Gene Diffusion

Philip Kenneweg[1] Rangaram Dandinasivara Raghuram[1] Alexander Schönhuth[1] Barbara Hammer [1]

[1] University Bielefeld

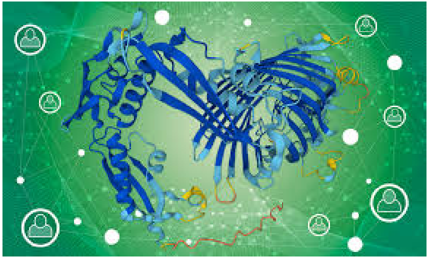
24-02-2023, Bielefeld

Introduction

Motivation

There is great promise in **disease prediction**, **functional understanding**, **drug synthesis** and other applications of artificial intelligence in the field of genetics.

(a) disease prediction

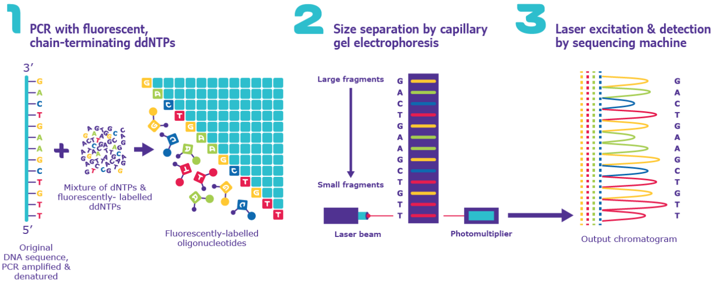


(b) protein folding

Challenges

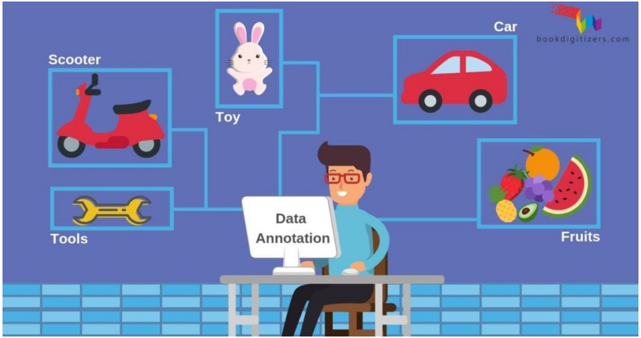
Very sensitive personal data, which should not be distributed.

Challenges


1 PCR with fluorescent, chain-terminating ddNTPs
2 Size separation by capillary gel electrophoresis
3 Laser excitation & detection by sequencing machine
3′
5′
G A C T G A A G C T G T T
Original DNA sequence, PCR amplified & denatured
Mixture of dNTPs & fluorescently- labelled ddNTPs
Fluorescently-labelled oligonucleotides
Large fragments
Small fragments
Laser beam
Photomultiplier
Output chromatogram

Genetic data is expensive to obtain, even though sequencing it is getting cheaper each year.

Challenges



Annotated data is costly to obtain and labels are unreliable, Especially true for medical data.

Idea

We want to adapt learning paradigms from the recently, successful area of natural language and image processing. I.e. processing large scales of unlabeled data which enable models to be trained with few labeled examples.

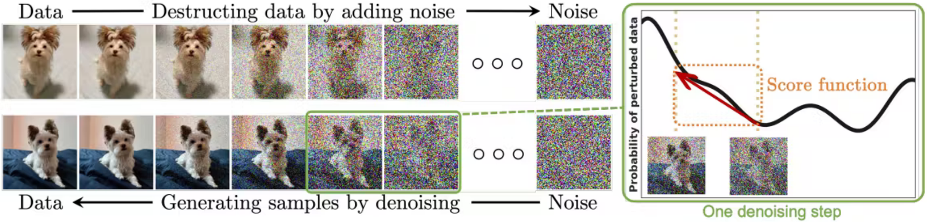
Idea

We are taking a closer look at **diffusion models.** These kind of models can help us with:

