arXiv:2507.00298v1 [stat.ML] 30 Jun 2025

Enhancing Interpretability in Generative Modeling: Statistically Disentangled Latent Spaces Guided by Generative Factors in Scientific Datasets

Arkaprabha Ganguli[1*], Nesar Ramachandra[2], Julie Bessac[3,4], Emil Constantinescu[1]

[1*]Mathematics & Computer Science Division, Argonne National Laboratory, Lemont, IL, 60439, USA.
[2]Computational Science Division, Argonne National Laboratory, Lemont, IL, 60439, USA.
[3]Computational Science Center, National Renewable Energy Laboratory, Golden, CO, 80301, USA.
[4]Department of Mathematics, Virginia Tech, Blacksburg, VA, 24061, USA.

*Corresponding author(s). E-mail(s): aganguli@anl.gov;
Contributing authors: nramachandra@anl.gov; julie.bessac@nrel.gov; emconsta@anl.gov;

Abstract

This study addresses the challenge of statistically extracting generative factors from complex, high-dimensional datasets in unsupervised or semi-supervised settings. We investigate encoder-decoder-based generative models for nonlinear dimensionality reduction, focusing on disentangling low-dimensional latent variables corresponding to independent physical factors. Introducing Aux-VAE, a novel architecture within the classical Variational Autoencoder framework, we achieve disentanglement with minimal modifications to the standard VAE loss function by leveraging prior statistical knowledge through auxiliary variables. These variables guide the shaping of the latent space by aligning latent factors with learned auxiliary variables. We validate the efficacy of Aux-VAE through comparative assessments on multiple datasets, including astronomical simulations.

Keywords: Disentangled generative factors, Posterior regularization, Representation learning, Variational AutoEncoder

1 Introduction

Semantic data representations are critical in artificial intelligence, significantly enhancing model performance in tasks like transfer and zero-shot learning (Lake et al., 2017). Central to this effort is to disentangle latent representations in generative models—representations where each latent dimension corresponds to an independent underlying factor of variation in the data. Disentanglement is achieved by leveraging statistical properties of the latent space and the dataset, enabling models where changes in one latent dimension affect only its corresponding factor without impacting others. This not only improves model interpretability but also enhances robustness against adversarial attacks (Yang et al., 2021). For a comprehensive review of disentanglement and its statistical underpinnings, see Wang et al. (2023).

Datasets encountered in scientific research are often heterogeneous in modalities, fidelities, and accuracy where a particular entity or a state may be simultaneously associated with multiple images, graphs, vectors, scalar parameters, or labels with various associated measurement uncertainties. Besides, many natural and non-natural phenomena exhibit stochasticity, increasing the problem complexity. In many scientific problems, domain experts aim to understand and characterize underlying patterns and associations of physical quantities in order to improve their predictability for instance, or elucidate on the underlying physical phenomena. However, due to the problem complexity and data diversity (modality, fidelity, accuracy) these patterns are typically hard to extract from traditional data exploration tools. Moreover, domain experts are often cognizant of "known knowns" and "known unknowns", whereas several research problems also have associated "unknown unknowns" Hatfield (2022). Classical data exploration rarely incorporates this type of partial or "unknown unknowns" information, hence the need for novel tools as proposed here to advance science. Finally, sensitivity analyses or computation of model response surfaces of input physical parameters are crucial for uncertainty quantification and forecasting Razavi and Gupta (2016); Raghavan Sathyan et al. (2018).

In Earth system science, for instance, information about a physical quantity can arise from numerical simulations, satellite imaging, or in situ sensors with various fidelities and uncertainties. For example, understanding the variation in weather patterns based on changing sea surface temperatures is essential for understanding the impacts of long-term environmental dynamics Deser et al. (2014); Maulik et al. (2020). However, capturing the complex multi-scale variability of atmospheric phenomena remains an open challenge Bauer et al. (2015), where data diversity along with partial expert knowledge is a typical setting, which remained un-leveraged by classical data science tools. Similar multi-modal, multi-fidelity datasets are also often encountered in astronomy: images and fluxes of galaxies are observed via telescopes, and a

subset of the associated physical parameters, such as the stellar mass or galactic distances, are calculated Bonvin and Durrer (2011). Relationships between these factors provide valuable information about the evolution of galaxies over cosmic time Newman and Gruen (2022), but they do not exhaustively explain all the physical processes and associations Somerville and Davé (2015). In such studies, researchers might find it intriguing to use a generative model, where the latent factors are clearly disentangled with the 'known knowns' generative factors. Latent factors representing 'known unknowns' or 'unknown unknowns', however, might remain entangled, collectively contributing to the overall generation. Understanding these associations could aid domain experts in gaining deeper insights into the underlying mechanisms driving their data generation processes. In this paper, we demonstrate the applicability of our method on a representative galaxy catalog that encompasses both the data-level complexities and the desired science goals mentioned above.

Recent disentanglement research primarily explores unsupervised learning methods, introducing inductive biases into the Variational Autoencoder (VAE) framework to structure the latent space without using known factors of variation. Works such as (Higgins et al., 2017; Chen et al., 2018; Kumar et al., 2018) have advanced these techniques, which are detailed in Section 2. However, these methods often overlook auxiliary information that may be crucial in scientific datasets. Emerging semi/weakly supervised methods (Chen and Batmanghelich, 2020; Mita et al., 2021) attempt to address this by leveraging observable ground-truth generative factors, but these approaches face challenges when auxiliary information is limited, requiring all ground-truth factors, which can be restrictive. To address this gap, we propose the ***Aux****iliary information guided* ***V****ariational* ***A****uto****E****ncoder (****Aux-VAE****)*, focusing on scenarios with available auxiliary information to disentangle representations with respect to known ground-truth generating factors while preserving data reconstruction capability.

Our contributions: (i) **Statistically Interpretable Latent Space with Preserved Reconstruction Ability**: We partition the latent space into two segments to disentangle known factors of interest using auxiliary information. The first d dimensions align with d ground truth generative factors, enhancing interpretability, while the remaining dimensions capture other unknown factors in an entangled state to preserve overall accuracy. To achieve this, we construct a targeted prior and enforce disentanglement through posterior regularization. This method balances the trade-off between accurate data reconstruction and improved disentanglement. (ii) **Enabling Control Over Ground Truth Factors for Understanding Dataset Characteristics**: Our disentanglement approach establishes a direct correspondence between the disentangled latent factors and the known generative factors of interest via the specification of the latent factor's distributions. This facilitates a precise understanding of specific physical characteristics of the data by reconstructing it with the corresponding latent factors adjusted to reflect the controlled level of the generative factors. This also enables a computationally efficient sensitivity study where one can compute model responses across the space formed by ground truth factors. (iii) **Introducing a Novel Disentanglement Metric**: We introduce a new metric that provides an intuitive, cost-effective way to measure disentanglement in the latent space relying on correlations between latent factors. This metric avoids the need for retraining separate

models and quantitatively assesses the qualitative aspects of disentanglement, scoring up to 1 for optimal separation of factors. The article outlines our methods and results as follows. Section 2 formalizes key definitions and reviews relevant literature; Section 3 introduces the Aux-VAE methodology; Section 4 presents experiments on both scientific and benchmark datasets; and Section 5 offers conclusions and future directions.

2 Variational Autoencoders (VAE) and related literature

We begin with a generative model for observed data, where a latent variable z is sampled from $p(z)$, and an observation x is generated by sampling from $p_\theta(x|z)$. The joint density of the latent variables and observations is denoted as $p_\theta(x,z) = p(z)p_\theta(x|z)$. The inference problem involves computing the posterior of the latent variables conditioned on the observations, i.e., $p_\theta(z|x) = \frac{p_\theta(x,z)}{\int p_\theta(x,z)dz}$. Given a finite set of samples (observations) from the true data distribution $p(x)$, the exact computation of the posterior is generally intractable and requires approximate inference. Variational inference addresses this by positing a family of approximate densities $\mathcal{Q}$ over the latent factors and minimizing the Kullback-Leibler (KL) divergence to the true posterior, i.e., $q_x^* = \min_{q\in\mathcal{Q}} \mathrm{KL}(q(z)||p_\theta(z|x))$ (Blei et al., 2017). A variational autoencoder (VAE) utilizes the amortized inference, a recognition model, parameterized by ϕ to encode an inverse map from observations to approximate posteriors. The recognition model parameters are learned by optimizing the problem $\min_\phi E_x[\mathrm{KL}(q_\phi(z|x)||p_\theta(z|x))]$, where the outer expectation is over the true data distribution $p(x)$ from which we have samples. This optimization is equivalent to maximizing the Evidence Lower Bound (ELBO):

$$\arg\min_\phi \mathbb{E}_x[\mathrm{KL}(q_\phi(z|x)||p_\theta(z|x))] = \arg\max_\phi \mathbb{E}_x\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x)||p(z)) \tag{1}$$

Often, the density forms of $p(z)$ and $q_\phi(z|x)$ are chosen such that their KL-divergence can be written in a closed-form expression (e.g., $p(z)$ is $\mathcal{N}(0,I)$ and $q_\phi(z|x)$ is $\mathcal{N}(\mu_\phi(x),\Sigma_\phi(x))$) (Kingma and Welling, 2022). This framework encourages the encoder to learn meaningful representations in the latent space while enabling the decoder to generate data samples that closely match the input data.

Unsupervised approaches: In the context of representation learning, all the unsupervised approaches follow the basic structure of the VAE and enforce the desired disentangling characteristic in the latent space with an additional term in the loss function. For example, $\beta-$VAE Higgins et al. (2017) imposes a weight $\beta > 0$ to the KL-term in the VAE-objective to ensure that each latent factor captures independent sources of variation in a disentangled manner. Building upon this, Kim and Mnih (2018) introduce an additional term in the VAE loss function, $\mathrm{KL}(q(\mathbf{z})||\prod_j q(z_j))$, encouraging the aggregate posterior $q(\mathbf{z})$ to align with the product of marginals $q(z_j)$, thereby achieving independent latent factors, originally referred to as Total Correlation (TC) in the literature. Numerous representation learning algorithms Kumar

et al. (2018); Kim and Mnih (2018) have been proposed in this unsupervised fashion, proving beneficial in scenarios where auxiliary information on the ground truth factors is unavailable. However, a plausible drawback of unsupervised methods is the identifiability issue, where different models may yield entirely different latent variables despite having the same marginal data and prior distributions due to potential transformations of the latent variable while preserving the marginal distribution. To address this, posterior regularization via the choice of variational family and the prior distribution has been investigated in the literature Mathieu et al. (2019); Kumar and Poole (2020). Nonetheless, a common limitation of unsupervised approaches is their tendency to exhibit high variance, making it challenging to identify well-disentangled models without supervision (Locatello et al., 2019). This is consistent with the theoretical findings of Locatello et al. (2019), suggesting that unsupervised learning of disentangled representations is unfeasible without appropriate inductive biases.

Disentanglement with auxiliary information: In response to certain limitations, a category of methods has emerged that leverages auxiliary information about ground truth factors within the traditional VAE framework. For example, existing semi-supervised approaches tackle the disentanglement of observed factors by utilizing limited supervised data on class levels (Reed et al. (2015); Cheung et al. (2014); Mathieu et al. (2016); Paige et al. (2017); Kingma et al. (2014)). State-of-the-art weakly-supervised disentanglement methods operate under the assumption that observations are grouped based on known relationships between images within the same group and their corresponding groups ((Bouchacourt et al., 2018; Hosoya, 2018; Chen and Batmanghelich, 2020; Locatello et al., 2020)). A concise overview of these methodologies can be found in Shu et al. (2019). While these approaches have demonstrated success in computer vision and other scientific domains, they face challenges in scenarios where generating factors are continuous and multiple sources of true generative factors remain unknown. This property is often seen in a wide range of scientific datasets, from astrophysics (both in observational datasets such as the COSMOS Scoville et al. (2007) or simulated datasets created for telescope surveys Korytov et al. (2019)) to earth system studies (e.g., Kaltenborn et al. (2023)) and medical sciences (e.g., Efron et al. (2004)). In the context of such datasets, clustering the images into distinct groups (or any other operation on the latent space of the VAE – such as classification, regression, or anomaly detection) is difficult if auxiliary information is not used along with a robust disentanglement scheme.

Taking a step towards a more general setting, Khemakhem et al. (2020) introduced the Identifiable-VAE (IVAE) framework. This framework learns a disentangled representation by employing a factorized prior from the exponential family, conditioned on auxiliary variables representing certain generative factors. Building upon this foundation, Mita et al. (2021) proposed an iterative training strategy 'IDVAE' utilizing two VAEs: one to capture latent representations from auxiliary information and the other to leverage these latent distributions for learning the data distribution. Despite the appealing theoretical guarantee of identifiability, Kim et al. (2023) observed that IVAEs may overlook observations in certain cases, potentially leading to posterior collapse in experiments. As discussed in Kumar et al. (2018), one potential remedy for this issue involves imposing regularization on the Expected variational

posterior. We extend this approach by incorporating limited auxiliary information on the ground-truth factors, as elaborated in Section 3.

