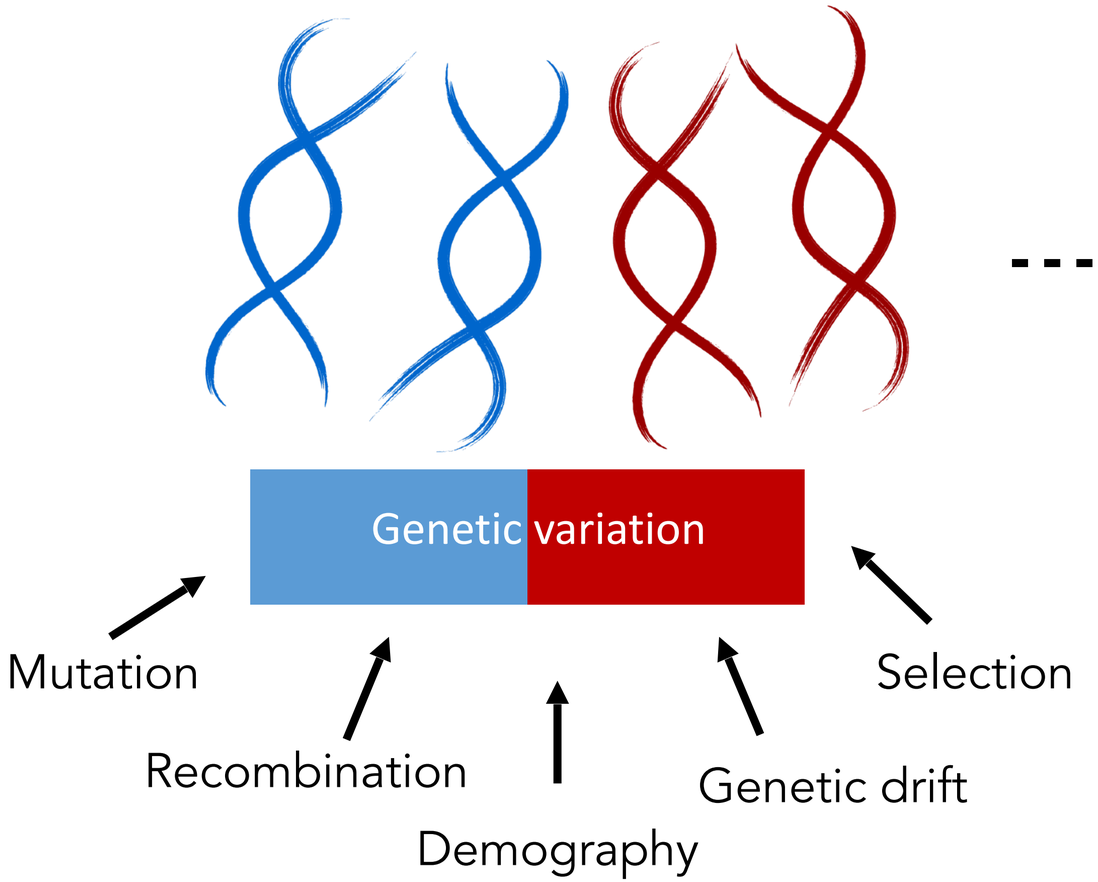


Population genomics

SNP: Single-nucleotide polymorphism, change of a single nucleotide at a specific position in the genome
Allele: A variant at particular position in the genome
Biallelic SNP: a SNP with two alternative alleles in the population (most human SNPs)



What is the source of variation?



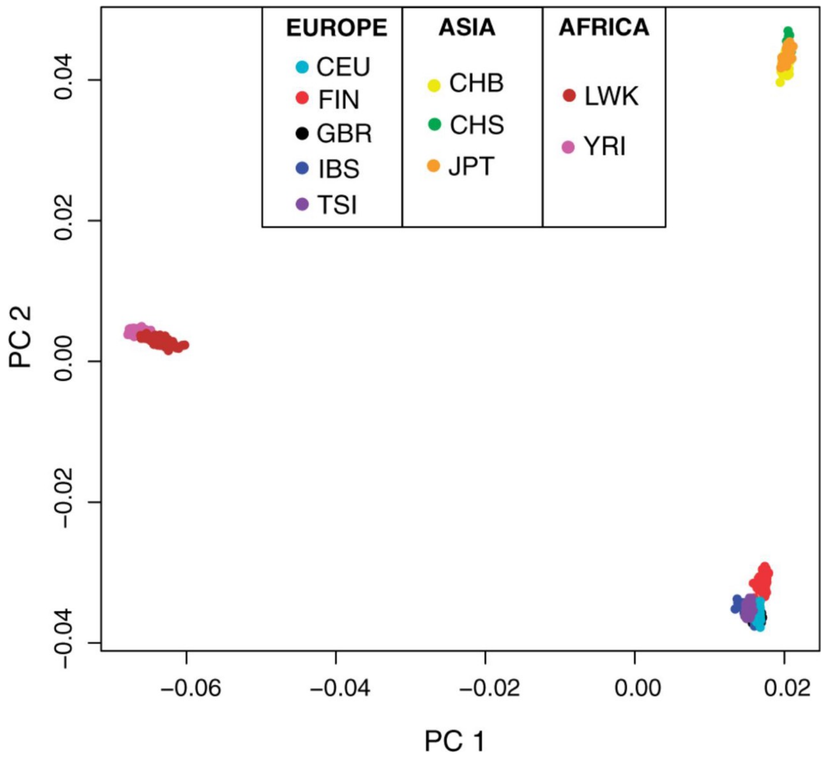
Genome1	A	A	T	G	A	…
Genome2	A	A	T	C	C	…
Genome3	A	C	T	C	C	…

Genome1	0	1	1	…
Genome2	0	0	0	…
Genome3	1	0	0	…

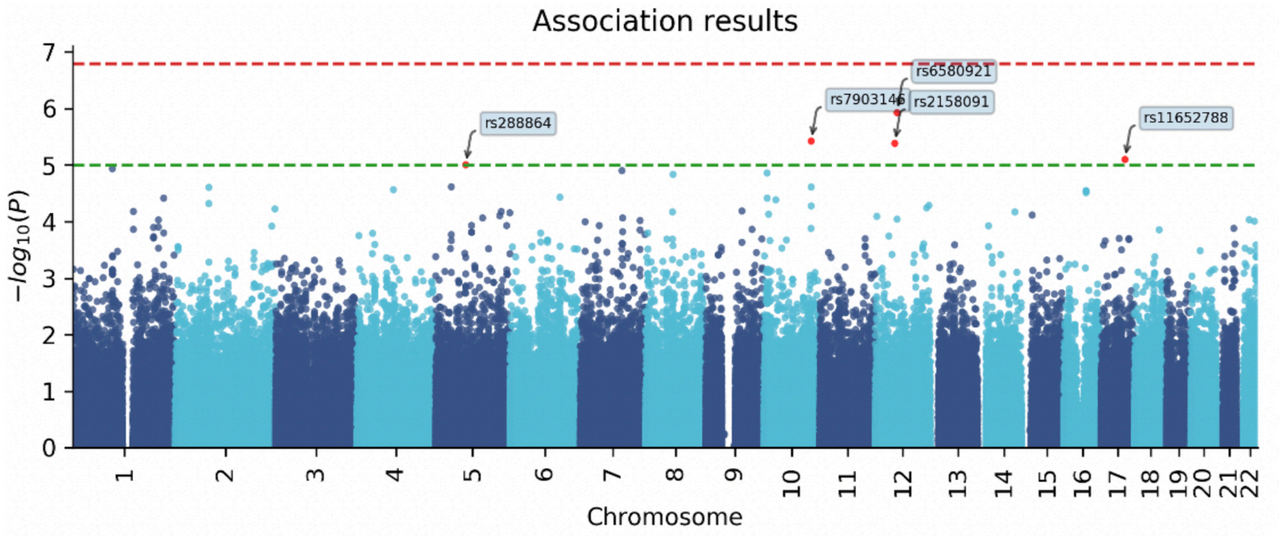
Why study genomes?

- Captures all the genetic information of an organism -> protein coding and regulatory regions
- Captures all genetic variation between individuals -> population structure and phenotypes (disease, height etc.)

PCA of 1000 Genomes data[1]



Type 2 diabetes genome-wide association study (GWAS) on Estonian genomes[2]



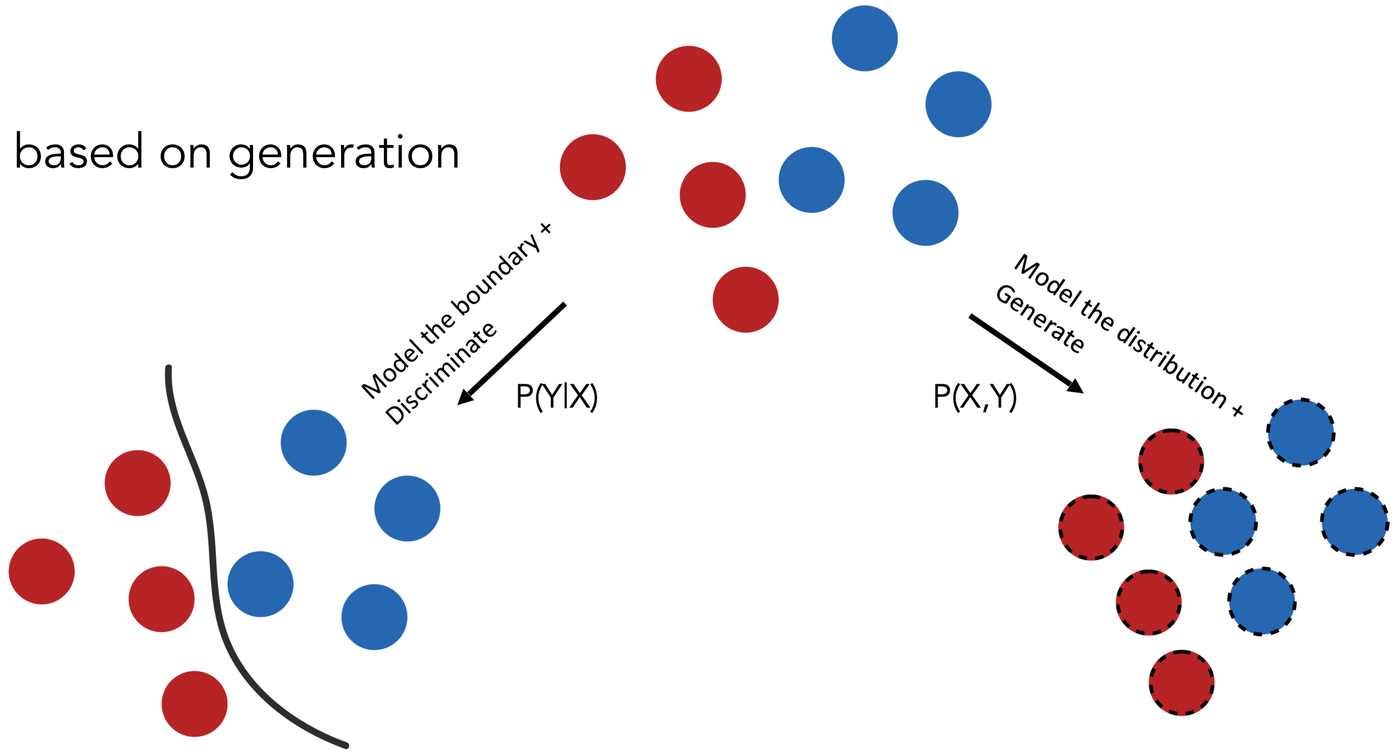
1. Duforet-Frebourg, Nicolas, et al. "Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data." *Molecular biology and evolution* 33.4 (2016): 1082-1093.
2. Kovalev, Gleb "Potential of Artificial Genomes in Genome-wide Association Studies" *University of Tartu Press* (2021).

Generative models

Definition 1 (Statistical):
Generative modelling -> joint probability P(X,Y), Discriminative modelling -> conditional probability P(Y|X)

Definition 2 (Task-oriented):
Any model which aims to generate partial or full data points

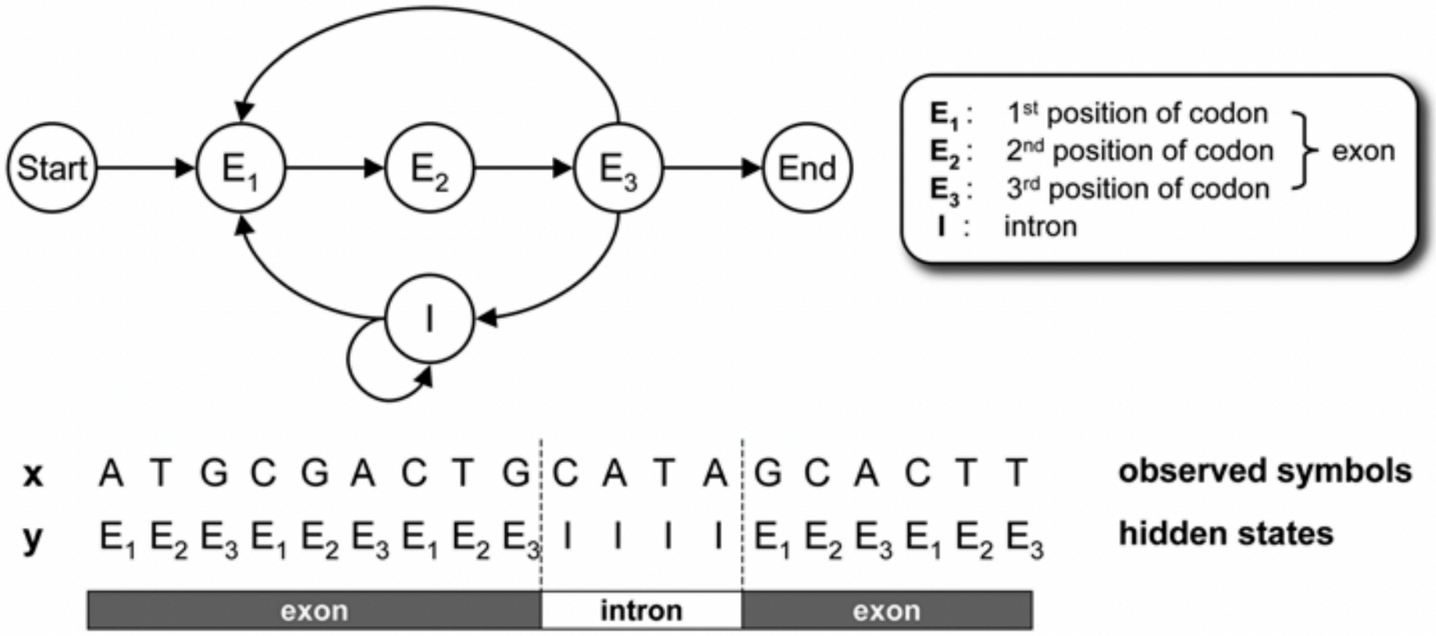
Definition 3 (Training-oriented):
Any model for which the training loss function is based on generation of partial or full data



Generative models in genomics

Generative modelling with biological sequences has a long history: An example is **Hidden Markov models** (HMMs)

HMM for modelling protein-coding eukaryotic genes[1]



1. Yoon, Byung-Jun. "Hidden Markov models and their applications in biological sequence analysis." *Current genomics* 10.6 (2009): 402-415.

Generative models in genomics

More recently, deep generative models such as **generative adversarial networks (GANs), variational autoencoders (VAEs)** and **large language models (LLMs)** for

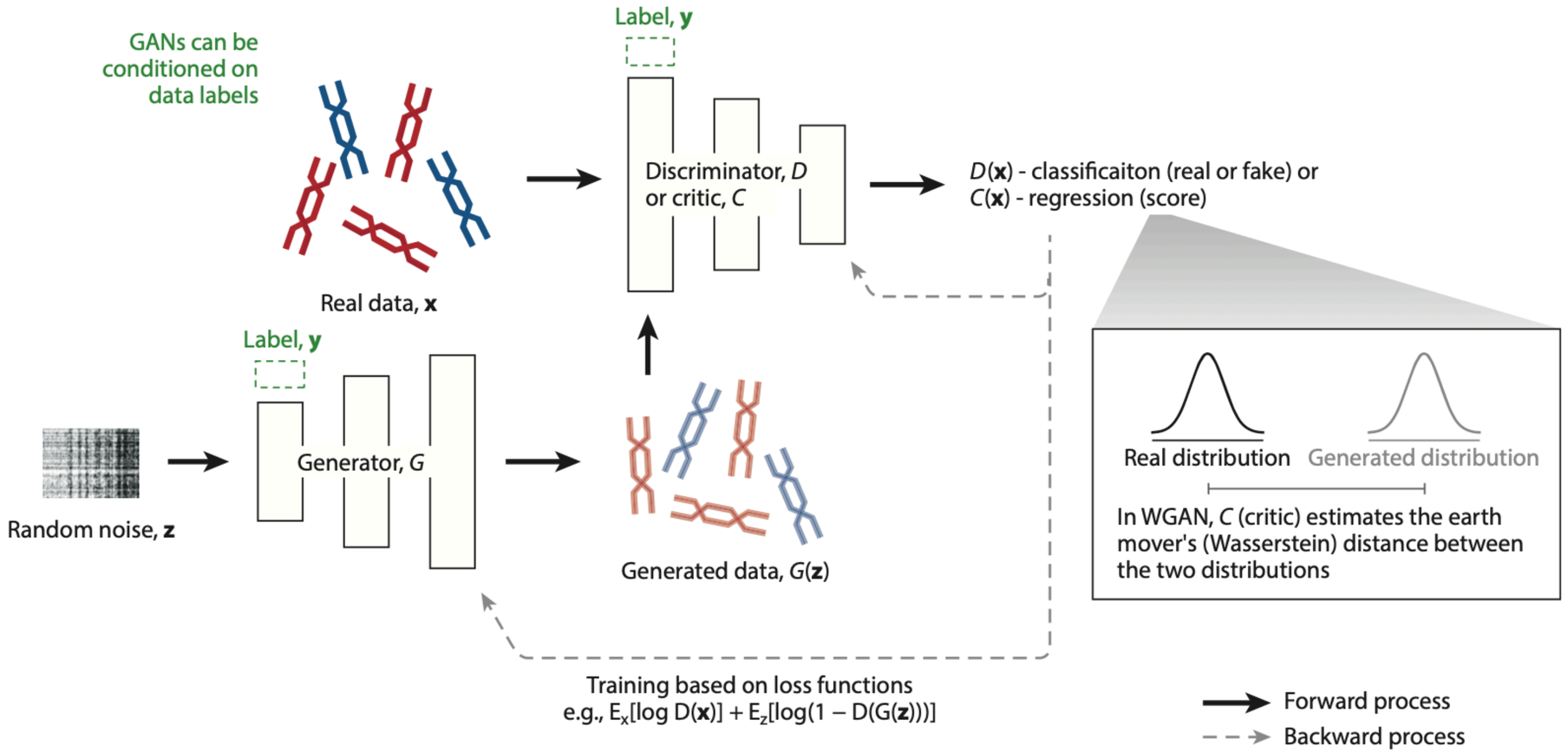
1. **Data generation:** Generation of realistic genomic data and design of functional sequences
2. **Dimensionality reduction:** From high-dimensional genomic space to low-dimensional latent space for characterisation of differences and downstream analyses
3. **Prediction:** Prediction of function, disease status or evolutionary parameters

Why deep generative models?