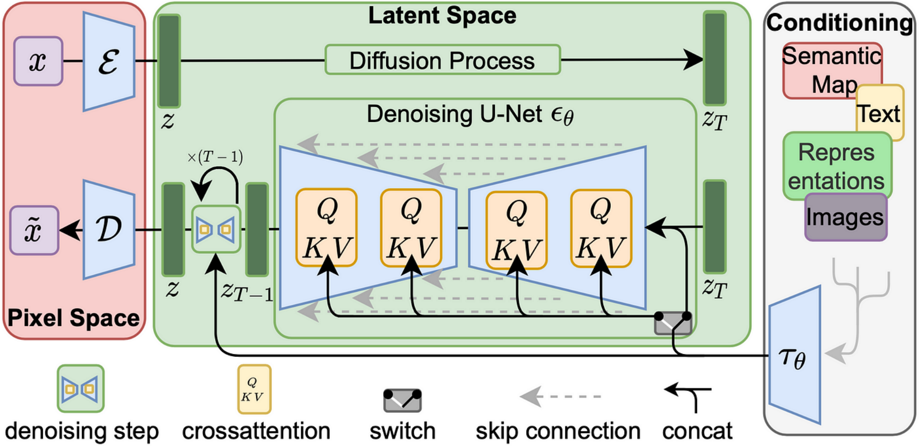
- **Semi-supervised learning**: Through the diffusion process we can learn on data without labels.
- **Privacy**: Generalizing from sensitive personal data to synthetic data.

Background

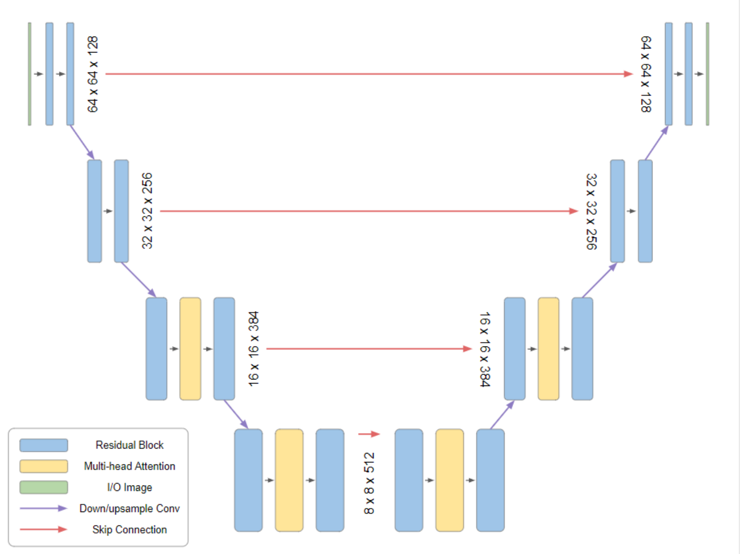
Diffusion Models


Data
Destructing data by adding noise
Noise
○ ○ ○
○ ○ ○
Data
Generating samples by denoising
Noise
Probability of perturbed data
Score function
One denoising step

