Author: Uğur Doruk Kırbeyi,
U.D.Kirbeyi@student.tudelft.nl

Responsible Professor: Prof. Dr. Thomas Abeel
Supervisor: Paul van Lent

OPTIMIZING STRAINS IN METABOLIC ENGINEERING: COMPARATIVE ANALYSIS OF Β-CONDITIONAL VARIATIONAL AUTO-ENCODER AND PROBABILISTIC PCA FOR SYNTHETIC DATA GENERATION

Background

- Metabolic engineering is the alteration of metabolic pathways often to produce valuable compounds [1].
- **The main difficulty:** To produce industrial strains and the cost to gather data to guide the engineering process [2].
- **Current solution:** Kinetic models, a set of Ordinary Differential Equations (ODEs), allow adjustments for optimizing parameters like product flux while minimizing other host organism functions.
- **Proposed solution:** Compression algorithms reduce dimensionality [3], providing an alternative to costly data generation from kinetic models. Generative models, like β-Conditional Variational Auto-encoder (β-CVAE), aim to capture data distribution and generate new samples.
- **The motivation:** Different machine learning models have produced encouraging outcomes [4]. Many models still remain to be explored. An example, the β-CVAE is implemented, tested and compared.
- **Objective:** After assessing the credibility of PPCA as baseline model, evaluate the viability of β-CVAE and compare it with PPCA as a data generation option for guiding metabolic strain optimization processes.
- **Evaluation metrics:**
  - KS Test:
    - Non-parametric statistical test assessing whether two sets of data follow the same distribution.
  - KL divergence:
    - A measure of how one probability distribution diverges from a second probability distribution.

Experimental Setup

- PPCA Implementation - Jupyter Notebook utilizing various methods from NumPy
- β-CVAE Implementation - Jupyter Notebook utilizing PyTorch library

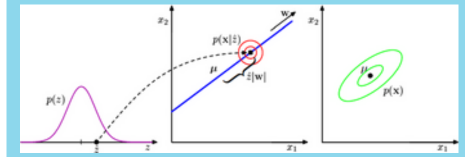

Figure: The basic workings of the PPCA model. From C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

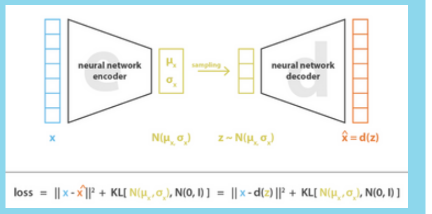
neural network encoder
μ_x
σ_x
sampling
neural network decoder
x
$N(\mu_x, \sigma_x)$
$z \sim N(\mu_x, \sigma_x)$
$\hat{x} = d(z)$
loss = $\|x - \hat{x}\|^2$ + KL[$N(\mu_x, \sigma_x)$, N(0, I)] = $\|x - d(z)\|^2$ + KL[$N(\mu_x, \sigma_x)$, N(0, I)]

Figure: The basic workings of the VAE model. From J. Rocca, "Understanding Variational Autoencoders (VAEs)," Medium, Mar.15, 2020. https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

PCA Model:

- Five sets of synthetic data generated using PPCA.
- Utilized only the first 10 principal components to match latent dimensions with β-CVAE.
- Validation based on statistical properties and distribution comparisons with the original dataset.

β-CVAE Model:

- Trained model used to generate 15 synthetic datasets for each hyperparameter configuration.
- Comparative analysis with PPCA, focusing on fidelity to the distribution and representation of the original dataset.

	β-CVAE
Number of Latent Dimensions	10
Batch Sizes	25 / 50 / 100
Optimizer	Adam optimizer
Weight Decay	1.0×10^{-3}
Learning Rate	1.0×10^{-4}
Number of epochs	1000
Loss function	$((1-\beta) * \text{MSE}) + (\beta * \text{KL-Divergence})$
Beta values	0.1 / 0.25 / 0.5 / 0.75 / 0.9

Table: Training hyperparameters of the β-CVAE model.

Research Question

- *How can β-Conditional Variational Autoencoders be effectively utilized to generate high-fidelity synthetic data for optimizing strains in metabolic engineering compared to the baseline model?*
  - What are the key parameters and features within β-CVAEs that significantly influence the fidelity and quality of data generated?
  - What quantitative metrics and qualitative benchmarks can be used to evaluate the fidelity and accuracy of synthetic data produced by β-CVAEs in comparison to the baseline?"
- **Hypothesis:** By fine-tuning hyperparameters, β-CVAE can achieve statistically significant improvement in the fidelity and accuracy of data generation compared to the baseline model.