1. Unsupervised and semi-supervised training
2. High-quality data generation
3. Meaningful non-linear mapping of high-dimensional genomic space to low-dimensional latent space

Generative adversarial network (GAN)[1]

Typical GAN and Wassertein GAN[2] models used in genomics[3]

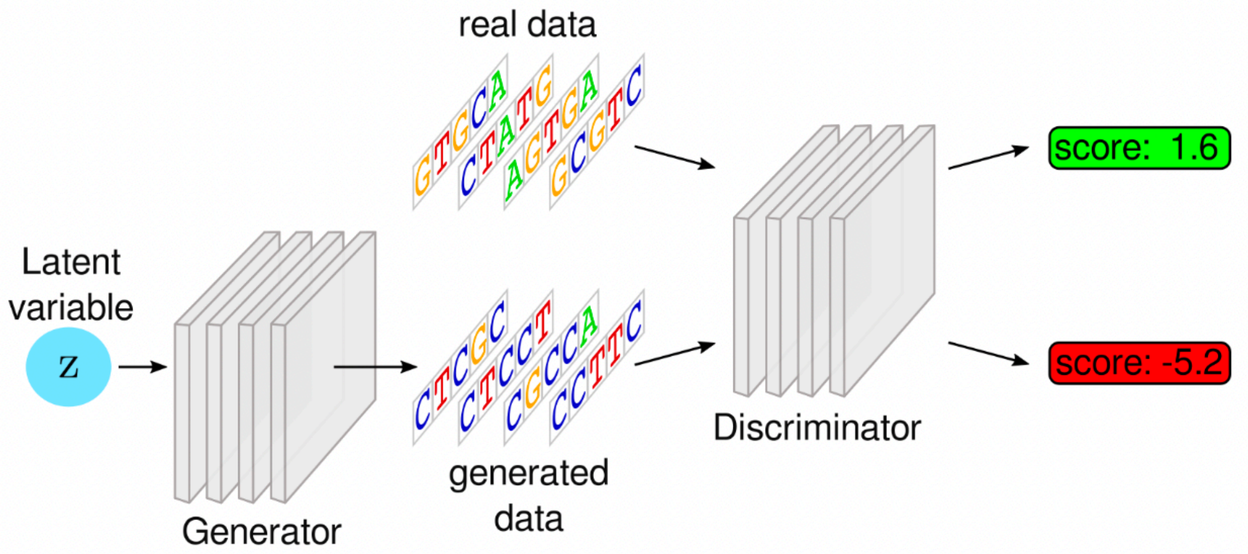


1. Goodfellow, Ian, et al. "Generative adversarial networks." *Communications of the ACM* 63.11 (2020): 139-144.
2. Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." *International conference on machine learning*. PMLR, 2017
3. Yelmen, Burak, and Flora Jay. "An Overview of Deep Generative Models in Functional and Evolutionary Genomics." *Annual Review of Biomedical Data Science* 6 (2023).

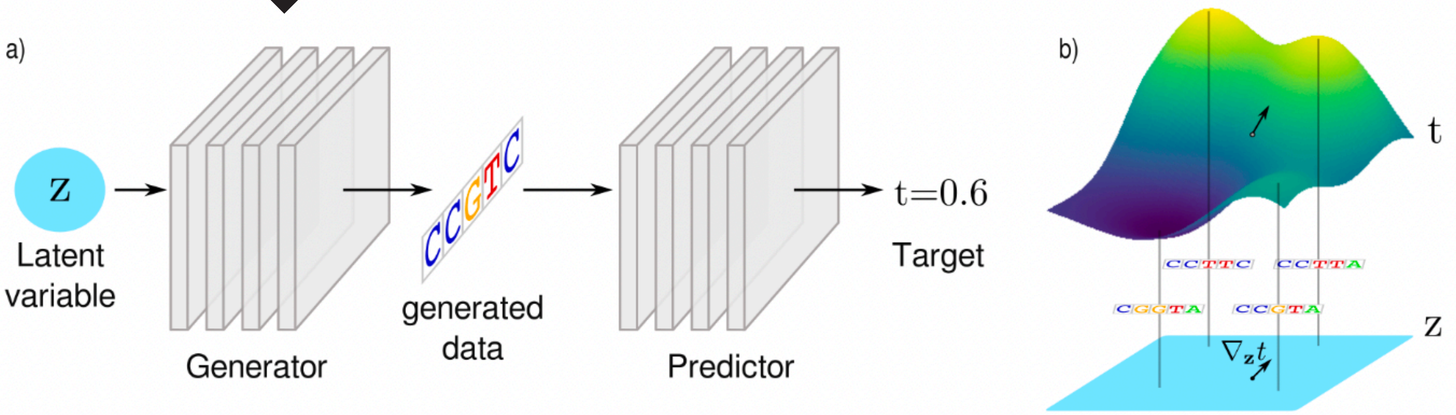
Relevant research:

GANs for designing DNA sequences[1]

Unsupervised GAN training



Using the trained generator combined with a predictor function to adjust the latent space for generating sequences with desired properties, such as higher protein binding

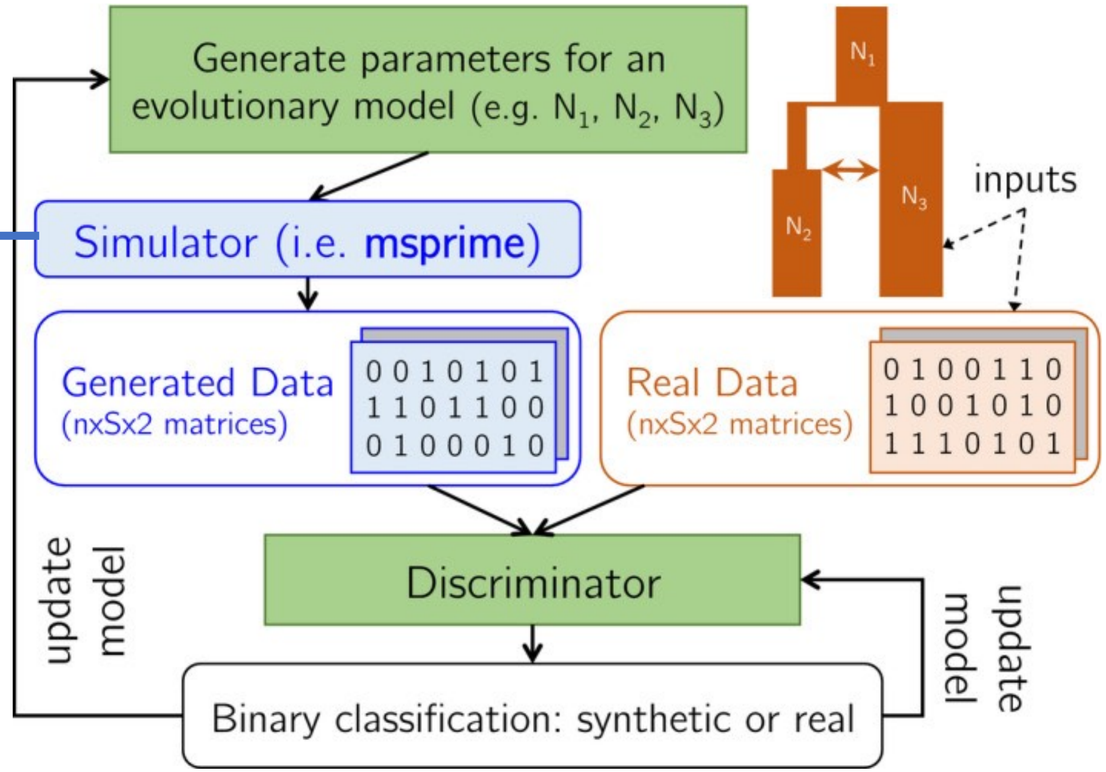


1. Killoran, Nathan, et al. "Generating and designing DNA with deep generative models." *arXiv preprint arXiv:1712.06148* (2017).

Relevant research:

GAN-like model for evolutionary parameter estimation[1]

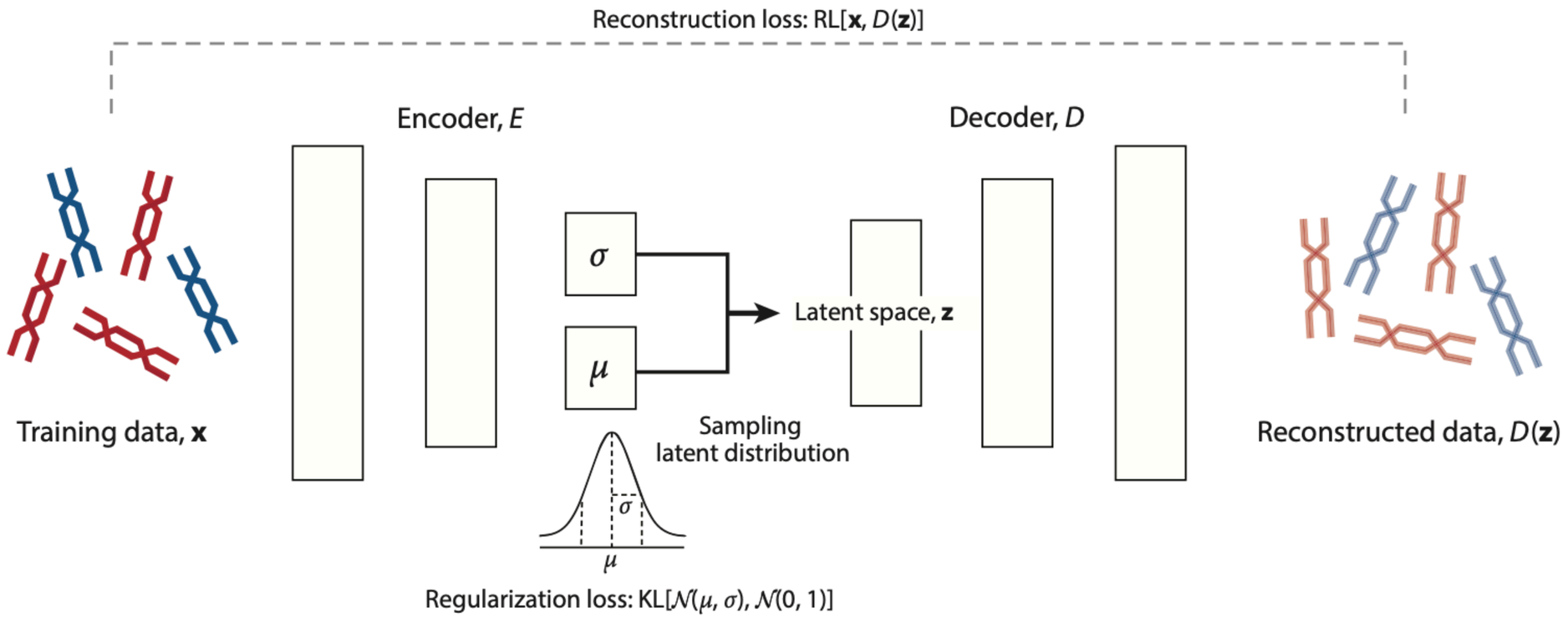
(Coalescent simulations)



1. Wang, Zhanpeng, et al. "Automatic inference of demographic parameters using generative adversarial networks." *Molecular ecology resources* 21.8 (2021): 2689-2705.

Variational Autoencoder (VAE)[1]

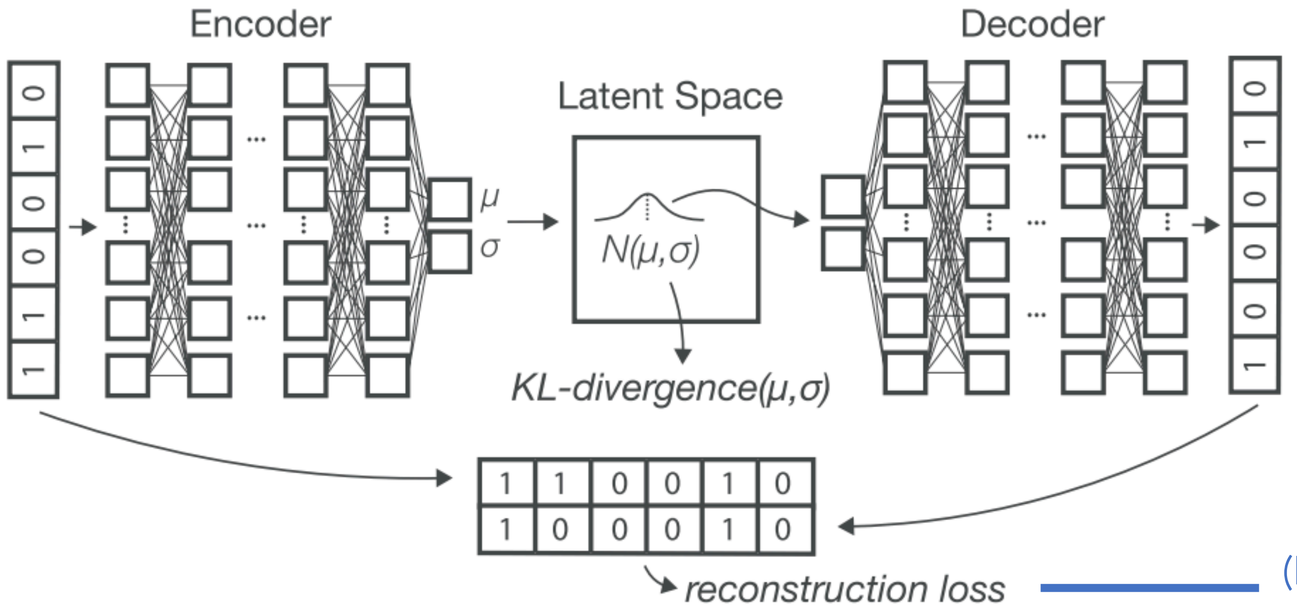
Typical VAE model used in genomics[2]



1. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).
2. Yelmen, Burak, and Flora Jay. "An Overview of Deep Generative Models in Functional and Evolutionary Genomics." *Annual Review of Biomedical Data Science* 6 (2023).

Relevant research:

VAE for dimensionality reduction[1]



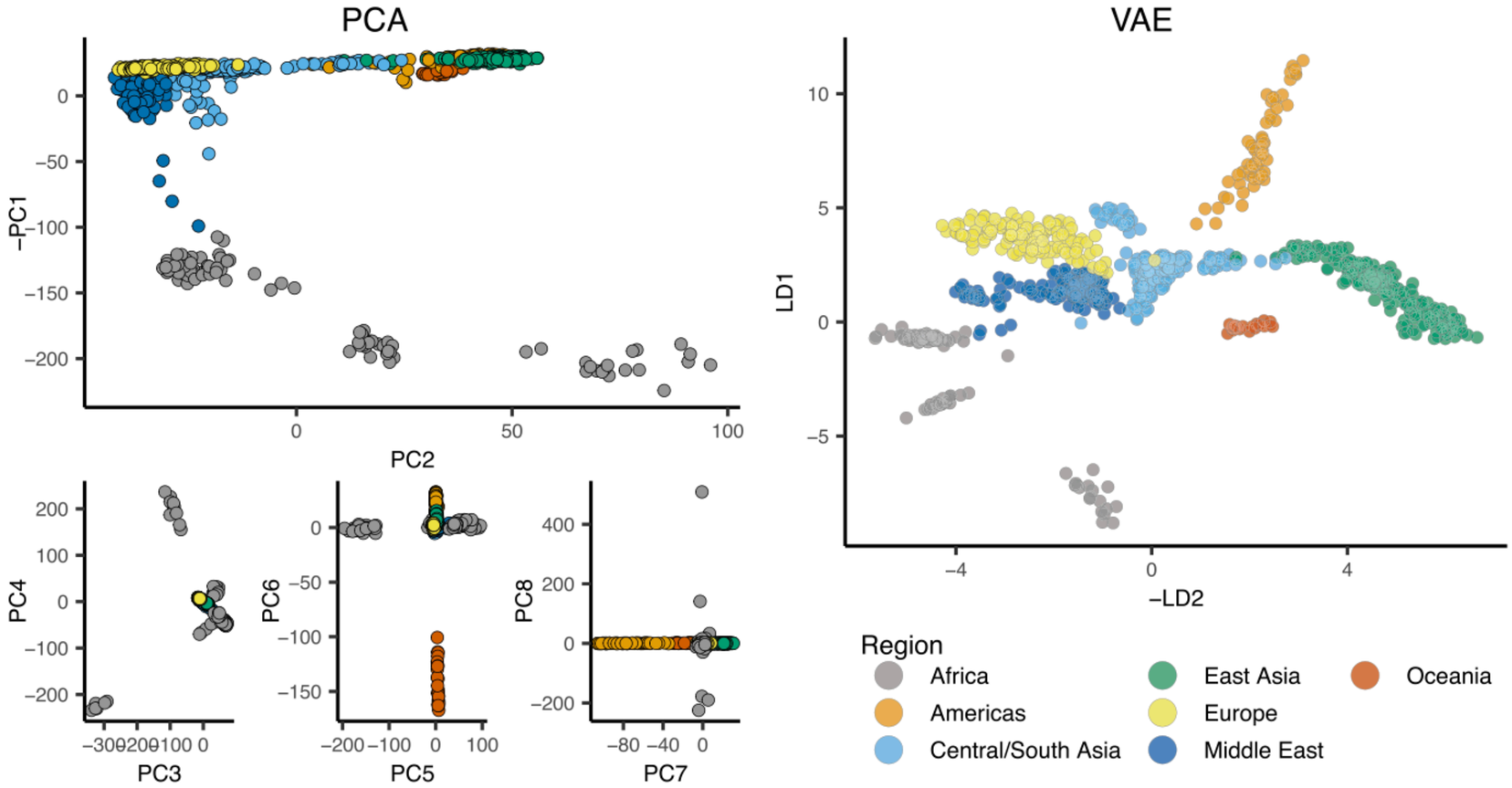
(Binary cross-entropy between input and decoded sequence)

Between latent and standard normal distribution N(0,1)

1. Battey, C. J., Gabrielle C. Coffing, and Andrew D. Kern. "Visualizing population structure with variational autoencoders." *G3* 11.1 (2021): jkaa036.

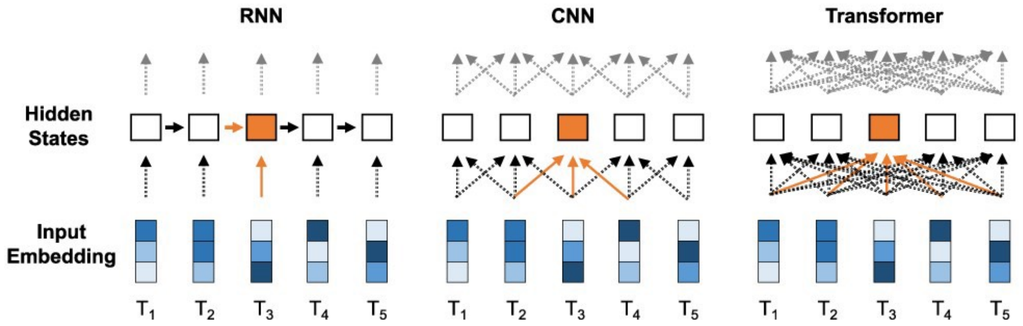
Relevant research:

VAE for dimensionality reduction[1]



1. Battey, C. J., Gabrielle C. Coffing, and Andrew D. Kern. "Visualizing population structure with variational autoencoders." *G3* 11.1 (2021): jkaa036.