3 Limited available auxiliary information and proposed approach

Observed database: Moving forward, let us presume access solely to a subset of the ground truth factors $S_{obs} \in \mathbb{R}^d$, conveyed through auxiliary variables $u \in \mathbb{R}^d$. Each auxiliary variable is intended to encapsulate a specific ground truth factor within S_{obs}. Our aim lies in disentangling the latent space concerning these identified ground truth factors S_{obs}, observable via u. We initiate this endeavor with an observed database $\mathcal{D}$ comprising n independent and identically distributed pairs of x and u, denoted as $\mathcal{D} = \{(x^{(1)}, u^{(1)}), (x^{(2)}, u^{(2)}), \ldots, (x^{(n)}, u^{(n)})\}$. We begin by reformulating the ELBO of a VAE:

$$L_{VAE} = \mathbb{E}_{z\sim q(z|x)} \left[\log p(x|z)\right] - \mathrm{KL}\left(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right), \tag{2}$$

where, $q(z|x)$ denotes the encoder, $p(x|z)$ represents the decoder, and $p(z)$ is the prior distribution. To incorporate information from auxiliary variables, we partition the latent space into two distinct components: auxiliary-informed latent factors z_{aux} and residual latent factors z_{recon}, represented as: $z^{1\times d_Z} = \left(z_{aux}^{1\times d}, z_{recon}^{1\times(d_Z-d)}\right)$. Here, the auxiliary-informed latent factors Z_{aux} signify the latent features associated with auxiliary variables in a disentangled fashion, while the residual latent factors Z_{recon} characterize the latent features necessary to capture the underlying factors not explicitly covered by the auxiliary variables u. Hence, conditional on the auxiliary variables u, we define the prior in the following way:

$$p_{z|u}(z) = \left(\prod_{j=1}^{d} p_{\mathcal{N}(u_j, \frac{1}{n})}(z_j)\right) p_{\mathcal{N}(0, I_{d_Z-d})}(z_{(d+1):d_Z}) = p_{\mathcal{N}(\mu_0, \Sigma_0)}(z) \tag{3}$$

where $\mu_0 = (u_1, u_2, \ldots, u_d, 0, \ldots, 0)$ and $\Sigma_0 = diag(\frac{1}{n} I_d, I_{d_Z-d})$ denotes the mean and variance of the Gaussian distribution in the prior. Now, while optimizing the ELBO in Eq. (2) (derivation detailed in the SM), the first term can be estimated by simple Monte-Carlo approximation: $\mathbb{E}_{z\sim q(z|x)} \left[\log p(x|z)\right] \hat{=} \frac{1}{J} \sum_{j=1}^{J} log(p(x|z^j))$.

Now, the KL part can be decomposed as follows utilizing the closed-form structure of the Gaussian distributions:

$$\begin{aligned} \mathrm{KL}\left(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{u})\right) &= \mathrm{KL}\left(\mathcal{N}(\mu_\phi, \Sigma_\phi)||\mathcal{N}(\mu_0, \Sigma_0)\right) \\ &= \frac{1}{2}\left[log\frac{|\Sigma_0|}{|\Sigma_\phi|} - d_Z + (\mu_\phi - \mu_o)'\Sigma_0^{-1}(\mu_\phi - \mu_o) + tr(\Sigma_0^{-1}\Sigma_\phi)\right]. \end{aligned} \tag{4}$$

Advancing Disentanglement via Expected Variational Posterior:

Achieving disentangled latent spaces requires more than aligning the encoder distribution $q_\phi(z|x)$ with the desired prior. Disentanglement should also be fostered in the *expected posterior* $p_\theta(z) = \int p_\theta(z|x)p(x)dx$. Its variational counterpart is expressed as the inferred prior or expected variational posterior: $q_\phi(z) = \int q_\phi(z|x)p(x)dx$. Utilizing the pairwise convexity property of KL-divergence, we can show that the distance between $q_\phi(z)$ and $p_\theta(z)$ is bounded by the objective of the variational inference Kumar et al. (2018):

$$\begin{aligned} \mathrm{KL}(q_\phi(z)||p_\theta(z)) &= \mathrm{KL}\left(E_{x\sim p(x)}q_\phi(z|x)||E_{x\sim p(x)}p_\theta(z|x)\right) \\ &\leq E_{x\sim p(x)}\mathrm{KL}(q_\phi(z|x)||p_\theta(z|x)). \end{aligned} \tag{5}$$

Hence, although, maximizing ELBO (2) would ideally decrease $\mathrm{KL}(q_\phi(z)||p_\theta(z))$, in many complex scenarios, the two sides of equation (5) might deviate at the stationary point of convergence (Kumar et al., 2018). Explicitly minimizing $\mathrm{KL}(q_\phi(z)||p(z))$ provides better control over the disentanglement. However, due to the intractable KL term, we implicitly enforce the following three main characteristics of the disentangled prior concerning auxiliary information u, such as: (1) ***Inter-independence*****:** $u \perp z_{recon}$, (2) ***Intra-independence*****:** $u_j \perp z_{aux,j'}$ for $j, j' = 1, 2, \ldots d, j \neq j'$, and (3) ***Explicitness*****:** $E_{z\sim p(z|u)}(z_{aux,j}) = u_j$ for $j = 1, 2, \ldots d$.

Quantifying the interdependency among z_{aux}, z_{recon}, and the auxiliary information u presents a challenge. To address this, we turn to polynomial regression, a technique that assesses nonlinear relationships among variables Rawlings et al. (1998). This approach measures the strength of dependency by aggregating correlations across various polynomial degrees and utilizes Monte Carlo samples of encoder outputs to compute these correlations. Specifically, to calculate these correlations, we utilize the expected latent factors μ_ϕ from the encoder $q_\phi(z|x)$. This choice is justified by the theorem of total variance. For instance, in the case of *Inter-independence*, the covariance between the polynomials u^k and $z^{k'}_{recon}$ can be expressed as:

$$\begin{aligned} Cov(u^k, z^{k'}_{recon}) &= E_{(x,u)\sim p(x,u)}Cov_{u,z\sim q_\phi(z|x)}(u^k, z^{k'}_{recon}) \\ &\quad + Cov_{(x,u)}(u^k, E_{z\sim q_\phi(z|x)}(z^{k'}_{recon}) \\ &= Cov_{(x,u)}(u^k, \mu^{k'}_{\phi,d+1:d_Z}). \end{aligned} \tag{6}$$

In this expression, under the outer expectation, the covariance in the first term becomes zero conditioned on u. Analogous properties apply to *Intra-independence* and *Explicitness* as well (detailed in Section 1 of the SM). For notational simplicity, let us denote, $\mu_{\phi,1:d} = \mu_{\phi,aux}$ and $\mu_{\phi,d+1:d_Z} = \mu_{\phi,recon}$.

With this approach, for two random vectors $v^{m_v\times 1}$ and $w^{m_w\times 1}, m_v \leq m_w$, we define the correlation matrix $Corr(v, w)$ as $diag(\Sigma_{v,v})^{-1/2}\Sigma_{v,w}diag(\Sigma_{w,w})^{-1/2}$, with $\Sigma^{m_v\times m_w}_{v,w} = \mathbb{E}[(v - \mathbb{E}(v))(w - \mathbb{E}(w))']$. Subsequently, we formulate the following two dependency metrics:

$$R_0^K(v,w) = \frac{1}{Km_v m_w} \sum_{k,k'=1,k\neq k'}^{K} \sum_{i=1}^{m_v}\sum_{j=1}^{m_w} |\left(Corr(v^k, w^{k'})\right)_{ij}|, \tag{7}$$

$$R_1^K(v,w) = \frac{1}{Km_v m_w} \sum_{k,k'=1,k\neq k'}^{K} \sum_{i=1}^{m_v} \left(1 - |\left(Corr(v^k, w^{k'})\right)_{ii}|\right). \tag{8}$$

In these metrics, the first summation aggregates associations from all possible polynomial combinations up to degree K, while the second sum separately considers various terms of the covariance matrix in $R_0(\cdot,\cdot)$ and $R_1(\cdot,\cdot)$. Consequently, $R_0(\cdot,\cdot)$ and $R_1(\cdot,\cdot)$ quantify the strength of pairwise nonlinear dependency by evaluating the association among the polynomials of the variables, enhancing our understanding of disentanglement.

These metrics are then incorporated into the loss function to enforce optimal disentanglement within and between Z_{aux} and Z_{recon}. Specifically, three regularization terms are introduced into the objective function:

$$\mathcal{L}_{Aux-VAE} = \mathcal{L}_{VAE} + \underbrace{\lambda_1 \sum_{j=1}^{d} \left(R_1^K(u_j, \mu_{\phi,aux,j}) + R_0^K(u_j, \mu_{\phi,aux,-j})\right)}_{\text{Intra-independence and explicitness regularizer}} + \underbrace{\lambda_2 \left(R_0^K(u, \mu_{\phi,rec})\right)}_{\text{Inter-independence regularizer}} \tag{9}$$

Here, the three regularizers play a crucial role and align with the intuitive logic of achieving disentanglement. The intra-group regularization includes two terms: the first ensures that each dimension of z_{aux} closely aligns with the auxiliary information u, while the second imposes a penalty on the dependency between any two latent factors in z_{aux} using the defined polynomial dependency metric. Similarly, the inter-group regularization aims to reduce the dependency between z_{aux} and z_{recon}. No restrictions are imposed on the dependency within z_{recon} to ensure good reconstruction quality.

4 Experiments

4.1 Experimental settings

We compare the proposed approach against two major alternative disentanglement methods: β-VAE (Higgins et al., 2017), and IDVAE (Mita et al., 2021). These two methods were chosen to represent different classes of existing disentangling approaches. β-VAE (Higgins et al., 2017) serves as a baseline for its simple yet effective unsupervised approach with minimal assumptions. It represents the class of unsupervised methods utilizing no ground-truth factor for disentanglement, but a regularization term is introduced to enforce disentanglement. On the other hand, IDVAE (Mita

et al., 2021) represents recent algorithms on auxiliary variable-informed methods, which is the most closely related approach in the literature analogous to Aux-VAE. We implemented these methods using the same hyperparameter settings as their publicly available repositories. All methods were implemented in PyTorch (Paszke et al., 2019), with code available at Ganguli et al. (2024).

4.1.1 Datasets

We have created a representative scientific dataset where measurements from instruments such as telescopes are associated with physical quantities. We simulate the galaxy images observed from telescopes using GalSim Rowe et al. (2015), a widely used in current and future space- and ground-based telescope missions Collaboration et al. (2021); Everett et al. (2022); Merlin et al. (2023). Each image is associated with 5 physical parameters, Apparent brightness of the galaxy ($flux$, in number counts), radius of the galaxy ($radius$, in arc-seconds), 2 reduced gravitational shear components ($g1$ and $g2$ in Cartesian coordinates), and the full width of half maximum of the Gaussian function (also called point-spread function, psf) used in the convolution. Further details of the experimental design are provided in the SM Section 1.

To maintain experiment realism, we refrain from using all ground-truth information as auxiliary variables. Instead, we compare results under a more practical setting. Specifically, we categorize the auxiliary variables into important and less important categories. The quantities $radius$, $g1$, and $g2$ as well as $flux$ are considered important, as they focus on essential physical characteristics, while psf is deemed less significant. This hierarchy of importance is illustrated in Figure 1. While this heuristic characterization is based on domain expertise, one may investigate correlation structures to identify auxiliary variables. In light of this, we examine the following three cases: **Case 1:** Using all five ground truth generating factors as auxiliary information, **Case 2:** Using only the important factors ($radius$, $g1$, and $g2$) as auxiliary information, excluding $flux$ and psf, and **Case 3:** Using mostly the less important factors ($flux$, and psf) as auxiliary information, omitting $radius$, $g1$, and $g2$. Evaluating these cases illustrates the interplay between the Z_{aux} and Z_{recon} factors, offering insights for datasets where exhaustive ground-truth factors are unknown. All ground-truth factors were normalized within the range [0, 1], assuming an implicit ordering for discrete factors before normalization.

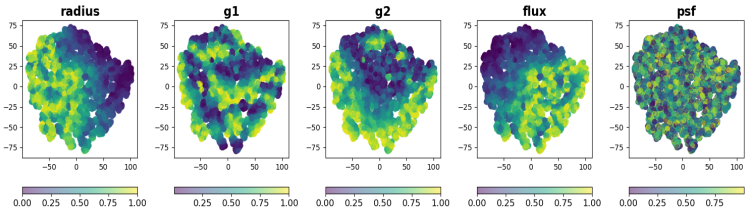


Fig. 1: t-SNE plot of the galaxy images colored by each of the generating factors, highlighting psf as the least important generator in this dataset.

