Optimizing strains in Metabolic Engineering: comparative analysis of β-Conditional Variational Auto-encoder and Probabilistic PCA for synthetic data generation

Uğur Doruk Kırbeyi[1]

Supervisor(s): Prof. Dr. Thomas Abeel[1]**, MSc. Paul van Lent**[1]

[1]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: Uğur Doruk Kırbeyi
Final project course: CSE3000 Research Project
Thesis committee: Prof. Dr. Thomas Abeel, Prof. Dr. Alan Hanjalic, MSc. Paul van Lent

Abstract

This research explores the landscape of dataset generation through the lens of Probabilistic Principal Component Analysis (PPCA) and β-Conditional Variational Auto-encoder (β-CVAE) models. We conduct a comparative analysis of their respective capabilities in reproducing datasets that mirror the distribution of the original data that comes from a hypothetical pathway kinetic model based on an *E.coli* strain using varied parameter settings falling within a specified range. The requirement of significant prior investment in acquiring accurate details about the distinct mechanisms governing each reaction and its parameters for the construction of these kinetic models push us to find alternative ways to generate data that guide metabolic engineering processes. This paper tries to find a viable option through compression algorithms that reduce dimensionality. The PPCA model demonstrates commendable fidelity in capturing overarching patterns, though areas for refinement in reproducing specific data points are identified. In contrast, the β-CVAE model exhibits higher fidelity, precision, and consistency, positioning it as a robust choice for data generation tasks. This study was constrained by both time and the specificity of the model architectures and the dataset. These limitations underscore the imperative for continual exploration and refinement within the dynamic landscape of generative modeling. Opportunities could be found in the refinement of both VAE, CVAE and β-CVAE models utilizing varied hyperparameters alongside different architectures, to increase applicability across diverse datasets within the realm of metabolic engineering.

Introduction

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, through a process named pathway optimization, often aimed at producing commercially valuable compounds such as fuels, vital chemicals, or pharmaceuticals [2].

A main challenge in metabolic engineering is the process of producing industrial strains that have a viable product flux value through combinatorial pathway optimization. One of the most important factors here is the amount of factorial space searches that need to be done for an application of the strain in order to guide this engineering process [2]. The pathway that we are testing for in this paper comes from a kinetic model, which is a set of Ordinary Differential Equations (ODEs) that is used to solve an initial value problem, where these values correspond to concentrations of metabolites. Kinetic models allow for detailed predictions of the global metabolic behavior that come up from dynamic concentration changes of cellular systems and components [3]. These models can be employed to compute adjustments required for optimizing parameters such as the product flux of desired compounds, all the while minimizing the other functions of the host organism to a fundamental yet crucial level [4]. However, the construction of kinetic models require a considerate amount of prior investment in acquiring accurate details about the distinct mechanisms governing each reaction and its parameters [5]. 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 [6]. Generative models aim to capture the inherent data distribution by reducing the dimensionality, however complex it may be, as is usually the case for the inner-workings within cellular systems. Thus, these models facilitate the generation of new data samples that closely mimic this underlying probability distribution.

In the realm of optimizing metabolic pathways, a spectrum of machine learning methodologies and models has emerged, some showcasing promising results in enhancing these intricate processes. An example is the review by Lawson et al., which explains various machine learning driven methods that have been carried out in order to achieve a higher product flux [7]. Being that maximizing the product flux is the main focus, 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 handwriting generation [8]. An example of these generative models, the β-Conditional Variational Auto-encoder (CVAE), a class of neural networks designed for unsupervised learning, 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. Alongside it, the Probabilistic PCA (PPCA) [9] model will also be compared as a baseline option.

Although Variational Auto-encoders of any kind are not commonly used in the field of generating floating-point numbers, we propose that with the right architecture and parameter selection, they can be a viable option. The research question of this paper is the following: "How can β-Conditional Variational Auto-encoders be effectively utilized to generate high-fidelity synthetic data for optimizing strains in metabolic engineering compared to the baseline model?". Throughout this paper, we will use the term "fidelity" to denote the degree of similarity between the generated dataset and the original dataset's shape, values, probability distribution and distribution function. Fidelity, in this context, reflects the accuracy with which the synthetic data captures the essential characteristics of the underlying distribution. Additionally, we introduce the term "disentanglement," which refers to the extent to which the learned representations in the synthetic data isolate distinct features, contributing to a clearer understanding of the underlying structure. Through answering the sub-questions, namely: "What are the key parameters and features within β-CVAEs that significantly influence the fidelity and quality of synthetic data generated?" and "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?", we will be finding the answer to our main question.

This paper conducts a comparative analysis of synthetic data generation for strain optimization in metabolic engineering, focusing on the effectiveness of the β-Conditional Variational Auto-encoder and Probabilistic PCA models. We systematically analyze methodologies, implementations, and architectures of the PPCA and β-CVAE models, including evaluation criteria. Empirical findings regarding data generation fidelity are presented, followed by a critical discussion of implications in metabolic engineering, contributions to related work, limitations, and potential avenues for future research. Responsible research principles observed throughout the process are discussed in the penultimate section. The concluding section summarizes key insights from the comparative study, outlining overarching contributions and implications for the field.