Methodology

- **The Data** – Simulated from hypothetical pathway kinetic model based on E.coli strain, 5000 items, each with 19 features and a product flux value. Combinatorial Nature, Continuous
- **Implementation** – Jupyter Notebook, PyTorch and NumPy Libraries
- **Baseline Model** – Probabilistic PCA
- **Main Model** – β-CVAE
- Experiment **Parameters** and **Features**
- **MSE** and **KL-Divergence** used for **training**
- **KL-Divergence, KS-test, PCA visualizations** and **MSE** between product fluxes used to **compare and evaluate** model performances
- Iterative Process

Results

PPCA

	1st Set	2nd Set	3rd Set	4th Set	5th Set	Average
KL-Divergence value:	1.512×10^{-2}	$\mathbf{1.064 \times 10^{-2}}$	1.242×10^{-2}	1.467×10^{-2}	1.451×10^{-2}	1.347×10^{-2}

Table: KL-Divergence values for 5 sets of data generated by 5 individually trained PPCA models. Best value is highlighted in bold.

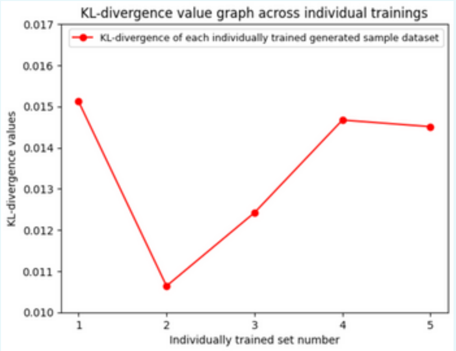

Table: Line graph comparison of KL-divergence values across individually trained generated sample datasets. Please note that this graph and its corresponding graph in the results of β-CVAE have different y-scales and x-values, necessitating caution in direct visual comparisons.

β-CVAE outperforms on every variation

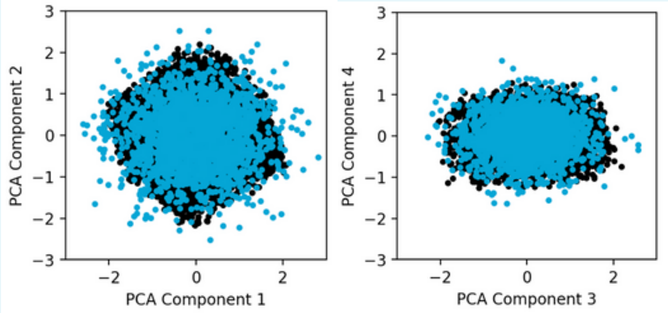

Figure: First four principal components from best performing data generated by PPCA (blue), compared to the same components in the real data (black)

Feature Number	KS Statistic
1	0.062
2	0.130
3	0.104
4	0.075
5	0.068
6	0.090
7	0.087
8	0.127
9	0.090
10	0.074
11	0.091
12	0.246
13	0.095
14	0.089
15	0.102
16	0.087
17	0.158
18	0.068
19	0.168
20	0.103

Table: KS test values for every feature from the best KL-Divergence value producing set of data generated by the PPCA model. (Lower values are better)

PPCA outperforms for some features and vice versa
Overall proves both models' viability

β-CVAE

Batch Size \ Beta Value	0.1	0.25	0.5	0.75	0.9
25	3.813×10^{-3}	3.002×10^{-3}	$\mathbf{1.558 \times 10^{-3}}$	2.248×10^{-3}	4.287×10^{-3}
50	2.058×10^{-3}	2.238×10^{-3}	$\mathbf{1.660 \times 10^{-3}}$	1.744×10^{-3}	2.091×10^{-3}
100	1.355×10^{-3}	2.025×10^{-3}	1.476×10^{-3}	$\mathbf{1.265 \times 10^{-3}}$	1.358×10^{-3}

Table: KL-Divergence values of the data generated from the β-CVAE model for the various tested batch size and beta values. Best values per row are highlighted in bold.

