Synthetic data generation for the optimization of strains in metabolic engineering using generative adversarial networks

Marcin Jarosz

Supervisors: Prof. Dr. Thomas Abeel, MSc. Paul van Lent

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 28, 2024

Name of the student: Marcin Jarosz
Final project course: CSE3000 Research Project
Thesis committee: Prof. Dr. Thomas Abeel, Prof. Dr. Alan Hanjalic, MSc. Paul van Lent

Abstract

This research investigates the application of Generative Adversarial Networks (GANs) and probabilistic Principal Component Analysis (PPCA) in generating synthetic data for pathway optimization in metabolic engineering. The study aims to compare the performance of these generative models, addressing key questions regarding their utilization, the quality of generated data compared to experimental data, and overall efficiency. The dataset comprises 5000 parameter configurations of kinetic models that simulate a hypothetical pathway. Constructing kinetic models traditionally involves obtaining complex scientific knowledge, a process that may be alleviated through a data-driven approach. Results indicate that both models, tried with different sizes of latent space, demonstrate good performance in modeling the underlying latent space of the data. However, GANs with the right set of parameters exhibit a better performance, evidenced by lower KL divergence and superior visual structure in the generated data. The findings highlight the potential of GANs to outperform probabilistic PCA, offering valuable insights for more cost-effective and streamlined strain optimization in metabolic engineering. Overall, this research advocates for further investigation of GANs capabilities in metabolic engineering as a potentially powerful tool for synthetic data generation.

Introduction

Enzymes catalyze virtually all cellular reactions along metabolic pathways [1]. Metabolic engineering involves the precise manipulation of those pathways to achieve specific system behaviors, such as higher product flux, typically for the production of economically significant substances like fuels, essential chemicals, or pharmaceuticals [2]. This process is often referred to as pathway optimization. To give some examples, pathway optimization has successfully been used in the optimization of lycopene biosynthesis in *E. Coli* [3] and xylose utilisation in *S. cerevisiae* [4].

Despite these successes, a key challenge in pathway optimization is the high costs associated with modifying strains to achieve economically viable outputs. The significant expense arises from the 'combinatorial explosion', where a multitude of possible combinations must be tested to determine the optimal configuration for a given use case [2]. Machine learning has been explored as a solution to these expenses, employing techniques like prediction-based neural networks [5], partial least square regressions [6], ensembles of different models [7], and recommender systems [8].

Another possible approach to address the problem of combinatorial explosion involves generative machine learning models, such as probabilistic principal component analysis (PPCA) [9], generative adversarial network (GAN) [10], or variational autoencoder (VAE) [11]. These models aim to capture the underlying probability distribution of data, enabling the generation of new data samples. GANs, for instance, work by training a generator neural network to create data samples that are indistinguishable from real data, while a discriminator neural network learns to differentiate between real and generated samples. The interplay between these networks and their simultaneous training results in a generator that produces increasingly realistic data.

Choudhury et al. [12] demonstrated the utilization of Generative Adversarial Networks (GANs) to generate kinetic models of *E. coli* metabolism. Kinetic models provide valuable insights into the temporal dynamics of cellular states, with each model representing a hypothetical pathway. This methodology offers a more detailed understanding of cellular metabolism compared to steady-state methods like flux balance analysis. However, the construction of kinetic models faces challenges in obtaining precise information about the specific mechanisms underlying each reaction and associated parameters, such as maximal velocities or Michaelis constants. These challenges arise due to the complexity involved in acquiring such detailed knowledge [12].

This paper builds upon Choudhury et al.'s work, investigating how generative adversarial networks can model the latent space essential for pathway optimization. The focus is on understanding the utility of GANs in generating kinetic models for optimizing strains in metabolic engineering and assessing the quality of the generated data. The research aims to answer two primary 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?

The remainder of the paper is organized as follows: the methodology section outlines the data used, model parameters and architecture, metrics for evaluating performance and code availability. Subsequently, the results are presented and analyzed. A section on responsible research follows to ensure the legitimacy and reproducibility of the conducted research. Finally, the conclusions section summarizes research findings and proposes recommendations for future contributions.

Methodology

This section addresses challenges in acquiring detailed information for kinetic models by using generative models. Principal Component Analysis (PCA) and Generative Adversarial Networks (GANs) are considered, both assuming an underlying latent distribution. The models are trained on a dataset of kinetic models, and implementation details are provided. Evaluation involves Kullback-Leibler divergence and visual inspection of generated data. The data and code are available on a GitHub repository.