Furthermore, to showcase the consistency and efficacy of our proposed method using auxiliary information, we conduct experiments on two other synthetic datasets: 'Cars3D' (Reed et al., 2015) and 'DSprites' (Burgess et al., 2018). These datasets provide explicit access to ground-truth factors, allowing us to assess our method against competing approaches.

4.1.2 Disentanglement metric

Assessing the degree of disentanglement among latent factors is crucial in understanding the effectiveness of a generative model, yet a nontrivial task in practice. While various disentanglement metrics exist in the literature Higgins et al. (2017); Kim and Mnih (2018); Kumar et al. (2018), many rely on fitting a supervised regression between the learned latent space and ground-truth factors. However, implementing additional regression models for evaluation can incur significant computational costs. Moreover, studies have shown that these model-based metrics may not always correlate well with qualitative disentanglement observed in latent traversal plots (Kumar et al., 2018). To address these challenges and efficiently evaluate disentanglement, we propose a novel metric called the Linear Disentanglement Score (LDS). For the auxiliary features u_j's and the latent factors z_l's, the LDS is defined as:

$$LDS = \frac{1}{d}\sum_{j=1}^{d}\frac{\max_{l} Corr(u_j, z_l)}{\sum_{l=1}^{d_Z}|Corr(u_j, z_l)|}. \tag{10}$$

The underlying idea stems from the concept that each generative factor u_j should ideally correlate with only one latent factor z_l. Therefore, for any u_j, a value closer to 1 for the term inside the summation indicates better disentanglement among the latent factors. Through a straightforward mathematical rationale, we establish that $LDS \in [\frac{1}{d_Z}, 1]$. While the idea of measuring the strength of linear dependence for this purpose is not new. The SAP score (Kumar et al., 2018) is computed as the average difference in prediction accuracy between the most and second-most predictive latent dimensions for each generative factor. A higher SAP indicates that a factor is captured predominantly by a single latent variable, reflecting strong disentanglement. For continuous factors, SAP ranges in $(0, 1]$. For categorical factors, balanced classification accuracy is used, allowing SAP to exceed 1. Our LDS metric offers an extension, providing a bounded version of the SAP-score for evaluating disentanglement. Moreover, it is crucial to recognize that a high SAP score might not exclude the possibility of one latent dimension effectively capturing multiple generative factors. Conversely, our proposed metric emphasizes evaluating each latent factor's capacity to represent individual generative factors distinctly or become entirely independent, thus offering a more comprehensive assessment of disentanglement.

4.2 Experimental results

In this section, we present the results of our numerical experiments, focusing on four main criteria: (1) Reconstruction accuracy, (2) The relative importance of Z_{aux} and Z_{recon} in the overall reconstruction, (3) Disentanglement among the latent factors,

(4) Latent traversal (generating from the learned decoder network by varying only one latent while keeping the others fixed) across Cases 1, 2, and 3 for the galaxy dataset mentioned in Section 4.1. To highlight the versatility of our method, we present results from Aux-VAE on both a galaxy simulation dataset and non-scientific datasets like Cars3D' and DSprites'. Primarily, we focus on the galaxy dataset due to space constraints, with further results available in the supplementary materials.

4.2.1 Reconstruction accuracy

It is common to encounter a trade-off between reducing reconstruction error and inducing disentanglement through posterior regularization. As a result, achieving disentanglement often comes at the cost of increased reconstruction error. Therefore, we use $\beta-$VAE as a baseline for assessing reconstruction accuracy. For evaluation, we employ the structural similarity index measure (SSIM) (Wang et al., 2004; Müller et al., 2020), a metric adept at quantifying image similarity by considering luminance, contrast, and structure. The SSIM between two image patches x and y is defined as

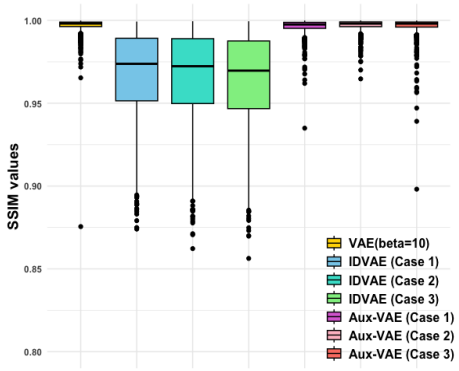


Fig. 2: Distribution of SSIM between the original and reconstructed test images.

$$\text{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)},$$

where μ_x, μ_y are the local means of x and y, σ_x^2, σ_y^2 are the local variances, σ_{xy} is the local covariance between x and y, L denotes the dynamic range of the pixel values (e.g. 255 for 8-bit images), $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$ are stabilizing constants (typically $K_1 = 0.01$ and $K_2 = 0.03$). In our experiments, SSIM is computed via the Python library `skimage.metrics.structural_similarity` (van der Walt et al., 2014). With SSIM scores ranging between -1 and 1, where 1 signifies perfect similarity, 0 denotes no similarity, and -1 implies perfect anti-correlation, it offers a nuanced assessment compared to pixel-wise methods. Figure 2 presents the SSIM scores for all methods across Cases 1, 2, and 3 using the galaxy image dataset.

Aux-VAE maintains strong reconstruction capabilities, on par with the baseline VAE, especially in Case 3, where key factors are missing from the auxiliary information, by effectively leveraging the residual latent factors Z_{recon}. Although IDVAE displays similar overall performance, it experiences a noticeable reduction in accuracy from Case 1 to Case 3, due to the exclusion of many crucial generative factors from

the auxiliary information and the absence of other latents in their implementation to compensate for missing information.

4.2.2 Disentanglement among the latent factors with respect to the ground-truth generating factors

Building on the fundamental concept of 'disentanglement' outlined in Section 1, where each latent factor is expected to correspond to a single underlying generative factor, we approach this comparison from two distinct angles. **Quantitatively**, we calculate the LDS metric (10) and SAP score Kumar et al. (2018) across all cases and datasets, as summarized in Table 1. Aux-VAE achieves notably high LDS scores, reflecting its strong ability to capture underlying relationships with robust disentanglement, as further supported by the reconstruction quality in Figure 2. While the SAP score shows a similar trend, it tends to be higher when multiple latent factors are related to the same generating factor, a known limitation (Kumar et al., 2018).

Table 1: Disentanglement Comparison Across Methods on Various Datasets and Cases Using the Proposed LDS Metric (SAP Scores in Parentheses).

Dataset		**β-VAE**	**IDVAE**	**Aux-VAE**
Galaxy image	Case 1	0.48 (0.39)	0.65 (0.72)	**0.88 (0.81)**
	Case 2	-	0.73 (0.74)	**0.94 (0.89)**
	Case 3	-	0.59 (0.68)	**0.81 (0.84)**
Card3D	-	0.21 (0.18)	0.67 (0.55)	**0.93 (0.85)**
DSprites	-	0.26 (0.27)	0.39 (0.43)	**0.83 (0.78)**

From a more **qualitative** standpoint, we examine each latent factor with respect to the underlying generative factors. Figure 3 showcases Case 2 for the galaxy image dataset, with similar plots for other datasets provided in the SM. We observe that Aux-VAE's auxiliary-informed latent factors Z_{aux} exhibit clear associations with the corresponding generating factors, remaining independent of the remaining latent factors. In Case 2, the residual latent factors Z_{recon} are less significant, since the auxiliary information encompasses most of the crucial ground-truth generating factors. Nonetheless, it is apparent that they are collectively attempting to represent the $flux$ in an entangled manner. Meanwhile, IDVAE also demonstrates some degree of one-to-one association with the associated factors. However, these relationships are often entangled, making it challenging to definitively attribute one particular latent factor to a specific generating factor. Despite IDVAE's factorized prior promoting independence in the latent space, the collective impact of entire auxiliary information on each dimension of the latent distribution may hinder complete disentanglement, especially for interdependent generative factors. A similar pattern is observed for basic $\beta-$VAE (relegated to the SM), where representations appear to be even more entangled.

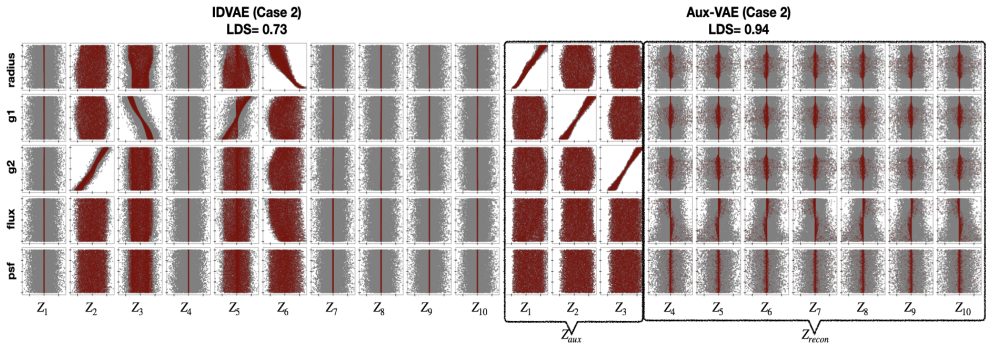


Fig. 3: Visualizing Disentanglement: Latent factors are displayed along the x-axis and generative factors along the y-axis. This scatterplot contrasts latent factors (Z, represented by grey dots) with the latent means (μ_ϕ, shown as maroon dots), alongside highlighting the LDS metric.

4.2.3 Relative importance between Z_{aux} and Z_{recon}

The proposed method, Aux-VAE, relies heavily on the interplay between two classes of latent factors, Z_{aux} and Z_{recon}. While the former aims to capture limited information from the auxiliary data in a disentangled manner, the latter endeavors to reconstruct the remaining unknown generative factors, not covered by the auxiliary data, in an entangled manner to achieve better reconstruction. Therefore, in scenarios where the majority of generative factors remain unknown and are absent from the auxiliary data, Z_{recon} becomes increasingly important in understanding the overall data generation process.

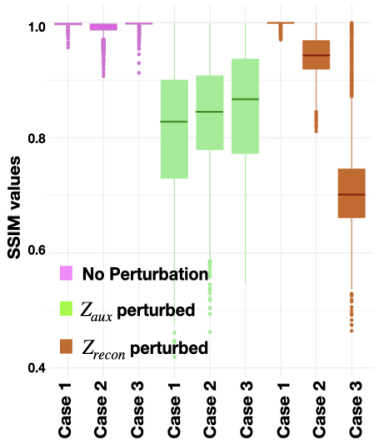


Fig. 4: Assessing Latent Factor Importance: Z_{aux} vs Z_{recon} Analysis with Gaussian Noise Perturbation.

To empirically demonstrate this, we consider three cases for the galaxy images dataset outlined in Section 4.1. To assess the relative importance of the latent factors, we conduct the following experiment: For each case, we utilize the encoder model on 1000 test images to obtain their latent representations $Z^{test} = (Z^{test}_{aux}, Z^{test}_{recon})$ and use the decoder to generate the outputs. Then, to understand the importance of Z_{aux}, we perturb only the factors in Z^{test}_{aux} with additive Gaussian noise, creating $Z^{perturbed}_{aux}$, and reconstruct the images using the decoder with $Z^{perturbed} = (Z^{perturbed}_{aux}, Z^{test}_{recon})$. Similarly, to assess the importance of Z_{recon}, we conduct a similar experiment but

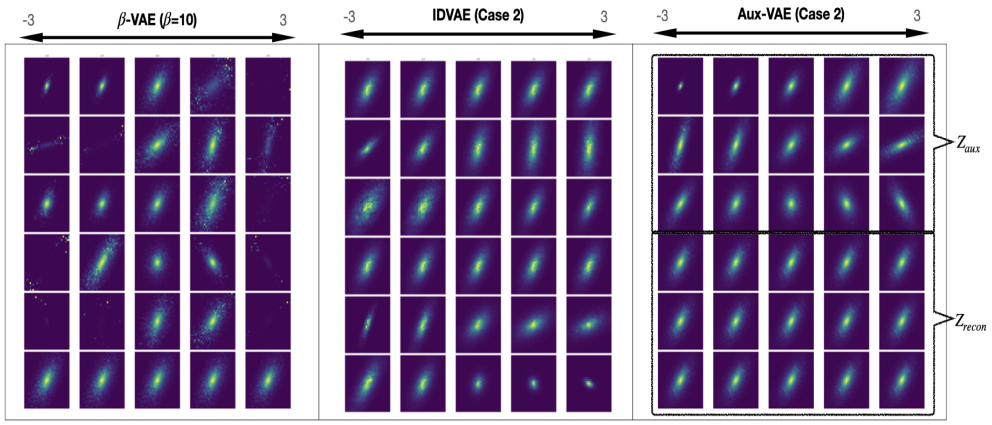


Fig. 5: Latent space traversal - Visualization of Latent Factor Adjustments in VAE, IDVAE, and Aux-VAE, showing how systematic changes to individual factors impact generated galaxy images.

perturb Z^{test}_{recon} instead. The SSIM boxplots for the 1000 test images are illustrated in Figure 4. We denote the first reconstruction with $Z^{test} = (Z^{test}_{aux}, Z^{test}_{recon})$ as 'No perturbation,' and the subsequent experiments as 'Z_{aux} perturbed' and 'Z_{recon} perturbed.' As expected, under the 'No perturbation' (magenta) setting, all three models perform similarly. However, with perturbations, a decrease in reconstruction quality due to the disruption of a significant latent factor will result in lower SSIM values, signaling the greater importance of that factor. In the 'Z_{aux} perturbed' (green) setting, we observe a gradual increase in SSIM from Case 1 to Case 3. In Case 1, where all true generating factors are included in the auxiliary information, Z_{aux} is most sensitive to perturbations compared to Cases 2 and 3. Conversely, under 'Z_{recon} perturbed' (maroon) setting, we observe no sensitivity in Case 1, while in Case 3, where most important generative factors are unavailable in the auxiliary data, Z_{recon} shows high sensitivity, resulting in a significant drop in SSIM. This experiment validates the intuition behind Aux-VAE's latent structure formation and underscores the importance of Z_{recon} in more realistic scenarios where most of the true generative factors are unknown.