Methodology

Data Collection and Preprocessing:

For the purpose of this research, the provided original dataset that is utilized to train the models comprises of 5000 items, each with 19 features and a product flux value as the last column, totaling to 20 columns, generated by a hypothetical pathway kinetic model based on a *E.coli* strain using varied parameter settings falling within a specified range, as explained further in [10] by van Lent et al. These settings are determined concerning a preset initial parameter configuration. The features in our dataset are the parameters of the ODEs (the kinetic model itself), that are defined w.r.t. an initial parameter configuration.

The data itself has a combinatorial nature and it needs to be preprocessed and formatted in order to suit the requirements of the β-CVAE and PPCA models. This preprocessing is done by first removing the indexing column from the data and splitting the total number of items to 60% training and 40% test sets. The conditioning label input for the sampling process of the β-CVAE is the test set itself, thus it conditions on creating results similar to the set of data that includes the product flux values (20 features in total). The distribution of the values of each feature in the dataset correspond moderately to a uniform distribution between a range of values for each feature, an example of this can be seen in Figure 1.

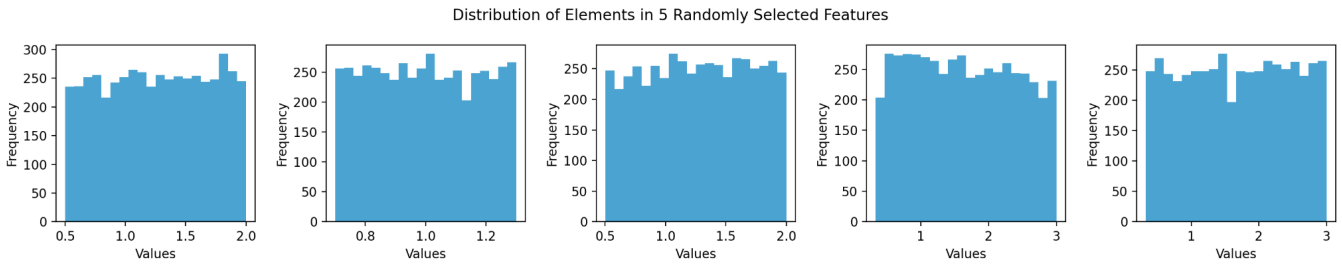

Figure 1: Histograms depicting the distribution of data across five randomly selected features within the dataset reveal a moderate correspondence to a uniform distribution between specific value ranges for each feature. The axes have been configured to enhance visualization clarity, with emphasis placed on the shape of the distributions rather than the precise value ranges.

Model Architectures:

PPCA Model

The Probabilistic PCA involves adopting a probabilistic model for the data and leveraging the latent representation derived from the principal components obtained through the Principal Component Analysis (PCA) process. PCA reduces the dimensionality of the data by encapsulating the factors that contribute the most to the variation in the dataset, effectively identifying the principal components [9].

β-CVAE Model

The β-CVAE model comes from the combination of two more commonly known and used models, namely the β-VAE and CVAE models. The β-VAE model, first introduced by Higgins et al., is an extension of the traditional VAE architecture with an additional hyperparameter called "beta." The beta-VAE introduces a balance between the fidelity of reconstructed data and the disentanglement of learned latent representations [11].

The latent space is where the model learns the hidden representations of the original dataset. In VAEs, the latent space represents a reduced-dimensional space where the essential features that explain the observed data are captured. The latent space is structured such that each point in the latent space corresponds to a probability distribution rather than a deterministic representation. This probabilistic nature allows for flexible and continuous representations [8, 12].

The standard VAE consists of an encoder network that maps input data to a probability distribution in the latent space, a decoder network that reconstructs the input data from samples in the latent space, and a loss function that encourages the learned latent space to follow a specific distribution, which is typically a Gaussian distribution [8]. In a beta-VAE, this loss function is modified to include a scaling factor called beta, which controls the emphasis on disentanglement [11]. The loss function usually has two components: a reconstruction loss, which ensures that the reconstructed data is similar to the input

data, and the Kullback-Leibler Divergence, which is a regularization term that encourages a more disentangled and structured latent space. The beta parameter adjusts the strength of the regularization term.

Kullback-Leibler Divergence (KL-Divergence), often denoted as D_{KL}, is a measure of how one probability distribution diverges from a second probability distribution. For two continuous probability distributions $p(x)$ and $q(x)$ over the same continuous random variable x, the KL Divergence is defined as:

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

Here, $D_{KL}(P \parallel Q)$ represents the KL-Divergence of distribution $p(x)$ from distribution $Q(x)$. The KL-Divergence term is non-negative and it equals to zero if and only if p and p are the exact same distribution. It is also non-symmetric, meaning the order of p and p matters [13, 14].