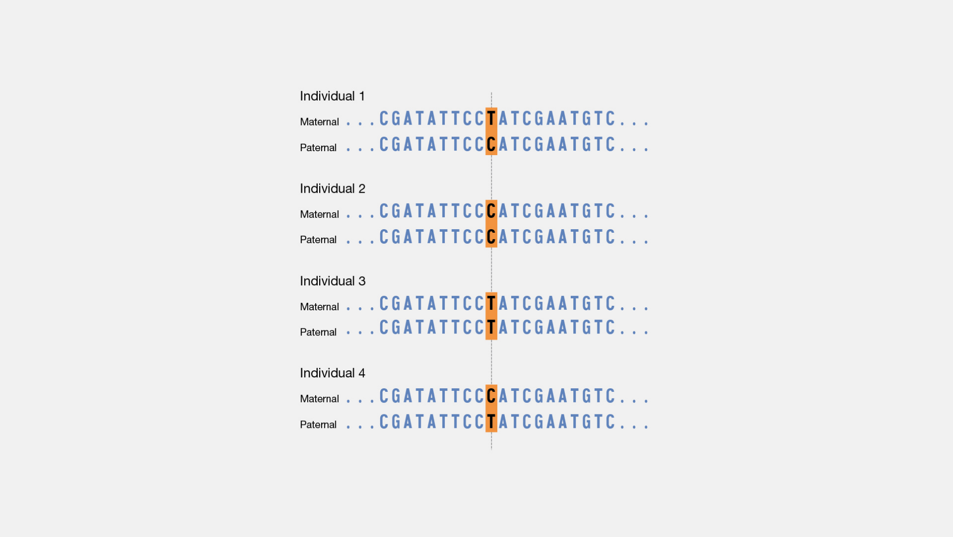
Diffusion Models



Architecture Unet 1D with Attention



Genetic Data SNPs



Individual 1

Maternal ...CGATATTCCTATCGAATGTC...

Paternal ...CGATATTCCCATCGAATGTC...

Individual 2

Maternal ...CGATATTCCCATCGAATGTC...

Paternal ...CGATATTCCCATCGAATGTC...

Individual 3

Maternal ...CGATATTCCTATCGAATGTC...

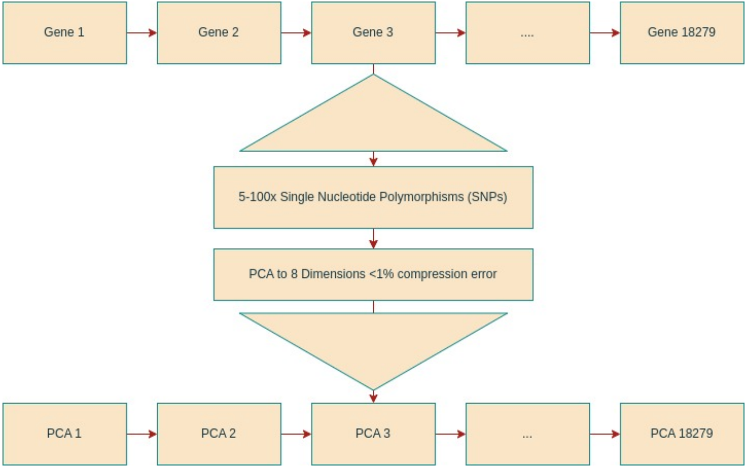
Paternal ...CGATATTCCTATCGAATGTC...

Individual 4

Maternal ...CGATATTCCCATCGAATGTC...

Paternal ...CGATATTCCTATCGAATGTC...

Genetic Data preprocessing



Gene 1
Gene 2
Gene 3
....
Gene 18279

5-100x Single Nucleotide Polymorphisms (SNPs)

PCA to 8 Dimensions <1% compression error

PCA 1
PCA 2
PCA 3
...
PCA 18279

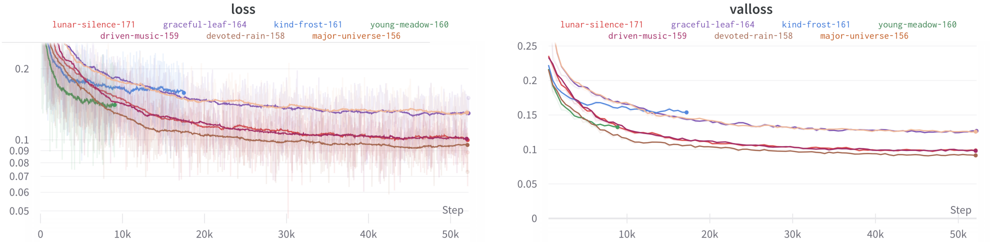
Latent Space Analysis

To evaluate the Diffusion Model it is paramount to look at a variety of metrics. Since high dimensional gene pca data is not understandable for humans we use a variety of other metrics.

- loss curves
- UMAP
- classification task performance (ALS disease detection)
- diversity measures (closest sample etc.)