Data used

Given the challenges associated with acquiring detailed information on the exact mechanisms and parameters essential for building comprehensive kinetic models of hypothetical pathways, generative models will be tried to overcome these limitations. The application of generative models presents a promising solution for the creation of kinetic models, as they offer a data-driven approach that can infer complex relationships and capture intricate patterns within cellular systems. The models will be trained on a dataset comprising of 5000 kinetic models, each consisting of 19 parameters, constructed by a traditional computer algorithm. By harnessing the power of generative models, we aim to address the gaps in knowledge hindering traditional kinetic modeling approaches, facilitating the development of more cost-effective methods of constructing representations of cellular behavior.

Models

Both models considered for the task assume an underlying latent distribution. They attempt to model it, and can then be used for sampling new data from the modeled distribution, and projecting the data onto the original space.

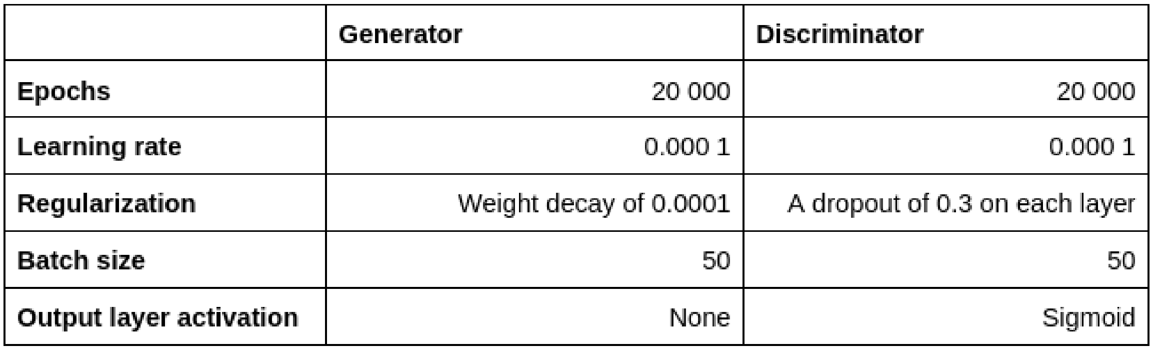
Principal Component Analysis (PCA) [13] stands out as a widely employed method for dimensionality reduction. It identifies the principal components within the data, effectively capturing the most noteworthy sources of variation. Probabilistic PCA [9], adopts a probabilistic model for the data and leverages the latent representation of the data derived from the principal components obtained through PCA. The model's implementation is guided by a Medium article by Oliver K. Ernst, Ph.D. [1], providing a more intricate explanation of the model's mechanics.

Generative Adversarial Networks (GANs) [10] have emerged as a powerful tool for generating realistic data samples. GANs consist of two neural networks, a generator and a discriminator, engaged in a competitive training process. The generator learns to produce synthetic data samples that are indistinguishable from real data, while the discriminator aims to differentiate between real and generated samples. Through adversarial training, GANs iteratively improve the generator's ability to generate realistic data, enabling it to capture complex data distributions. The latent space of the data in this case corresponds to the input layer of the generator. The model can be sampled from through generating random noise vectors and feeding it to the generator.

Model parameters and training procedure

Both models were implemented in Python 3.11. The tried latent dimensions will be between 1 and 18, as these are all possible latent dimensions compatible with the PPCA model. GAN could also be implemented with more latent dimensions, but since we are interested in comparing the two models, we will not consider higher latent dimensions in this study.

The GAN was implemented using deep learning framework PyTorch 2.1.2 and trained on a computer with CUDA enabled GPU, which significantly reduces the training time. The architecture and hyperparameters of the model were chosen empirically due to a large number of possible combinations. A more systematic selection of hyperparameters, such as grid search, should be performed in future research to better determine the optimal values for the. Both generator and discriminator consist of one hidden layer of 1024 neurons with ReLU activation function and use Adam optimizer with mini-batch and binary cross entropy loss as the objective function. Hyperparameters of both networks are summarized in Table 1.



