Noname manuscript No.
(will be inserted by the editor)

A Case Study of Beta-Variational Auto-Encoders, disentanglement impacts of input distribution and beta variation based upon a computational multi-modal particle packing simulation

Jason R. Hall[a,b] · **Taylor D. Sparks**[b]

Received: date / Accepted: date

Abstract We present work that quantifies the disentanglement of the reconstruction of β-VAEs varying the hyper-parameter β for three different input distributions[1]. Currently the majority use of VAEs are for image processing and little work has been done in the field of material science using this ML technique to create reconstructions to explore the search for new designs. This work highlights the importance of the distribution shape can be more important than the quantity of data in creating neural network reconstructions such as β-VAEs which has been used for this effort. Furthermore, this work shown highlights that the best disentangled reconstruction doesn't necessarily create the best reconstruction.

Keywords Auto Encoders · Particle packing · ballistic deposition · Packing fraction

1 Introduction

1.1 Motivation

Variational Auto-Encoders (VAEs) is a type of neural network system that utilizes a probabilistic approach in the latent space (mean and standard deviation vectors). One of the purposes of the creation of this tool was to perform

Corresponding Author
Jason Hall
[a]Northrop Grumman Corporation
E-mail: jason.hall@ngc.com

[b]University of Utah
Department of Materials Science & Engineering
Tel.: +1-801-581-8632
E-mail: sparks@eng.utah.edu

efficient learning with probabilistic models with intractable posterior distributions and large datasets[2]. Key applications of image analysis have shown to be promising with the general application towards recognizing handwritten images[3]. VAEs were an improvement over typical Auto-Encoders by enforcing a regularization constraint on the latent space by forming a normal distribution.

An improvement to the VAE model has been done by incorporating a regularization coefficient, β, which constrains the capacity of the latent information and puts implicit independence pressure on the learnt posterior [4]. This regularization coefficient is a weight on the Kullback–Leibler divergence term, D_{KL}. Weighting D_{KL} is an attempt to force the learning process to drive the entropy between the two distributions to zero. A general sense of increasing this coefficient has the effect of encouraging the model to learn the most efficient representation of the data. However, it is theorised that driving β too high can result in a poor reconstruction due to the loss of the high frequency details when compressing the data to the latent space[4].

1.2 Particle Packing Application

Typically Auto-Encoder models have been generally used for image processing and recognition systems[5]. In the recent years these VAEs have been finding applications into research of materials areas such as generative chemistry[6], transient fluid flow[7], molecule optimization[8]. Additionally, there is little work in the area of data distribution inputs into a β-VAE machine learning model[9]. The work by Alam and Shehu apply a variation between input sizes of contact maps (64 square pixels vs 72 square pixels). The data distribution in this work is a much larger type of variation.

This effort compares two types of distributions of data. The first type of distribution is understanding the concepts of packing fractions. A single datapoint used in the learning is made of the composition of three different compounds. The volume occupied by all of the compounds over the total volume is the packing fraction, since the compounds are made up of spheres this results in a value less than one due to the voids between the particles. Each of these three compounds have their own variation to capture realistic manufacturing methods. Creating one of the compounds results in a variation in particle size which can be controlled. This results in a set of three distributions for each data point of learning, an example is shown in Figure 1.

The second type of distribution evaluated in this effort is the distribution of the packing fraction generated from the data set. The prior work[10] from the author indicated a highly Gaussian data set resulted in poor model performance. The author has taken the compilation of the results and removed data points to force the distribution of packing fraction to create three different distributions to see if the distribution of the data can be more of a influence over the number of data points used in learning. The distributions of packing fractions are shown in Figure 2.

Since VAEs utilize a probabilistic scheme in the latent space the authors believe this could see an improvement for problems involving distributions. The importance of understanding the relationship to particle size and distributions relate to the mechanical and ballistic performance of the solid fuel that is comprised of similar compounds. Understanding the geometric variation present is the first step in a series of steps to measure the material performance. The optimal ballistic performance comes from the most densely packed solution and the optimal structural performance comes from the least packed solution. There exist a configuration where both sets of requirements is satisfied. This effort is a series of studies to understand the variables of importance that should be controlled to obtain this optimum configuration for a set of conditions based upon the polymer binder, structural loads, and ballistic requirements.