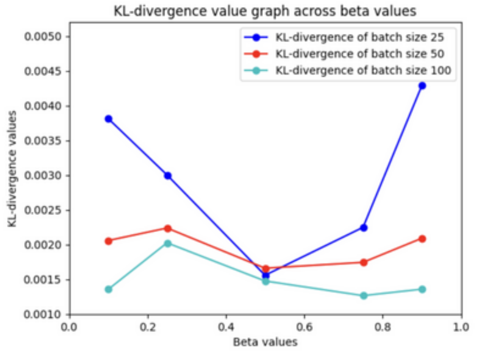

Table: Line graph comparison of KL-divergence values across beta values per batch size. Please note that this graph and its corresponding graph in the results of PPCA have different y-scales and x-values, necessitating caution in direct visual comparisons.

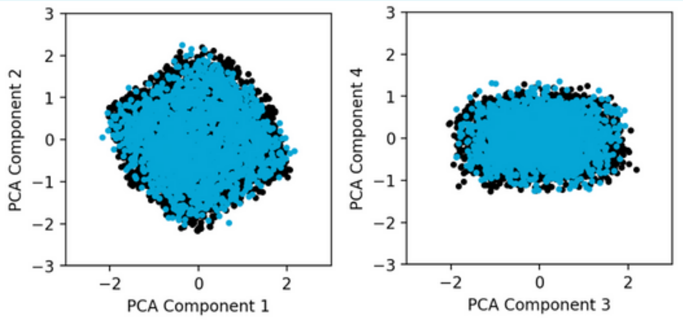

Figure: First four principal components from best performing data generated by β-CVAE (blue), compared to the same components in the real data (black)

Feature Number	KS Statistic
1	0.061
2	0.133
3	0.132
4	0.077
5	0.094
6	0.110
7	0.104
8	0.112
9	0.068
10	0.038
11	0.410
12	0.460
13	0.420
14	0.038
15	0.383
16	0.051
17	0.419
18	0.274
19	0.434
20	0.080

Table: KS test values for every feature from the best KL-Divergence value producing set of data generated by the β-CVAE model. (Lower values are better)

β-CVAE outperforms again

Model	MSE score
PPCA	4.297×10^{-1}
β-CVAE	1.862×10^{-2}

Table: MSE scores calculated between the product flux columns of best-performing datasets from each model and the resulting product flux column from running the kinetic model with the parameter values from each generated dataset. (Lower values are better)

Conclusions & Limitations

- **Main Findings:**
  - PPCA serves as an adequate baseline model.
  - β-CVAE demonstrates superiority in fidelity, robustness, and accuracy.
- **Hypothesis Confirmation:**
  - Fine-tuning hyperparameters in β-CVAE yields higher-fidelity data generation compared to PPCA.
  - Supported by both visualizations and quantitative metrics.
- **β-CVAE Model Evaluation:**
  - Exhibits higher fidelity, precision, and consistency.
  - Capable of generating datasets closely mirroring the original distribution.
  - Minimal variation and absence of outliers make it a robust choice for data generation tasks.
- **Potential Implications:**
  - β-CVAE could be a viable alternative to kinetic models in metabolic engineering.
  - Opens new possibilities for synthetic data generation.
- **Acknowledgment of Limitations:**
  - Study limitations: focus on specific models and dataset.
  - Performance metrics provide a snapshot; more exhaustive evaluation could involve a broader spectrum of metrics.
  - Testing hyperparameters and architectures were limited by time constraints.

Future Work

- **Continuous Exploration and Refinement:**
  - Despite limitations, promising performance of β-CVAE suggests its potential for synthetic dataset generation in metabolic engineering optimization.
  - Future investigations could focus on refining existing models, exploring novel architectures, and extending applicability to diverse datasets.
- **Noteworthy Aspect:**
  - The relative newness and underutilization of Variational Autoencoders and Conditional Variational Autoencoders in this field highlight untapped potential for advancing data generation methodologies, especially for floating-point number generation.

References

1. B. Alberts et al., Molecular biology of the cell, 6th ed. New York, Ny: 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, "VAESNE: 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. C. E. Lawson, J. M. Marti, T. Radivojevic, S. V. R. Jonnalagadda, R. Gentz, N. J. Hillson, S. Peisert, J. Kim, B. A. Simmons, C. J. Petzold, S. Singer, A. Mukhopadhyay, D. Tanjore, J. G. Dunn, and H. G. Martin, "Machine learning for metabolic engineering: A review.," Metabolic Engineering, vol. 63, pp. 34–60, 2021.C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.