DNA Language Models

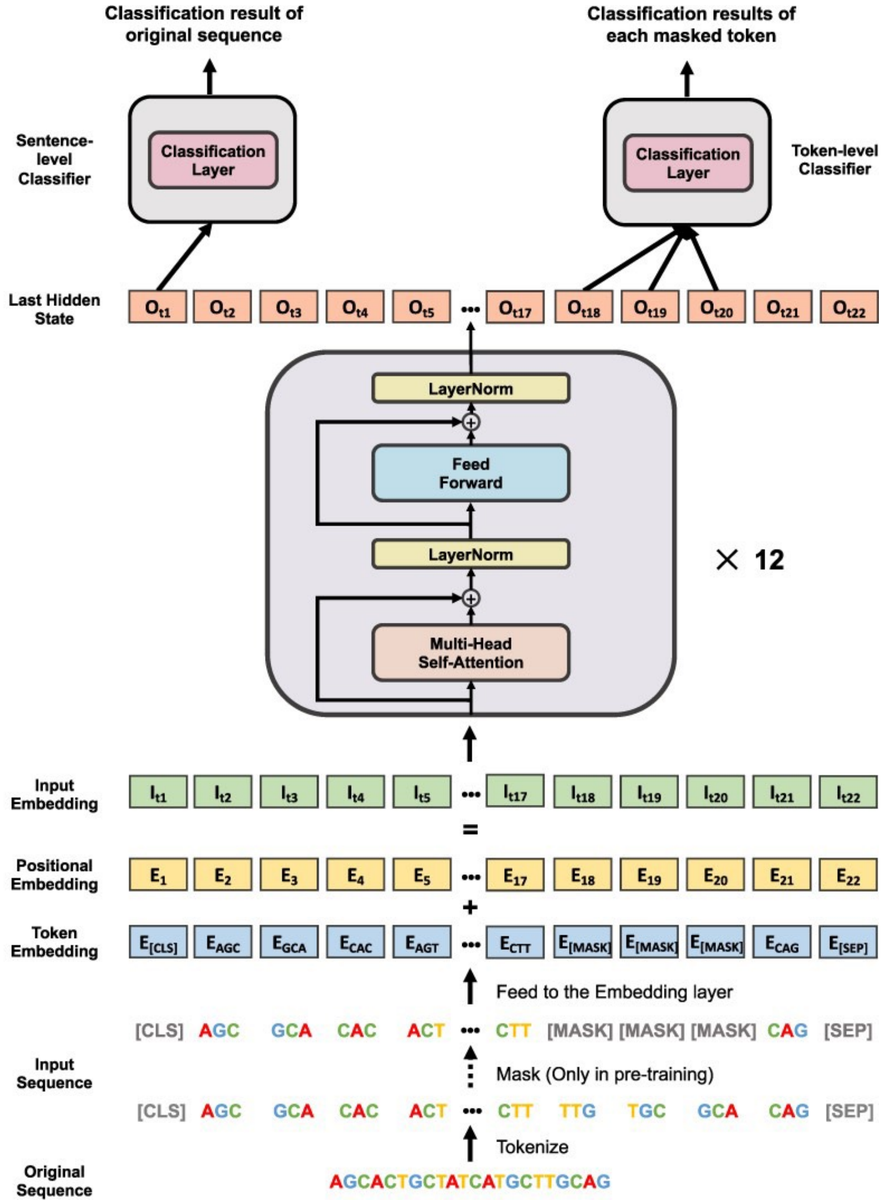


Contextual difference in using different architectures:
Global (Attention) vs Local (Convolution)[1]

Cross-entropy loss for pre-training:

$$L = \sum_{i=0}^{N} -y'_i \log(y_i)$$

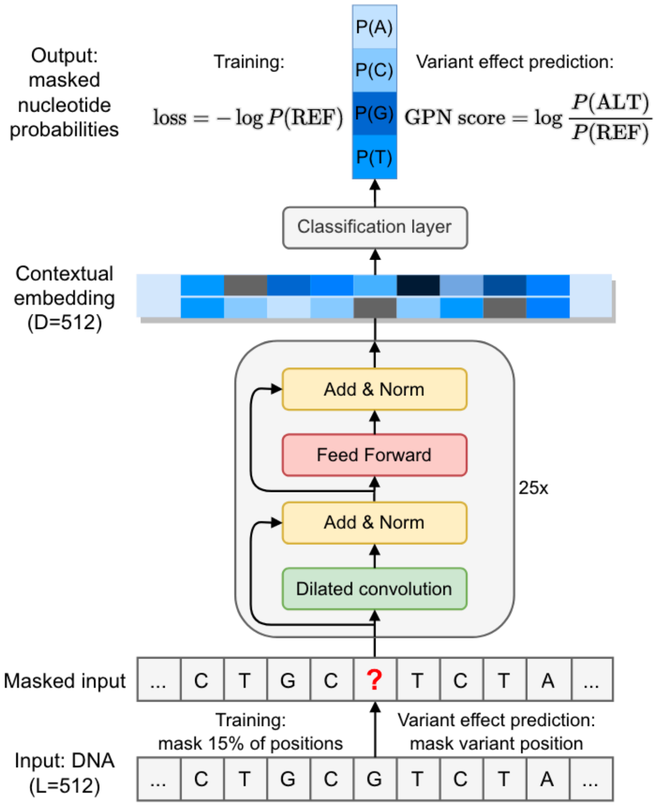
DNABERT model[1]



1. Ji, Yanrong, et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." *Bioinformatics* 37.15 (2021): 2112-2120.

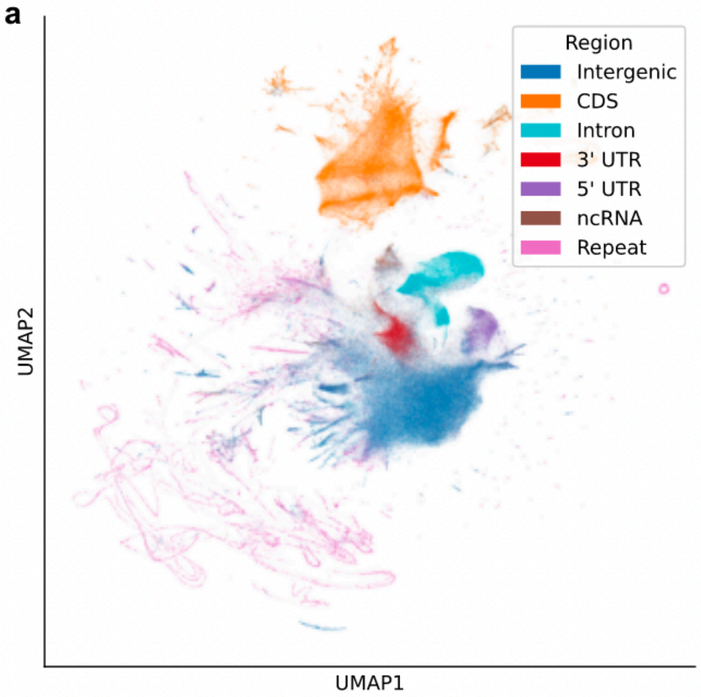
Relevant research:

DNA language models for functional sequence prediction[1]

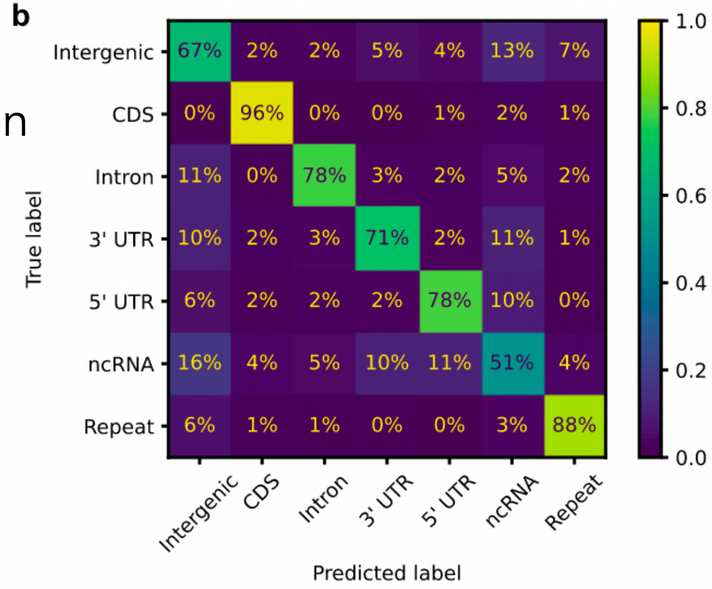


Genomic Pre-trained Network (GPN)

UMAP clustering of GPN embeddings



Prediction with logistic regression on embeddings



1. Benegas, Gonzalo, Sanjit Singh Batra, and Yun S. Song. "DNA language models are powerful predictors of genome-wide variant effects." *Proceedings of the National Academy of Sciences* 120.44 (2023): e2311219120.

Generation of artificial human genomes

RESEARCH ARTICLE

Creating artificial human genomes using generative neural networks

Burak Yelmen[1,2,3]*, Aurélien Decelle[3,4], Linda Ongaro[1,2], Davide Marnetto[1], Corentin Tallec[3], Francesco Montinaro[1,5], Cyril Furtlehner[3], Luca Pagani[1,6], Flora Jay[3]*

1 Institute of Genomics, University of Tartu, Tartu, Estonia, **2** Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, **3** Laboratoire de Recherche en Informatique, CNRS UMR 8623, Université Paris-Sud, Université Paris-Saclay, Paris, France, **4** Departamento de Física Téorica I, Universidad Complutense, Madrid, Spain, **5** Department of Biology-Genetics, University of Bari, Bari, Italy, **6** APE Lab, Department of Biology, University of Padova, Padova, Italy

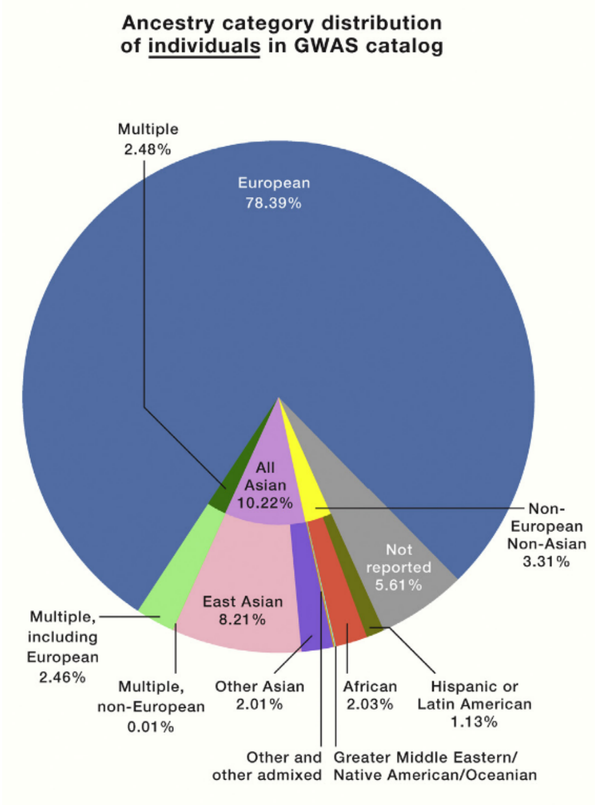
* burakyelmen@gmail.com (BY); flora.jay@lri.fr (FJ)

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., ... & Jay, F. (2021). Creating artificial human genomes using generative neural networks. *PLoS genetics*

Why generate genomic data?

- **Data accessibility**: Substantial amount of genomic data held by companies and institutions are not easily accessible due to privacy issues

- **Underrepresented populations in research**



From the GWAS catalogue through Jan 2019
Sirugo et al. 2019

Overarching research goals

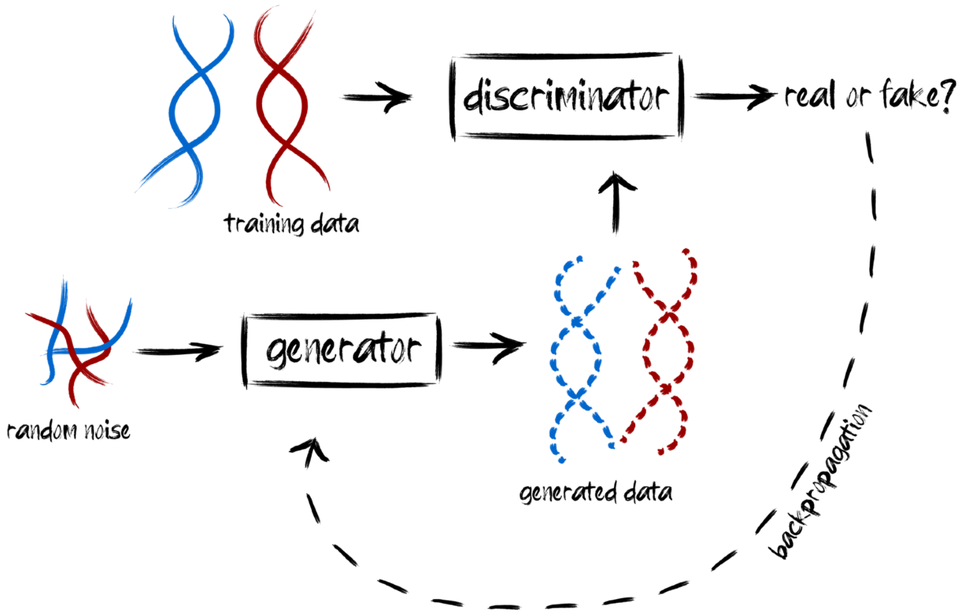
- Creating artificial genomes (AGs) which cannot be traced back to the original genomes yet bear all important characteristics of them

- Making high quality AG datasets as surrogates of private genome banks which can be accesses publicly

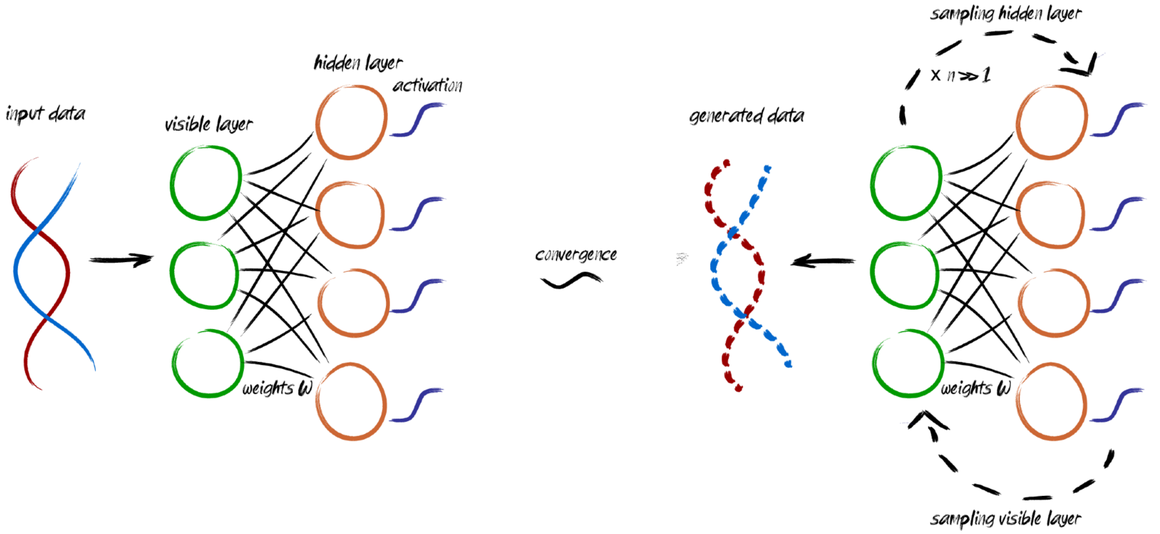
Methods

- **Generative neural networks**

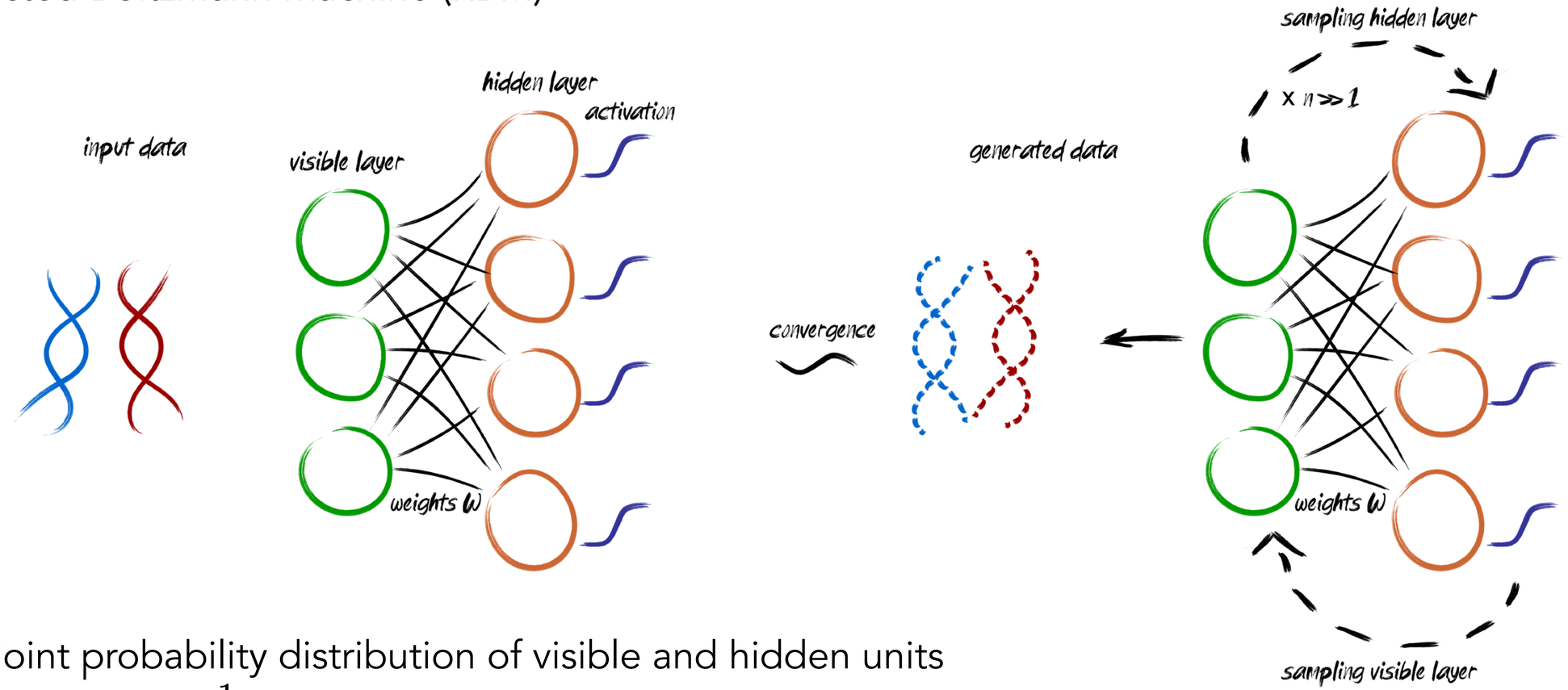
Generative adversarial network (GAN)



Restricted Boltzmann machine (RBM)



Restricted Boltzmann machine (RBM)*



Joint probability distribution of visible and hidden units

$$p(\underline{s}, \underline{\tau}) = \frac{1}{Z} \exp\left(\sum_{ia} w_{ia} s_i \tau_a + \sum_i \theta_i s_i + \sum_a \eta_a \tau_a\right)$$

Optimize weights and biases to maximise likelihood

$$L = \sum_{m=1}^{M} \left[\sum_i \theta_i s_i^{(m)} + \sum_a \log\left(1 + \exp\left(\sum_i w_{ia} s_i^{(m)} + \eta_a\right)\right)\right] - M \log(Z)$$

