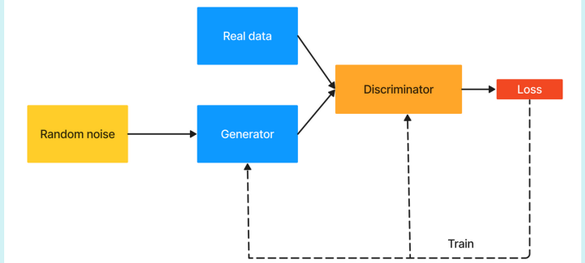
SYNTHETIC DATA GENERATION FOR THE OPTIMIZATION OF STRAINS IN METABOLIC ENGINEERING USING GENERATIVE ADVERSARIAL NETWORKS

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 Questions

- How can the performance of a generative model be measured to compare data generated by it with data obtained using traditional, more costly methods?
- What is the comparative performance of the PPCA model (baseline) and the GAN model in the context of pathway optimization, and what are the best performing latent dimensions for each model?

3.Methodology

- Data used to train the models comprises of 5000 kinetic models of a hypothetical pathway
- Models are implemented in Python, using PyTorch
- The comparison is conducted based on KL divergence, as well as visual inspection
- Both generator and discriminator of GAN are neural networks with 1 hidden layer of 1024 neurons
- Hyperparameters:

	Generator	Discriminator
Epochs	20 000	20 000
Learning rate	0.000 1	0.000 1
Regularization	Weight decay of 0.0001	A dropout of 0.3 on each layer
Batch size	50	50
Output layer activation	None	Sigmoid

5.Conclusions

- PPCA exhibited steady increase of performance with the increase of latent dimensions, best at 18
- In latent dimensions 6-15, GAN outperformed PPCA with lower KL divergence. PPCA performed better in other (1-5 and 16-18)
- Visual inspection revealed that GAN better captured the features of real distribution
- 8 was determined to be the optimal choice of latent dimensions for GAN, due to proximity of KL divergence to global minimum and best visual structure of data

6.Future work & Limitation

- Try different architectures and hyperparameters of the neural networks in GAN to further improve its performance
- Find a better way analyze the quality of the generated data, rather than just comparing the distribution to real data.
- Further investigate performance of PPCA at latent dimensions 1-5 and 16-18, where it outperformed GAN in terms of KL divergence

4.Results

Latent size		1	2	3	4	5	**6**	7	8	9	10	11	12	13	14	15	16	17	18
KL divergence	PPCA	3.50	2.89	2.03	1.97	1.98	**1.85**	1.82	1.68	1.59	1.36	1.21	0.88	0.68	0.54	0.31	0.20	0.17	0.08
	GAN	2.58+e07	3.22+e04	644.94	19.33	2.63	**1.30**	1.46	0.35	0.85	0.28	0.41	0.45	0.44	0.31	0.19	0.21	0.20	0.22

Figure: KL divergence calculated for data generated with differente latent space size

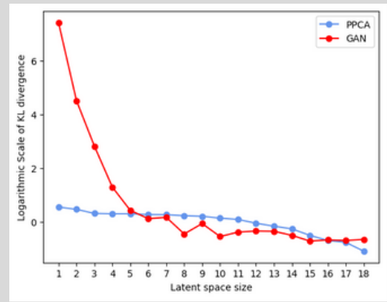
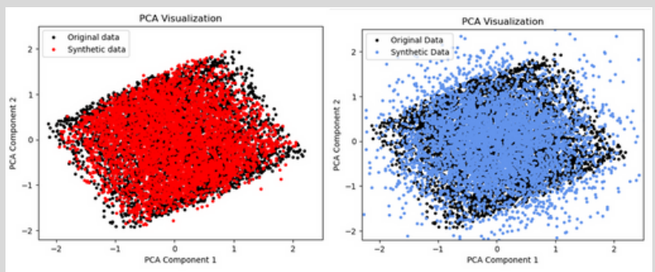
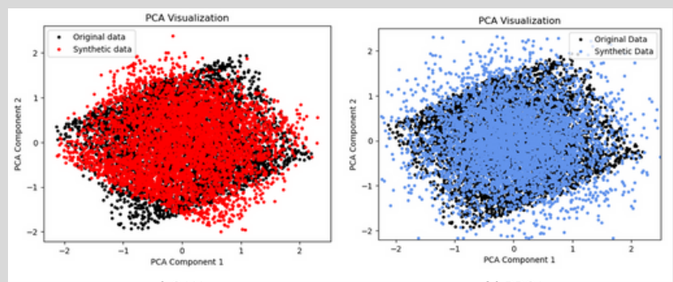


Figure: Trend line of the KL divergence across different latent dimensions (log scaled)



a) GAN **b)** PPCA

Figure: PCA visualization of data generated by with 8 latent dimensions



a) GAN **b)** PPCA

Figure: PCA visualizations of data generated with 15 latent dimensions

[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