By tuning the beta parameter, one can control the trade-off between faithful reconstruction and disentangled representation. A higher beta value promotes a more disentangled representation but may sacrifice some fidelity in the reconstruction, while a lower beta value prioritizes faithful reconstruction at the potential expense of disentanglement. Adjusting beta allows the exploration of different levels of representation learning in the latent space.

A Conditional Variational Auto-encoder (CVAE), proposed by Sohn et al., is also an extension of the traditional (VAE) that incorporates conditional information during the encoding and decoding processes. The key idea behind a CVAE is to condition the model on additional information, such as class labels or other relevant attributes, to generate more controlled and specific latent representations. By conditioning the model on additional information, the CVAE encourages the learned latent space to be influenced by the provided conditions. This allows for controlled generation or modification of data based on specified attributes. CVAEs are particularly useful when there's a need to generate data with specific characteristics such as the case in our research for generating high-fidelity data [15].

The β-CVAE combines the improvements from both models in order to both be able to control the trade-off between the reconstruction loss and the regularization term through the beta parameter, while also being able to condition the data generation process on a given input label set in order to maximize the fidelity of the generated data to the kinetic model data.

Implementation of the β-CVAE and PPCA Models:

Jupyter Notebook Setup:

We have created separate structured Jupyter Notebook environments (Python 3.11) for implementing the β-CVAE and PPCA models using the latest stable versions at the time of the *PyTorch* (2.1.2) and *NumPy* (1.26.3) libraries respectively. We structured the data using the *pandas* (2.1.4) library methods and utilized several methods from *matplotlib* (3.8.2) and *scikit-learn* (1.4.0) libraries for evaluation purposes.

Probabilistic Principal Component Analysis Model:

We implemented and trained the PPCA model based on the implementation by Oliver K. Ernst [1] using various methods from the *NumPy* library. We have also adjusted the number of principal components to 10 in order to both align with the dataset characteristics, and to keep the latent dimensionality consistent in between the models.

β-Conditional Variational Auto-encoder Model:

We implemented and trained the β-CVAE model on a computer with a CUDA enabled GPU using various methods from the *PyTorch* library based on the implementation from the article by Konstantin Sofeikov [2]. The architectural design and hyperparameters of the model were determined through empirical selection after iterative experimentation. The encoder comprises two hidden layers: the first layer decreases the input dimensionality from 20 to 15, and the second layer reduces it again to the chosen latent dimension of 10, with both layers incorporating a Rectified Linear Unit (ReLU) activation function. The decoder follows a similar pattern with two layers, increasing the dimensionality to 15 in the first layer, and back to the desired output dimensionality of 20 in the second layer, with a ReLU activation function between them. The training employed the Adam optimizer, utilizing a combination of Mean Squared Error (MSE) and Kullback-Leibler divergence as the loss function. The specific hyperparameters for the model are summarized in Table 1 in the following section.

Training Hyperparameters:

In determining the hyperparameters for our model, we employed a pragmatic yet thoughtful approach due to time constraints and limited computing resources. While an exhaustive grid search was unfeasible within the given timeframe, the chosen hyperparameter values resulted in practical and effective outcomes for our research objectives. The selection process, although empirical, was not arbitrary; it was informed by a systematic consideration of the model's requirements and a preliminary

[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. 20, 2024)

[2] K. Sofeikov, "Implementing conditional variational auto-encoders(cvae) from scratch", medium.com. https://medium.com/@sofeikov/implementing-conditional-variational-auto-encoders-cvae-from-scratch-29fcbb8cb08f (Accessed Jan. 20, 2024)

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

Table 1: Training hyperparameters of the β-CVAE model

exploration of hyperparameter spaces. A balance was struck between the need for meaningful results and the constraints imposed by the available resources. This approach not only allowed us to produce usable results within our constraints but also makes the research more accessible and reproducible for individuals without access to powerful computational resources.

In addition to the aforementioned considerations, we intentionally maintained consistency in the latent dimensionality across our two models, fixing it at 10. This decision aimed to facilitate a direct and fair comparison between the models, ensuring that differences in performance were attributed to variations in the model architectures and training processes.

We chose the learning rate and weight decay parameters to mitigate potential overfitting issues. The learning rate dictates the step size during optimization, impacting the convergence speed, while weight decay regulates the model's tendency to overemphasize specific features during training, serving as a regularization mechanism.

The primary focus of our exploration was on different beta values, as they play a crucial role in the disentanglement of latent representations. This decision aligns with our research objectives, prioritizing the investigation of beta values to understand their effects on model behavior.

While batch size was considered a secondary variable, we explored its impact on training dynamics. In the context of the β-CVAE model, different batch sizes influence the stochastic gradient descent process. Smaller batches introduce more noise but can help escape local minima, while larger batches offer more stable gradients but may converge to suboptimal solutions. This exploration allowed us to gauge the sensitivity of the model to variations in batch size and its implications for training efficiency and generalization.