4.2.4 Latent space traversal - systematically changing each dimension of Z_{aux} at a time

In our exploration of the latent space traversal, we aim to unravel the influence of individual latent factors on the generated outputs. Beginning with the extraction of latent factors from a sample image using the encoder network, we initiate the traversal by incrementally adjusting one latent factor's values at a time while keeping the others constant. Leveraging the decoder network, we then produce the corresponding output for each adjustment, visualizing the results along a row in Figure 5. This systematic exploration allows us to gain insight into how variations in latent factors affect the generated images. Figure 5 features three models—β-VAE, IDVAE, and our proposed Aux-VAE, used in case 2 of the galaxy image dataset. Due to space limitations, we

present the latent space traversal for only the top six most sensitive latent factors out of ten for each method. Detailed traversal results for this and additional datasets are provided in the SM.

We note that Aux-VAE's auxiliary latent factors Z_{aux} effectively adapt to the corresponding generating factors like $radius$, $g1$, and $g2$, thereby maintaining the geometric significance of these auxiliary factors. Given the minimal importance of Z_{recon} in this scenario, no noticeable changes are observed in the generated images during latent space traversal for Z_{recon}. This underscores the model's capacity to encapsulate relevant auxiliary information in a disentangled fashion. In contrast, for VAE and IDVAE, the latent factors exhibit more entangled relationships, as observed in the scatterplot shown in Figure 3. However, in the latent space traversal experiment, IDVAE performs better than VAE, likely due to the encompassing of auxiliary information. Specifically, the fifth, second, and eighth latent factors in IDVAE align well with the underlying generative factors $radius$, $g1$, and $g2$, respectively.

5 Conclusion

The proposed Aux-VAE method introduces a novel approach to variational autoencoder architecture, effectively integrating auxiliary (potentially non-exhaustive) information to enhance latent space disentanglement while preserving data generation quality. Demonstrated through extensive experiments across diverse datasets, Aux-VAE surpasses traditional VAEs and other disentanglement techniques, showcasing robustness and versatility due to its additional latents compensating for non-exhaustive auxiliary information, which is typical in scientific applications.

Looking ahead, we will validate Aux-VAE on real-world datasets and deepen its theoretical foundations. In particular, we plan to move beyond our current polynomial, pairwise dependency measures by adopting mutual-information–based metrics that more efficiently capture nonlinear relationships between latent factors and auxiliary variables. We also intend to investigate potential downstream uses of Aux-VAE—for example, determining whether a candidate measurement truly contributes independent variation by training without it and then assessing its alignment with the learned latent dimensions. We believe these efforts will both refine Aux-VAE's performance and broaden its applicability across diverse scientific domains.

Declarations

- Funding: This work is supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics, Office of Advanced Scientific Computing Research through the Scientific Discovery through Advanced Computing (SciDAC) program and through the FASTMath Institute, under contracts DE-AC02-06CH11357, DE-AC05-06OR23177, and DE-SC0023472, in collaboration with Argonne National Laboratory, Thomas Jefferson National Laboratory, National Renewable Energy Laboratory, and Virginia Tech. Work at Argonne National Laboratory was supported by the U.S. Department of Energy, Office of High Energy Physics. Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated by UChicago Argonne LLC under contract no.

DE-AC02-06CH11357. The training is carried out on Swing, a GPU system at the Laboratory Computing Resource Center (LCRC) of Argonne National Laboratory. This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308.

- Conflict of interest/Competing interests: The authors have no competing interests to declare that are relevant to the content of this article.
- Ethics approval and consent to participate: Yes
- Consent for publication: Yes
- Data availability: We have provided the code in the supplementary material to reproduce the galaxy simulation data used in our experiments. Other used datasets are publicly available, we cited the original papers that introduced these datasets.
- Materials availability: Not applicable
- Code availability: The code for the proposed method is available at Ganguli et al. (2024) and further mentioned in the supplementary material.
- Author contribution: All authors contributed to the conception and design of the study. Methodology development and data analysis were conducted by Arkaprabha Ganguli, while Nesar Ramachandra generated the galaxy simulation dataset. Arkaprabha Ganguli drafted the initial version of the manuscript, and all authors provided critical feedback and revisions to earlier versions. All authors read and approved the final manuscript.

References

Bonvin, C., Durrer, R.: What galaxy surveys really measure. Physical Review D **84**(6), 063505 (2011)

Balaprakash, P., Egele, R., Salim, M., Maulik, R., Vishwanath, V., Wild, S., et al.: "DeepHyper: A Python Package for Scalable Neural Architecture and Hyperparameter Search" (2018). https://github.com/deephyper/deephyper

Block, D.L., Freeman, K.C.: A Walk with Dr Allan Sandage—Changing the History of Galaxy Morphology, Forever: A Conference in honour of David Block and Bruce Elmegreen. In: Lessons from the Local Group: A Conference in Honor of David Block and Bruce Elmegreen, pp. 1–20 (2015). https://doi.org/10.1007/978-3-319-10614-4_1

Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in beta-vae. arXiv preprint arXiv:1804.03599 (2018)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American statistical Association **112**(518), 859–877 (2017)

Bauer, P., Thorpe, A., Brunet, G.: The quiet revolution of numerical weather prediction. Nature **525**(7567), 47–55 (2015)

Bouchacourt, D., Tomioka, R., Nowozin, S.: Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

Collaboration, L.D.E.S., *et al.*: The lsst desc dc2 simulated sky survey. Astrophysical Journal, Supplement Series **253**(1), 34 (2021)

Chen, J., Batmanghelich, K.: Weakly supervised disentanglement by pairwise similarities. Proceedings of the AAAI Conference on Artificial Intelligence **34**(04), 3495–3502 (2020) https://doi.org/10.1609/aaai.v34i04.5754

Cheung, B., Livezey, J.A., Bansal, A.K., Olshausen, B.A.: Discovering hidden factors of variation in deep networks. arXiv preprint arXiv:1412.6583 (2014)

Chen, R.T.Q., Li, X., Grosse, R., Duvenaud, D.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems (2018)

Deser, C., Phillips, A.S., Alexander, M.A., Smoliak, B.V.: Projecting north american climate over the next 50 years: Uncertainty due to internal variability. Journal of Climate **27**(6), 2271–2296 (2014)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression (2004)

Everett, S., Yanny, B., Kuropatkin, N., Huff, E., Zhang, Y., Myles, J., Masegian, A., Elvin-Poole, J., Allam, S., Bernstein, G., *et al.*: Dark energy survey year 3 results: measuring the survey transfer function with balrog. The Astrophysical Journal Supplement Series **258**(1), 15 (2022)

Ganguli, A., Ramachandra, N., Bessac, J., Constantinescu, E.: Aux-VAE: Variational AutoEncoder with Auxiliary Variables for Disentangled Representations. GitHub (2024). https://github.com/ArkaStatistics/Aux-VAE

Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

Hatfield, P.: Quantification of Unknown Unknowns in Astronomy and Physics. arXiv e-prints, 2207–13993 (2022) https://doi.org/10.48550/arXiv.2207.13993 arXiv:2207.13993 [astro-ph.IM]

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=Sy2fzU9gl

Hosoya, H.: Group-based learning of disentangled representations with generalizability for novel contents. arXiv preprint arXiv:1809.02383 (2018)

Korytov, D., Hearin, A., Kovacs, E., Larsen, P., Rangel, E., Hollowed, J., Benson, A.J., Heitmann, K., Mao, Y.-Y., Bahmanyar, A., *et al.*: Cosmodc2: A synthetic sky catalog for dark energy science with lsst. The Astrophysical Journal Supplement Series **245**(2), 26 (2019)

Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: International Conference on Artificial Intelligence and Statistics, pp. 2207–2217 (2020). PMLR

Kaltenborn, J., Lange, C., Ramesh, V., Brouillard, P., Gurwicz, Y., Nagda, C., Runge, J., Nowack, P., Rolnick, D.: Climateset: A large-scale climate model dataset for machine learning. Advances in Neural Information Processing Systems **36**, 21757–21792 (2023)

Kim, Y.-g., Liu, Y., Wei, X.-X.: Covariate-informed representation learning to prevent posterior collapse of ivae. In: International Conference on Artificial Intelligence and Statistics, pp. 2641–2660 (2023). PMLR

Kim, H., Mnih, A.: Disentangling by factorising. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2649–2658. PMLR, ??? (2018). https://proceedings.mlr.press/v80/kim18b.html

Kingma, D.P., Mohamed, S., Jimenez Rezende, D., Welling, M.: Semi-supervised learning with deep generative models. Advances in neural information processing systems **27** (2014)

Kumar, A., Poole, B.: On implicit regularization in β-VAEs. In: International Conference on Machine Learning, pp. 5480–5490 (2020). PMLR

Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, ??? (2018). https://openreview.net/forum?id=H1kG7GZAW

Kingma, D.P., Welling, M.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (2019) https://doi.org/10.1561/2200000056

Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes (2022)

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. Advances in neural information processing

systems **32** (2019)

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: International Conference on Machine Learning, pp. 4114–4124 (2019). PMLR

Lackner, C.N., Gunn, J.E.: Astrophysically motivated bulge–disc decompositions of Sloan Digital Sky Survey galaxies. Monthly Notices of the Royal Astronomical Society **421**(3), 2277–2302 (2012) https://doi.org/10.1111/j.1365-2966.2012.20450.x https://academic.oup.com/mnras/article-pdf/421/3/2277/5838204/mnras0421-2277.pdf

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., Tschannen, M.: Weakly-supervised disentanglement without compromises. In: International Conference on Machine Learning, pp. 6348–6359 (2020). PMLR

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, 253 (2017)

Merlin, E., Castellano, M., Bretonnière, H., Kuchner, U., Tuccillo, D., Buitrago, F., Peterson, J., Conselice, C., Caro, F., Dimauro, P., *et al.*: Euclid preparation-xxv. the euclid morphology challenge: Towards model-fitting photometry for billions of galaxies. Astronomy & Astrophysics **671**, 101 (2023)

Müller, M.U., Ekhtiari, N., Almeida, R.M., Rieke, C.: Super-resolution of multispectral satellite images using convolutional neural networks. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **V-1-2020**, 33–40 (2020) https://doi.org/10.5194/isprs-annals-V-1-2020-33-2020

Mita, G., Filippone, M., Michiardi, P.: An identifiable double vae for disentangled representations. In: International Conference on Machine Learning, pp. 7769–7779 (2021). PMLR

Maulik, R., Fukami, K., Ramachandra, N., Fukagata, K., Taira, K.: Probabilistic neural networks for fluid flow surrogate modeling and data recovery. Physical Review Fluids **5**(10), 104401 (2020) https://doi.org/10.1103/PhysRevFluids.5.104401 arXiv:2005.04271 [physics.flu-dyn]

Mathieu, E., Rainforth, T., Siddharth, N., Teh, Y.W.: Disentangling disentanglement in variational autoencoders. In: International Conference on Machine Learning, pp. 4402–4412 (2019). PMLR

Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.)

Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc., ??? (2016). https://proceedings.neurips.cc/paper_files/paper/2016/file/ef0917ea498b1665ad6c701057155abe-Paper.pdf

Newman, J.A., Gruen, D.: Photometric redshifts for next-generation surveys. Annual Review of Astronomy and Astrophysics **60**, 363–414 (2022)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

Paige, B., Van De Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., Torr, P., et al.: Learning disentangled representations with semi-supervised deep generative models. Advances in neural information processing systems **30** (2017)

Razavi, S., Gupta, H.V.: A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. theory. Water Resources Research **52**(1), 423–439 (2016)

Rowe, B.T.P., Jarvis, M., Mandelbaum, R., Bernstein, G.M., Bosch, J., Simet, M., Meyers, J.E., Kacprzak, T., Nakajima, R., Zuntz, J., Miyatake, H., Dietrich, J.P., Armstrong, R., Melchior, P., Gill, M.S.S.: GALSIM: The modular galaxy image simulation toolkit. Astronomy and Computing **10**, 121–150 (2015) https://doi.org/10.1016/j.ascom.2015.02.002 arXiv:1407.7676 [astro-ph.IM]

Rawlings, J.O., Pantula, S.G., Dickey, D.A. (eds.): Polynomial Regression, pp. 235–268. Springer, New York, NY (1998). https://doi.org/10.1007/0-387-22753-9_8 . https://doi.org/10.1007/0-387-22753-9_8

Raghavan Sathyan, A., Funk, C., Aenis, T., Winker, P., Breuer, L.: Sensitivity analysis of a climate vulnerability index-a case study from indian watershed development programmes. Climate Change Responses **5**, 1–14 (2018)

Reed, S.E., Zhang, Y., Zhang, Y., Lee, H.: Deep visual analogy-making. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc., ??? (2015). https://proceedings.neurips.cc/paper_files/paper/2015/file/e07413354875be01a996dc560274708e-Paper.pdf

Scoville, N., Aussel, H., Brusa, M., Capak, P., Carollo, C.M., Elvis, M., Giavalisco, M., Guzzo, L., Hasinger, G., Impey, C., *et al.*: The cosmic evolution survey (cosmos): overview. The Astrophysical Journal Supplement Series **172**(1), 1 (2007)

Shu, R., Chen, Y., Kumar, A., Ermon, S., Poole, B.: Weakly supervised disentanglement with guarantees. arXiv preprint arXiv:1910.09772 (2019)

Somerville, R.S., Davé, R.: Physical models of galaxy formation in a cosmological

framework. Annual Review of Astronomy and Astrophysics **53**, 51–113 (2015)

Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

Walt, S., Schönberger, J.L., Núñez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., contributors: scikit-image: Image processing in python. PeerJ **2**, 453 (2014) https://doi.org/10.7717/peerj.453

Wu, X., Balaprakash, P., Kruse, M., Koo, J., Videau, B., Hovland, P., Taylor, V., Geltz, B., Jana, S., Hall, M.: ytopt: Autotuning scientific applications for energy efficiency at large scales. Concurrency and Computation: Practice and Experience **37**(1), 8322 (2025) https://doi.org/10.1002/cpe.8322 https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.8322

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) https://doi.org/10.1109/TIP.2003.819861

Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled Representation Learning (2023)

Yang, S., Guo, T., Wang, Y., Xu, C.: Adversarial robustness through disentangled representations. Proceedings of the AAAI Conference on Artificial Intelligence **35**(4), 3145–3153 (2021) https://doi.org/10.1609/aaai.v35i4.16424

A Theoretical justification on regularizing the Expected Variational Posterior

In this section, we formalize the problem setting and outline the underlying assumptions. Suppose that our true generative model is $x = g^*(s) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ denotes the random fluctuations. We aim to learn a latent-variable model with prior $p(z)$ and generator g, where $g(z) \overset{d}{=} g^*(s)$. We also assume our access solely to a subset of the ground truth factors $s_{obs} \subset s, s_{obs} \in \mathbb{R}^d$, conveyed through auxiliary variables $u \in \mathbb{R}^d$. Each auxiliary variable is intended to encapsulate a specific ground truth factor within s_{obs}. Hence, our observed database contains n independently and identically distributed pairs of x and u, denoted as $\mathcal{D} = \{(x^{(1)}, u^{(1)}), (x^{(2)}, u^{(2)}), \ldots, (x^{(n)}, u^{(n)})\}$.

Defining the VAE objective:

Under this setting, the generative process can be written as:

$$z \sim p(z) \tag{11}$$
$$x \sim p_\theta(x|z, u) = p_\theta(x|z) \tag{12}$$

as, the latent factors z collectively represents the whole ground-truth generative factors, conditional on z, x and u are independent. Similarly, to develop the VAE framework, our first pathway is the inference process, denoted $q_\phi(x, z, u) = q_\phi(z|x, u)q(x, u) = q_\phi(z|x)q(x, u)$. Now, to obtain a sample (z, x, u) from this joint distribution, one would simply consider:

$$\begin{aligned} x, u &\sim q(x, u) \\ z &\sim q_\phi(z|x) \end{aligned} \tag{13}$$

where $q(x, u)$ denotes the ground truth data distribution, and $q_\phi(z|x)$ is the learnable variational posterior. The inference process aims to extract latent representations from actual samples from the data distribution $q(x, u)$. Hence, as illustrated in Kingma and Welling (2019), one feasible approach to optimize wrt the KL distance:

$$\begin{aligned} argmax_{\theta,\phi} &- KL\Big[q_\phi(x, z, u)||p_\theta(x, z, u)\Big] \\ &= \mathbb{E}_{q_\phi(x,z,u)}\Big[log\frac{p_\theta(x, z, u)}{q_\phi(x, z, u)}\Big] \\ &= \mathbb{E}_{q_\phi(z|x)}\Big[log\frac{p_\theta(x|z, u)p(z, u)}{q_\phi(z|x)}\Big] - \mathbb{E}_{q_\phi(z|x)}logq(x, u)) \\ &= \mathbb{E}_{q_\phi(z|x)}\Big[logp_\theta(x|z)\Big] + \mathbb{E}_{q_\phi(z|x)}\Big[log\frac{p(z, u)}{q_\phi(z|x)}\Big] - \text{constant} \\ &= \mathbb{E}_{q_\phi(z|x)}\Big[logp_\theta(x|z)\Big] + \mathbb{E}_{q_\phi(z|x)}\Big[log\frac{p(z|u)}{q_\phi(z|x)}\Big] - \text{constant} \\ &= \mathbb{E}_{q_\phi(z|x)}\Big[logp_\theta(x|z)\Big] - KL\Big[q_\phi(z|x)||p(z|u)\Big] - \text{constant} \end{aligned}$$

$$= \text{likelihood} - KL\Big[q_\phi(z|x)||p(z|u)\Big] - \text{constant} \tag{14}$$

Regularization over Expected Variational Posterior:

As discussed in Section 2 of the main manuscript, enforcing disentanglement can be approached by directly minimizing the KL divergence between the expected variational posterior $q_\phi(z) = \int q_\phi(z|x)p(x)dx$ and the prior $p(z)$. However, this KL term lacks a closed-form expression, which complicates optimization efforts. As an alternative, we propose to promote the major structural properties of $p(z)$ within $q_\phi(z)$. Specifically, we concentrate on the following three disentangled properties of $p(z)$ which can be expressed through its conditional distribution wrt u:

- ***Inter-independence*:** $u \perp z_{recon}$

  Proof:

  As we observe that, for $1 \leq j \leq d, 1 \leq j^{'} \leq d_Z - d$,

$$\begin{aligned}
\mathbb{E}(u_j z_{j'}) &= \int u_j z_{j'} p(u_j, z_{j'}) du_j dz_{j'} \\
&= \int u_j z_{j'} p(z_{j'}|u_j) p(u_j) du_j dz_{j'} \\
&= \int u_j z_{j'} p(z_{j'}) p(u_j) du_j dz_{j'} \\
&= \mathbb{E}(u_j)\mathbb{E}(z_{j'})
\end{aligned} \tag{15}$$

- ***Intra-independence*:** $u_j \perp z_{aux,j'}$ for $j, j' = 1, 2, \ldots d, j \neq j'$

  Proof:

$$\begin{aligned}
\mathbb{E}(u_j z_{j'}) &= \int u_j z_{j'} p(u_j, z_{j'}) du_j dz_{j'} \\
&= \int u_j z_{j'} p(z_{j'}|u_j) p(u_j) du_j dz_{j'} \\
&= \int u_j z_{j'} p(z_{j'}) p(u_j) du_j dz_{j'} \\
&= \mathbb{E}(u_j)\mathbb{E}(z_{j'})
\end{aligned} \tag{16}$$

- ***Explicitness*:** $Corr(u_j, z_j) \to 1$, for $1 \leq j \leq d$ as $n \to \infty$.

  Proof:

  By the theorem of total probability,

$$cov(u_j, z_j) = E_u cov_{u,z\sim p(z|u)}(u_j, z_j) + cov_u(u_j, E(z_j|u)) = cov(u_j, u_j) = var(u_j)$$
$$var(z_j) = E_u var(z_j|u) + var_u(E(z_j|u)) = \frac{1}{n} + var(u_j)$$

$$corr(u_j, z_j) = \frac{cov(u_j, z_j)}{\sqrt{var(u_j)var(z_j)}} = \frac{var(u_j}{\sqrt{(\frac{1}{n} + var(u_j))var(u_j)}} \to \infty$$

as $n \to \infty$

Quantitatively measuring these non-linear dependencies to measure Inter and Intra-independence strength under $q_\phi(z)$ is also non-trivial, and we turn our attention to polynomial regression here.

- ***Inter-independence*:** To measure the Inter-independence strength, we aggregate the correlation between different degrees of polynomials of u and Z_{recon}. e.g.

$$\begin{aligned} Cov(u^k, z_{recon}^{k'}) =& E_{(x,u)\sim p(x,u)} Cov_{u,z\sim q_\phi(z|x)}(u^k, z_{recon}^{k'}) + \\ & Cov_{(x,u)}(u^k, E_{z\sim q_\phi(z/x)}(z_{recon}^{k'}) \\ & = Cov_{(x,u)}(u^k, \mu_{\phi,d+1:d_Z}^{k'}). \end{aligned} \tag{17}$$

  Hence, we simply use a running estimate of these correlations between the u and the means of the latent factors to create the summary statistics $R_0(\cdot,\cdot)$ and $R_1(\cdot,\cdot)$which are informative in assessing the non-linear dependency. Similarly,

- ***Intra-independence*:**

$$\begin{aligned} Cov(u_j^k, z_{aux,j'}^{k'}) =& E_{(x,u)\sim p(x,u)} Cov_{u,z\sim q_\phi(z|x)}(u_j^k, z_{aux,j'}^{k'}) + \\ & Cov_{(x,u)}(u_j^k, E_{z\sim q_\phi(z|x)}(z_{aux,j'}^{k'}) \\ & = Cov_{(x,u)}(u_j^k, \mu_{\phi,j'}^{k'}), \text{ for } j, j' = 1, 2, \dots d, j \neq j'. \end{aligned} \tag{18}$$

  and

- ***Explicitness*:** This property also indicates that under $q_\phi(z|u)$, the correlation between $Z_{aux,j}$ and $u_j, j = 1, 2, \dots, d$ should be strong. Hence, we calculate

$$\begin{aligned} Cov(u_j, z_{aux,j}) =& E_{(x,u)\sim p(x,u)} Cov_{u,z\sim q_\phi(z|x)}(u_j, z_{aux,j}) + \\ & Cov_{(x,u)}(u_j, E_{z\sim q_\phi(z|x)}(z_{aux,j}) \\ & = Cov_{(x,u)}(u_j, \mu_{\phi,j}), \text{ for } j = 1, 2, \dots d. \end{aligned} \tag{19}$$

  Hence,

$$\begin{aligned} Corr(u_j, z_{aux,j}) &= \frac{Cov(u_j, z_{aux,j})}{\sqrt{var(u_j)var(z_{aux,j})}} \\ &= \frac{Cov(u_j, \mu_{\phi,j})}{\sqrt{var(u_j)(E(var_{z\sim q_\phi(z|x)}(z_{aux,j})) + var(\mu_{\phi,j})}} \end{aligned} \tag{20}$$

This incorporates $Corr(u_j, z_{aux,j}) < Corr_{(x,u)}(u_j, \mu_{\phi,j})$, for $j = 1, 2, \dots d$. While it is theoretically feasible to regularizing $Corr(u_j, z_{aux,j})$ in eq. 20 towards one, we

observed through our experiments that regularizing $Corr_{(x,u)}(u_j, \mu_{\phi,j})$ is computationally much more efficient and achieves similar levels of accuracy. For datasets with higher complexity, however, one may need to directly regularize $Corr(u_j, z_{aux,j})$ using eq. 20 using eq. 20 to capture the finer relationships within the data.

Hence, the final objective function of Aux-VAE incorporates the concepts of non-linear dependency between the latent factors and the auxiliary information into the optimization of the ELBO, thus effectively achieving the desired disentanglement. When constructing the main loss for Aux-VAE, we focus on Pearson's correlation coefficient due to its bounded nature and use running estimates over the mini-batch for the correlation.