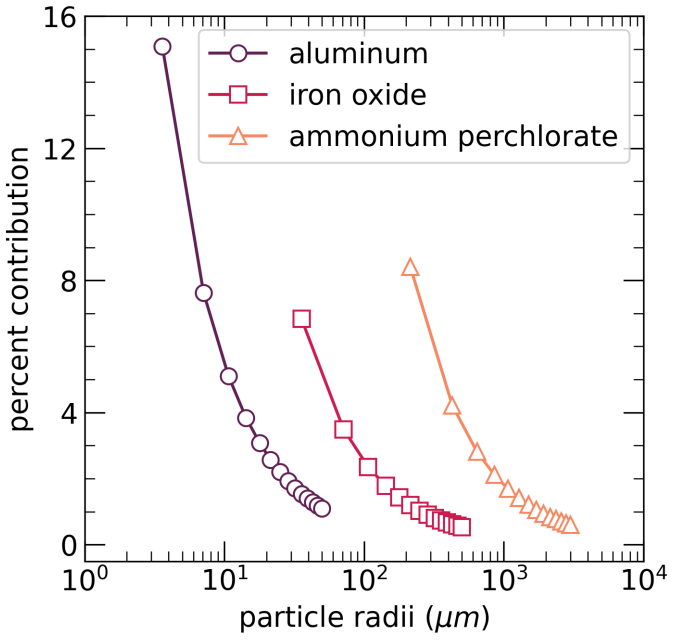

Fig. 1 Example Multi-modal Distribution of a singe packing fraction

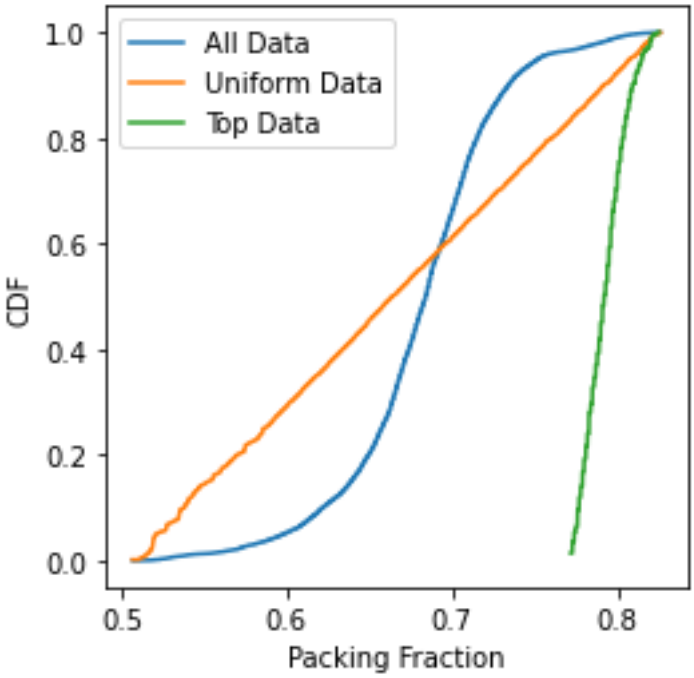

Fig. 2 Cumulative Distribution Functions of different combinations of particle distributions used as variations of input data

The focus of this study is to examine and determine the following:

1) Determine the prediction accuracy of the various input distributions for $\boldsymbol{\beta}$-VAEs
2) Examine the relationship in disentanglement metrics based on a wide range of $\boldsymbol{\beta}$

2 Methods

2.1 Vanilla $\boldsymbol{\beta}$-VAE

The machine learning model used in this study is not a new novel and is an impact assessment of what a basic βVAE model defined by Higgins et. al. [4] and adapted from MNIST applications[3]. Typically the total loss in computing the difference in VAE models is the sum of a reconstruction loss and the D_{KL}. This summation is called the Evidence Lower Bound (ELBO). β is a scalar value that enhances the loss metric based upon the D_{KL} term. The complete ELBO for this situation is shown below in Equation . Main changes from the basic application for this work includes changing the reconstruction loss function from the standard Binary Cross Entropy (BCE) to Mean Squared Error

(MSE). This is due to the order of magnitude of difference in the parameters used in the data set resulted in driving BCE negative, which is technically impossible. BCE is commonly used in creating the loss function for VAEs, however initially stability concerns became evident when applying BCE that were only fixed when normalizing all of the data respective to itself. Even with applying a normalization it was evident that MSE resulted in an significant reduction in magnitude of the loss value. The model was a two layer neural network that scaled down to a two feature latent vector for the mean and standard deviation. Additionally, after decoding the data the mass continuity had to be enforced by scaling the mass fractions to ensure the sum of the three mass fractions equaled one hundred percent. This is a natural law of the physics problem that was needed for a good reconstruction and required for calculating the theoretical via ballistic deposition code.

$$ELBO = \beta(L_{KLD}) + L_{MSE} \tag{1}$$

2.2 Input Data

Initial working on VAEs with this data set resulted in acceptable convergence but poor prediction accuracy. Some of the concerns from the authors' previous work indicated that the accuracy was impacted by the input data distribution being highly Gaussian [10]. Part of this study is to show the impact of various forms of input data. Figure 2 graphically shows three different cumulative distribution functions (CDFs) that are used through the remainder of this work. Table 2.2 quantifies the data within each distribution. The uniform and top data sets are subsets from the original data set constructed in such a way to demonstrate variation in the input data used.