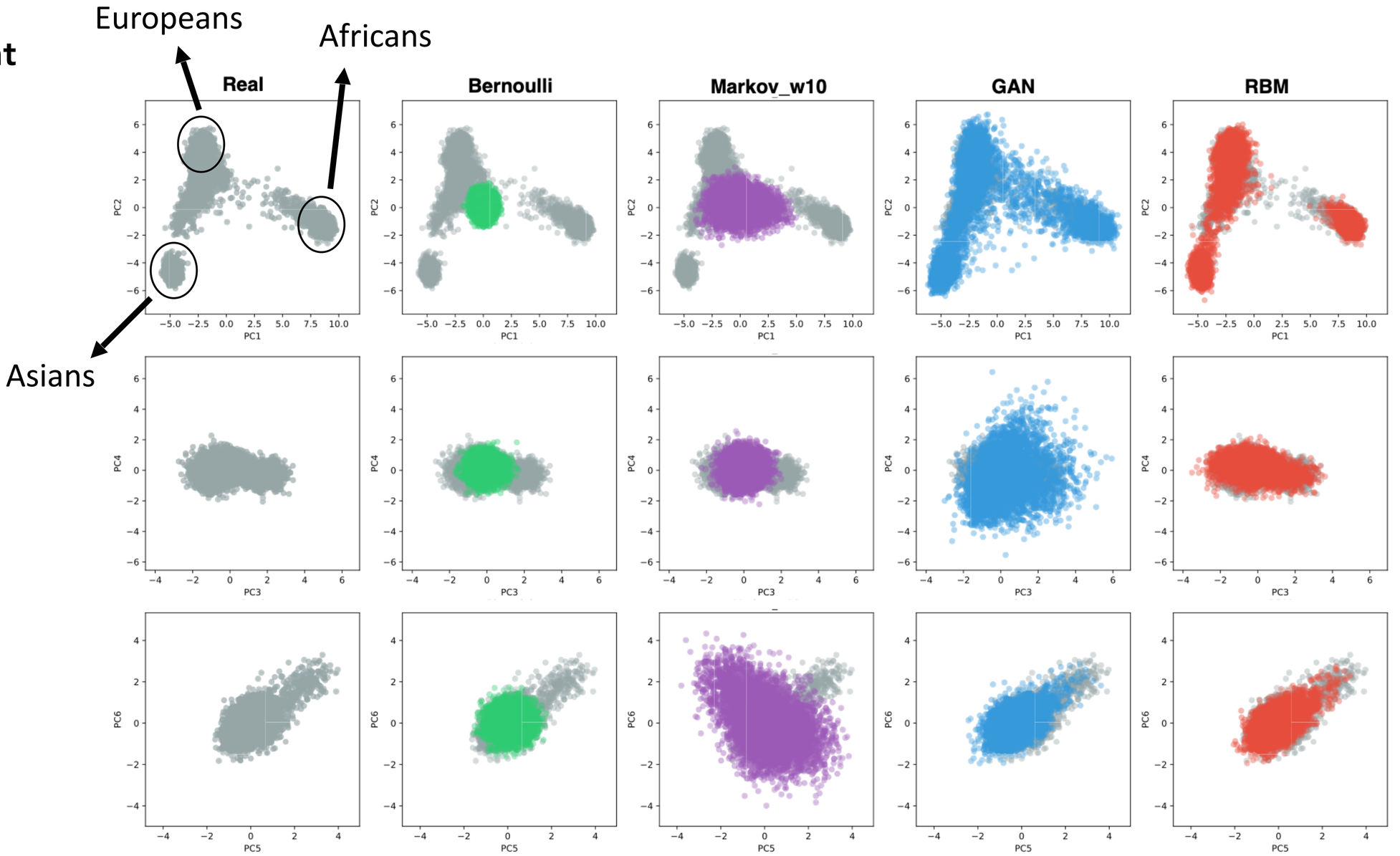
Aurélien Decelle

*Smolensky 1986; Teh and Hinton 2001

Model assessment

PCA

1000G dataset

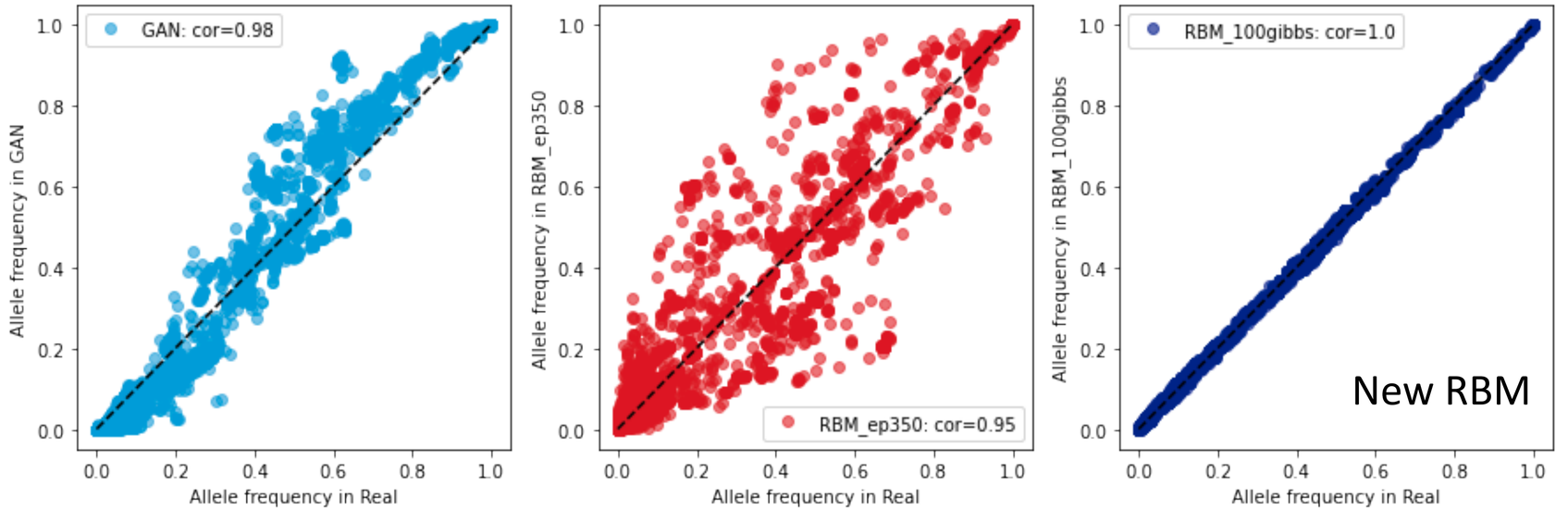


Model assessment

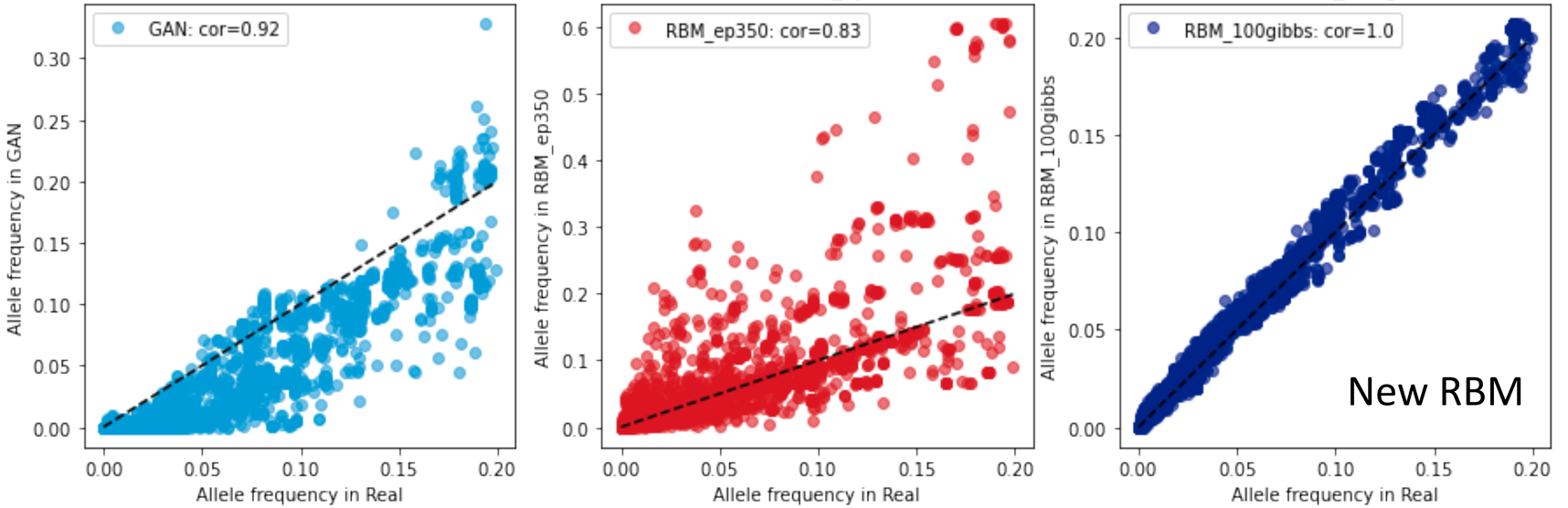
Allele frequencies

Estonian dataset

Genome1 0 1 1 …
Genome2 0 0 0 …
Genome3 1 0 0 …



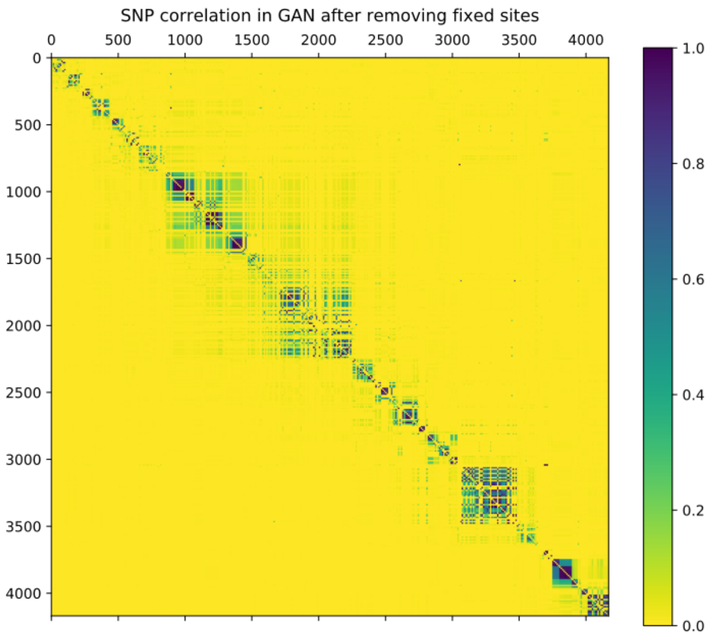
Zoom on low frequency features

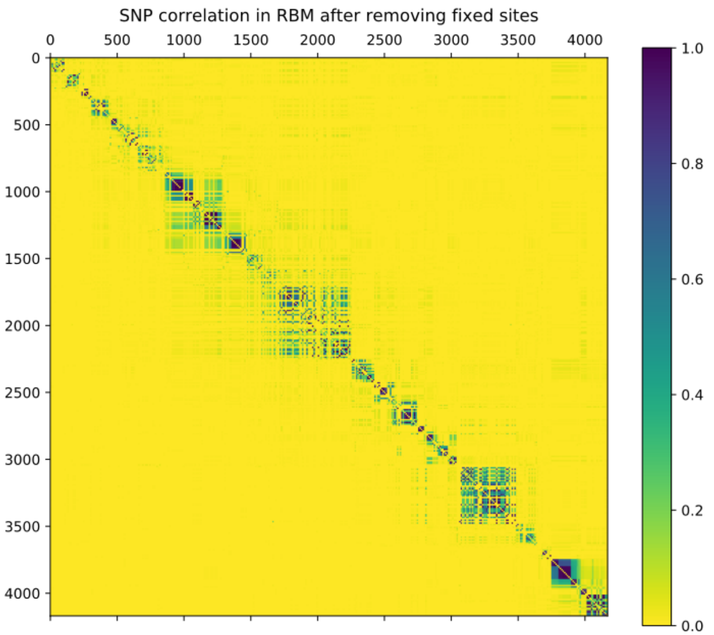


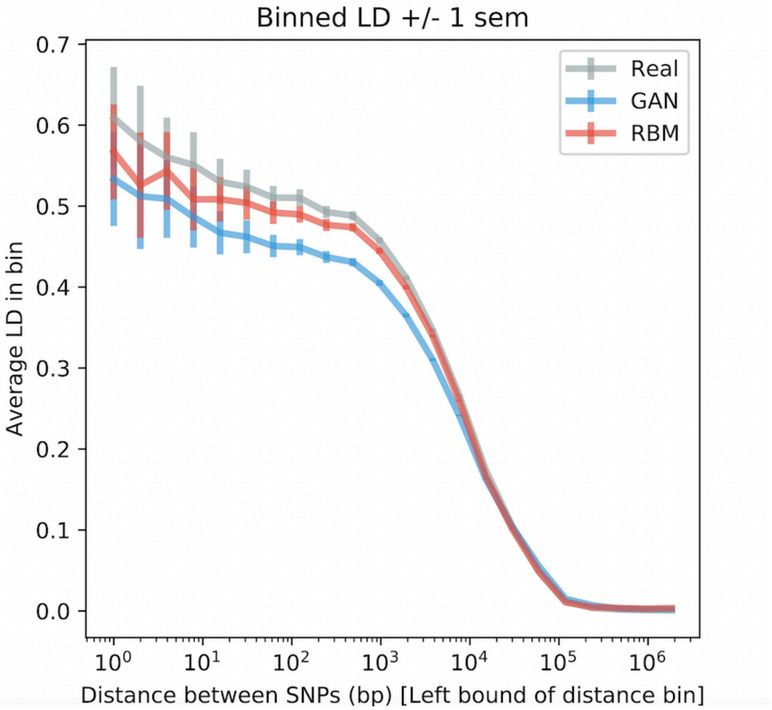
Model assessment

Linkage Disequilibrium (non-random association of alleles at different positions)

Estonian dataset







Model assessment

Overfitting/underfitting[1]

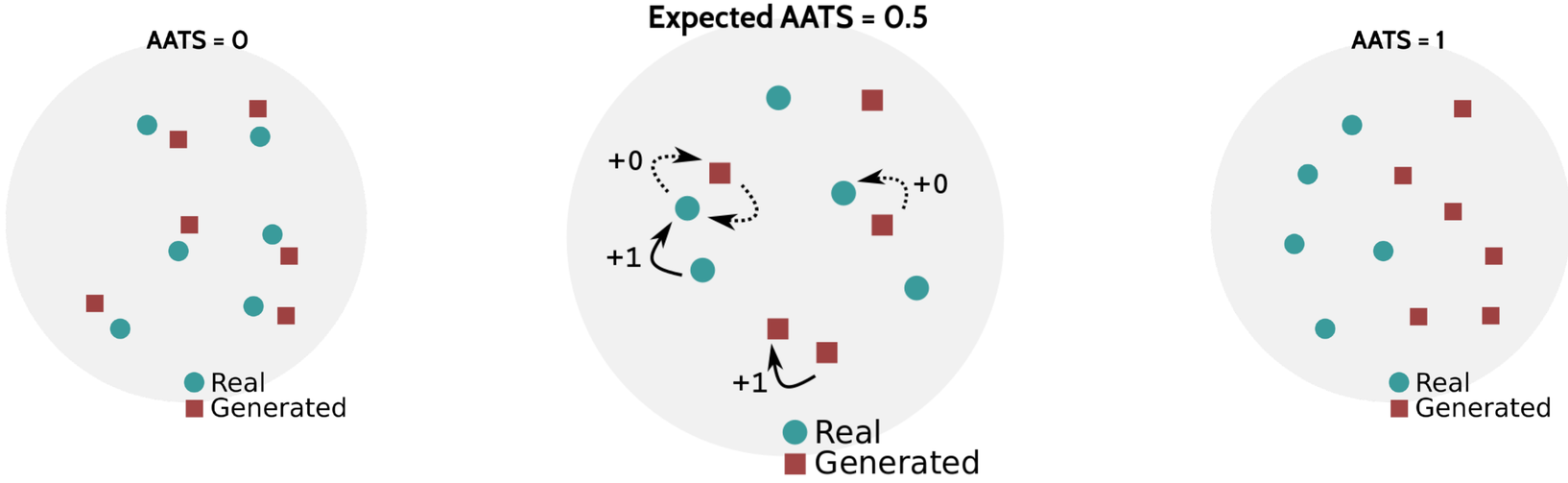
Estonian dataset

Nearest Neighbour Adversarial Accuracy (AA_{TS}) below 0.5 -> overfitting; above 0.5 -> underfitting

$$AA_{syn} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(d_{ST}(i) > d_{SS}(i)) \qquad AA_{truth} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(d_{TS}(i) > d_{TT}(i))$$

$$AA_{TS} = \frac{1}{2}(AA_{truth} + AA_{syn})$$

AATS: sum[is the closest neighbor of each **Real** or **Fake** individual of the same type (+1) or not (+0) ?]



1. Yale, Andrew, et al. "Privacy preserving synthetic health data." *ESANN 2019-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2019.