Synthetic Data Generation and Model Evaluation:

PPCA synthetic data generation

Utilizing the PPCA model, we have generated five sets of synthetic data. As the only alterable hyperparameter in this model is the number of principal components, and we aim to keep both models in the same amount of latent dimensions, we utilized only the first 10 principal components of the dataset, effectively reducing the dimensionality of the dataset to its half. The quality and fidelity of the synthetic datasets are validated by comparing statistical properties and distributions with the original dataset, which are further explained in the following subsection.

β-CVAE synthetic data generation

We employed the trained β-CVAE model to generate a total of 15 synthetic datasets. A comparative analysis of the synthetic data produced by the β-CVAE and PPCA are conducted, focusing on aspects such as fidelity to the distribution and representation of the original dataset.

Evaluation of the generated datasets

Alongside visual examination for evaluation, both newly generated datasets are applied to the kinetic model in strain optimization. The efficiency and accuracy of strain optimization results obtained from β-CVAE generated synthetic data versus PPCA generated synthetic data are evaluated and compared to the results from the kinetic model. Model performance is assessed through KL divergence and two sample Kolmogorov-Smirnov (KS) test, or sometimes colloquially referred to simply as the Smirnov test, value comparisons. We have already explained what KL divergence is, as it was also used in the training process of the β-CVAE model.

The KS test is a statistical test used to assess whether two sets of data appear to follow the same distribution. The test begins with the null hypothesis, which is an assumption that two samples X and Y are drawn from the same theoretical distribution. It employs the test statistic, denoted as D, which represents the maximum vertical deviation between the empirical distribution functions (EDFs) of the samples being tested. If the calculated test statistic D is greater than the chosen value, the null hypothesis is rejected. Smaller test statistic values indicate that the distributions are indeed very close and the opposite for higher values. The KS test is non-parametric, therefore it makes minimal assumptions about the underlying distribution of the data [16]. For two EDFs $F_n(x)$ and $G_m(x)$ for the two samples X and Y respectively, the KS test statistic for two samples is defined as:

$$D = max[sup_x|F_n(x) - G_m(x)|, sup_x|G_m(x) - F_n(x)|]$$

Where sup_x denotes the supremum, which is the smallest upper bound.

Additionally, a form of visual inspection is carried out by presenting generated data in lower dimensions utilizing principal component analysis. This type of evaluation aims to determine the capability of the models in capturing and replicating the data distribution of the original dataset.

Regrettably, due to the unavailability of access to knowledge and equipment, a direct assessment of the success of metabolic engineering fabrication with the generated data is unfeasible.

Data and Code Availability:

The used and produced datasets for this study, alongside the code with instructions are available in the following GitHub repository: https://github.com/Abeellab/RP2023_kirbeyi

Results

To address the research question, we first examine the performance of PPCA as a baseline model. The motivation behind the research question is to reduce the dependency on kinetic models in metabolic engineering by utilizing generative models. PPCA, being a traditional generative model, is chosen as the baseline for comparison.

Moving on to the main comparison, we investigate the performance of the β-CVAE model compared to the PPCA model. We hypothesize that, by fine-tuning hyperparameters, β-CVAE can achieve higher-fidelity data generation than the PPCA model. This hypothesis again stems from the motivation to explore alternatives to kinetic models in metabolic engineering and the relative lack of exploration of β-CVAE in this specific field for generating floating-point numbers.

We present the results obtained from evaluating the fidelity of the generated datasets from the PPCA model to the original dataset from the kinetic model using two distinct metrics: KL-divergence and the KS test, as explained in the previous section. These metrics serve as critical measures for assessing how well the generated datasets capture the underlying distribution of the original data in this subsection.

KL-divergence quantifies the dissimilarity between probability distributions, providing insights into the information lost during the generation process. Lower KL-divergence values suggest a more faithful representation of the original dataset. As calculating the KL-divergence requires us to know the probability distribution functions (PDFs) of the datasets, we have chosen Kernel Density Estimation (KDE) to estimate the respective PDFs of the datasets. We chose KDE as it is a non-parametric estimation method that allows flexibility towards the inherent data characteristics.

Additionally, the KS test assesses the agreement between the empirical distribution function of the generated datasets and that of the original dataset. Smaller KS test statistics indicate a closer match between the generated and original datasets.

Lastly, we compare the Mean Squared Error (MSE) values between the product flux column of the β-CVAE generated dataset and the resulting product flux column from running the kinetic model with the β-CVAE generated feature values.

PPCA as a baseline model

We present KL-divergence values in Table 2, for generated datasets from five different training instances of the PPCA model, highlighting the divergence between each generated dataset and the original dataset.