Dataset #	**Data Name**	**Data size**	**Packing Fraction Range**
1	Gaussian Dataset	14,765	0.507-0.826
2	Uniform Dataset	600	0.507-0.826
3	Top Packing Dataset	500	0.750-0.826

2.3 Disentanglement Impact

The key hyper-parameter of this study is β, which the impact of this parameter is more than a convergence term. β controls the amount of disentanglement directly by forcing the distributions in attempt to drive the relative entropy between the datasets to be zero. One would immediately try to drive the model to use high values of beta. However, there is a drawback in this approach, high frequency details are lost by passing through a constrained latent bottleneck. This results in an optimal β that can be measured theoretically with the best disentanglement metric. [4]. Figure 3 graphically shows the impact of β. In

short, an appropriate amount of overlap is needed in the latent space to have interpretability and enough constraint to exhibit the desired structure.

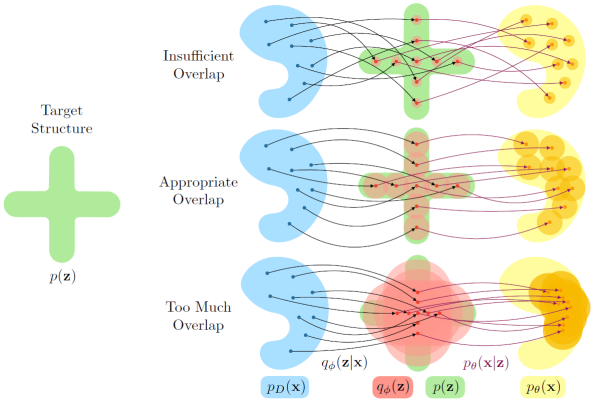

Fig. 3 Illustration of decomposition with varying levels of overlap of the likelihood and prior information[11]. (Top) Too large of a value for β (Middle) Perfect value for β (Bottom) Too small of a value for β.

3 Results and Discussion

3.1 Disentanglement Metric Scoring

Part of the learning process of machine learning (ML) is training the model efficiently. The better the model can segregate the data into groups of the defined outcome the better predictive capabilities of the model. Figure 4 gives insight into this grouping via TSNE of various values of β used in the learning process alone. The TSNE plots here were generated with the sklearn module using the final reconstruction after 250 epochs. This is what condition that was used to compare to the actual that the ballistic deposition code generates. The initial data (top left) has not undergone any learning to show how poor the grouping is based upon the feature vector's values. There are plenty of subgroups, however, this lumps high and low values of packing fraction not able to create good regression alone. The other subplots highlight the β-VAE's reconstruction with the same TSNE procedure. The reconstruction data appears to do a better job of sorting the data, initially into gradients and eventually into groupings based upon their packing fraction. However, these values do not encompass the entire range of values present in the original data. This highlights the trade-off with the β-VAE process, increasing learning with the lack of model reconstruction. It has been shown prior with this data set that the tails of the data are hard to capture from the learning process[10]. It is of note that that each of the shown configurations in Figure 4 has a different range in the legend which highlights a portion of the variability in the reconstruction process. As it is shown in Figure 9 the variability of the output can be small

or large and it is likely that data is lost in the reconstruction process. This is a flaw in reducing feature count to a low level such as this case being reduced down to two features.

