Research Plan for CSE3000 Research Project

Synthetic Data Generation for the Optimization of Strains in Metabolic Engineering using Variational Autoencoders

Uğur Doruk Kırbeyi

January 28, 2024

Background of the research

Metabolism, which is the progression of cellular reactions and thus life itself, is guided by enzymes through so-called pathways [1]. Metabolic engineering entails the precise alteration of these pathways to attain particular system functionalities, often aimed at producing commercially valuable compounds such as fuels, vital chemicals, or pharmaceuticals [2].

The main difficulty in metabolic engineering is that it is hard to produce industrial strains and one of the most important factors is the cost to gather data in order to guide the engineering process [2]. Instead of going through the costly process of generating and analyzing the space of data, the need for finding a better way to understand the data and answer the scientific questions could be quenched through compression algorithms that reduce dimensionality [3]. Generative models aim to characterize the fundamental data distribution, enabling the creation of fresh data from the identical probability distribution. Diverse machine learning methods and models have been suggested to optimize the metabolic pathways, with some demonstrating encouraging outcomes [4]. This has motivated the research that our group is conducting, as numerous generative machine learning models still remain unexplored or untested, generally being used for other purposes such as image and writing generation. An example of this, the Variational Autoencoder will be implemented and tested in this project in order to establish whether it is a viable option as to generate data that could guide the strain optimization processes.

Research Question

"How can Variational Autoencoders be effectively utilized to generate high-fidelity synthetic data for optimizing strains in metabolic engineering compared to the baseline model?" is the research question. As long as I have data to compare results from the VAE (Variational Autoencoder) to, it is reasonable to implement a VAE and experiment with it in the time frame of the project. The implementation should not take too much time, and I am planning to spend more time on the experimentation part. I expect to show that it should be possible to at least utilize synthetic data in some way for the optimization of strains, even if there are weaknesses along with it.

Some sub-questions could be: "What are the key parameters and features within VAEs that significantly influence the fidelity and quality of synthetic data generated?", "What quantitative metrics and qualitative benchmarks can be used to evaluate the fidelity and accuracy of synthetic data produced by VAEs in comparison to the baseline?", "Can generated data adequately represent the complexity and variability found in real experimental data, impacting the accuracy of strain optimization algorithms?", "Under which conditions is utilizing VAE's better than real experimental data/other models?" Through finding the answers to these questions, I will be able to also try to answer the research question as they are the individual parts of the main question.

Method

Continuing on researching all throughout the project is the main task; not only to gain knowledge on various methods that can be utilized to generate synthetic data for the process of metabolic engineering, but also for the data generation process of generative models. The secondary task is to implement a baseline model that would be comparable to the VAE in the way that they can generate synthetic data from a set of given data. For this, we have chosen to work with an implementation for Probabilistic PCA. After this task, the next one would be to implement a VAE that works with the given data, which is a set of synthetic data that has been simulated and has a combinatorial nature, and thus could complicate the process of implementation. From there on, experimenting on the parameters and features of VAE's, along with finding out metrics and benchmarks to compare the generated results with the given data and the baseline would be the next task. The implementations will be done on a Jupyter notebook, utilizing the PyTorch library. The results of these tasks could impact the answers to the other sub-questions, meaning it will be an iterative process until the posed questions will be answered.

We as a group intend to collaborate on the implementation of a base Probabilistic PCA algorithm together with my group. One group member is also working with VAEs, therefore we will collaborate on the implementation of the VAE itself and share findings about best and worst parameters. The papers will be completely individual, along with the questions that we will be searching the answers for. Still, the members will be sharing findings and relevant papers in order to make the process both more efficient and effective. There are not any real dependencies between the group members, as even the implementations could be done separately if wanted to, but working together could very much expedite the process.

Planning of the research project

The workload of the project process is approximately 38-42 hours per week. There will be meetings with our supervisor every week, and in these meetings we plan to discuss updates and how the process is going, along with small presentations on relevant literature that we have found helpful for the whole group. The weekly plan, including deadlines will be as follows:

Week 1: Read literature on subjects such as: Synthetic data generation, Variational Autoencoders and their use cases along with strengths/weaknesses, Probabilistic PCAs and their use cases along with their strengths/weaknesses, Metabolic engineering (for improving both my and the paper's background). Create general outline for paper.

- Deadlines: Research Plan (19th November), Research Plan Presentation slides (19th November).

Week 2: Continue reading/researching. Start introduction for the paper. Begin the implementation of Probabilistic PCA.

- Deadlines: Research Plan Presentation (23rd November).

Week 3: Finish implementation in terms of having a working model for the Probabilistic PCA and start the implementation of the VAE. Start experimenting. Attend Responsible Research lectures. Complete ACS assignment 1.

- Deadlines: ACS assignment 1 (1st December).

Week 4: Attend Responsible Research lectures. Complete ACS assignments 2a-b. Continue reading. Continue experimenting and writing (Aim for 30-40% of paper). Start on midterm presentation.

- Deadlines: ACS assignments 2a-b (8th December).

Week 5: Complete ACS poster. Prepare for midterm presentation. Receive feedback and improve upon feedback. Continue experimenting, reading and writing (Aim for 50% of paper).

- Deadlines: ACS poster (11th December), Midterm Progress Presentation (13th December).

Week 6: Complete ACS assignment 3 and ACS paper. Continue experimenting, reading and writing (Aim for 70% of paper).

- Deadlines: ACS assignment 3 (19th December), ACS paper (22nd December).

Week 7: Finish paper and peer draft v1. Receive feedback from draft and improve upon it. Continue experimentation, reading and writing (Aim for 80% of paper).

- Deadlines: Paper Draft v1 (9th January).

Week 8: Finish peer repair and paper draft v2. Create draft on final poster. Start finalizing results from experiment. Continue reading and writing (Aim for 90% of paper).

- Deadlines: Paper Draft v2 (17th January).

Week 9: Receive feedback from paper v2. Finish finalizing results from experiment. Start presentation, continue on poster. Finish the paper and submit.

- Deadlines: Paper submission (28th January).

Week 10: Attend ACS session on final poster. Submit poster. Submit presentation and present.

- Deadlines: Final poster submission (29th January).

References

1. B. Alberts et al., *Molecular biology of the cell*, 6th ed. New York, Garland Science, 2015, pp. 43–88.

2. M. Jeschek, D. Gerngross, and S. Panke, "Combinatorial pathway optimization for streamlined metabolic engineering," *Tissue, cell and pathway engineering*, vol. 47, pp. 142–151, 2017, doi: https://doi.org/10.1016/j.copbio.2017.06.014.

3. J. M. Graving and I. D. Couzin, "VAE-SNE: a deep generative model for simultaneous dimensionality reduction and clustering," *bioRxiv*, p. 2020.07.17.207993, Jan. 2020, doi: https://doi.org/10.1101/2020.07.17.207993.

4. A. V. de Kleut, "Variational AutoEncoders (VAE) with PyTorch," *Alexander Van de Kleut*, May 14, 2020. https://avandekleut.github.io/vae/