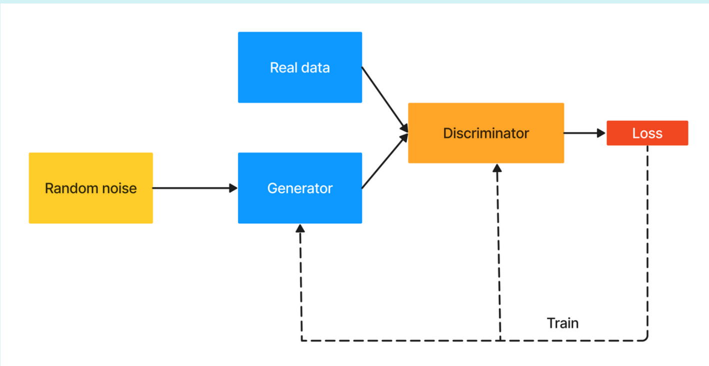
# SYNTHETIC DATA GENERATION FOR THE OPTIMIZATION OF STRAINS IN METABOLIC ENGINEERING USING GENERATIVE ADVERSARIAL NETWORKS

**Marcin Jarosz**
**m.w.jarosz@student.tudelft.nl**

**Supervisor: Paul van Lent**
**Responsible professor: Thomas Abeel**

## 1.Background

- Metabolic engineering [1] involves the precise manipulation of those pathways to achieve specific system behaviors, such us higher product flux, typically for the production of economically significant substances like fuels, essential chemicals, or pharmaceuticals
- Generative adversarial network [2] comprise of two neural networks, generator and discriminator, trained simultanouesly. The generator is the model that tries to capture the distribution of data, while the discriminator distinguishes between real and fake data and is required to compute the loss of the generator and minimize it.




## 2.Research Question

- How can Generative Adversarial Networks be utilized to generate synthetic data for optimizing strains in metabolic engineering, and what is the quality of the generated data compared to experimental data?
- How can performance of a generative model be measured, in order to compare the data generated by it to experimental data and determine its overall efficiency?
- How efficient is the probabilistic PCA model?
- How efficient is the GAN model and how does it compare to probabilistic PCA (baseline)?

## 3.Methodology

- Data used to train the models is synthetic, coming from a kinetic model
- Models are implemented in Python, using PyTorch
- Probabilistic PCA model uses 1 component to generate new data
- Both generator and discriminator of GAN are neural networks with 1 hidden layer of 1024 and latent (input) size of 15 neurons.
- Generated data is visualized using 2 PCA components explaining the largest variance, as well as original features. It is then compared to the real data.
- The comparison is conducted based on statistical properties (mean, variance), as well as visual inspection

## 4.Results

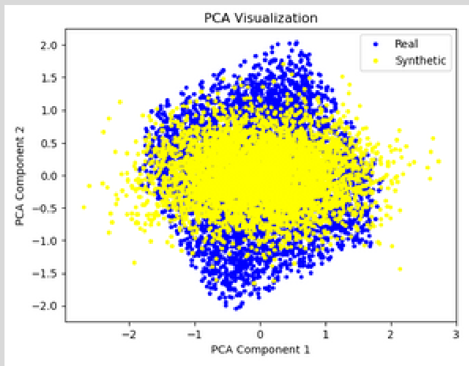



**Figure**: PCA visualization of data generated by **PPCA**, compared to real data

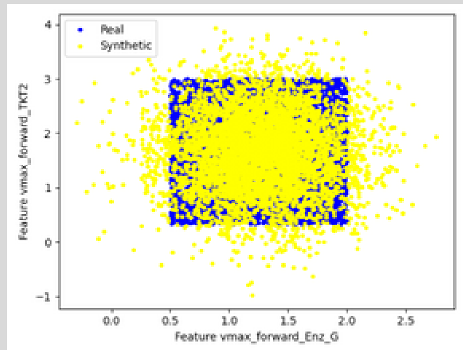



**Figure:** 2 features of data generated by **PPCA**, compared to real data

PPCA statistical properties

| Distance measure | Mean difference | Variance difference |
|---|---|---|
| Manhattan | 0.0924 | 1.7596 |
| Euclidean | 0.0263 | 0.6277 |

**Figure:** Differences of statistical properties of **PPCA** and real data

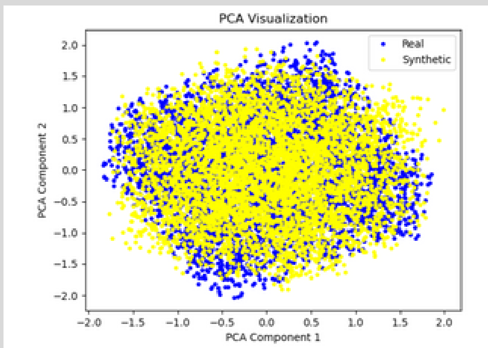



**Figure**: PCA visualization of data generated by **GAN**, compared to real data

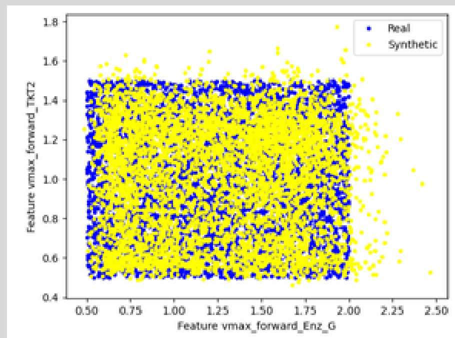



**Figure:** 2 features of data generated by **GAN**, compared to real data

GAN statistical properties

| Distance measure | Mean difference | Variance difference |
|---|---|---|
| Manhattan | 0.2788 | 0.2533 |
| Euclidean | 0.0809 | 0.0840 |

**Figure:** Differences of statistical properties of **GAN** and real data

## 5.Conclusions

- Data generated by probabilistic PCA has similar mean to real data, but significantly larger variance, and often generates unrealistic samples.
- GAN is able to accurately model the probability distribution of real data, both in terms of mean and variance, as well as the overall shape of the distribution
- GAN, however, has to be trained for a very long time (>10000 training epochs) to achieve these results

## 6.Future work & Limitation

- Try different architectures and hyperparameters of the neural networks in GAN to further improve its performance
- Find a way to better analyze the quality of the generated data, rather than just comparing the distribution to real data

[1] Markus Jeschek et al. (2017). Combinatorial pathway optimization for streamlined metabolic engineering. Current Opinion in Biotechnology, 47, 142-151.
[2] Ian J. Goodfellow et al. (2014). Generative Adversarial Networks

TU Delft