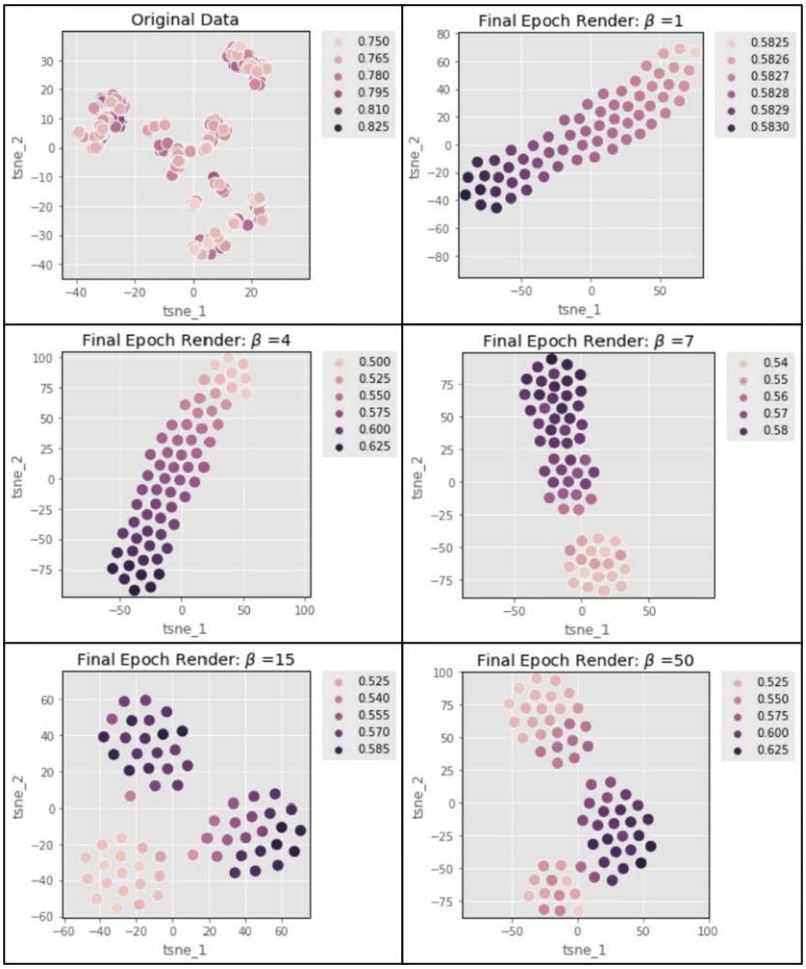

Fig. 4 TSNE plot of various β from the Uniform input data. This graphically shows higher β results in a higher separation or grouping of like results. This is due to the large β forces increased learning of the input parameters.

Many different metrics have been proposed to determine the degree of disentanglement such as: β-VAE, FactorVAE, DCI, SAP, and MIG. Recently, a comparison of these proposed methods to show that most of them do not satisfy the basic requirements of being disentangled[12]. Using this information the MIG will be the metric to compare the disentanglement. Characteristics of a disentangled representation have the following fundamental properties:

Property 1. *A metric gives a high score to all representations that satisfy the characteristic that the metric reflects.*

Property 2. *A metric gives a low score for all representations that do not satisfy the characteristic that the metric reflects.*

Mutual information Gain (MIG) has been shown to be the only real metric to quantify the degree of disentanglement of the methods that are commonly used[12]. MIG comes from probability and information theory and is the the joint entropy of both separate features. A perfect MIG would be composed of identical subsets. MIG will be used to examine this metric in this work in comparing the hyper-parameter β. For two completely random variables the MIG and D_{KL} are the same by definition, however this is only for a limited scenario and does not hold true for all values[13]. MIG used in this effort is calculated after the fact and compares the reconstruction against the computed data. Since the theory of MIG and D_{KL} is of similar approach it's reasonable to train with D_{KL} and to compare with MIG as a metric-like property since it satisfies the disentangled representation. MIG is not a true metric by definition since it does not satisfy the identity of indiscernibles. Additionally, MIG has been shown to be a practical measure in β-VAEs [14]. This effort could not get a fully disentangled data-set with the training (MIG=1, in bits).

3.2 β Variance Results

The β-VAE model described herein was run for 250 epochs for all 27 combinations, this was chosen to ensure that each simulation fully converged. The convergence value of the total loss was approximately 0.1 - 0.4 and behaved similarly to what is shown in Figure 5. This level of loss is expected for a VAE[15]. Then the author took the last reconstruction for comparison to actual computed values. Then 25 actual computations of each of the 27 reconstructions were performed to compare against the true value.

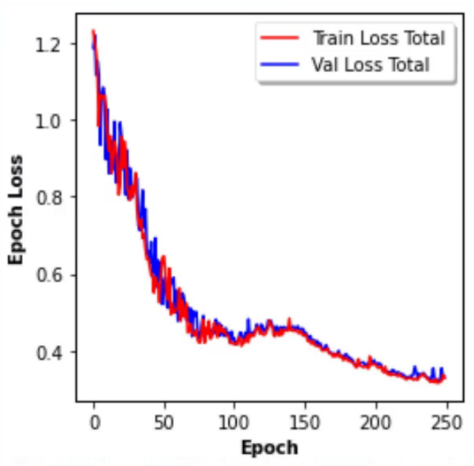

Fig. 5 Typical epoch loss for each run

Validating the three different distributions as inputs for different values of β is found in the parity plots in Figure 6. In general, these model predictions are quite horrible for most conditions with key exceptions of β equal to 10 and 15. This chart represents the key motivation of this effort since this hyper-parameter is a key impact to the reconstruction independently.