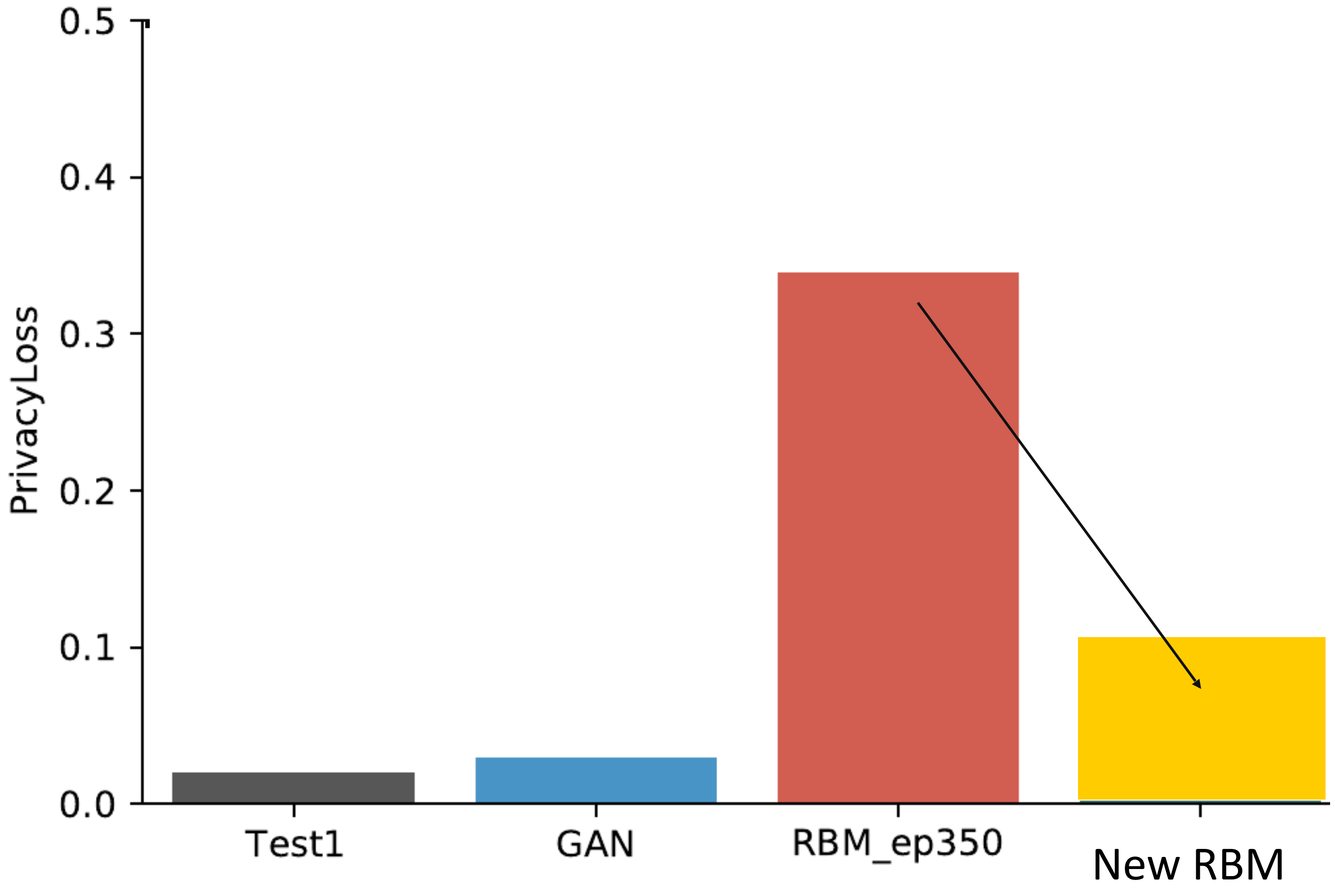
Model assessment

Privacy Scores[1]

Estonian dataset

Privacy loss -> higher values more information leakage

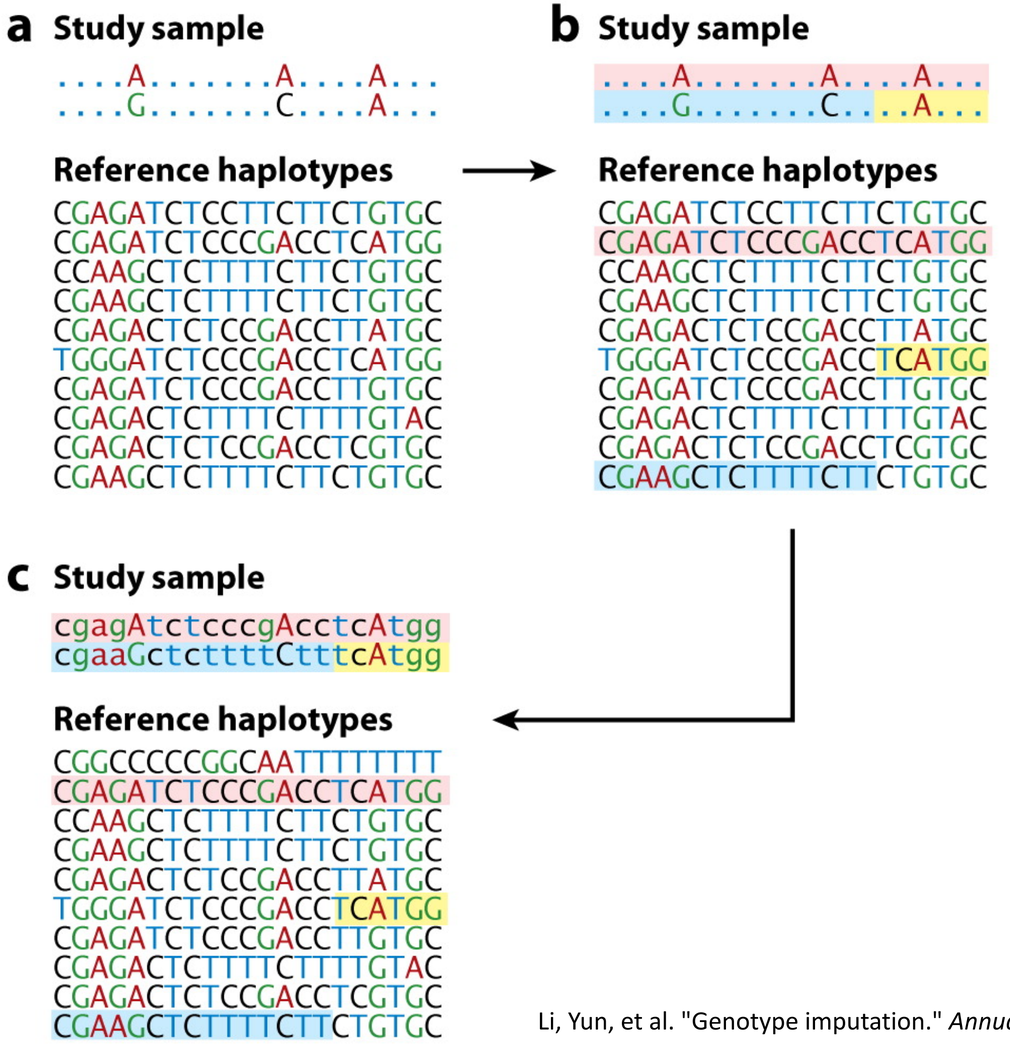
$$PrivacyLoss = TestAA_{TS} - TrainAA_{TS}$$



1. Yale, Andrew, et al. "Privacy preserving synthetic health data." *ESANN 2019-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2019.

Applications

Imputation (statistical inference of missing genotypes)

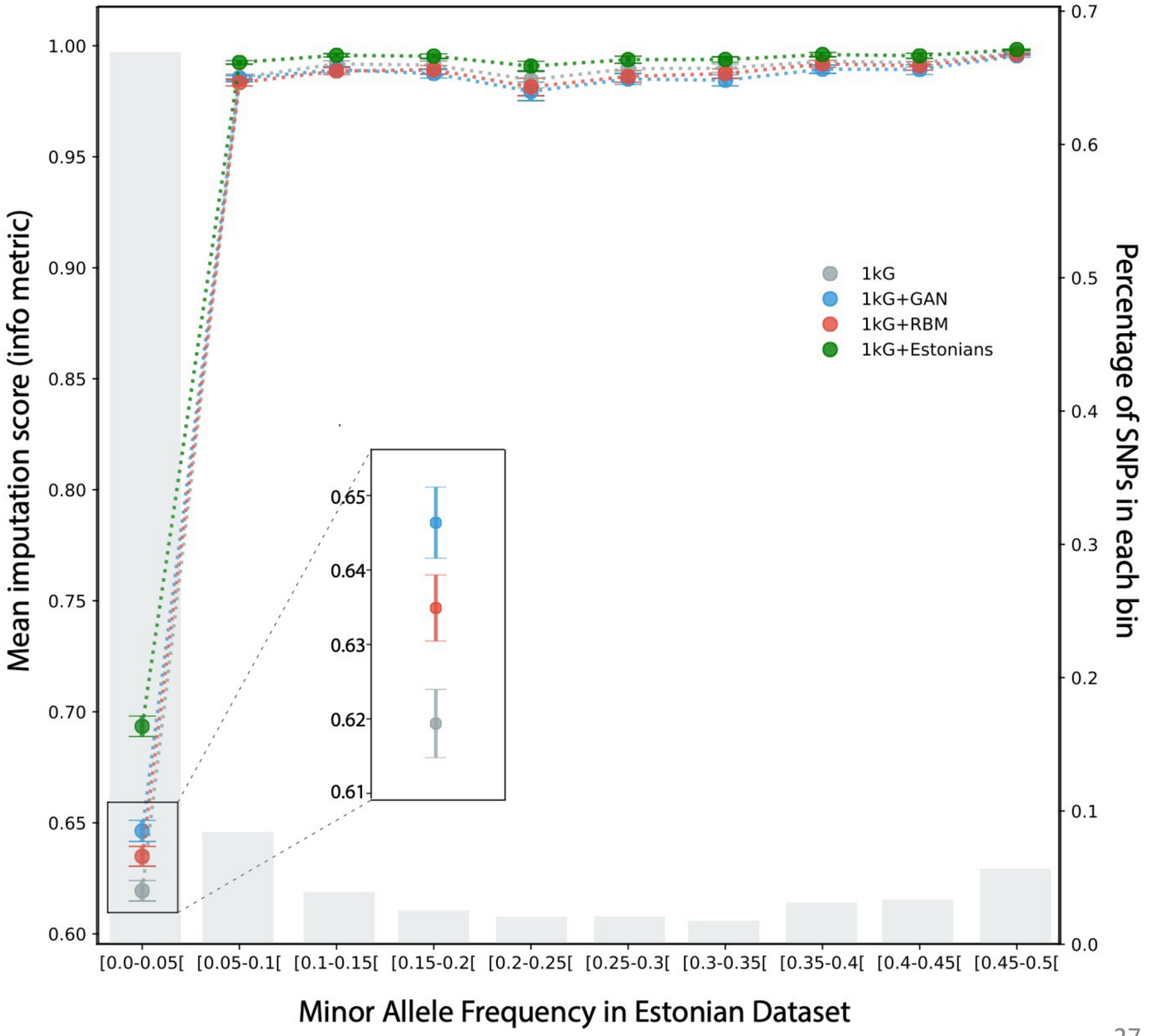


Li, Yun, et al. "Genotype imputation." *Annual review of genomics and human genetics* 10 (2009): 387-406.

Applications

Imputation (statistical inference of missing genotypes)
Estonian dataset

Imputation scores can be improved using additional population specific reference panels*



*Gurdasani et al. 2015; Mitt et al. 2017

Applications

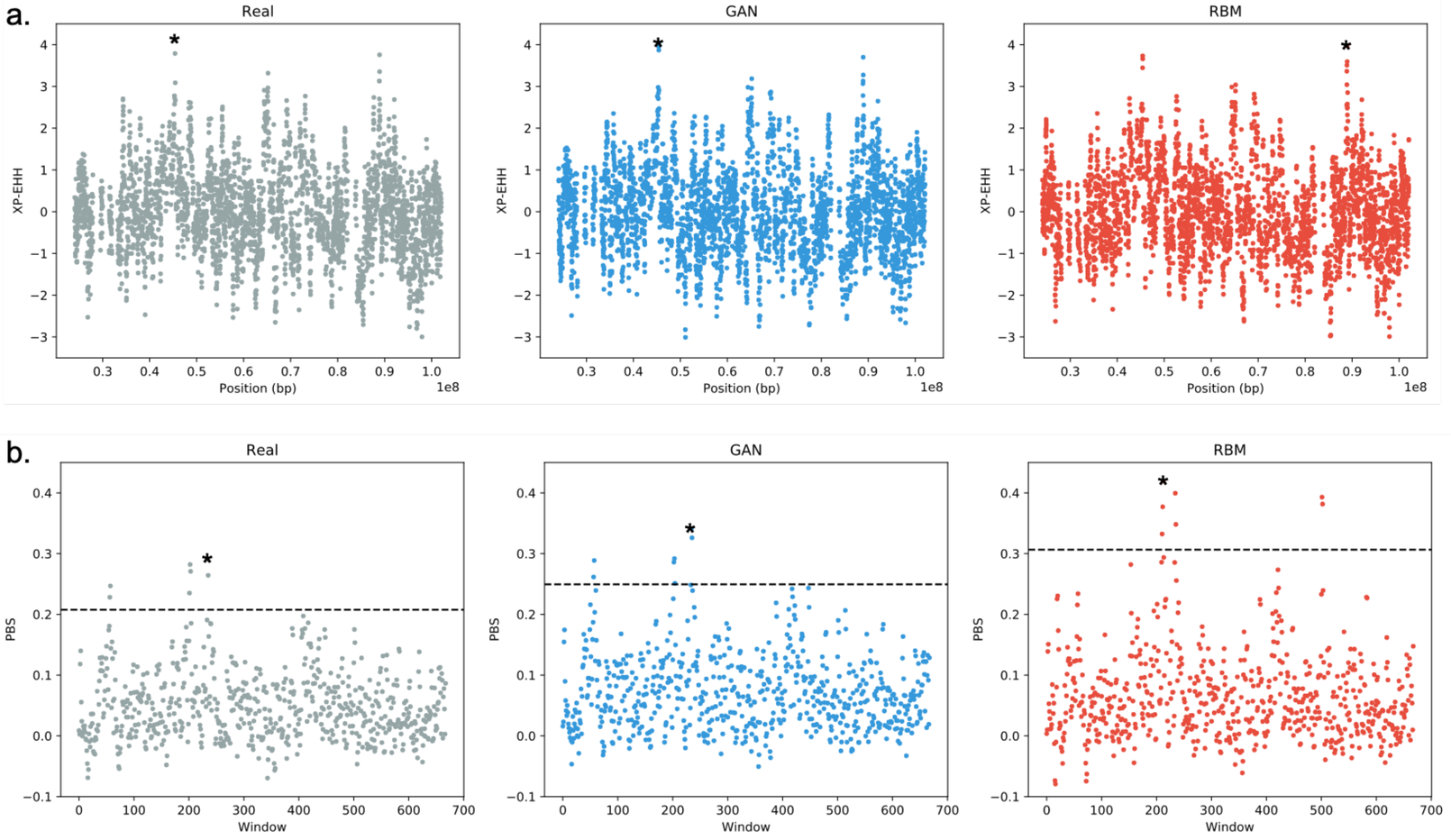
Selection scans

Estonian dataset

Artificial genomes preserve selection signals in real genomes detected by allele frequency-based (PBS) and haplotype-based (XP-EHH) methods.

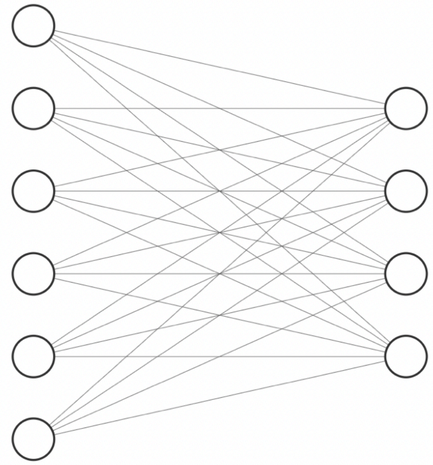
XP-EHH

PBS



Major obstacles for large sequence generation

Fully connected neural network



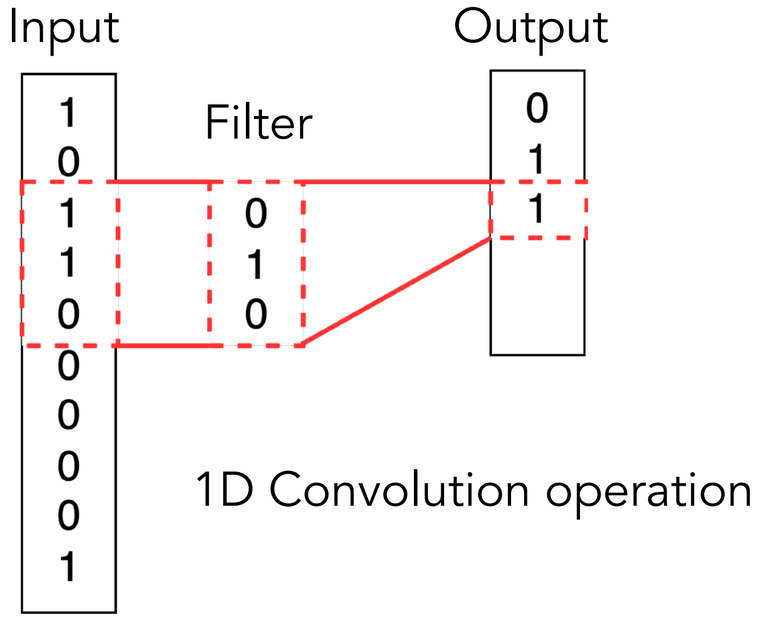
- **Computational complexity**
  - Number of parameters in fully connected GAN model for 10K SNP dataset:
    238 million

```
>>> 600*(10e3//1.2) + (10e3//1.2)*(10e3//1.1) + (10e3//1.1)*10e3 + 10e3*(10e3//2) + (10e3//2)*
(10e3//3) + (10e3//3)*1 + (10e3//1.2) + (10e3//1.1) + (10e3//2) + (10e3//3)
238340859.0
```

- **Training instability**
  - GAN training heavily depends on data, architecture, hyperparameters and even chance on rare occasions due to stochastic nature of the models

Possible solutions

- **Computational complexity**
  - **Convolutional architecture for GAN**
    - Around 7M parameters for 65K SNP data
    - Can learn local structures
  - **Conditional RBM**
    - Train the RBM for multiple chunks with shared regions



1D Convolution operation

- **Training instability**
  - **Wasserstein GAN (WGAN)**
    - Instead of “discriminator” (real or fake prediction) -> “critic” (realness score)
    - Critic tries to estimate Wasserstein distance (Earth-Mover distance) between real and generated distributions

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\, \|x - y\| \,\right]$$

RESEARCH ARTICLE

Deep convolutional and conditional neural networks for large-scale genomic data generation

Burak Yelmen[1,2]*, **Aurélien Decelle**[1,3], **Leila Lea Boulos**[1,4], **Antoine Szatkownik**[1], **Cyril Furtlehner**[1], **Guillaume Charpiat**[1], **Flora Jay**[1]

1 Université Paris-Saclay, CNRS, INRIA, LISN, Paris, France, **2** University of Tartu, Institute of Genomics, Tartu, Estonia, **3** Universidad Complutense de Madrid, Departamento de Física Teórica, Madrid, Spain, **4** Université d'Évry Val-d'Essonne, Évry-Courcouronnes, France

* burakyelmen@gmail.com

Yelmen, B., Decelle, A., Boulos, L. L., Szatkownik, A., Furtlehner, C., Charpiat, G., & Jay, F. (2023). Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLoS Comp. Bio*

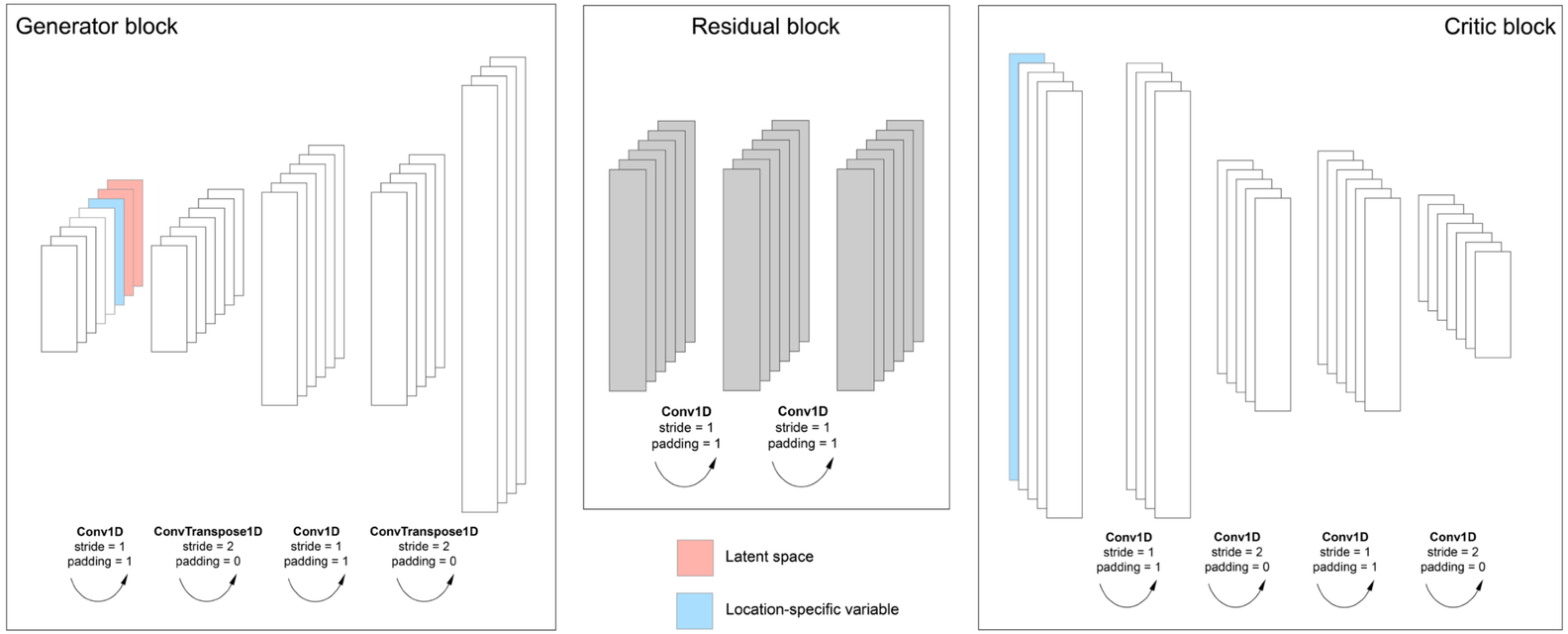
WGAN Model

WGAN-GP with **multiple noise** inputs at different resolutions for the generator, **trainable location-specific vectors** for preserving the positional information, **residual blocks** to prevent vanishing gradients and **packing** for the critic to eliminate mode collapse