	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

Table 1: Hyperparameters used in the training of GAN

Evaluation metrics

Model performance will be assessed through computing Kullback-Leibler (KL) divergence [14] of real and synthetic data for different latent space dimensions. KL divergence is a commonly used measure of distance between two distributions:

$$D_{\mathrm{KL}}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

where p and q are the two considered, continuous probability distributions. The choice of KL divergence as the evaluation metric stems from its capability to quantify the difference between two probability distributions, making it particularly suitable for assessing the performance of generative models. KL divergence measures the information lost when one distribution is used to approximate another. In the context of this study, it provides a meaningful measure of how well the generative models capture the underlying distribution of real data.

First, the distribution functions will be approximated to multivariate normal distributions, and then fed into the KL formula for evaluation. Additionally, for the best performing latent space sizes, visual inspection will be performed, presenting generated data in both the original dimensions and in lower dimensions (the first two principal components), using scatterplots. The visual inspection will be done to check whether low KL divergence indeed corresponds to similar data distribution.

Data and code availability

The data and code for this study are available in a GitHub repository. The repository with datasets and implementation details can be found at **`https://github.com/AbeelLab/RP2023_Jarosz`**.

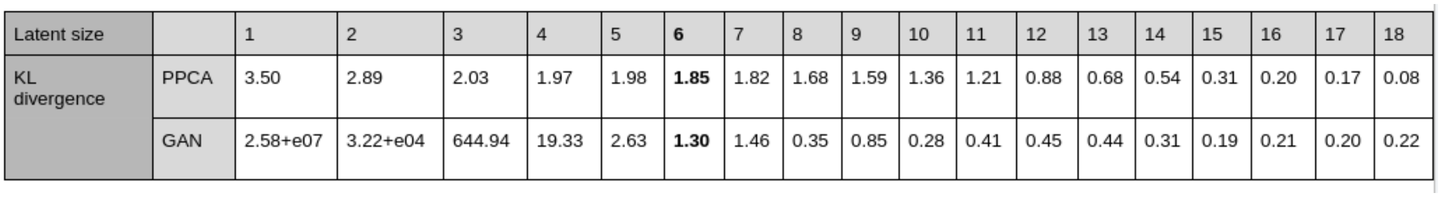
[1] O. K. Ernst, "The simplest generative model you probably missed", medium.com. https://medium.com/practical-coding/the-simplest-generative-model-you-probably-missed-c840d68b704 (Accessed Jan. 11, 2024)

Results and discussion

This chapter presents the similarity between real and synthetic data, from both PPCA and GAN, through analysis of KL divergence between the distributions generated with different latent sizes as well as visual inspection of the data and comparison of the two models.

KL divergence values

To quantitatively assess the performance of Generative Adversarial Networks (GANs) and probabilistic Principal Component Analysis (PPCA), we calculated the Kullback-Leibler (KL) divergence for latent space sizes ranging from 1 to 18. Table 2 presents the KL divergence values for GAN and PCA. A trend line visualization (Figure 1) illustrates the comparative trend of KL divergence across different latent space sizes.



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

Table 2: KL divergence computed between synthetic and real data, per latent size

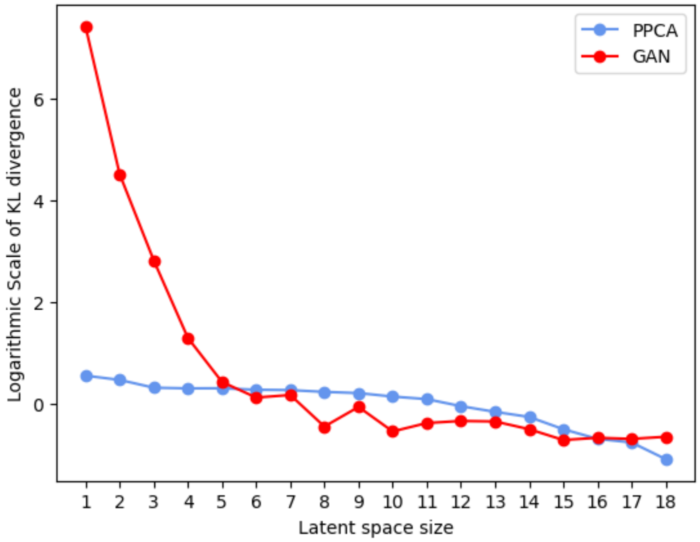


Figure 1: Trend lines of logarithmic scale of KL divergence, across different latent dimensions for PPCA and GAN