The response of β values of zero and one drives the reconstruction values essentially the same in under predicting. However, when β is equal to zero no matter the input distribution configuration both the reconstruction and actual values generated resulted in the same, significantly under-predictive. Changing β to unity the model is still under-predictive but now the input distribution starts to show some impact.

At β equal to four, as mentioned by Sikka, the reconstruction starts to show some feasibility in prediction, however, with only the uniform distributed data. The other two systems are still quite under-predictive but have some variation in actual values. Skipping to β of ten and fifteen the trend of increasing model reconstruction accuracy holds true for everything but the largest most Gaussian data-set. After values of fifteen and the near random response at seven β seems to start having reconstructions that are failing to be representative of the actual values. Additionally, at a β value of 7 the response seems to highly under predictive to the actual and has no understanding on the true variation. This intermediate value could be a key point where the reconstruction is missing significant information and yet does not benefit from the disentanglement. There are a few key takeaways from the parity plots:

1) Only changing this hyper-parameter (β) to change the result drastic enough to go from no correlation to one overlaying with the ideal response.
2) Regardless of the value of β for "all data" provided the predictive nature of the model was terrible. The best fits were β of 10 and 15.

3) Intermediate and high values of β show poor reconstruction, this is likely from the competition of poor reconstruction quality and improved disentanglement resulting in a net negative response. Some intermediate values do show good correlation and this could be when the competition results in a net positive response.

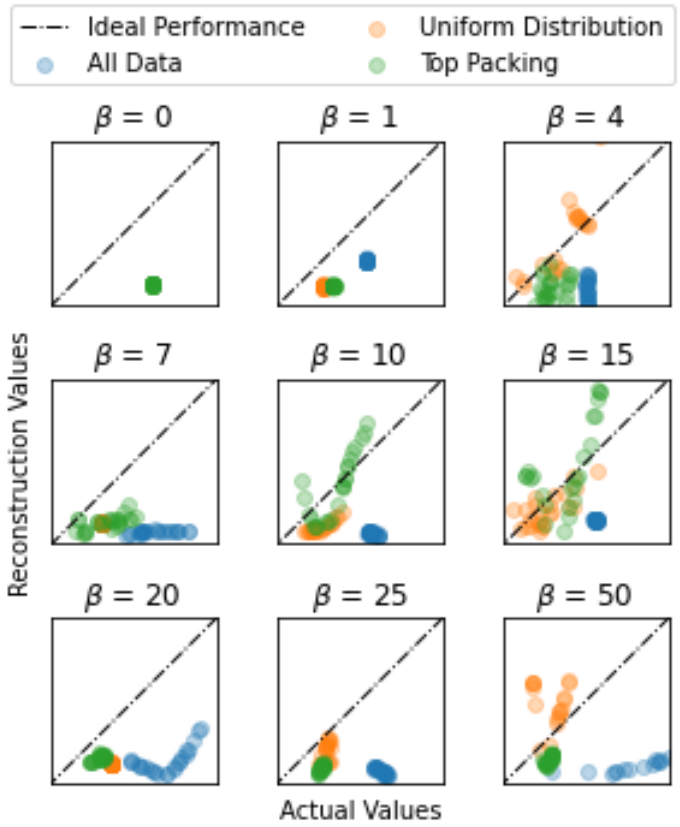

Fig. 6 Parity plots for all distributions and varying β, (Blue) All input data, (Orange) Uniform input data (Green) Top input data

One of the main points about this effort is to document the model's ability to reproduce a similar distribution with the hyper parameter β. Figure 7 shows a general trend of the MIG with increasing β with MIG in units of bits, possible range of 0-1. In general the β appears to have no direct impact on the MIG with such a poor correlation. However, second order polynomials were fit to each distribution separately and in which some conclusion can be considered from this that agrees with Welling and Kingma determining the improved disentangling prior to the reconstruction loss dominating. The "uniform data"

is evenly spread and shows more of the general nature of what is expected of the impact of varying β. For $\beta > 1$ MIG improves with increasing β until a reconstruction loss exists, although the accuracy of this trend is quite poor.