To assess the PPCA model as a baseline, we explore how the PPCA model fares in terms of fidelity to the original dataset. The important factor here is that the results should first be able to be tested for fidelity. This means that the model should be able to produce results that at the very least resemble the original dataset in terms of shape and values. When we examine Table 2, we are able to see that PPCA is not only able to generate results that can be utilized to test for fidelity; but also produce results in a consistent and robust way, signified by the relatively close KL-divergence values from each training. There is no set bound that we have chosen to consider values low or high here, but as explained before, the lower these values are the better. We cannot classify the results here before we do the comparison with the β-CVAE model. However, we thus address the question of whether the PPCA model is a viable option for being used as a baseline in the synthetic data generation process.

β-CVAE outperforms the PPCA

For the β-CVAE model, we again present KL-divergence values for five distinct beta values (0.1, 0.25, 0.5, 0.75, and 0.9) and three different batch sizes (25, 50, and 100). These values are organized in Table 3, and visualized in Figure 2 to provide a comprehensive overview of how different model configurations influence the dissimilarity between the generated and original datasets.

To see that the β-CVAE model outperforms the baseline, which is the PPCA model, we first take a look at the KL-divergence values. Although on their own these values would not be giving us much information as we have not set lower or upper bounds for these values, the comparison between the best PPCA value and the best β-CVAE value explains quite a bit. There is a factor of 10 in between the values, but that does not directly equate to being 10 times better than the other. The KL-divergence value itself is not a direct measure of fidelity but rather a measure of information lost during the generation process. A 10 times lower KLD suggests less information loss. Thus in the way that we have defined the term fidelity in the introduction, the results from the β-CVAE model indicate higher-fidelity to the original dataset in terms of similarity in probability distributions.

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

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 3: 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.

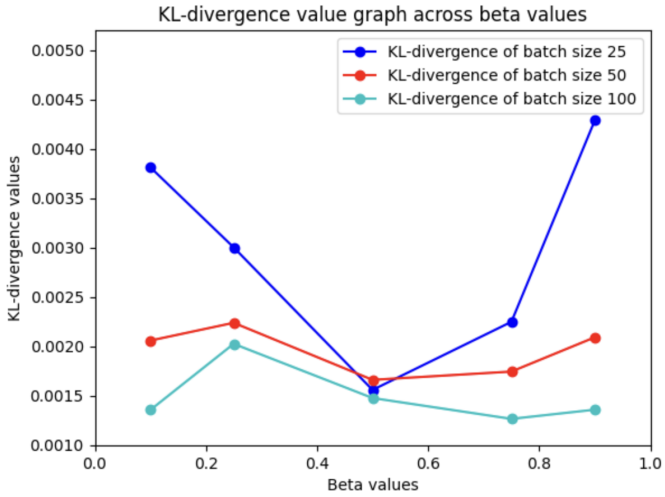

Figure 2: Line graph comparison of KL-divergence values across beta values per batch size.

The influence of hyperparameters, specifically the beta and batch size, on the KL-divergence performance of the model reveals insightful patterns. The beta hyperparameter, which controls the weight assigned to the disentanglement term in the loss function, demonstrates an expected impact on KL-divergence values. As anticipated, higher beta values result in lower KL-divergence values, indicating a stronger emphasis on disentanglement and improved fidelity to the original dataset. However, this trend only holds up to a certain threshold, with 0.75 as the optimal beta value in our case for the overall best performer, surpassing the performance of 0.9 in all batch sizes. This suggests a diminishing return on disentanglement beyond a certain point, highlighting the need for a balanced weighting in the loss function.

Furthermore, the interplay between beta values and batch sizes provides additional insights. For batch sizes of 25 and 50, the KL-divergence values are lowest when using a beta of 0.5. This observation suggests that, for these relatively smaller batch sizes, a balanced emphasis on the disentanglement during training achieves better KL-divergence performance. This could be attributed to the balance between preserving global structure and capturing local variations within smaller batches. Notably, this could suggest that high beta values might lead to over-regularization, hindering the model's ability to capture intricate details present in the dataset.

Moreover, the association between beta values and batch sizes prompts consideration of dataset characteristics. The choice of a balanced beta (0.5) for smaller batch sizes may indicate that these batches are not large enough to fully benefit from a higher disentanglement emphasis. However, as batch sizes increase, the model's capacity to capture more complex relationships also expands. This is reflected in the trend where, with batch size 100, a higher beta (0.75) becomes optimal, suggesting that a larger batch size allows the model to benefit more from increased disentanglement.

It's essential to note that while even larger batch sizes may lead to better results, there is a trade-off. Beyond a certain batch size, computational efficiency and model accuracy may be compromised. Increasing the batch size too much can result in longer training times and potentially hinder the model's ability to generalize well. Therefore, the choice of an optimal batch size involves a careful consideration of the dataset size, computational resources, and the desired balance between accuracy and efficiency. Overall, the observed interdependence between beta values and batch sizes highlights the importance of fine-tuning

hyperparameters based on the specific characteristics and requirements of the dataset in question.