The PPCA model exhibited notably superior performance at lower latent dimensions, specifically below 6. Beyond 6 dimensions (highlighted in bold in the table), GAN tended to outperform PPCA. Remarkably, the GAN model reacheed its peak performance at 15 latent dimensions, achieving a KL divergence of 0.19. However, the KL divergence was already relatively close to the minimal value (0.35) at 8 dimensions. PPCA on the other hand demonstrated a consistent improvement starting from 6 dimensions onward. It reached its optimal performance at latent dimensions of 18, approaching the dimensions of the original data. Consequently, determining the most effective latent dimensions for PPCA proved challenging based solely on the provided tables and trend lines.

To delve deeper into the analysis, we will visualize the data generated by both models at latent dimensions 8 and 15. The choice of 8 is particularly of interest for GAN, given its proximity to the minimum, while 15 represents the global minimum for GAN.

Performance with 8 latent dimensions

The choice of 8 latent dimensions for visualizations is justified based on the observed performance trends and the balance between model efficiency and effectiveness. As evidenced in the KL divergence analysis, GAN demonstrated notable performance at 8 latent dimensions, approaching the minimal divergence. This suggests that GAN captures the essential features of the underlying data distribution well at this dimensionality. We present visualizations of the data generated by both GAN and PCA. Figures 2 and 3 showcase the data in PCA-reduced dimensions, while Figure 4 and 5 depict the original data.

GAN visibly outperformed PCA in capturing the intricate features of the data, demonstrating superior visual structure and lower KL divergence. It was able to precisely model the distribution, generating fewer data points outside the original boundaries of the distribution than the PPCA. Thus, we conclude that GAN performed significantly better than PPCA at latent dimensions 8 and generated data from distribution similar to real. We will now examine how it compares to latent dimensions of 15.

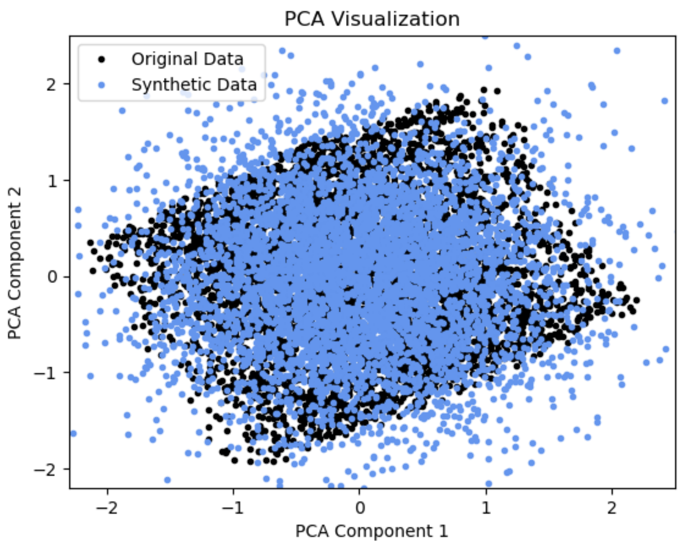


Figure 2: PCA visualization with first 2 principal components for data generated by PPCA, using 8 latent dimensions

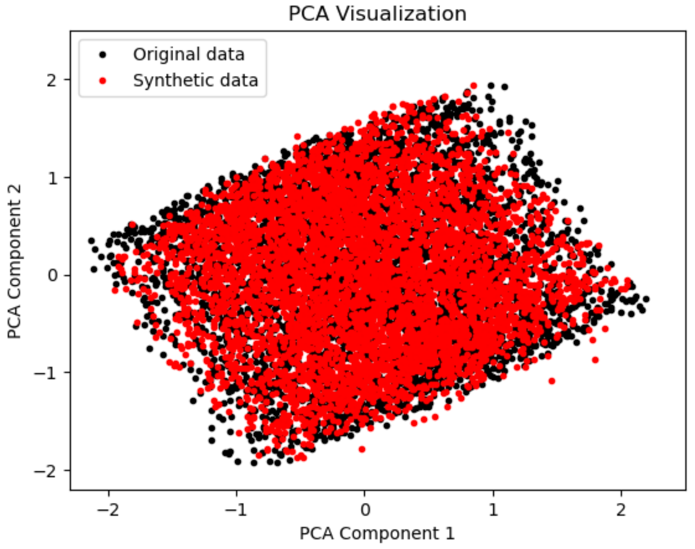


Figure 3: PCA visualization with first 2 principal components for data generated by GAN, using 8 latent dimensions