Both datasets "all data" and "top data" are of Gaussian nature which has the inverse trend for the first few incremental values of β until the value is sufficiently high that the MIG increases to something of value. This could suggest that Gaussian distributed data has the best reconstruction with higher values of β than compared to a uniform distribution. Data-set "all data" differs by approximately 30 times the number of data-points and still have a similar trend with β which is another likely outcome; the amount of data is not the largest impact and where the data lies is a first order effect.

Furthermore, it is expected that a poor MIG would at a $\beta = 0$ since this criteria forces the KL divergence to zero which is what attempts to force the learning to create the same reconstruction. This point is quite shown quite well with all β values between $0.4 - 0.5$. The highest MIG was obtained with $\beta = 1$ or the vanilla VAE, however it is evident that the best response in a pairty plot occurs with a MIG ranging from 0.5-0.9 (β of 10 and 15). It is this response that shows the best model isn't necessarily maximum disentanglement due to poor reconstruction loss.

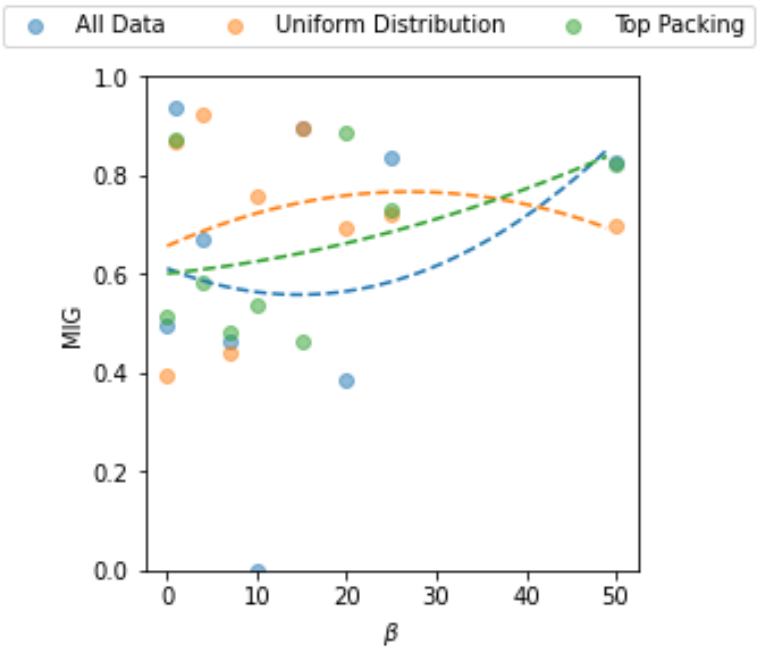

Fig. 7 MIG vs β. Best fit lines are second order polynomials

The scatter found in Figure 7 was concerning with such a poor fit. This fit is also comparable to the pairity plots in Figure 6 where most predictions were quite horrible. An examination of the reconstruction data found an interesting, yet expected trend. Figure 8 highlights the MIG value of all β against the normalized standard deviation of the reconstructed values. This chart shows that the higher the spread in data (standard deviation) the higher the MIG. Also, nearly all values are ≥ 0.4 showing there is some similarity in the actual vs reconstruction with the exception of a single data-point that had a reconstruction of the exact same value. This implies a standard deviation of zero which also gave a MIG value -10^{16}. Additionally, Figure 8 graphically shows the uniform distribution results in a higher MIG score for near all values of reconstruction. This figure highlights part of the goal of this study, having well placed data can lead to better prediction than having high amounts of data.

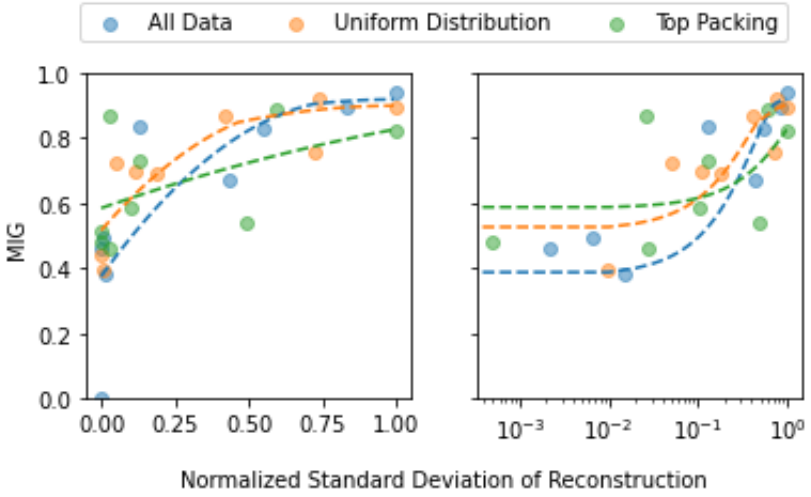

Fig. 8 MIG vs normalized reconstruction standard deviation (Left) Linear scale (Right) Log scale. All MIG units are in bits. The MIG of zero data point had a reconstruction data set all of the same value and had a numerically zero standard deviation. Best fit lines are second order polynomials.