To further assess the comparative performance of the β-CVAE model against the PPCA, we present and examine the KS test statistic values obtained from the lowest KL-divergence value producing dataset from both the PPCA and β-CVAE models in Tables 4 and 5. These statistics represent the maximum vertical deviations between the empirical distribution functions of the generated datasets and the original dataset. The focus on these KS test values allows us to pinpoint highlights in the performance of the model to yield the closest match to the original data distribution, signifying the fidelity.

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 4: 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)

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 5: 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)

The KS test statistic values show that, while the β-CVAE does not consistently outperform the PPCA across all features, there are instances where the β-CVAE demonstrates superior performance, showcasing strengths in certain features within the dataset. It is crucial to recognize that the KS test statistic values indicate nuanced differences, with some features exhibiting better fidelity in the β-CVAE generated datasets and others showing less consistent results compared to the PPCA.

The observation of varying performance suggests that the β-CVAE has specific features that are better captured, indicated by lower test statistic values, and others where improvement is needed, such as features 11, 12, 13, 17 and 19. Despite not exhibiting uniform superiority in all aspects, the β-CVAE model's overall performance is close to that of the PPCA on many features. This mixed performance underscores the potential of the β-CVAE as a promising option for data generation, with room for refinement and optimization in areas such as model architecture and hyperparameter selection. Further investigation into parameter sensitivity, trade-offs, and potential optimizations could unlock the full capabilities of the CVAE model in generating high-fidelity datasets for metabolic engineering applications.

PCA Visualizations Confirm Performance Gap

To further explore the gap in performance in fidelity between the models, we employ PCA visualization to present scatter plots illustrating the distribution of data points in the reduced-dimensional space defined by the first 10 principal components. Specifically, we organize these visualizations into five plots, each capturing a distinct pair of principal components: the first and second, third and fourth, and so forth, with the fifth plot visualizing the last pair of components, as can be seen in Figure 3.

We utilized PCA as it is a powerful dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space, capturing the most significant sources of variation. By visualizing the data in the reduced-dimensional space, we gain valuable insights into the structure and relationships within the dataset.

The importance of PCA visualizations in our context lies in their ability to uncover patterns and anomalies in the generated datasets when compared to the original dataset. Each scatter plot represents a unique perspective on the distribution of data points, enabling us to identify potential discrepancies, shifts, or distortions introduced during the generation process. Furthermore, PCA visualizations offer an intuitive means of assessing the alignment of data points between the generated and original datasets. A faithful representation of the original dataset would exhibit similar clustering across corresponding principal com-

ponents. Deviations or disparities observed in the visualizations may indicate areas where the generated datasets diverge from the original data distribution.

In summary, the PCA visualizations presented in this subsection provide a comprehensive and visual assessment of the fidelity of the generated datasets. Their interpretation offers valuable insights into the success of different model variants in preserving the essential structure and relationships present in the original dataset.

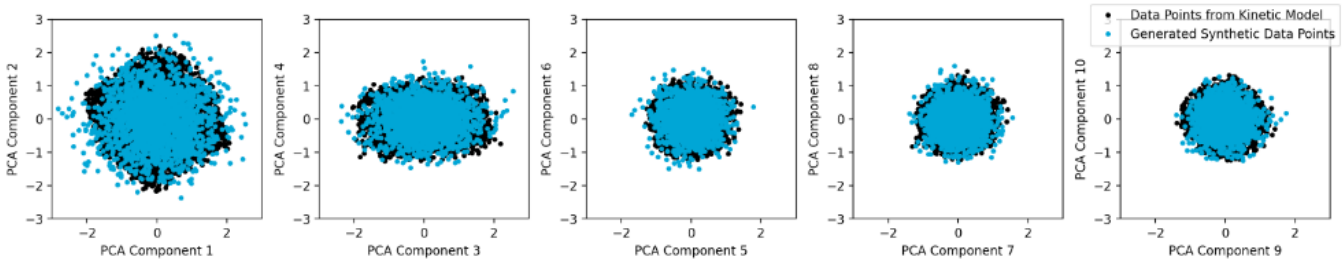

Figure 3: PCA visualizations of the best performing dataset from the PPCA model, up to 10 principal components.

In examining the PPCA model's PCA visualization plots, several key observations come to light. The plots demonstrate a visually adequate overall coverage, with points occupying a similar region as the original dataset. This indicates that the PPCA model is able to capture the general distribution of the data. However, the plots show visible variation and outliers in especially the first two plots that cover the first four principal components. This suggests that the model may not precisely replicate the full extent of inherent data variability and these outliers warrant further investigation to understand the nature of the discrepancies.

While the majority of generated points are in close proximity to the original points, there are instances where the distances are larger and the space of points are not covered. This suggests that, while the model captures the general trend, there are areas where it might struggle to reproduce specific data points accurately.

The orientation of principal components in the plots aligns with the major axes of variability in the original dataset. This suggests that the model captures the primary sources of variation adequately. The observed characteristics from the plots align well with the objective of capturing the essential features of the dataset, highlighting the PPCA model's success in representing key data patterns.

