Research Plan for CSE3000 Research Project

Synthetic data generation for the optimization of strains in metabolic engineering using Variational Autoencoders (VAEs)

Neil Alwani

November 19, 2023

Background of the research

Microorganisms have been widely used for the production of many products that people like to use, such as beer, wine, cheese, yogurt. Metabolic engineering, the process of optimizing the metabolism of a microorganism using DNA technologies such as CRISPR-CAS, has allowed researchers to expand to a variety of new products, such as nutraceuticals (Liu et al. 2019), pharmaceuticals (Zhang et al. 2022), and biofuels (Keasling et al. 2021). One problem in metabolic engineering is the costs involved in engineering a strain such that its production yield is high enough to be economically viable on an industrial scale. While the reasons for this being so expensive are many, one important factor is the cost of generating data that is used to guide the engineering (Jeschek, Gerngross, & Panke, 2017). We will look into generative models for producing synthetic data for metabolic engineering strategies. Generative models are statistical models that are used to model a joint probability distribution of an observable variable and target variable. The samples drawn from this model are very similar to the distribution that the model has trained.

Research Question

1. What current ways of generating synthetic data are commonly used in metabolic engineering, and how does generative modeling fit into this picture?
2. What methods could be used to generate the specific data type that we want?
3. Why is a generative model specifically for the data we want often used in Metabolic engineering?
4. How well does the probabilistic PCA perform as a good baseline generative model?
5. What kind of new designs can we sample using the latent space modeled by an implementation of a Variational Autoencoder (VAE) using dimensionality reduction?

Method

Identifying the various methods of generating synthetic data in metabolic engineering will require delving into the relevant literature on this topic. Establishing the link between generative modeling and synthetic data generation in metabolic engineering necessitates thorough research to identify the specific data requirements for training generative models. Implementing a robust baseline machine learning (ML) model entails a comprehensive understanding of Probabilistic PCA. Quantifying the performance of generative models for metabolic engineering involves grasping concepts such as KL divergence and exploring optimization strategies to minimize it. In my particular case, establishing benchmarking for Variational Autoencoder (VAE) against Probabilistic PCA will provide a meaningful comparison.

For the implementation of the baseline model, a framework like Pytorch and VAE machine-learning libraries will be used. The data utilized in this study will consist of synthetic data simulated from a kinetic pathway model.

For this project, Doruk Kirbeyi and I will focus on VAE models. Although we will be writing separate papers exploring different research directions, we plan to collaborate on the implementation aspect of the model. All members of this project are required to implement Probabilistic PCA as the baseline model, making collaborative implementation efforts efficient. Our assigned professor will provide a single grade for the presentation's common context portion, meaning the best presenter will present this part on behalf of all members.

The tasks for other team members will involve assisting with implementation-specific tasks and sharing relevant papers to enhance understanding of the context or model. My responsibilities include writing the final paper, creating a poster for the presentation, and conducting experiments to address my research question. It is crucial to note that the implementation of the VAE model and the execution of experiments are highly interdependent.

The process is crucial for our supervisor, so finishing a task means we can move on to the next step. Alongside this, there's a grading rubric for the project, and feedback from the supervisor and peers will help guide me through the process.

Planning of the research project

On a weekly basis, we plan to spend approximately 41 hours on this project for the next 10 weeks. Additionally, we have scheduled weekly meetings with our supervisor to discuss relevant literature and provide project updates.

Week 2.1: Write a draft plan, complete the literacy assignment, and watch section videos.

- Deadline: Research Plan (19th November)

Week 2.2: Do readings to bridge knowledge gaps. Begin the implementation of the decided-upon baseline model.

- Deadline: Research Plan Presentation to the responsible supervisor (19th November)

Week 2.3: Follow lectures on responsible research, complete ACS assignments, do readings, and write the introduction.

Week 2.4: Start implementing the other model, continue writing, follow the responsible research course, complete ACS assignments, and conduct further readings.

Week 2.5: Midterm presentations, refine planning, continue writing, use feedback for process and writing. Decide on metrics for comparing results.

- Deadline: Midterm progress presentation to the responsible supervisor (13th December)

Week 2.6: Start experimentation, continue writing, and do readings.

Week 2.7: Finish draft 1, get feedback, continue experimentation, and do readings.

- Deadline: Paper submission (13th December)

Week 2.8: Submit draft for peer review, do readings, draft the poster, and finalize results.

Week 2.9: Get feedback, follow coaching sessions, finish experimentation, and deliver the final paper.

Week 2.10: Create presentation, poster, and have the final presentation.

References

- Keasling, J. D., Renninger, N. S., Singh, A. (2021). Microbial production of advanced biofuels. Nature Reviews Microbiology.
- Liu, Q., Steinebach, F., GÃ€tgens, J., Klein, T., BÃ¶lker, M., Bolten, C. J. (2019). Rewiring carbon metabolism in yeast for high level production of aromatic chemicals. Nature Communications, 10(1), 1-11.
- Zhang, J., Barajas, J. F., Burdu, M., Wang, G., Baidoo, E. E., Keasling, J. D. (2022). A microbial supply chain for production of the anti-cancer drug vinblastine. Nature, 609(7926), 1-7.
- Jeschek, M., Gerngross, D., Panke, S. (2017). Combinatorial pathway optimization for streamlined metabolic engineering. Current Opinion in Biotechnology, 47, 142-151.