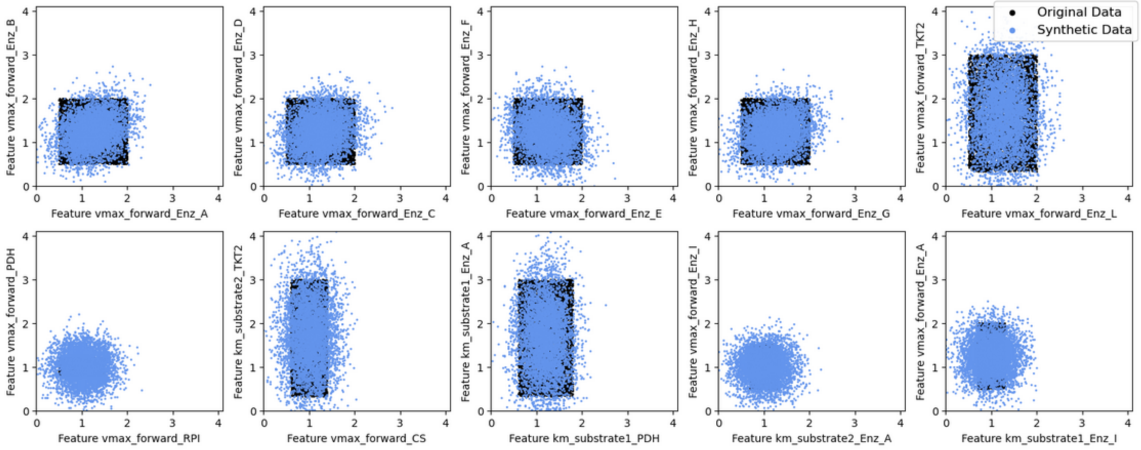


Figure 4: Visualization of data generated by PPCA, using 8 latent dimensions

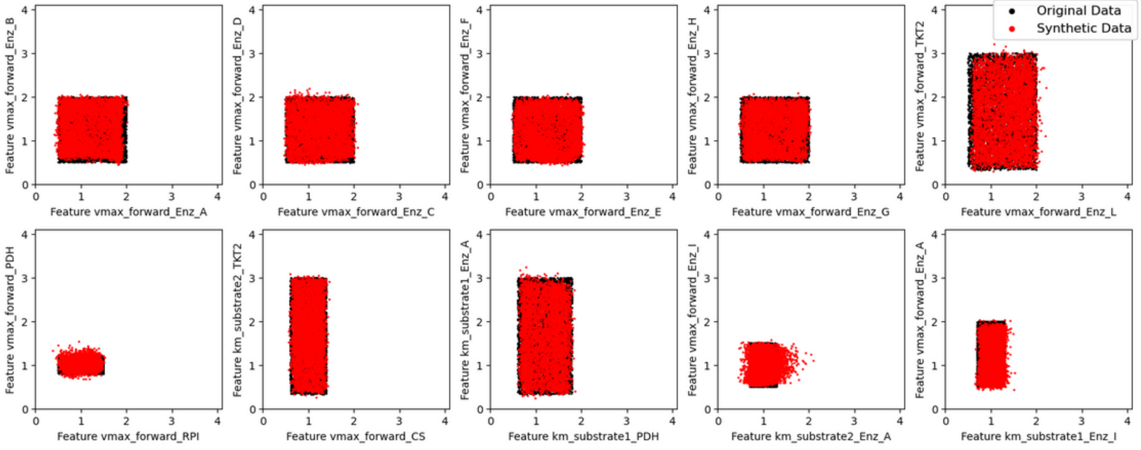


Figure 5: Visualization of data generated by GAN, using 8 latent dimensions

Performance with 15 latent dimensions

Similarly, we present visualizations of data generated by GAN and PCA for latent space size 15. Figures 6 and 7 exhibit data in PCA-reduced dimensions, while Figures 8 and 9 illustrate the original data.

Consistent with the latent space size of 8 results, GAN continued to outperform PCA in capturing the nuanced features of the data. Looking at the data in original dimensions, there is, however, no noticeable increase of performance compared to the lower dimensions for both of the models, perhaps there are actually more outliers generated. The PCA visualization of data generated by GAN seems to be a rotation of the real data. Further investigation is required on what causes this rotation at latent size of 15 and not at latent size of 8.