The PPCA model demonstrates a satisfactory level of fidelity, effectively capturing the distribution and structure of the original dataset. Attention to outliers and variations will be critical for refining the model and ensuring more precise replication of specific data points.

Adding on the knowledge acquired from the PPCA model, our focus shifts to the PCA visualizations of the data generated by the β-CVAE. This section offers a thorough assessment of how well the model captures the distribution of the original dataset. We again organize these visualizations into five plots, each capturing a distinct pair of principal components: the first and second, third and fourth, and so forth, with the fifth plot visualizing the last components, showcased in Figure 4.

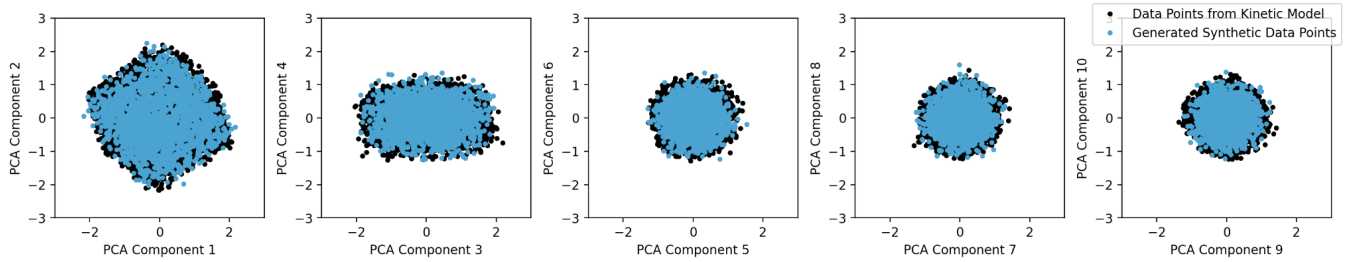

Figure 4: PCA visualizations of the best performing dataset from the β-CVAE model, up to 10 principal components.

The plots exhibit a visually improved overall coverage, with points closely aligning with the distribution of the original dataset. This suggests that the β-CVAE model not only successfully captures the general structure of the data, but does it better than the PPCA model. Outliers and variations from the original dataset's clustering are scarce in the plots. This signifies that the β-CVAE model successfully avoids introducing data points that deviate significantly from the original distribution. The tightly clustered nature of the points indicates that the β-CVAE model reproduces data points with higher precision and consistency.

The proximity of generated points to the original points is remarkably close, underlining the precision with which the β-CVAE model reproduces the actual data points. This proximity suggests a higher degree of fidelity. The orientation of principal components in the plots aligns with the major axes of variability in the original dataset. This suggests that the model successfully captures the primary sources of variation, similarly to the PPCA model.

However, there are small but still visible spaces of data points that are not covered by the generated dataset, especially around the edges of the clustering. Although these spaces are significantly smaller than the PPCA model, this shows that there is still room for improvement in the model architecture. Identifying and addressing these improvement points could further improve the model's precision.

In contrast to the PPCA model, the β-CVAE plots demonstrate a more refined and faithful representation of the original dataset. The minimal variation, absence of outliers, proximity to original points and an enhanced overall coverage of the space distinguish the β-CVAE model's superior performance and potential as an option for synthetic data generation.

MSE Scores Reinforce Performance Gap

To further reinforce the findings, we calculate MSE scores for the product flux values between the best-performing generated dataset from the PPCA model and the resulting product flux column from running the kinetic model with the parameter values from the same PPCA generated dataset. We did the same for the β-CVAE model's best-performing generated dataset and the resulting product flux column from running the kinetic model with the parameter values from the β-CVAE generated dataset. These values are organized in Table 6

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

Table 6: 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)

Analysing the MSE scores, the score associated with the β-CVAE model's dataset is more than a factor of 10 smaller than the corresponding score from the PPCA model. This difference suggests that the β-CVAE model excels in producing both parameters and product fluxes that are not only more effective in the kinetic model in comparison but also exhibit a closer alignment with the actual values derived from the kinetic model.

We suggest that the values produced from the β-CVAE model are more effective in the kinetic model because of a crucial observation, which is the presence of blank product fluxes in the resulting dataset from the kinetic model of the dataset coming from the PPCA model, hinting at potential challenges in generating datasets that lead to successful kinetic model simulations. This discrepancy further underscores the capacity of the β-CVAE model to produce more robust and reliable datasets, demonstrating its superiority in capturing the underlying structure of the metabolic system.

In essence, the pronounced disparity in MSE scores reaffirms the β-CVAE model's superior robustness, fidelity, and performance in the data generation process compared to the PPCA model. The ability of the β-CVAE model to yield datasets that not only produce viable kinetic model results but also closely approximate the actual values underscores its efficacy in advancing the state-of-the-art in synthetic data generation for metabolic engineering applications.