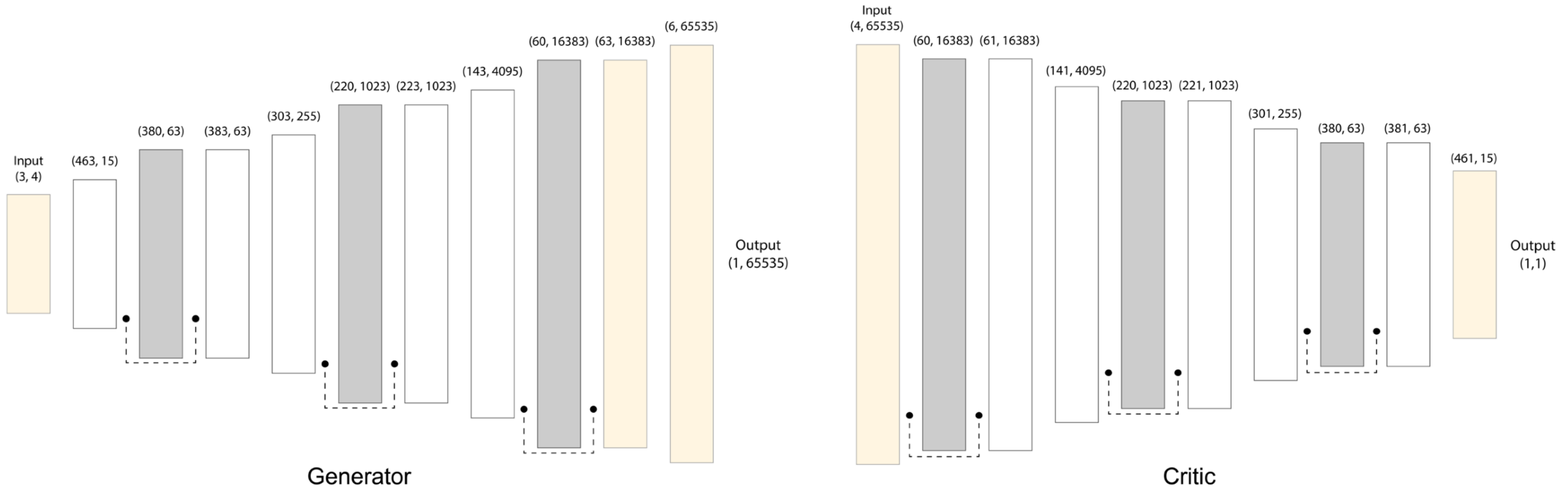
Objective function to be minimised by the generator and maximised by the critic:

$$E_x[C(x)] - E_z[C(G(z)]$$

a)

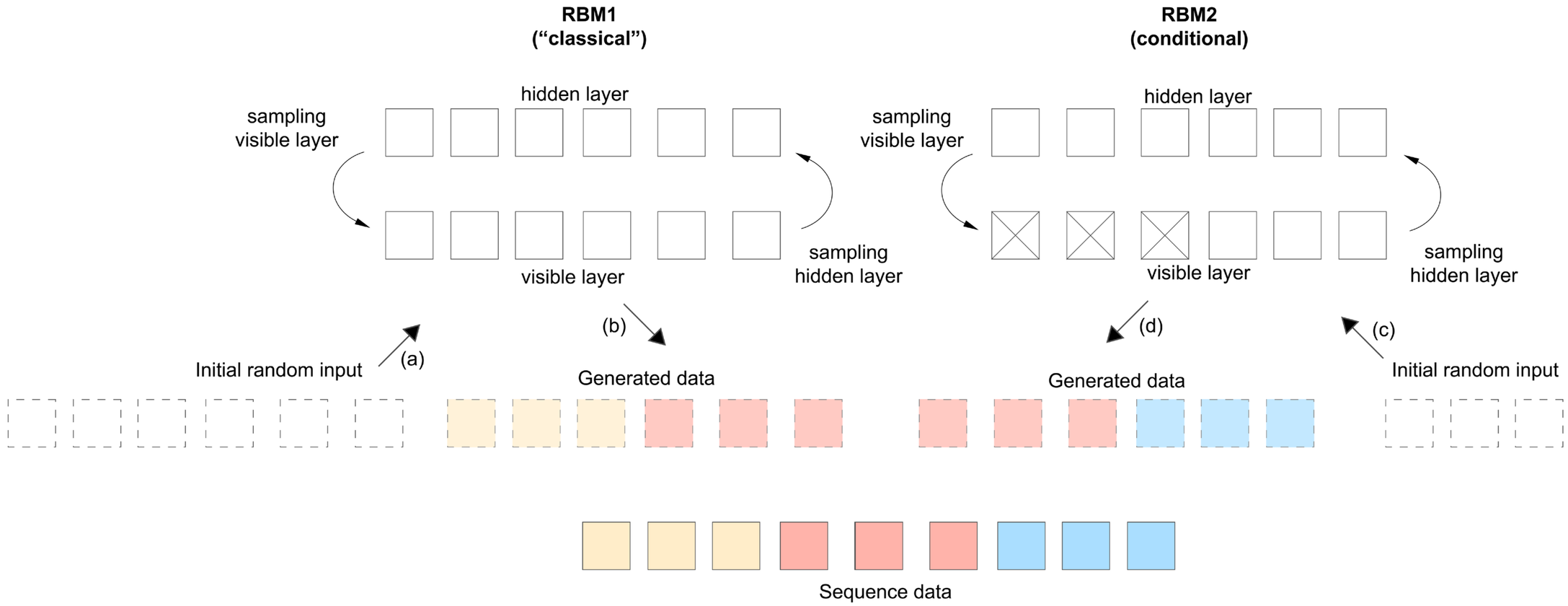


b)



CRBM Model

Conditional training multiple RBMs (**CRBM**) based on **shared genomic regions** with **out-of-equilibrium training** scheme



Aurélien Decelle

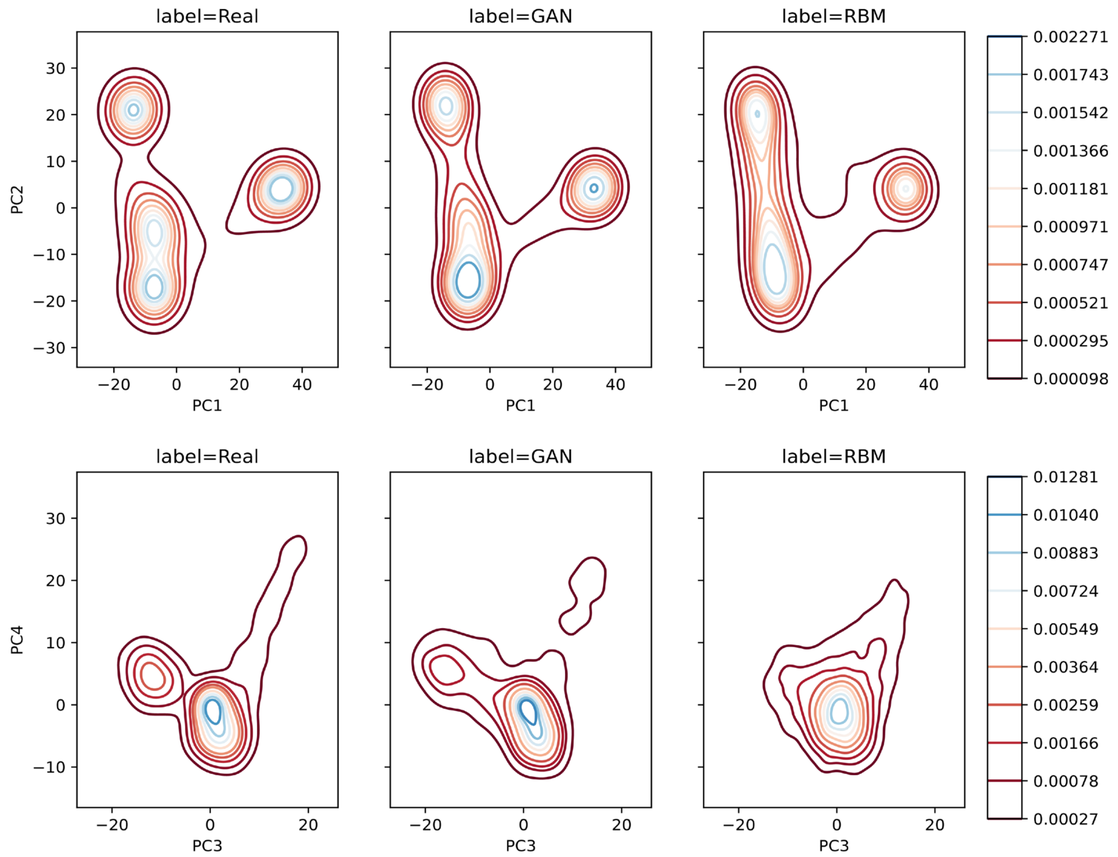
Model assessment

1000G dataset

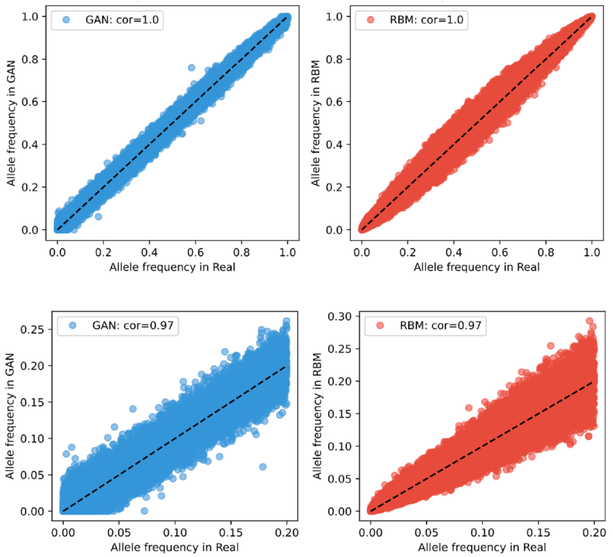
65,535 SNPs

Principal component **(a)**, allele frequency **(b)** and linkage disequilibrium (LD) decay **(c)** analyses of artificial genomes with 65,535-SNP size.

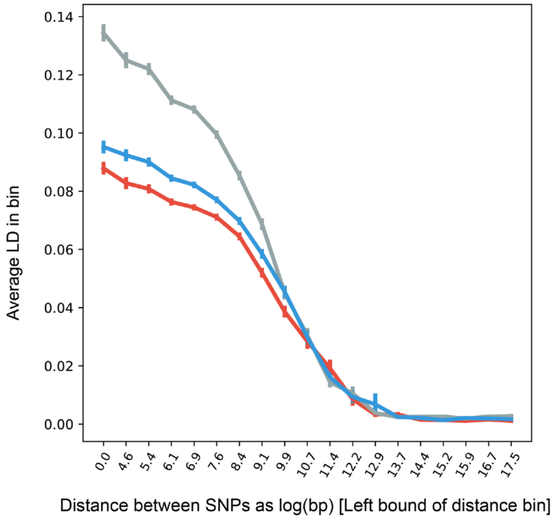
a)



b)



c)



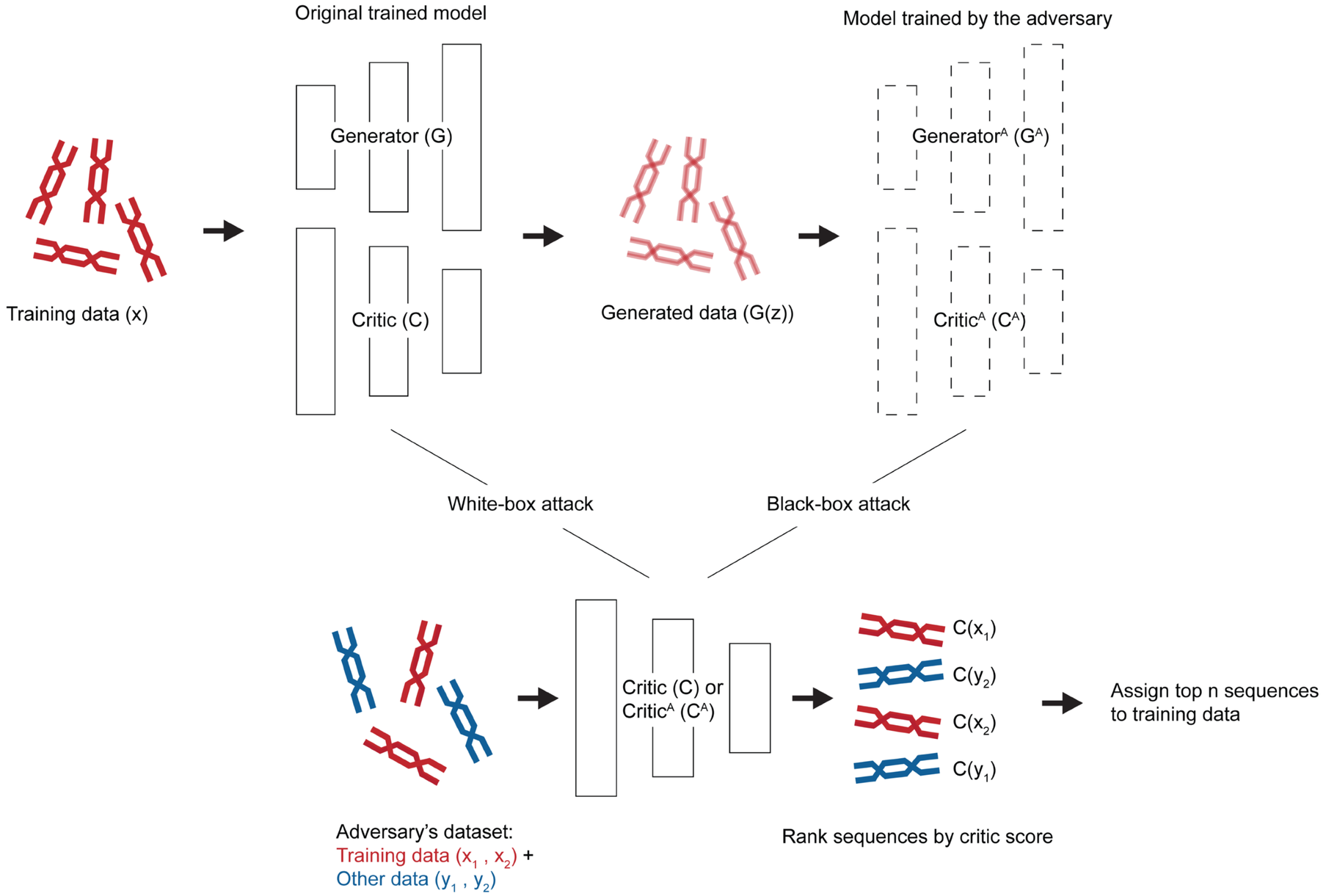
Privacy checks

1000G dataset
10,000 SNPs

Membership inference attacks: An adversary holds a collection of samples some of which are thought to be from the training data. The adversary tries to detect these sequences.

White-box attack: Adversary has full access to the model

Black-box attack: Adversary has access to the model architecture and generated samples but not the weights



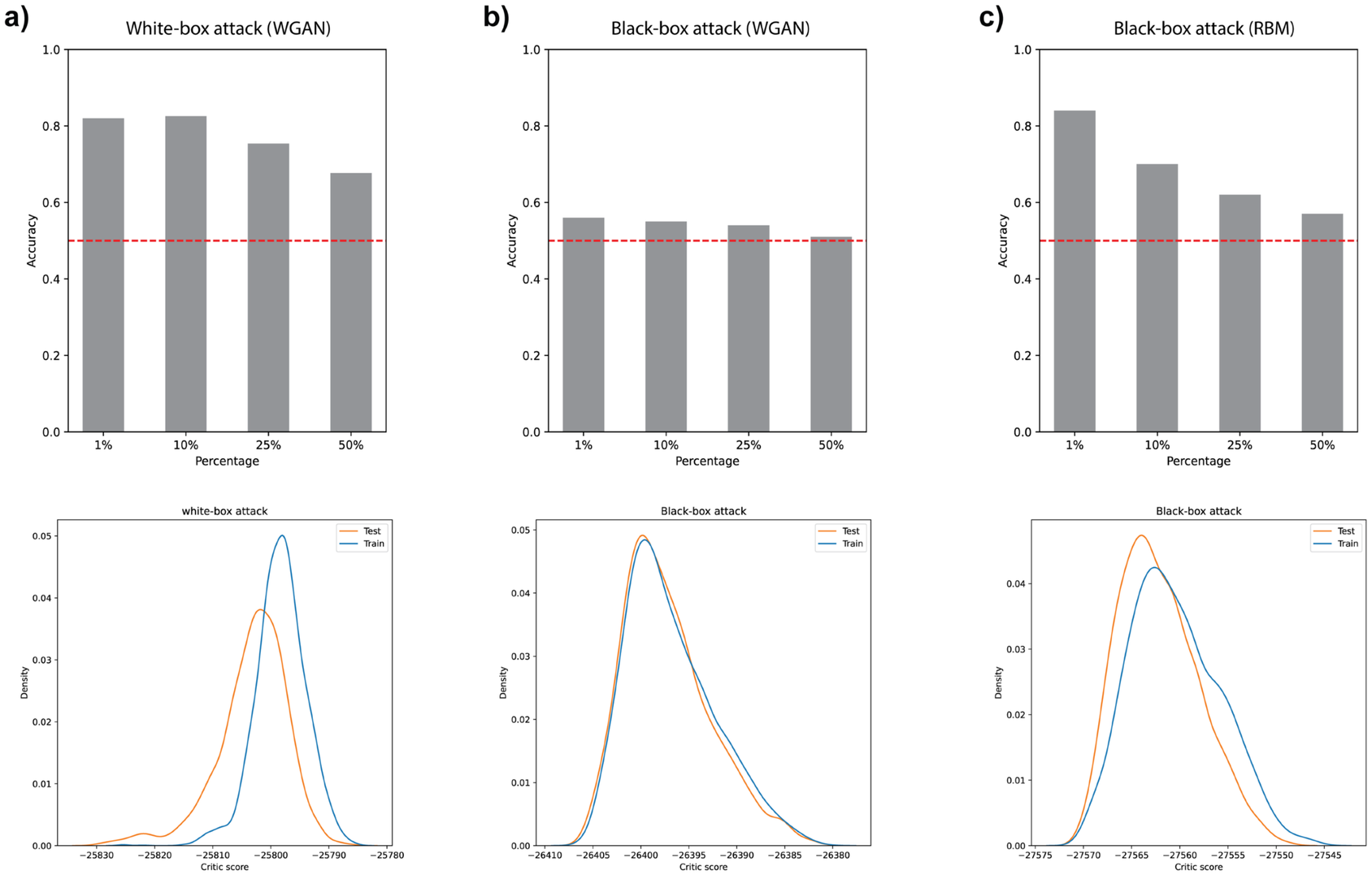
Hayes, J., Melis, L., Danezis, G., & De Cristofaro, E. (2017). Logan: Membership inference attacks against generative models.