B Additional Experimental Details

B.1 Brief description of the galaxy simulation data

In simulated and experimental datasets in scientific research, a subset of the auxiliary variables may be 'controlled' in the experimental design. Whereas in observational research domains such as astronomy, such quantities may just be measured or inferred either directly or with complementary studies. The dataset we have utilized in this effort is a representative simulation of telescopic observations created using GalSim Rowe et al. (2015). We assume that the light profile of each galaxy can be approximated as an exponential disk, which is known to be a good description of the outer, star-forming regions of spiral galaxies Lackner and Gunn (2012). In reality, galaxies can exhibit spiral, elliptical, barred spirals, irregulars, and other diverse morphologies Block and Freeman (2015). Telescopes are often systematically sensitive to certain types of galaxies, depending on which stage of galaxy evolution is probed based on instrumental specifications. For creating the synthetic galaxy image dataset, we first consider 5 physical parameters of varying importance. Ranges of these parameters are heuristically determined based on real observations, a Latin-Hypercube sampling over the range is performed to select 16,384 simulation points. The details of the ranges of the parameters are shown in Table 2. Each galaxy image is created with 33x33 pixels, with galaxies at the centers. We also note that while the galaxies here are in grayscale, the majority of the modern telescopes observe the Universe in multiple bands of channels of the light spectrum.

Table 2: Parameter descriptions and ranges of galaxy image dataset

Parameters	**Short description**	**Range**
$flux$	Apparent brightness of the galaxy (Number counts)	10^4 – 10^5
$radius$	Radius of the galaxy (Arc-seconds)	0.1 – 1
$g1$, $g2$	Reduced gravitational shear components (Cartesian coordinates)	-0.5 – 0.5
psf	Full width at half maximum of the point-spread function	0.2 – 0.4

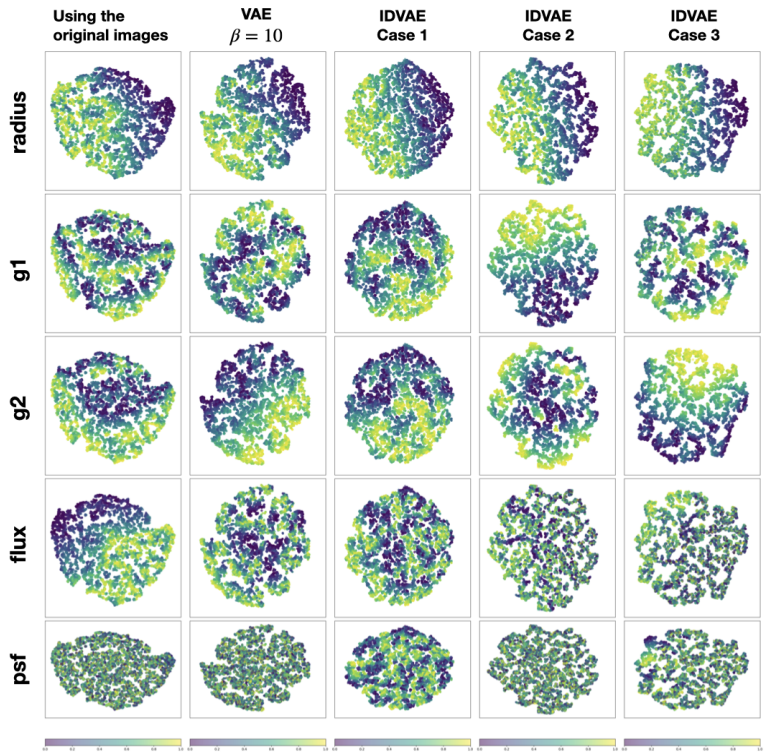


Fig. 6: Comparative t-SNE Visualization of Latent Spaces from VAE and IDVAE Across Multiple Scenarios with Respect to Generative Factors for the Galaxy Simulation Dataset

B.1.1 Relative importance between Z_{aux} and Z_{recon}

To elucidate the roles of Z_{aux} and Z_{recon}, we conducted a t-SNE analysis (van der Maaten and Hinton, 2008), illustrated in Figures 6 and 7. This analysis visualized the 2D components of t-SNE representations derived from original test images and latent factors from VAE, IDVAE, and Aux-VAE models across three scenarios outlined in Section 3 of the main manuscript. The plots are organized in a grid, with each column representing different configurations and each row color-coded by one of five generative factors, providing a method to assess each model's factor representation.

Key Observations:

1. **From t-SNE representation of latent spaces of competing methods in Figure 6**:
   - The first column, featuring t-SNE plots of original images, reveals the minimal impact of the psf factor on generative modeling, as shown by the absence of distinct clustering for psf.

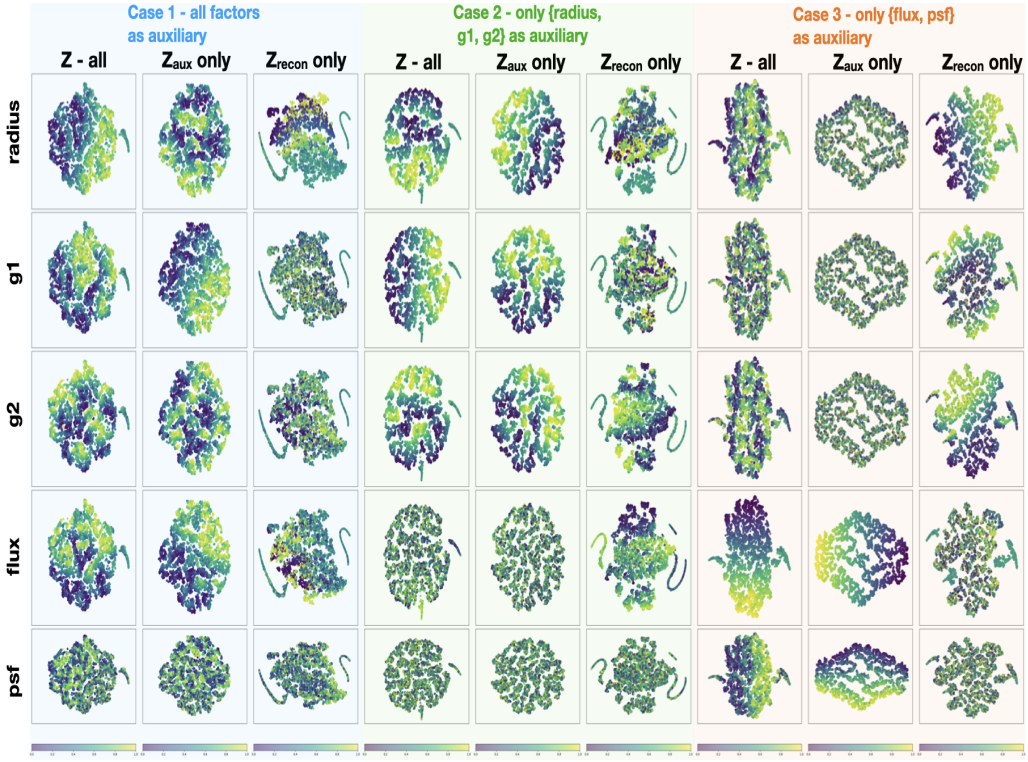


Fig. 7: Distinct Impact of Z_{aux} and Z_{recon} on Generative Factor Representation in Aux-VAE's Latent Space: A t-SNE Visualization

- The second column displays t-SNE plots of VAE's 10-dimensional latent factors from 1000 test images, indicating VAE's ability to recognize underlying generative factors, albeit with entangled representations. This suggests that while VAE identifies different factors, it has difficulty clearly separating them in the latent space.
- Columns 3-6 present t-SNE plots from IDVAE's application on cases 1, 2, and 3 (detailed in Section 3). Despite incorporating $flux$ and psf as auxiliary information in Case 3, IDVAE did not effectively represent these factors, struggling to distinctly segregate them in the latent space, which points to shortcomings in its auxiliary data integration.

2. **From the t-SNE representation of latent space of Aux-VAE in Figure 7**: This analysis was segmented into three parts, each examining the t-SNE representations of all latent factors, as well as the separate contributions of Z_{aux} and Z_{recon} to the preservation of generative factor information in the latent space.
   - **Case 1:** Here, all generative factors are encapsulated by the auxiliary features, and thus by Z_{aux}. The t-SNE plots show similar patterns for $Z-$all' and Z_{aux} only', indicating that Z_{recon} does not significantly contribute additional information regarding the generative factors in this scenario.

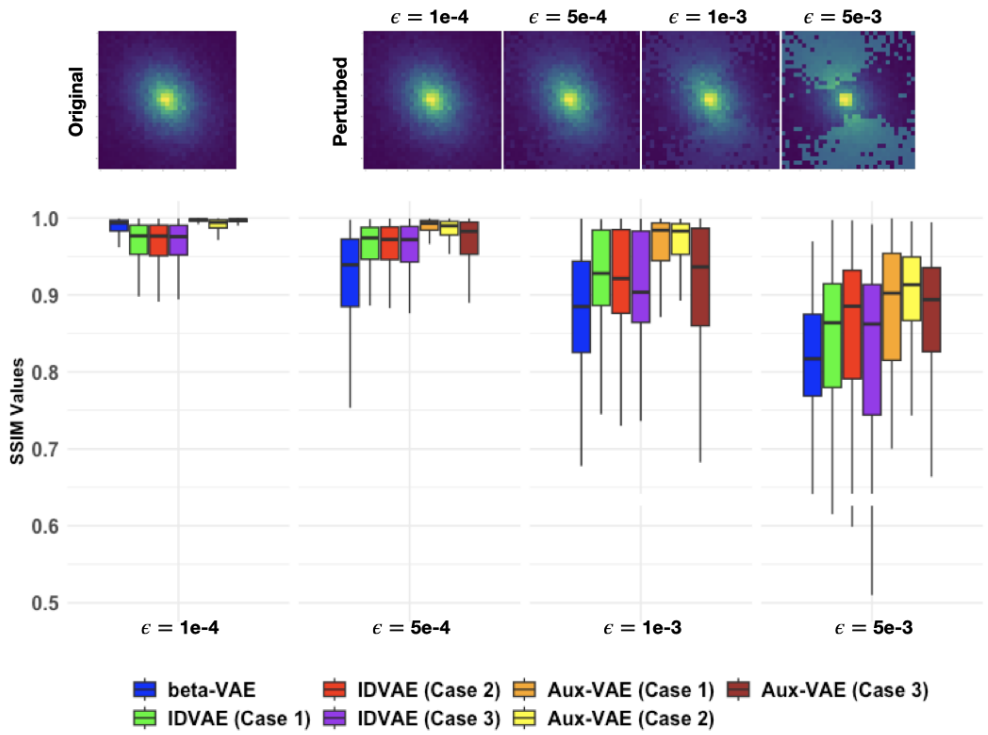


Fig. 8: Perturbed images and SSIM boxplots evaluating adversarial robustness across Beta-VAE, IDVAE, and Aux-VAE models under FGSM attack.

- **Cases 2 and 3:** A consistent pattern emerges across these cases. The generative factors that Z_{aux} covers are clearly depicted by the corresponding latent factors, while Z_{recon} effectively captures the remaining factors. For instance, in case 3, factors like $flux$, $g1$, and $g2$ are distinctly represented in the t-SNE plots of Z_{recon}, demonstrating its effectiveness in portraying uncovered generative aspects.

This analysis underscores how Aux-VAE dynamically adjusts the relative importance of Z_{aux} and Z_{recon}, from scenarios with comprehensive auxiliary information (case 1) to those with limited auxiliary data (case 3). This flexibility demonstrates Aux-VAE's capability to adapt and effectively utilize the available information to maintain accurate representation of underlying generative factors.

B.1.2 Adversarial robustness

In the adversarial robustness comparison experiment, we subjected the models to a Fast Gradient Sign Method (FGSM) attack, a common technique for testing model robustness against adversarial examples (Goodfellow et al., 2014). For an input image, the FSGM method uses the gradients of the loss with respect to the input image to create a perturbed new image that maximizes the loss. The perturbation strength

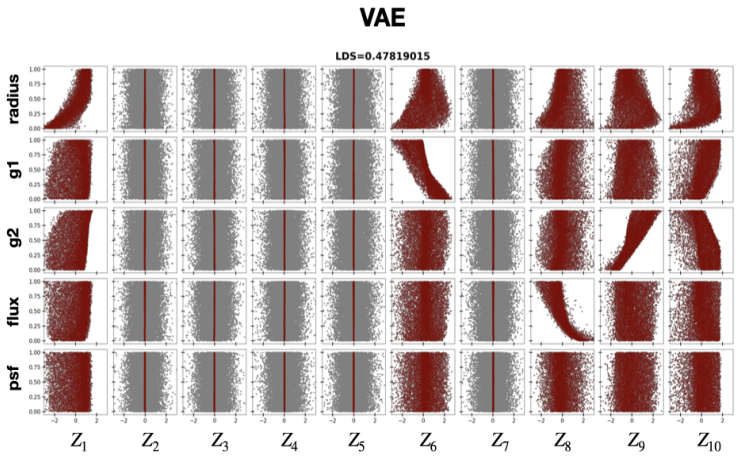


Fig. 9: Visualizing Disentanglement for VAE: This scatterplot contrasts latent factors (Z, represented by grey dots) with the latent means (μ_ϕ, shown as maroon dots), alongside highlighting the LDS metric.

is controlled by a parameter epsilon (ϵ). These perturbed images were then used to evaluate the models' performance under attack. In Figure 8, an illustration of the perturbed images under varying perturbation strength ϵ is presented. Additionally, for a test set of 1000 images, we calculate the SSIM metric between the input images and the reconstructed images after the FGSM attack, and the boxplots are presented in Figure 8. Naturally, we see a decline in SSIM-metric with the increase in perturbation strength ϵ. As anticipated, β-VAE, lacking proper disentanglement, exhibited heightened vulnerability to the adversarial attack. Conversely, both IDVAE and Aux-VAE, which demonstrate some level of disentanglement, exhibited comparatively greater robustness. Notably, in Case 3, where the auxiliary information lacks representation of several crucial generative factors, both IDVAE and Aux-VAE displayed relatively poor performance, suggesting the importance of comprehensive auxiliary information for enhanced adversarial robustness.