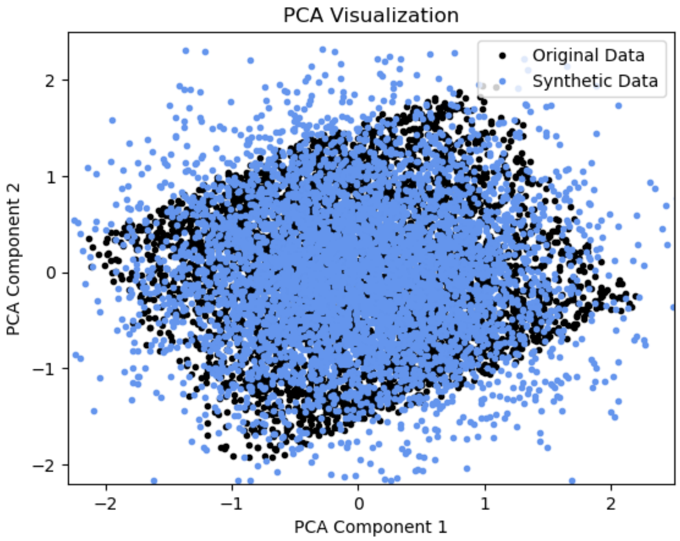


Figure 6: PCA visualization with first 2 principal components for data generated by PPCA, using 15 latent dimensions

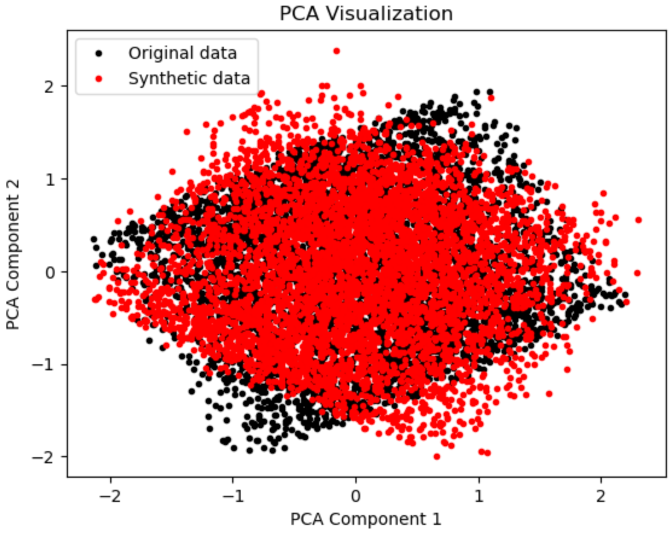


Figure 7: PCA visualization with first 2 principal components for data generated by GAN, using 15 latent dimensions