Privacy checks

1000G dataset
10,000 SNPs

White-box attack: Adversary has full access to the model

Black-box attack: Adversary has access to the model architecture and generated samples but not the weights



Uncertainty quantification

Why measure uncertainties?

1. **Genome-wide uncertainty** for model improvement and selection of generated genomes
2. **Position-specific uncertainty** for model improvement and potential discovery
3. Further evidence for **ethical** and **regulatory compliance** in real life applications

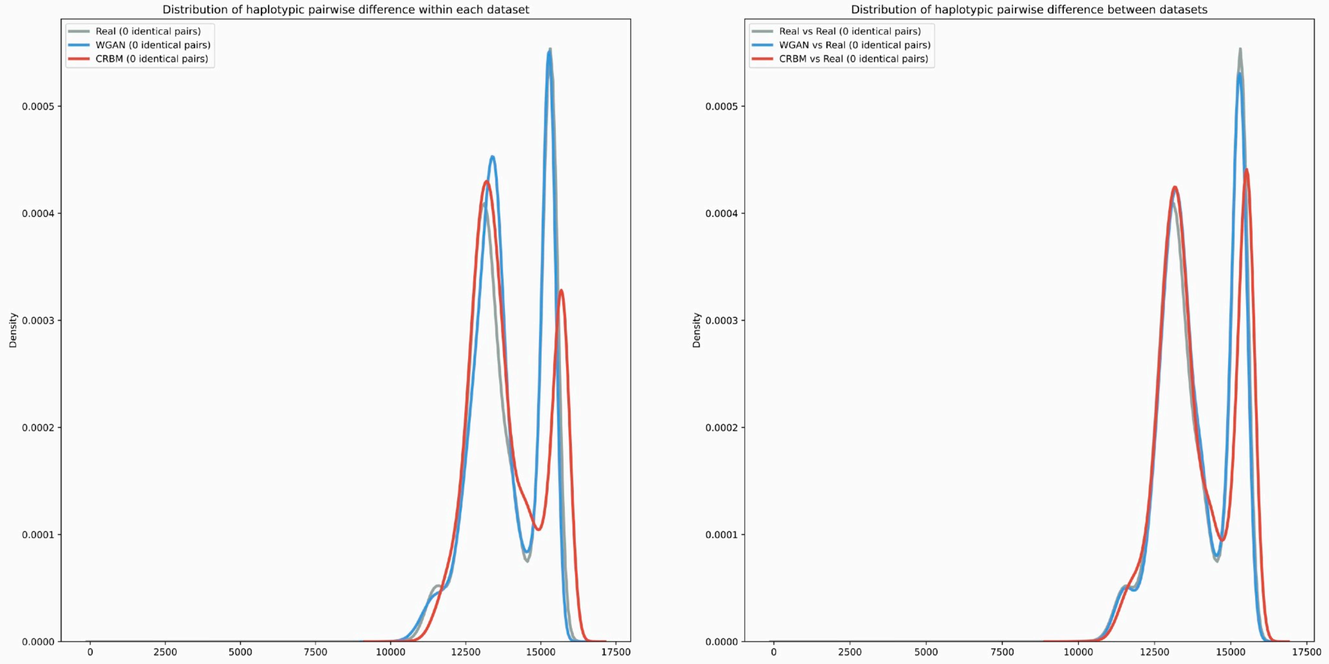
Unique challenges in uncertainty quantification for artificial genomics

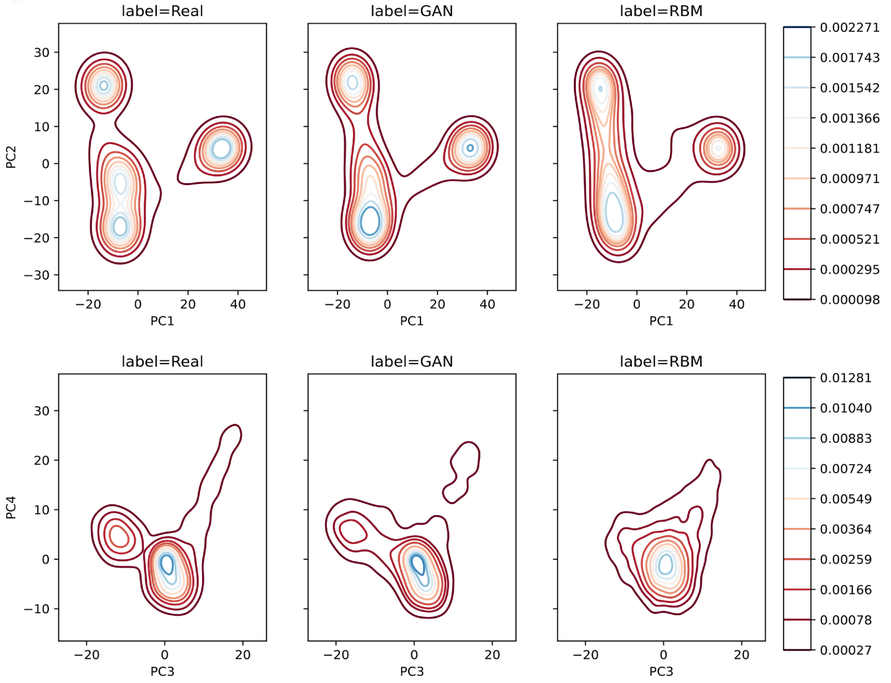
1. Highly correlated features
2. Very high-dimensional data
3. Potential trade-off between **privacy preservation** and **data uncertainty**
4. Potential trade-off between **novel haplotypes** and **data uncertainty**

Uncertainty quantification

1000G dataset

Indirect assessment of **genome-wide** data uncertainty by model evaluation:
Models capture the inherent genomic variability well

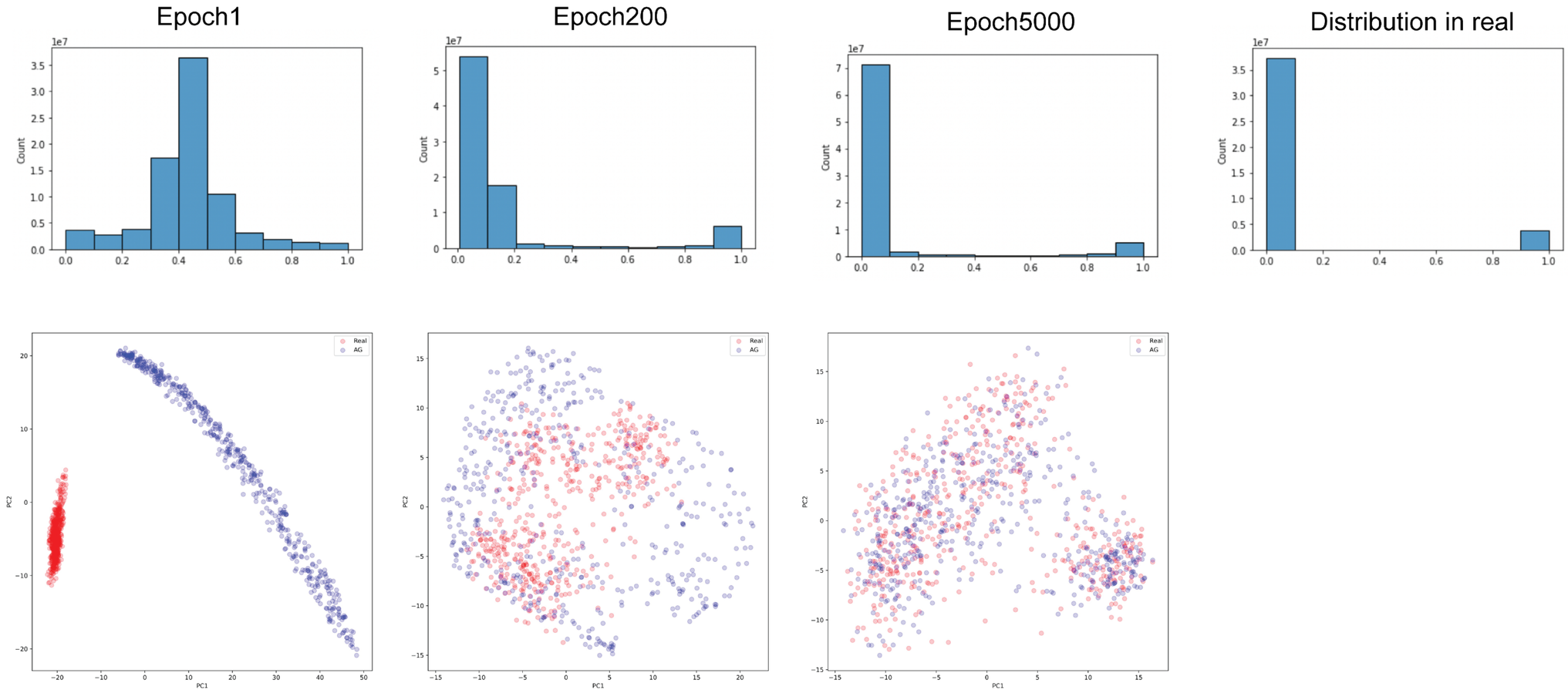




Uncertainty quantification

1000G dataset

A preliminary assessment of **position-specific** data uncertainty for the GAN model:
Distribution of output probabilities over epochs

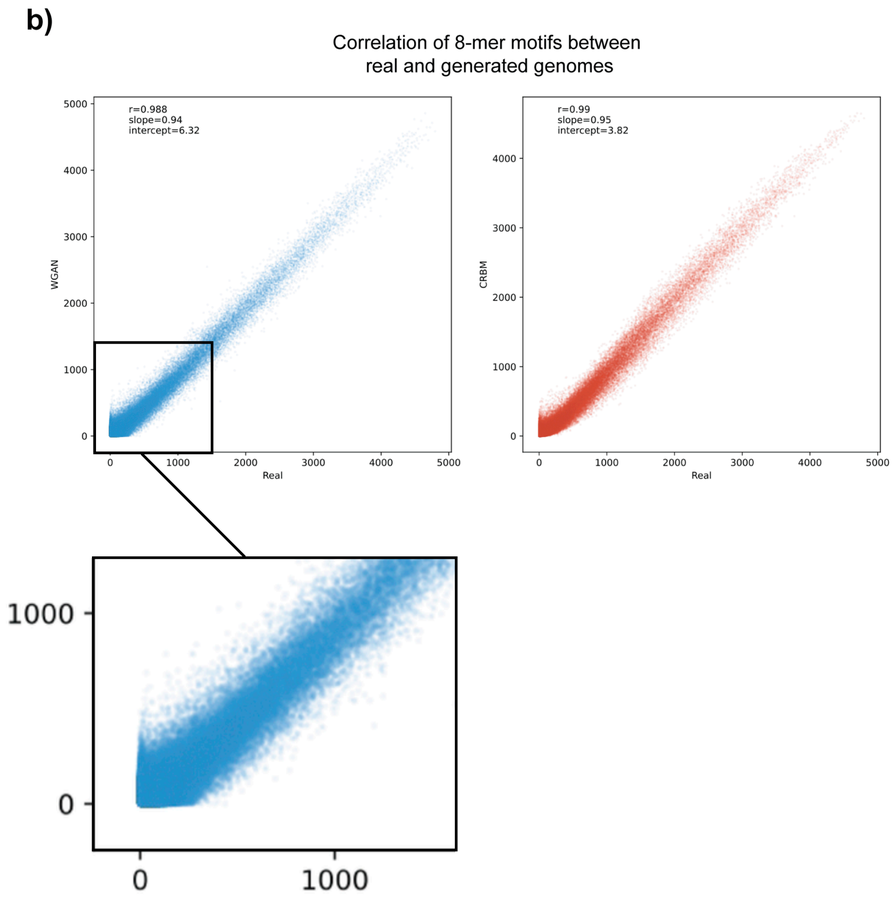


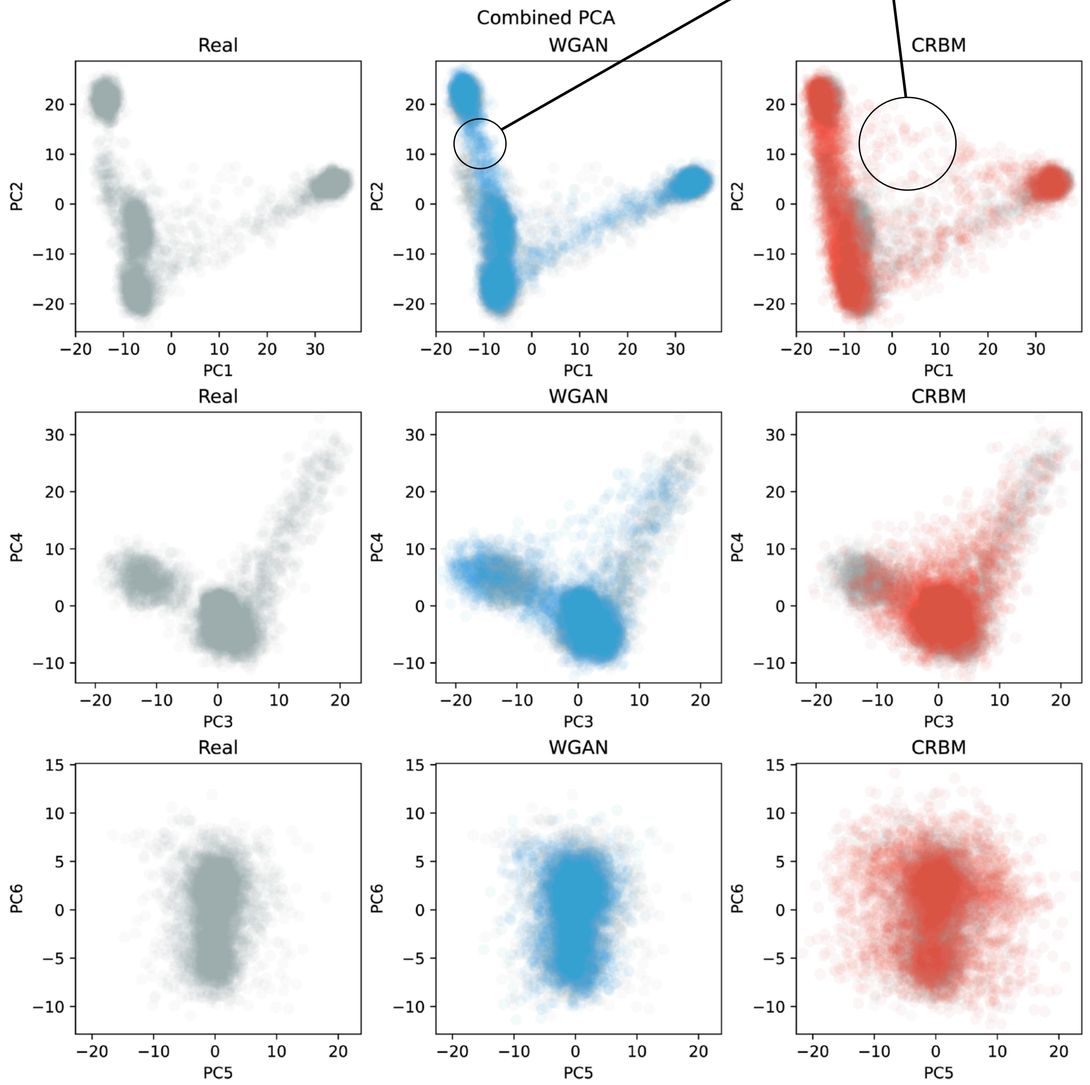
Unlike prediction tasks, we cannot assess out-of-distribution examples for the generator

Uncertainty quantification

Potential trade-off between novel haplotypes and uncertainty

Potentially plausible admixed artificial genomes?





Closing remarks

- **Artificial Genomics**: Newborn field with many promising applications in the future? (from functional sequence design to whole-genome generation)

- **Artificial genome banks** can soon become a reality with improved **haplotype quality** and **privacy guarantees**, increasing **data accessibility**

- For uncertainty quantification, possible future routes are **Bayesian** methods (might not be feasible), **ensembles/bagging, MC dropout** or **variational inference (checking variability in VAE latent space?)**

- Many computational and algorithmic challenges remain for modelling **high-dimensional space** with **complex interactions** -> generative modelling in reduced space?*

* **Towards creating longer genetic sequences with GANs: Generation in principal component space**

Antoine Szatkownik[1], **Cyril Furtlehner**[1], **Guillaume Charpiat**[1], **Burak Yelmen**[1,2,*], **Flora Jay**[1,*]
[1] Université Paris-Saclay, CNRS, INRIA, LISN, Paris, France
[2] University of Tartu, Institute of Genomics, Tartu, Estonia
* These authors contributed equally.
Corresponding author: Antoine Szatkownik <szatkownik@lisn.fr>

Acknowledgements

- Flora Jay
- Aurélien Decelle
- Guillaume Charpiat
- Cyril Furtlehner
- Antoine Szatkownik
- Leila Léa Boulos
- Linda Ongaro
- Davide Marnetto
- Luca Pagani
- Francesco Montinaro
- Corentin Tallec
- Inria TAU team
- University of Tartu HPC Center

cnrs dépasser les frontières

Inria

Paris-Saclay Center for Data Science

anr agence nationale de la recherche

UNIVERSITY OF TARTU Institute of Genomics

université PARIS-SACLAY

LISN

UNIVERSITY OF TARTU Institute of Molecular and Cell Biology

REALMENT

CoMorMent

The CoMorMent project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant agreement 847776

Supplementary slides

Uncertainty quantification

Potential trade-off between privacy preservation and data uncertainty: A superficial connection with differential privacy

A randomised algorithm ***G*** is ϵ-differentially private if for all datasets T_1 and T_2 differing on at most one element and for all sets ***S*** of possible outputs, the following holds:

$$P(G(T_1) \in S) \leq e^{\varepsilon} \times P(G(T_2) \in S)$$

Since differentially privacy is typically achieved by adding noise and added noise increases data uncertainty, we can write uncertainty as a function of epsilon:

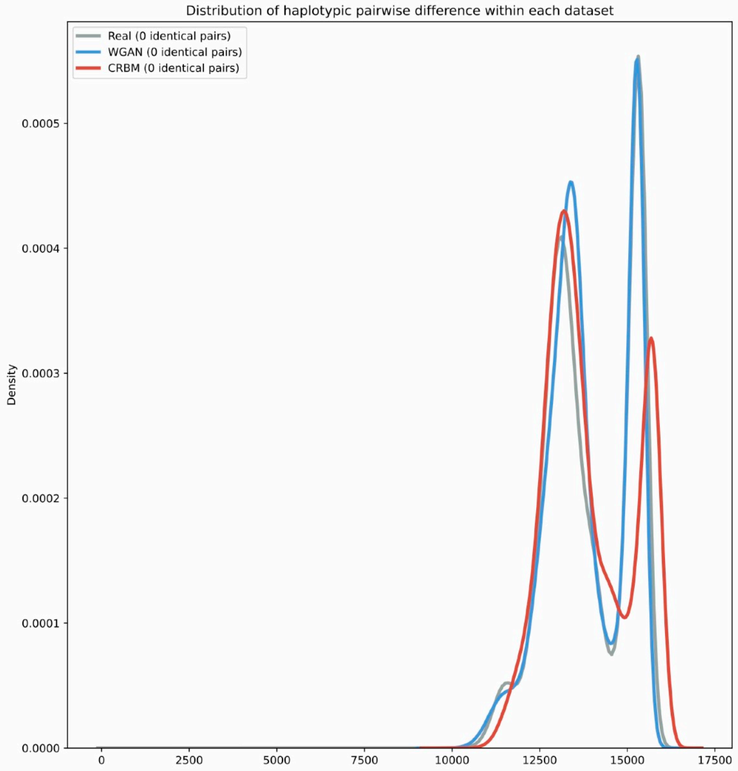
$$U(\widehat{Y}|Y, \varepsilon) = f(\varepsilon)$$