B.1.3 Disentanglement among the latent factors wrt the ground-truth generating factors - On the galaxy simulation data, Cars3D and Dsprites datasets

Following Section 3 of the main manuscript, here we present the remaining results on Disentanglement among the latent factors wrt the ground-truth generating factors across VAE, IDVAE and Aux-VAE.

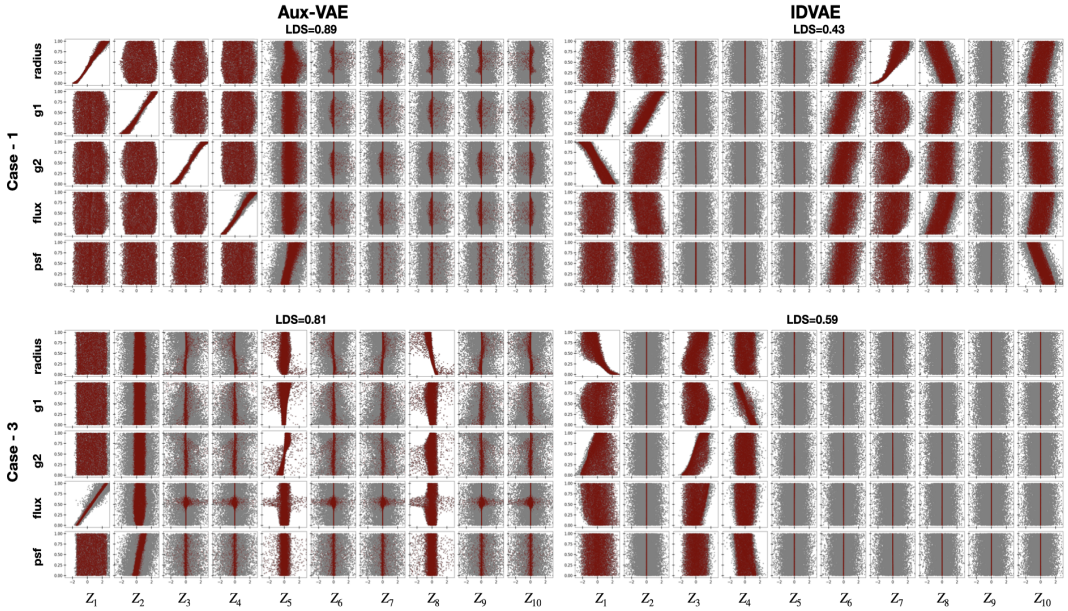


Fig. 10: Disentanglement Visualization for IDVAE and Aux-VAE: This scatterplot illustrates the comparison between latent factors (Z, depicted as grey dots) and latent means (μ_ϕ, represented as maroon dots), with an emphasis on the LDS metric. The plot includes results for cases 1 and 3 of the galaxy simulation dataset, with case 2 detailed in Section 3 of the main manuscript.

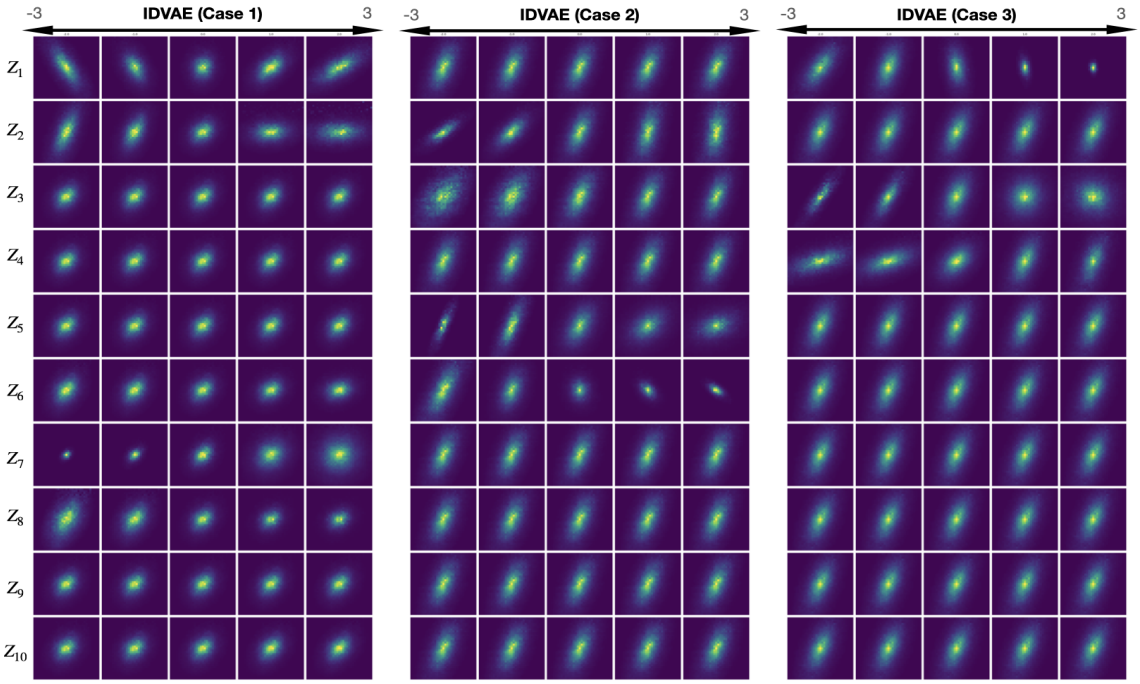


Fig. 11: Latent space traversal for IDVAE on the three cases considered for the galaxy simulation dataset.

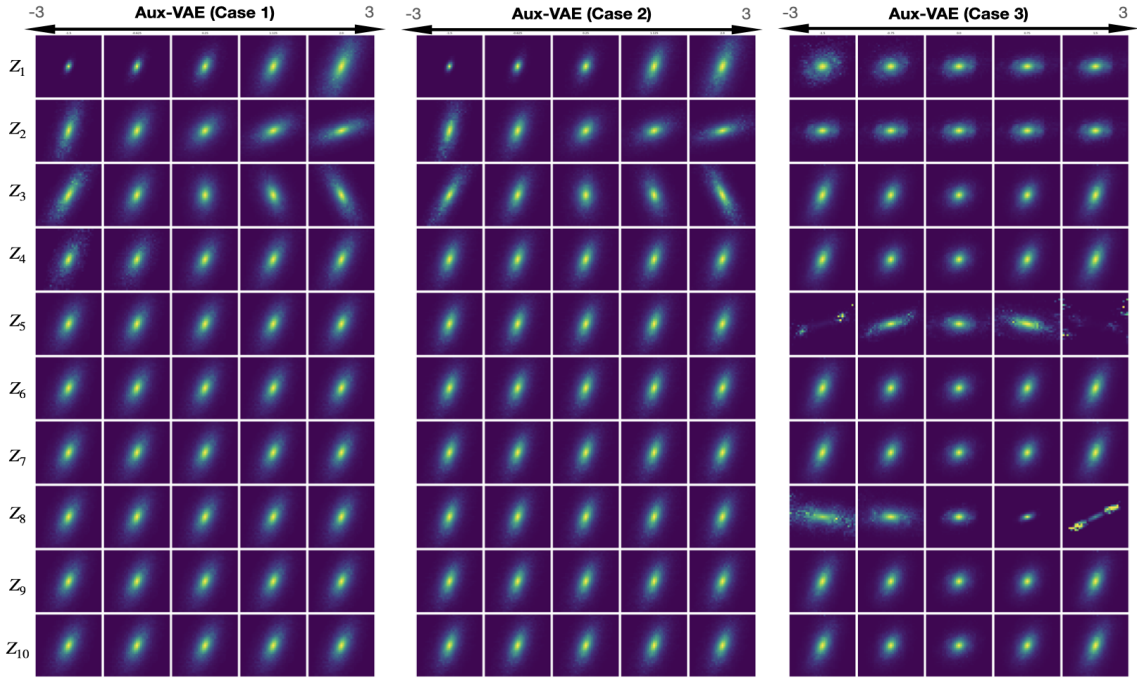
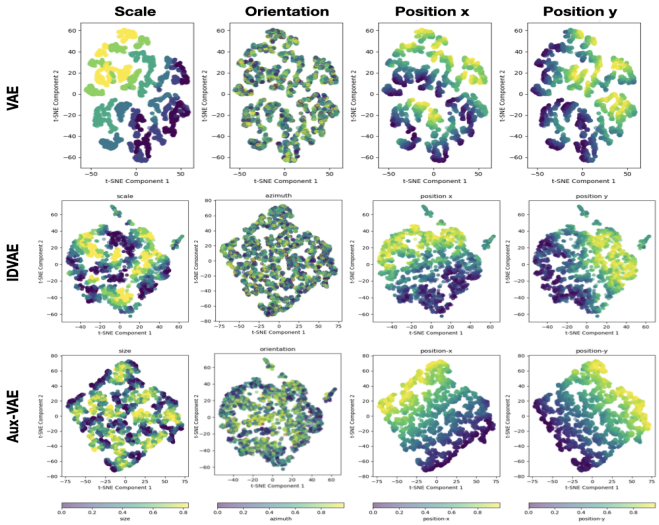
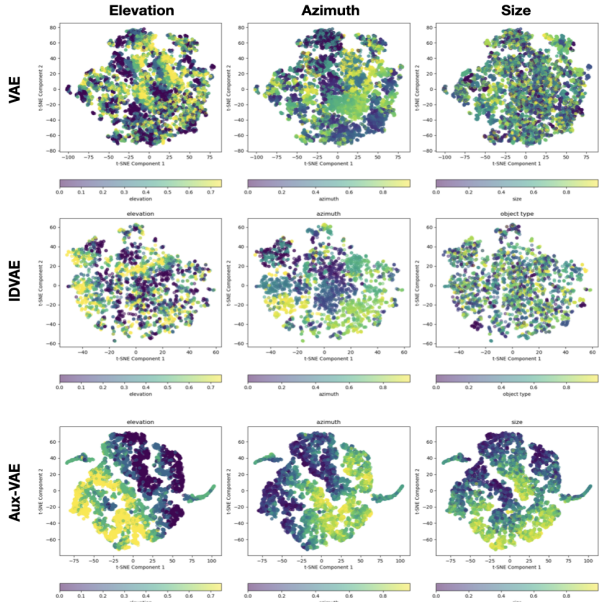


Fig. 12: Latent Space Exploration for Aux-VAE Across Three Scenarios in the Galaxy Simulation Dataset. In case 1, Z_{aux} includes $Z_{1:5}$; in case 2, Z_{aux} comprises $Z_{1:3}$; and in case 3, Z_{aux} consists of $Z_{1,2}$. The results demonstrate that in each scenario, Z_{aux} effectively adapts to the associated generative factors.

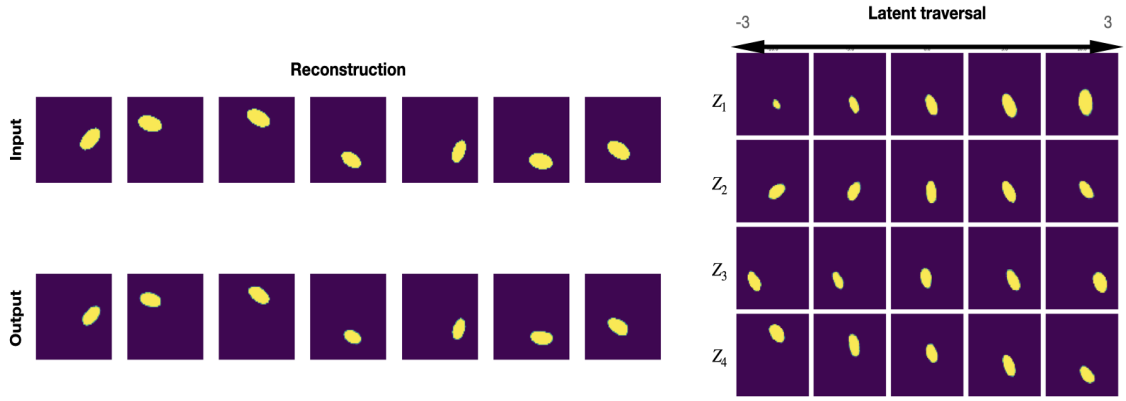


(a) Latent space representation on **DSprites** dataset by t-SNE: four columns represent the generative factors 'Scale', 'Orientation', 'Position x', and 'Position y'. Aux-VAE achieves better disentanglement compared to the other competing methods.