In conclusion, the results collectively highlight PPCA as an adequate baseline model but underscore β-CVAE's superiority in terms of fidelity, robustness, and accuracy. The hypothesis that fine-tuning hyperparameters in β-CVAE leads to higher-fidelity data generation compared to PPCA is supported by both visualizations and quantitative metrics. This research could potentially pave the way for β-CVAE as a viable alternative to kinetic models in the field of metabolic engineering, opening new possibilities for synthetic data generation.

Responsible Research

Minimization of Bias:

To ensure the integrity of our research, efforts were made to mitigate potential sources of bias at various stages. The utilized dataset from the kinetic model was constructed by randomly selecting 5000 items from the overall list of items from running the model. This randomness in the data construction ensures the mitigation of bias while transparent reporting of preprocessing steps, which does not include any changes to the items in the dataset, and model configurations being prioritized aid this mitigation process.

Reproducibility:

Detailed documentation of experimental setups and model architectures, and access to the code repository containing both models is provided in the Data Availability section to facilitate the replication of our work. Researchers and practitioners are encouraged to replicate our experiments and validate the findings, thereby strengthening the soundness and dependability of the research outcomes.

Open Data Sharing:

In line with the commitment to open science, datasets used in the research, including the original dataset and generated datasets, are made available for public access, again through the given link in the Data Availability section.

Conclusions and Future Work

In conclusion, this research has undertaken a comparative analysis of dataset generation for the sake of optimizing strains for metabolic engineering, employing PPCA and β-CVAE models. The analysis has shown distinctive attributes of each model, providing insights.

The Probabilistic Principal Component Analysis model showcased adequate fidelity, successfully capturing overarching data patterns. However, the discerned areas for improvement, particularly in accurately reproducing specific data points, underscore the ongoing quest for refinement in generative methodologies.

Conversely, the β-Conditional Variational Auto-encoder model emerged as a standout performer, exhibiting higher fidelity, precision, and consistency. Its capability to generate datasets closely mirroring the original distribution, coupled with minimal variation and absence of outliers, positions the β-CVAE model as a robust choice for data generation tasks.

The insights gained from this study could motivate for further research in the domain of generative modeling. Future investigations could delve into refining existing models, exploring novel architectures, and extending the applicability of generative models to diverse datasets and domains.

As with any research, it is essential to acknowledge this one's limitations. The study focused on specific generative models and a singular dataset, and the findings may not be universally applicable. Additionally, the performance metrics employed provide a snapshot of model efficacy, but a more exhaustive evaluation could involve a broader spectrum of metrics, such as Quantile-Quantile plots to compare the quantiles of the generated and original datasets and Frechet Distance comparisons to measure the similarity between two shapes of data distributions. Although we tested a set of hyperparameters and architectures, it was not a possibility in the given timeframe to exhaustively search for any and all possible combinations. Therefore, a significantly better result could be achieved especially from the β-CVAE model through the trial of other hyperparameters and architectures. External factors such as computational resources and data complexity also influence the generalizability of the results.

In light of these limitations, it is evident that continuous exploration and refinement are imperative. Nevertheless, the promising performance of the β-CVAE model, especially with room for improvement, suggests its potential as a viable option for generating synthetic datasets in the optimization of strains for metabolic engineering. This is particularly noteworthy, as Variational Autoencoders and Conditional Variational Autoencoders remain relatively new and unused in this field, especially for floating-point number generation, holding untapped potential for advancing data generation methodologies.

References

[1] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell*. Garland Science, 6 ed., 2015.

[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.

[3] A. F. Villaverde, S. Bongard, K. Mauch, E. Balsa-Canto, and J. R. Banga, "Metabolic engineering with multi-objective optimization of kinetic models," *Journal of Biotechnology*, vol. 222, pp. 1–8, 2016.

[4] J. Almquist, M. Cvijovic, V. Hatzimanikatis, J. Nielsen, and M. Jirstrand, "Kinetic models in industrial biotechnology–improving cell factory performance," *Metabolic engineering*, vol. 24, pp. 38–60, 2014.

[5] A. Khodayari and C. D. Maranas, "A genome-scale escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains," *Nature communications*, vol. 7, no. 1, p. 13806, 2016.

[6] J. M. Graving and I. D. Couzin, "Vaesne: a deep generative model for simultaneous dimensionality reduction and clustering," *bioRxiv*, p. 2020.07.17.207993, 01 2020.

[7] 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.

[8] C. Doersch, "Tutorial on variational autoencoders," 2021.

[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 1999.

[10] P. van Lent, J. Schmitz, and T. Abeel, "Simulated design–build–test–learn cycles for consistent comparison of machine learning methods in metabolic engineering," *ACS Synthetic Biology*, vol. 12, no. 9, pp. 2588–2599, 2023. PMID: 37616156.

[11] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[12] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent space for multi-label classification," vol. 31, Feb. 2017.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[14] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, pp. 183–233, 1999.

[15] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[16] J. W. Pratt, J. D. Gibbons, J. W. Pratt, and J. D. Gibbons, "Kolmogorov-smirnov two-sample tests," *Concepts of nonparametric theory*, pp. 318–344, 1981.