Figure 9 highlights how the variation in reconstruction of the data set changes with both β and MIG. No real apparent trend is evident in with increasing β on the variation of the reconstruction's output. One desire when starting this effort was a highly disentangled representation would result in a highly variable reconstruction. This desire seems to be more fit for a CVAE where one could force that condition over a β-VAE. The MIG comparison shows that a higher variation in the reconstruction results in a higher MIG value. What is known from this is figure is that the hyper-parameter is not the

manner of enforcing the variation and that the higher the variation the more disentangled the result.

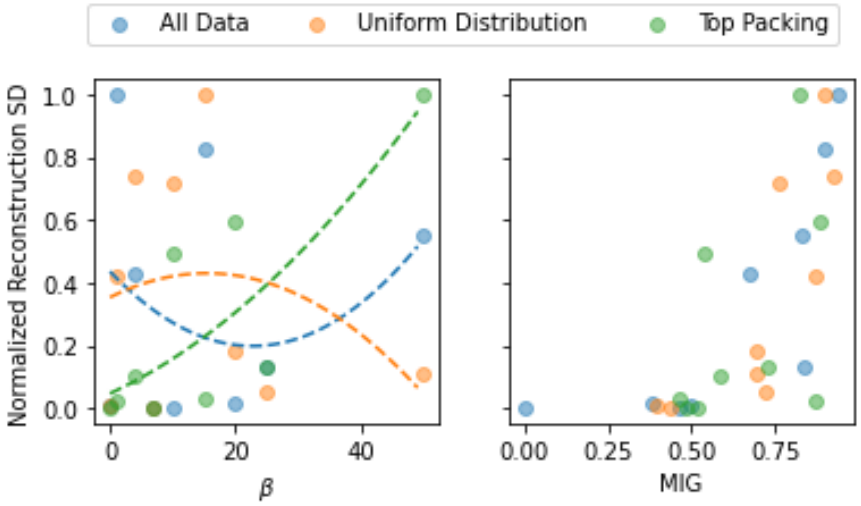

Fig. 9 (Left)Normalized reconstruction standard deviation v β (Right)Normalized reconstruction standard deviation v MIG.

Typical VAE models, like most machine learning, have increasing prediction accuracy with increasing data. Stein et. al. had typical results that followed a nearly doubling of R2 with an order of magnitude increase in data size[16]. However, the results within this effort show that the distribution has a larger effect in driving accuracy. This result is consistent with Alam and Shehu who has showed that the data set used for training impacted the magnitude of the p-value changing the statistical significance[9].

Figure 10 shows the variation of the input data-set and hyper-parameter resulted in creating the potential to have a statistical significant comparison of the reconstruction data. This figure highlights the comparison of the reconstruction prediction against the actual computation. Each point in Figure 10 comprises 25 different inputs to the ML model and actual computations. The most discouraging trait that is evident is that the t-test values don't trend in a way consistent with increasing β. However, one thing is aware when plotting all the data against MIG, this is an intersecting of the critical student-t value and the best fit line of the MIG. The intersection occurs at a MIG of approximately 0.9. Using this correlation, it can be applied to datasets generated with ML that cannot reject the null hypothesis and will also likely fulfill the properties for being disentangled.