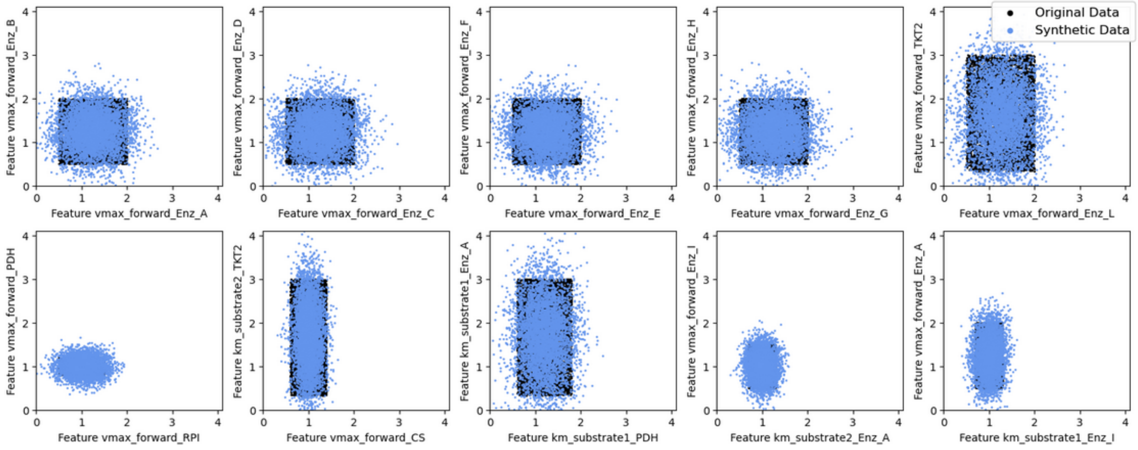


Figure 8: Visualization of data generated by PPCA, using 15 latent dimensions

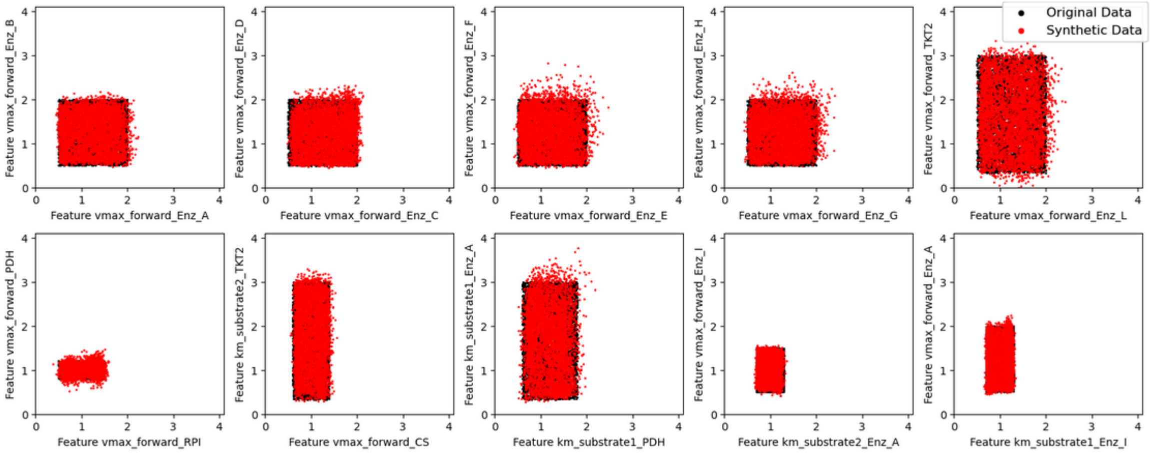


Figure 9: Visualization of data generated by GAN, using 15 latent dimensions

Although in terms of KL divergence PPCA performed better with low (fewer than 6) latent dimensions, these results collectively highlight the superior performance of GAN in both quantitative KL divergence metrics and qualitative visualizations for best performing latent dimensions. After visual inspection of data, we determined that the model performed best with 8 latent dimensions, however further visual investigation should be performed to evaluate the models with different sets of parameters, especially at low latent dimensions where PPCA might have performed better. In summary, the results emphasize the potential for generating synthetic data in metabolic pathway optimization using GANs.

Responsible research

This section critically addresses the ethical considerations inherent in the research and elucidates the adopted measures to bolster result reproducibility.

Bias

While efforts were made to minimize bias, it's crucial to acknowledge certain inherent limitations. The dataset used in this study was generated through a computer algorithm, introducing a potential source of bias associated with the algorithm's design and assumptions. Although no selective data curation was applied, the algorithm itself might inadvertently introduce biases. In the visualization of results, though every generated point was inclusively depicted across all features, it's essential to recognize that the very nature of generative models introduces a level of bias based on the learned patterns from the training data.

Reproducibility

To uphold standards of transparency and reproducibility, all code utilized in this research is openly accessible online, at **`https://github.com/AbeelLab/RP2023_Jarosz`**. While Jupyter notebooks enhance understanding, it's crucial to acknowledge the challenges inherent in reproducing complex models and results. The intricacies of hyperparameter tuning, algorithmic nuances, and the stochastic nature of certain processes may pose challenges in precisely replicating the study. Despite our commitment to openness, the inherent complexity of generative models may limit the ability of others to precisely reproduce the results, urging caution in generalizing findings.

Limitations

It's important to acknowledge the limitations of this study. The dataset, albeit generated algorithmically, may not fully capture the complexity and diversity of real-world biological systems. Additionally, the choice of evaluation metrics, such as KL divergence, provides valuable insights but may not comprehensively represent the performance of generative models in all aspects. The decision to focus on specific latent dimensions for visualizations, while justified, does not account for potential variations in optimal dimensions under different conditions. These limitations underscore the need for continued research, exploration of alternative methodologies, and cautious interpretation of the results in the broader context of metabolic pathway optimization.