Loss Curves

echnology



(a) train

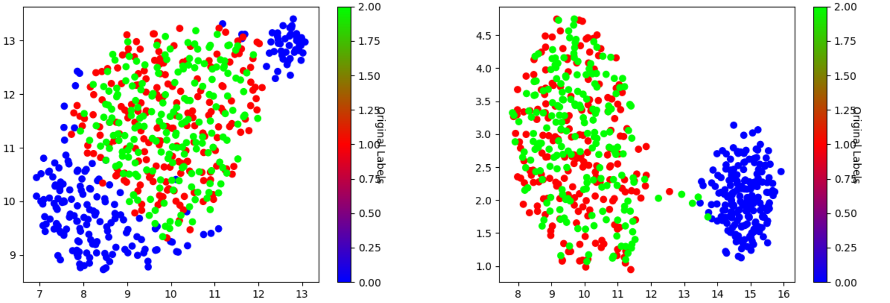
(b) validation

Figure: Performance differences for different network architectures

In general:

- Bigger = Better (no overfitting observed)
- some additional preprocessing beneficial (a custom dense layer for each pca)

UMAPs

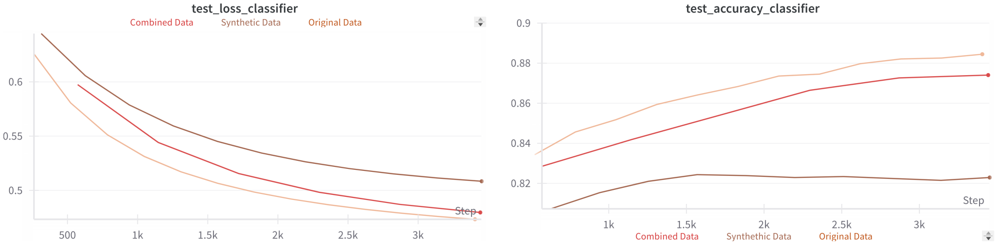


(a) MLP

(b) CNN

Figure: UMAPs for different network backbones using cosine similarity as a distance metric. Euclidean distance UMAPs are not informative.

ALS classifier performance

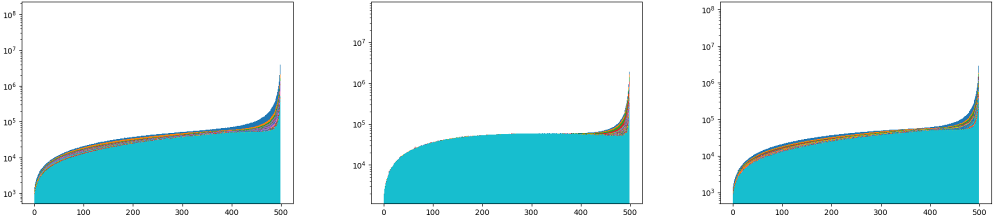


(a) test loss

(b) test acc

Figure: Comparison between synthetic and real training data. Performance is measured on a hold out test set.

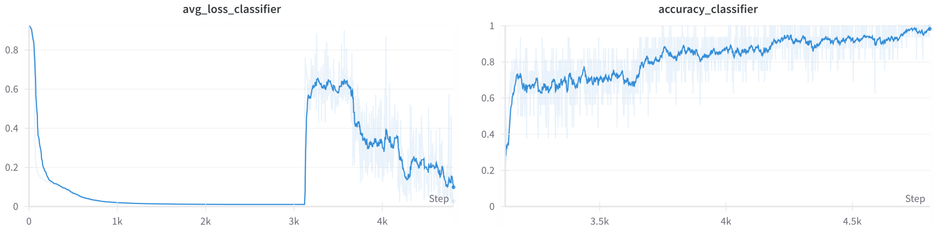
Training Curves



(a) CNN (b) MLP (c) CNN + MLP

Figure: Loss curves during training for different Model Architectures.

Transformer Masked Token Modeling



(a) loss

(b) train accuracy

Figure: Loss and Accuracy for Transformer during pre-training and fine-tuning.

Discussion

Improvements?

- More data
- Different validation task (ALS seems problematic)
- ... your ideas?

Thank you!