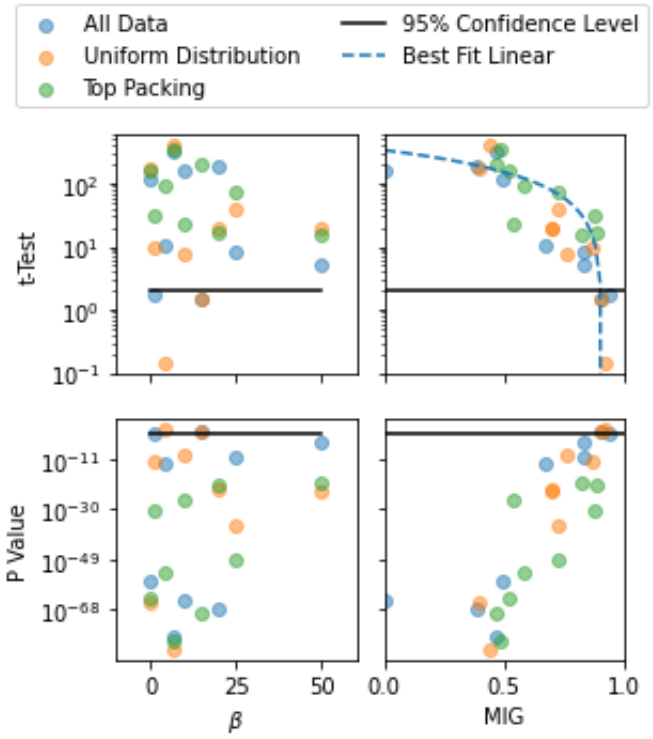

Fig. 10 (Top Left) t-Test v β (Top Right) t-Test v MIG (Bottom Left) P-value v β (Bottom Right) P-value v MIG.

4 Conclusions

This work has shown the common working knowledge of increasing β can directly increase the MIG or disentanglement of the reconstruction. However, the relationship is variable alone in this model and likely to be true in other models until a Monte Carlo like assessment can be done to understand the impact of random states in the latent space since this is likely the cause of the poor accuracy of the polynomial fits.

Key takeaways:

1. The best disentangled reconstruction (MIG) isn't the best prediction (parity).
2. The trend in increasing β isn't linear due to the competition of the reconstruction fidelity and increased disentanglement.
3. The data distribution is more important solely than the number of datapoints.
4. Disentangled reconstruction (MIG) trends well with reconstruction standard deviation and the student t-test.

5. It's likely to assume that a data-set that passes the student t-test will also be highly disentangled.

The problem used to examine these traits is limited due to the small amount of features in the learning process impacts the layering of the networks and is why the author performed this on a vanilla VAE being the simplest for the data at hand. If this process was applied to a larger feature set additional variables could be investigated in such a way to see if the latent space could make up for the concerns with β-VAE providing any meaningful progress to certain datasets (i.e., large and Gaussian).

5 Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. J. Hall. Beta-variational auto-encoder disentanglement with different input distributions for computational multi-modal particle packing. *https://zenodo.org/record/8003522*, 2023.
2. M. Welling D. P. Kingma. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.
3. S. R. Rath. Getting Started with Variational Autoencoders using PyTorch. `https://debuggercafe.com/getting-started-with-variational-autoencoders-using-pytorch/`, 2020. [Online; accessed 29-Jan-2023].
4. I. Higgins et al. "β-vae: Learning basic visual concepts with a constrained variational framework. *ICLR Conference Paper*, 2017.
5. H. S. Stein et. Al. Machine learning of optical properties of materials – predicting spectra from images and images from spectra. *Chemical Science*, 10(47), 2019.
6. A. M. Groener R. J. Richards. Conditional β-vae for de novo molecular generation. *arXiv:2205.01592*, 2022.
7. K. Gundersen et al. Semi-conditional variational auto-encoder for flow reconstruction and uncertainty quantification from limited observations. *Phys. Fluids*, 33(017119), 2021.
8. J. Yu et al. Structure-aware conditional variational auto-encoder for constrained molecule optimization. *Pattern Recognition*, 125(108581), 2022.
9. Fardina Fathmiul Alam and Amarda Shehu. Data size and quality matter: Generating physically-realistic distance maps of protein tertiary structures. *Biomolecules*, 12(7), 2022.
10. J. R. Hall S. K. Kauwe, T. D. Sparks. Sequential machine learning applications of particle packing with large size variations. *Integr Mater Manuf Innov*, 10:559–567, 2021.
11. E. Mathieu et al. Disentangling disentanglement in variational autoencoders. *arXiv: 1812.02833v3*, 2019.
12. M. de Rijke A. Sepliarskaia, J. Kiseleva. How to not measure disentanglement. *arXiv: 1910.05587*, 2021.
13. R. Gray. Entropy and information theory, 1st edition. *Springer Verlag*, 1990.
14. R. T. Q. Chen et al. Isolating sources of disentanglement in vaes. *arXiv:1802.04942v5*, 2019.
15. H. Sikka et al. A closer look at disentangling in β-vae. *arXiv:1912.05127v1*, 2019.
16. H. S. Stein et al. Machine learning of optical properties of materials - predicting spectra from images and images from spectra. *Chemical Science*, 2019.