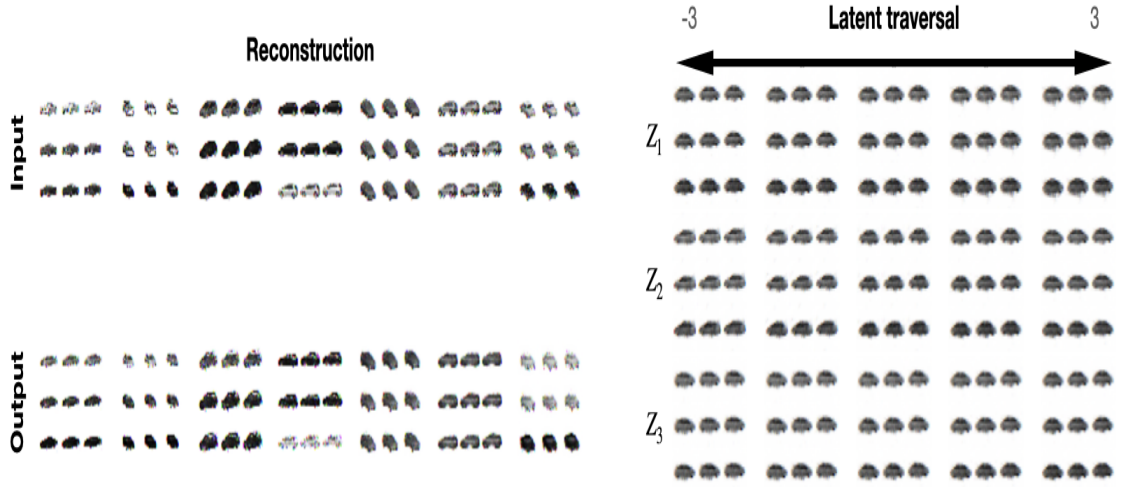


(b) Latent space representation on **Cars3D** dataset by t-SNE: three columns represent three generative factors 'height', 'azimuth', and 'size'. Aux-VAE achieves better disentanglement than the other competing methods.

Fig. 13: Latent space representation on **Cars3D** and **DSprites** dataset by t-SNE

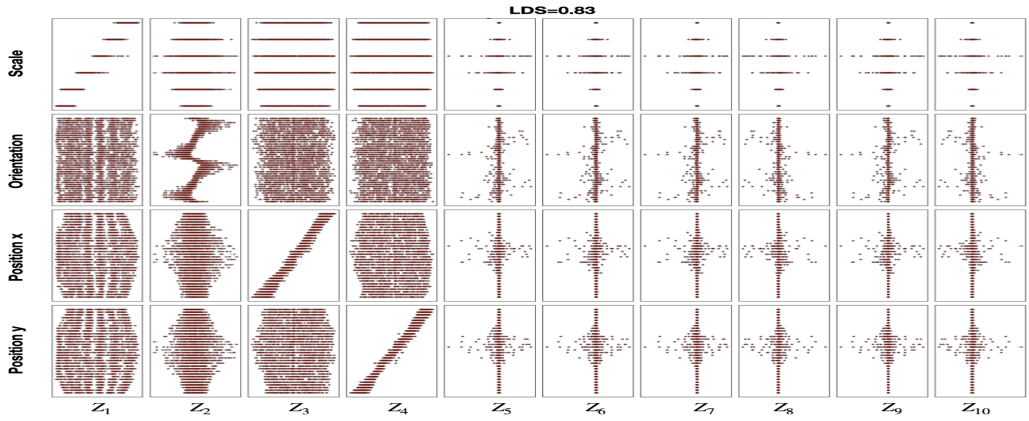


(a) Evaluation of Aux-VAE on the **DSprites** Dataset: Four columns display the generative factors 'Scale', 'Orientation', 'Position x', and 'Position y', demonstrating the model's reconstruction accuracy and latent space traversal. This analysis focuses exclusively on circular shapes within the dataset. Aux-VAE exhibits superior disentanglement relative to competing models, aligning with findings from previous studies Chen et al. (2018); Mita et al. (2021). Notably, Aux-VAE shows enhanced precision in capturing the 'Orientation' factor, a challenge where other models have shown limitations.

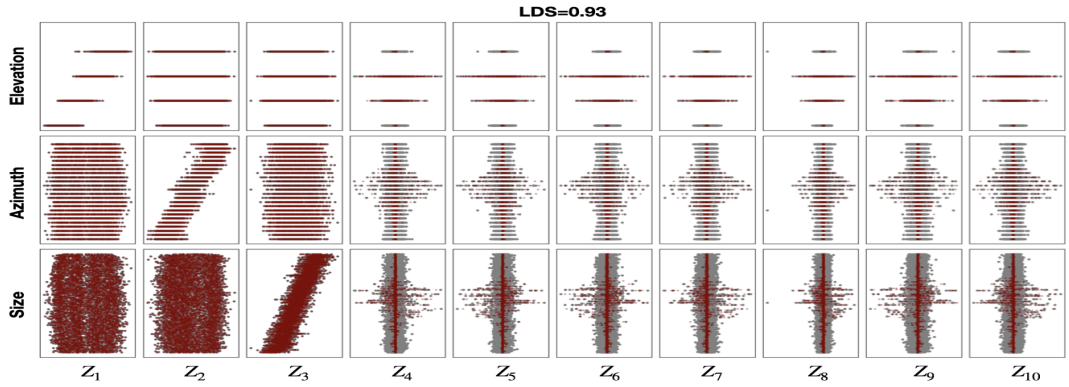


(b) Demonstration of Aux-VAE's reconstruction accuracy and latent space traversal on **Cars3D** dataset: four columns represent the generative factors 'Scale', 'Orientation', 'Position x', and 'Position y'. Aux-VAE achieves better disentanglement compared to the other competing methods.

Fig. 14: Demonstration of Aux-VAE's reconstruction accuracy and latent space traversal on **DSprites** and **Cars3D** dataset



(a) Visualizing Disentanglement in **'DSprites'** dataset: This scatterplot contrasts latent factors (Z, represented by grey dots) with the latent means (μ_ϕ, shown as maroon dots), alongside highlighting the LDS metric.



(b) Visualizing Disentanglement in **'Cars3D'** dataset: This scatterplot contrasts latent factors (Z, represented by grey dots) with the latent means (μ_ϕ, shown as maroon dots), alongside highlighting the LDS metric.

Fig. 15: Disentanglement in **'DSprites'** and **'Cars3D'** dataset: Each latent factor is plotted against the underlying generative factors available. It shows the first d latent factors (where d= no. of generative factors) are properly disentangled wrt the corresponding generative factors.

C Details on model evaluation, hyperparameter tuning, and code repository

C.1 Architectures

The architecture for the competing methods, $\beta-$VAE and IDVAE, follows the specifications from their respective repositories: $\beta-$VAE-repo and IDVAE repo. Each of the training is carried out on a single-NVIDIA A100 GPU. For Aux-VAE (code available at Ganguli et al. (2024)), we employed a basic grid-search approach for hyperparameter tuning, evaluating the mean squared error (MSE) and the disentanglement score (LDS) across different configurations. Specifically, we split the data into 7:2:1 as train, validation and test split. On the validation split, we tested d_r different combinations of hyperparameters $(\beta, \lambda_1, \lambda_2)$, calculating the test MSE and test LDS for each configuration, denoted as $MSE_1, MSE_2, \ldots, MSE_{d_r}$ and $LDS_1, LDS_2, \ldots, LDS_{d_r}$. We standardized these criteria and calculated their product $MSE(1 - LDS)$ to identify the optimal hyperparameter setting that jointly optimizes both reconstruction and disentanglement. Figure 16 illustrates the experiment and the hyperparameter values to select for the experiment. Table 4 shows the final selected values of the hyperparameters for each dataset. For other architecture details, we set batch size=64, and learning rate=1e-3 with Adam as the optimizer. However, we recognize that with an increasing number of hyperparameters, grid-search becomes impractical. In such cases, one would adopt stochastic hyperparameter optimization algorithms, such as Deephyper (Balaprakash et al., 2018; Wu et al., 2025), for a more efficient search. Table 3 outlines the basic configuration of Aux-VAE used for the galaxy simulation dataset. For the Cars3D and DSprites datasets, the configurations are adjusted to align with the IDVAE settings discussed in Mita et al. (2021).

For $\beta-$VAE and IDVAE, we checked the MSE and LDS scores to find the optimal level of regularization, in the case of galaxy simulation data analysis. For our β-VAE implementation, we grid-searched $\beta \in \{1, 5, 10, 20\}$ and found $\beta = 10$ optimal for the galaxy simulation dataset. For Cars3D and dSprites, we adopted $\beta = 5$ following Mita et al. (2021), given its proven balance of reconstruction fidelity and disentanglement. To be consistent with the settings of IDVAE, our experiments primarily focused on convolutional layers. A simpler MLP-based configuration is also provided in the Aux-VAE code repository for more general usage.

Table 3: Architecture of the Aux-VAE for the Galaxy Simulation Dataset

Encoder	**Decoder**
Input: $33 \times 33 \times 1$	Input: R^{10}
Conv2d(1, 32, 4, stride 2, padding 1), ReLU	ConvT2d(32, 128, 4, stride 2, padding 1), ReLU
Conv2d(32, 64, 4, stride 2, padding 1), ReLU	ConvT2d(128, 64, 4, stride 2, padding 1), ReLU
Conv2d(64, 128, 4, stride 2, padding 1), ReLU	ConvT2d(64, 32, 4, stride 2, padding 1), ReLU
Conv2d(128, 256, 4, stride 2, padding 1), ReLU	ConvT2d(32, 1, 5, stride 4, padding 0), Sigmoid
FC 256, FC 2×10	

β	λ_1	λ_2	Test MSE	LDS	1-LDS	Z_{MSE}=Standardized (MSE)	$Z_{1\text{-}LDS}$=Standardized(1-LDS)	Z_{MSE}*$Z_{1\text{-}LDS}$
5.00	0.10	0.10	0.000010	0.82	0.18	1.18	1.12	1.32
		1	0.000003	0.83	0.17	0.70	1.98	1.38
		2.00	0.000005	0.87	0.13	0.22	1.79	0.39
	1.00	0.1	0.000018	0.88	0.12	0.50	0.03	0.02
		1.00	0.000013	0.90	0.10	1.02	0.58	0.59
		2.00	0.000014	0.80	0.20	1.57	0.46	0.72
	2.00	0.1	0.000020	0.86	0.14	0.09	0.36	0.03
		1.00	0.000016	0.75	0.25	3.08	0.25	0.76
		2.00	0.000012	0.85	0.15	0.15	0.85	0.13
10.00	0.10	0.10	0.000029	0.83	0.17	0.82	1.46	1.20
		1	0.000009	0.87	0.13	0.31	1.21	0.37
		2.00	0.000007	0.83	0.17	0.75	1.52	1.14
	1.00	0.1	0.000021	0.87	0.13	0.25	0.46	0.11
		1.00	0.000022	0.85	0.15	0.23	0.62	0.14
		2.00	0.000017	0.88	0.12	0.71	0.09	0.06
	2.00	0.1	0.000021	0.87	0.14	0.16	0.43	0.07
		1.00	0.000022	0.88	0.12	0.52	0.60	0.31
		2.00	0.000019	0.84	0.16	0.47	0.20	0.10
20.00	0.10	0.10	0.000008	0.92	0.08	1.69	1.37	2.33
		1	0.000027	0.84	0.16	0.61	1.27	0.77
		2.00	0.000021	0.83	0.17	0.76	0.50	0.37
	1.00	0.1	0.000025	0.89	0.11	0.74	1.01	0.75
		1.00	0.000024	0.87	0.13	0.30	0.83	0.25
		2.00	0.000020	0.89	0.11	0.86	0.31	0.26
	2.00	0.1	0.000027	0.90	0.10	1.22	1.30	1.58
		1.00	0.000023	0.90	0.10	1.16	0.69	0.80
		2.00	0.000026	0.88	0.12	0.58	1.16	0.67

Fig. 16: Hyperparameter Tuning for Aux-VAE on Galaxy Simulation Data: Grid Search Approach. The final scores are rounded up to two decimal places. The selected configuration, highlighted in the table, effectively balancethe s optimization of MSE and disentanglement in the latent factors.

Dataset	β	λ_1	λ_2
Galaxy simulation dataset	5	1	0.1
Cars3D	5	2	1
DSprites	10	2	2

Table 4: Hyperparameter values used for Aux-VAE on each evaluated dataset.