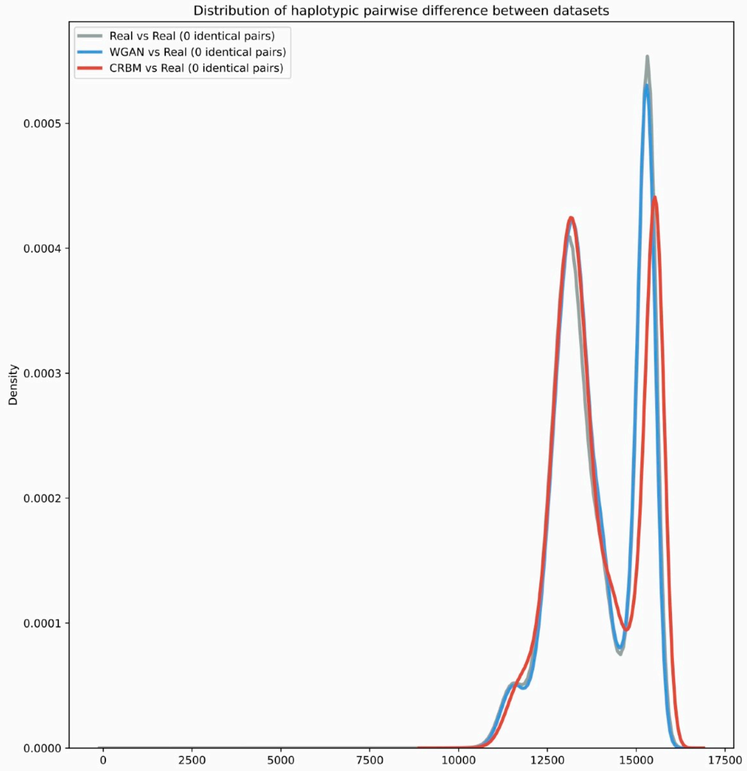
Privacy checks

1000 Genomes

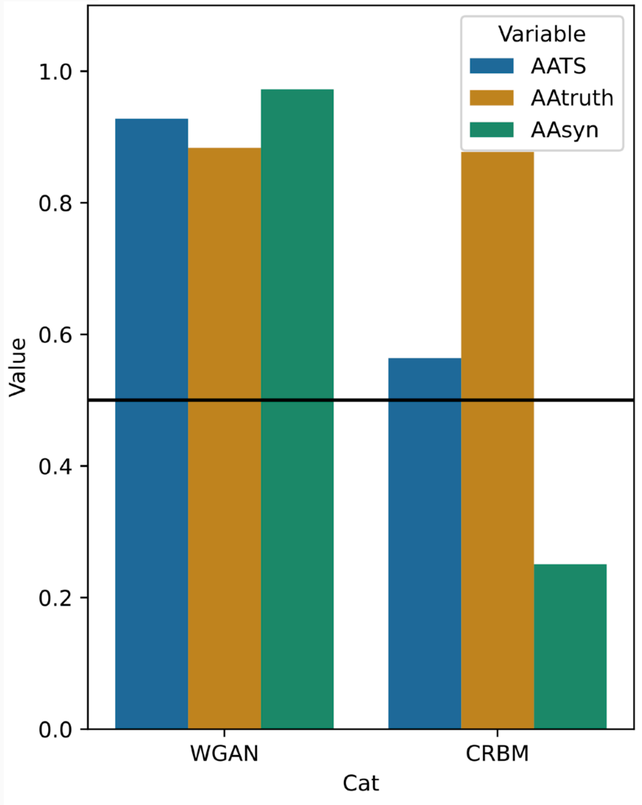
65,535 SNPs

Distribution of **haplotypic pairwise difference** within (left) and between (right) 65,535-SNP datasets.

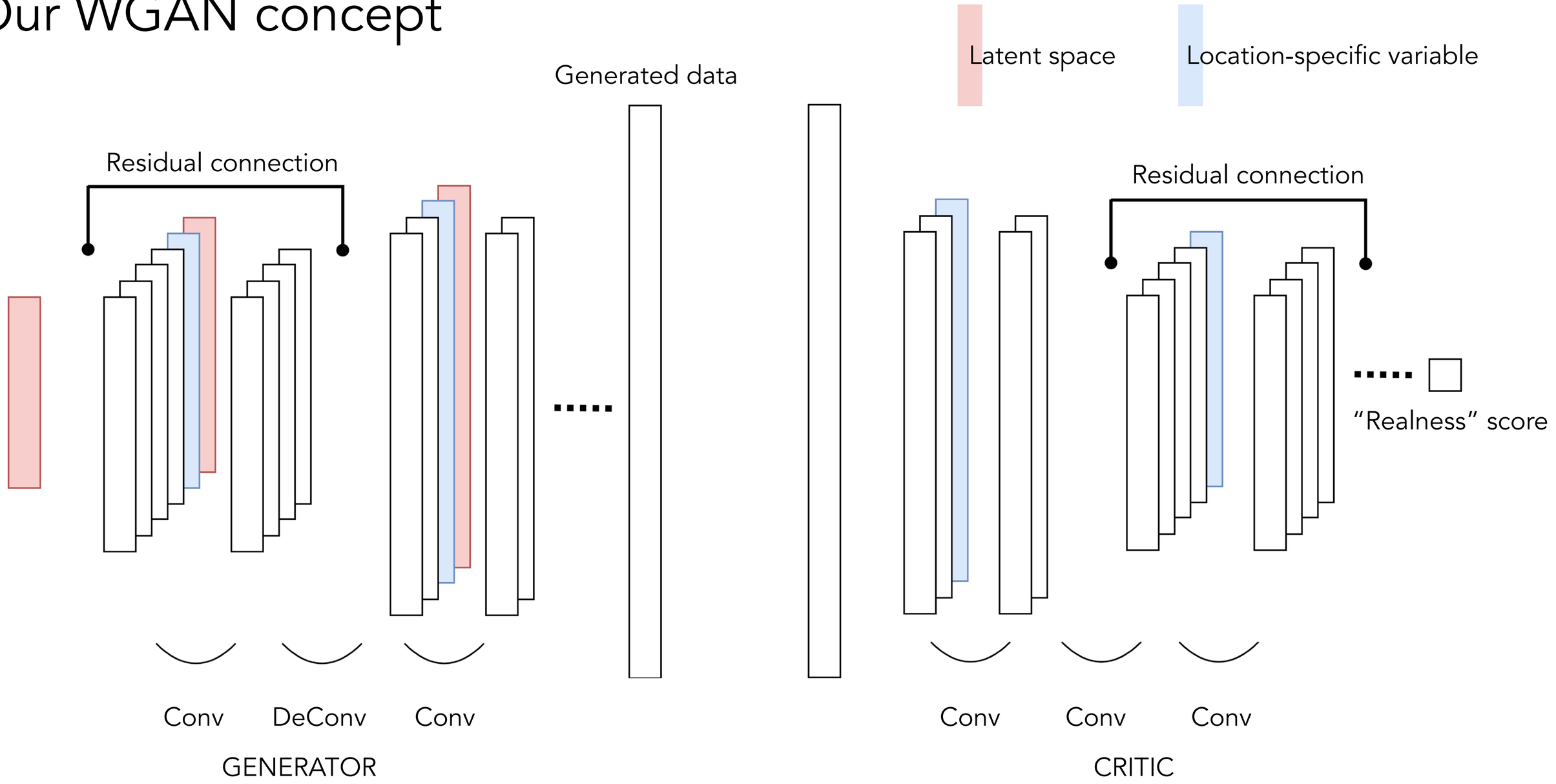




Nearest neighbour adversarial accuracy (AATS)



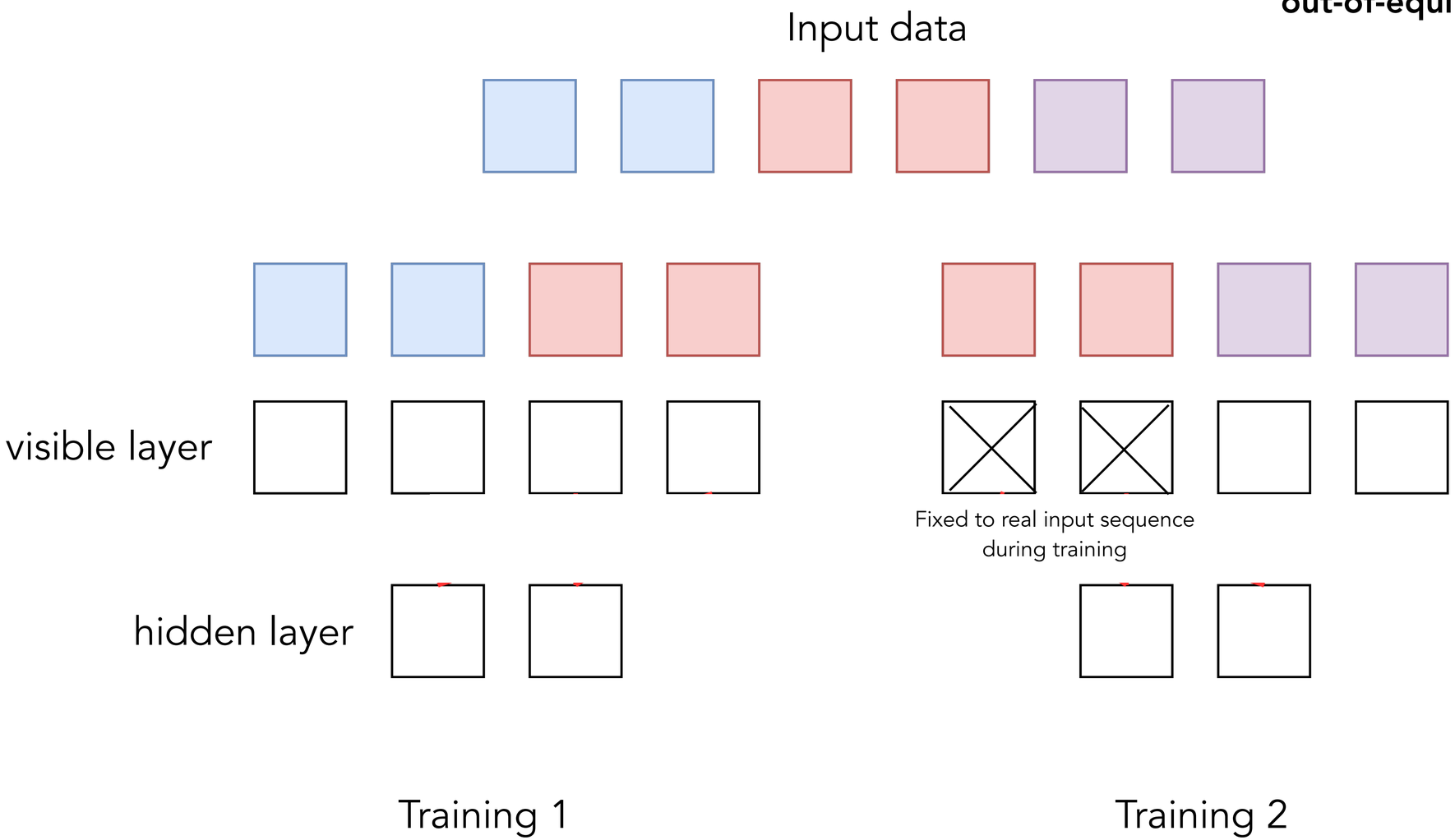
Our WGAN concept



Ausmees, K., & Nettelblad, C. (2022). A deep learning framework for characterization of genotype data. G3

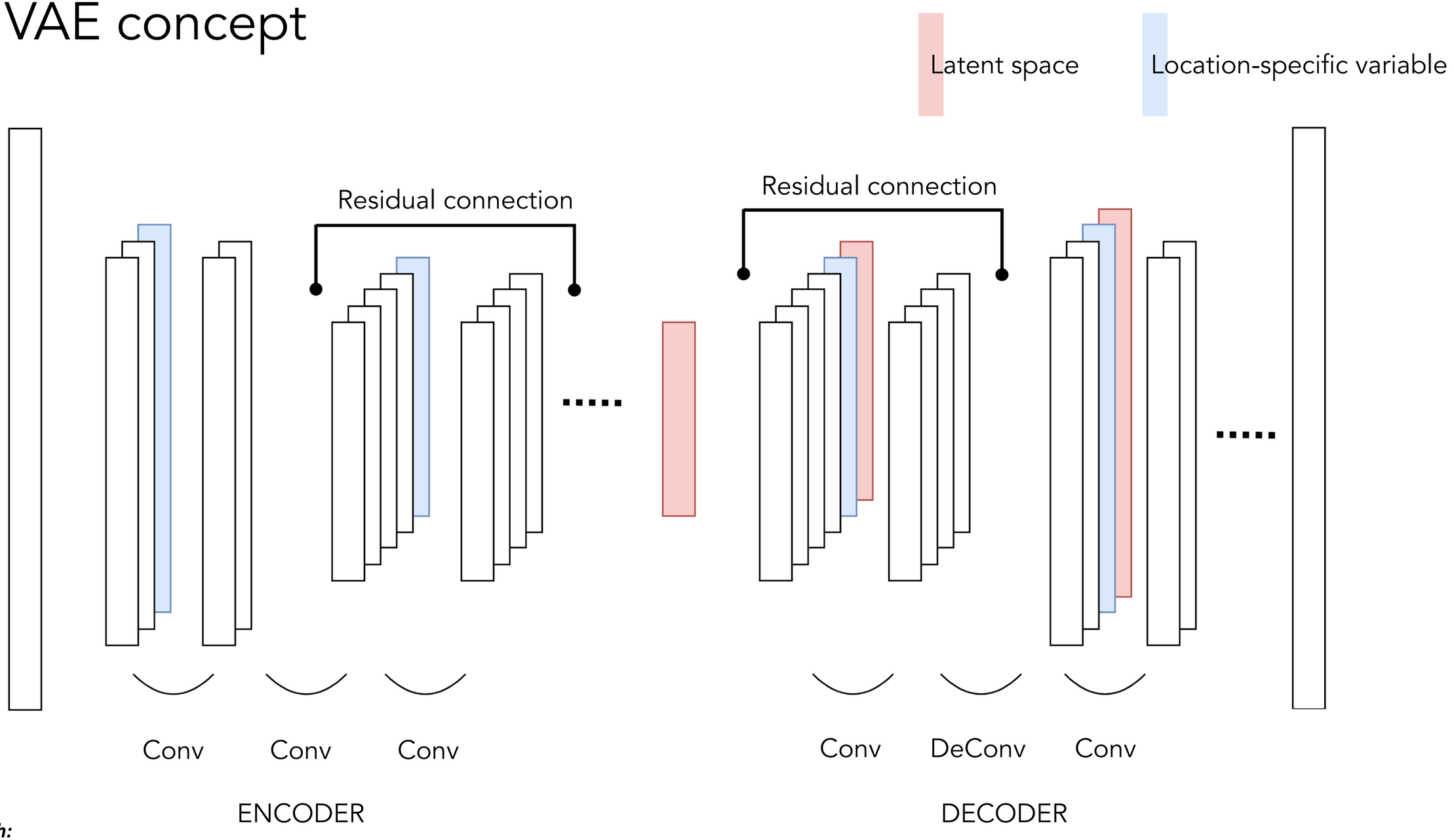
Our conditional RBM (CRBM) concept

Conditional training multiple RBMs (**CRBM**) based on **shared genomic regions** with **out-of-equilibrium training** scheme



Aurélien Decelle

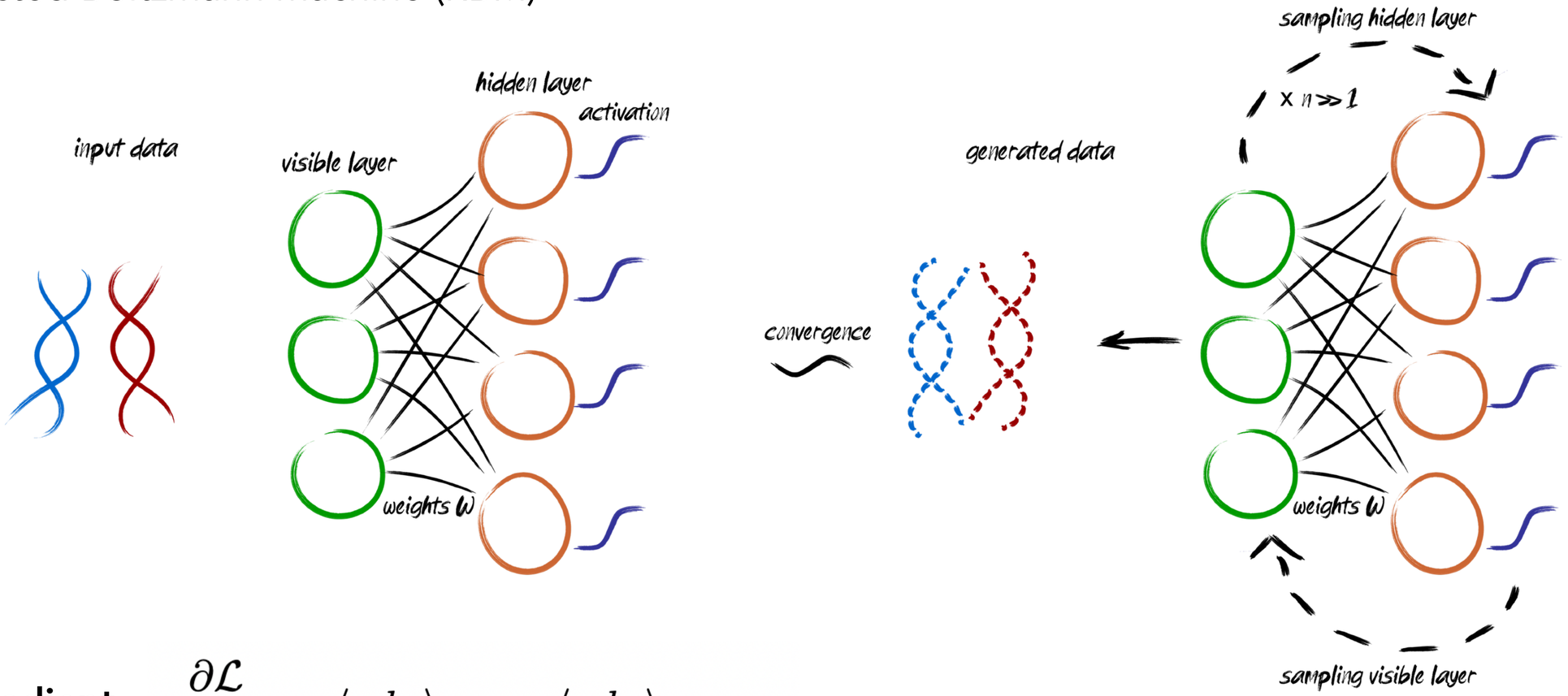
Our VAE concept



Related research:

Battey et al. 2021 - Visualizing population structure with variational autoencoders Ausmees et al. 2021 - A deep learning framework for characterization of genotype data

Restricted Boltzmann machine (RBM)*



Gradient $\frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{RBM}$

Monte Carlo to estimate

PCD-k

MC chains start from previous state

Rdm-k (**New RBM**)

MC chains start from random initial conditions (some fixed distribution)

*Smolensky 1986; Teh and Hinton 2001