Conclusions

This research has provided insights into the comparative effectiveness of Generative Adversarial Networks and probabilistic PCA in generating synthetic data for metabolic pathway optimization in metabolic engineering.

The comprehensive analysis demonstrated that GANs outperformed probabilistic PCA across multiple dimensions. This superiority was evident in quantitative metrics, specifically lower KL divergence values. However, other statistical tests could be performed to further validate this observation.

The visual inspections of generated data corroborated the quantitative results, confirming that GANs captured nuanced features of the real data distribution more effectively than probabilistic PCA. This aligns with the goal of not only good performance in the statistical test but also accurately modeling the intricate patterns inherent in metabolic pathways.

The determination of optimal latent space sizes revealed that both models performed best at latent dimensions of 8, better than at 15. Although 15 latent dimensions exhibited lower KL divergence, the visual structure of generated data looked more realistic for 8 latent dimensions. This insight highlights the importance of carefully considering dimensionality in the application of generative models for synthetic data generation in metabolic engineering, and that more latent dimensions in a model does not always lead to better performance, even though KL divergence or other statistical measure might suggest that.

While these findings contribute to the understanding of generative models in the context of pathway optimization, it's important to acknowledge the inherent limitations and the need for further exploration. Future research could delve into investigating whether the generated kinetic models correlate with elevated flux values and explore alternative model architectures or hyperparameter configurations to enhance performance. The results presented here offer a foundation for building upon and advancing the application of generative models to optimize metabolic pathways, potentially leading to more efficient and cost-effective strategies in the field.

References

[1] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell*. Garland Science, 2017.

[2] M. Jeschek, D. Gerngross, and S. Panke, “Combinatorial pathway optimization for streamlined metabolic engineering.,” *Current Opinion in Biotechnology*, vol. 47, pp. 141–151, 2017.

[3] X. L. Chen, P. Zhu, and L. M. Liu, “Modular optimization of multi-gene pathways for fumarate production.,” *Metabolic Engineering*, vol. 33, pp. 76–85, 2016.

[4] L. N. Latimer, M. E. Lee, D. Medina-Cleghorn, R. A. Kohnz, D. K. Nomura, and J. E. Dueber, “Employing a combinatorial expression approach to characterize xylose utilization in saccharomyces cerevisiae.,” *Metabolic Engineering*, vol. 25, pp. 20–29, 2014.

[5] H. Meng, J. Wang, Z. Xiong, F. Xu, G. Zhao, and Y. Wang, “Quantitative design of regulatory elements based on high-precision strength prediction using artificial neural network,” *PLOS ONE*, vol. 8, pp. 1–9, 04 2013.

[6] A. J. Jervis, P. Carbonell, S. Taylor, R. Sung, M. S. Dunstan, C. J. Robinson, R. Breitling, E. Takano, and N. S. Scrutton, “Selprom: A queryable and predictive expression vector selection tool for escherichia coli,” *ACS Synthetic Biology*, vol. 8, pp. 1478–1483, Jul 2019.

[7] Y. Zhou, G. Li, J. Dong, X. hui Xing, J. Dai, and C. Zhang, “Miya, an efficient machine-learning workflow in conjunction with the yeastfab assembly strategy for combinatorial optimization of heterologous metabolic pathways in saccharomyces cerevisiae,” *Metabolic Engineering*, vol. 47, pp. 294–302, 2018.

[8] T. Radivojević, Z. Costello, K. Workman, and H. Garcia Martin, “A machine learning automated recommendation tool for synthetic biology,” *Nature Communications*, vol. 11, p. 4879, Sep 2020.

[9] M. E. Tipping and C. M. Bishop, “Probabilistic Principal Component Analysis,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 61, pp. 611–622, 01 2002.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” vol. 27, 2014.

[11] C. Doersch, “Tutorial on variational autoencoders,” 2021.

[12] S. Choudhury, M. Moret, P. Salvy, D. Weilandt, V. Hatzimanikatis, and L. Miskovic, “Reconstructing kinetic models for dynamical studies of metabolism using generative adversarial networks,” *Nature Machine Intelligence*, vol. 4, pp. 710–719, Aug 2022.

[13] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[14] S. Kullback and R. A. Leibler, “On Information and Sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79 – 86, 1951.