RWTH Aachen University
Software Engineering Group

An Investigation of a Multimodal Variational Autoencoder Framework for Physics Data

Master Thesis

presented by

Effen, Moritz Bernhard

1st Examiner: Prof. Dr. Abigail Morrison

2nd Examiner: Prof. Dr. Bastian Leibe

Advisor: Dr. Zhuo Cao, Dr. Hanno Scharr

The present work was submitted to the Chair of Software Engineering

Aachen, September 29, 2025

Abstract

Many scientific domains, such as physics, provide multimodal data when observing complex phenomena or when doing experiments. Understanding individual contributions of each modality can help to optimise experimental setups and sensors, thereby potentially increasing accuracy on domain-specific tasks that rely on such data. This thesis examines the role of multimodal data in (downstream) prediction tasks, with a focus on the unique and shared contributions of the respective modalities. Disentangled representation learning is a paradigm that aims to extract the independent, underlying factors from data. We employ this approach for multimodal data, proposing an extension to the disentangled multimodal variational autoencoder (DMVAE) by incorporating an additional optimisation objective to enforce minimal redundancy between shared and unique latent representations extracted by the DMVAE. Based on these representations, we train and evaluate several downstream tasks to study their contributions to the task. We compare this method to the traditional DMVAE and VAE across multimodal and single-modal configurations and also compare it directly to regression models. In our experiments, this approach is applied to the Multimodal Universe (MMU) astronomical dataset, which includes both imagery and spectral data. We also evaluate the impact of a physical-model-based differentiable image decoder model for extracting meaningful parameters into the latent space. Additionally, the setup is applied to HyPlant hyperspectral remote sensing data, which consists of airborne measurements of Earth's surface, to study it as a source of multimodal data to test how much information images and spectra contain about hyperspectral data.

Acknowledgment

This thesis was written at the Forschungszentrum Jülich, Institute of Advanced Simulation 8: Data Analytics and Machine Learning (IAS-8), which provided resources and facilities. I want to thank Dr. Hanno Scharr for the opportunity to write my thesis in a supportive and inspiring research environment. I also want to thank Dr. Zhuo Cao for his support throughout the work.

Task description

The aim of this thesis is to investigate multimodal disentangled representation learning (DRL) on physics data and to study whether incorporating multimodal information can increase task performance. The tasks considered may include classification, regression or detection in the physics domain. Moreover, this thesis will study the unique and shared information encoded across different modalities and investigate how they affect task performance. To achieve this, existing architectures, such as multimodal extensions of VAEs [KW$^+$13] like DMVAEs [LP21], may be adapted or improved, and a suitable framework for multimodal DRL will be implemented. This approach can be applied to the recently published Multimodal Universe dataset [AAB$^+$24], which contains multimodal data in the astronomical domain with potential tasks such as physical property prediction or galaxy morphology classification. Also, other physics datasets may be evaluated.

A small complementary aspect of the thesis is to explore physical model-guided disentangled representation learning. As disentangled representation learning in an unsupervised way may be unstable, we want to investigate how to incorporate existing physical models to guide the learning of meaningful disentangled representations while also incorporating unmodeled features. This requires implementing a differentiable physical model and integrating it into the framework, followed by an evaluation of its impact on the latent space and task performance.

Overall, the goal is to develop and investigate a multimodal, variational autoencoder-based framework for physics data that can study the impact of shared and unique information in multimodal data on a downstream task and apply this framework to the Multimodal Universe dataset to study which modalities are suited for the task of physical property prediction of galaxies and also apply this framework to other physics datasets.

Contents

Chapter 1

Introduction

In many scientific domains, complex processes are investigated from multiple complementary perspectives. Combining information from various measurements of the same process or object into multimodal datasets can potentially allow for a more accurate and comprehensive view of the underlying information. This complementary information contained in the data can then be leveraged to potentially improve the predictive performance on downstream tasks, predicting meaningful properties more accurately. This is especially relevant when measurements are noisy and single modalities alone cannot fully represent the process or object studied. This can happen in scientific domains, such as physics, where vast amounts of data are collected using several sensors optimised for different purposes. This multimodal data can, for example, include images, spectra or other forms of data. However, as not all modalities can make an equal contribution to a given prediction task, it is also important to understand which modalities can effectively help extract the desired information and provide useful and unique content, and which modalities do not provide additional helpful information or only make redundant contributions ot the task. This information can help optimise a sensor setup to capture complementary data that targets observations that are most informative for the desired task.

To address this, we require a suitable method that can process large amounts of data and can accurately and efficiently extract an informative, independent (disentangled) latent variable representation of the data, also known as disentangled representation learning (DRL) [WCT$^+$24], based on which we can evaluate downstream task performance. Deep learning has been around for decades, but has gained traction in recent years through the introduction of novel probabilistic methods, such as the Variational Autoencoder (VAE) [KW$^+$13], faster algorithms, and hardware for training at larger scales. This has made machine learning a standard tool for analysing physics data. Using deep models designed for handling multimodal data enables the extraction of shared and unique representations, suitable for studying where essential information is distributed across modalities.

This thesis explores the potential of a (model-agnostic) multimodal variational autoencoder framework to study downstream prediction tasks on the unique and shared contribution of each modality. We investigate whether and to what extent the usage of multimodal data, compared to individual modalities, can influence the accuracy of physical property predictions and other downstream tasks. The approach used is based on an extension of Variational Autoencoders called the Disentangled Multimodal Variational Autoencoder (DMVAE) [LP21]. These are both generative models that can be used both to generate novel data and to extract meaningful underlying generative features of the data, which can serve as an embedding usable for downstream tasks. The DMVAE is suited for multimodal data and is designed to extract modality-specific unique features and features shared between multiple modalities into separate latent spaces. This is useful for investigating unique and shared information between the modalities. To ensure a clear information-theoretic separation between shared and unique features and to reduce redundancies between them, we extend this model by an additional penalty on the mutual information between the shared and unique representations. This penalty on mutual information is the Contrastive Log-ratio Upper Bound on mutual information (CLUB) [CHD+20]. Using this framework, we can directly determine the contribution of each latent space by training downstream models on the respective latent spaces.

Our approach combines several advantages of embeddings, VAEs and DMVAEs. It enables cross-reconstruction from one modality, provides an embedding that strictly separates unique and shared features usable for downstream tasks and allows for a clearer analysis of where task-relevant information is encoded compared to the traditional DMVAE. These separated representations could also be used for further processing.

Astronomical data is a natural choice for applying this framework, as it provides vast amounts of multimodal data. Astronomical objects are often captured through multiple sensors, including telescope images, spectra, hyperspectral measurements, time-series observations, and tabular data, which include inferred physical properties or morphologies of galaxies. In this thesis, the focus lies on images and spectra as two modalities that capture different views of an object, containing crucial physical information about galaxies. We utilise these modalities for the downstream task of physical property prediction, encompassing properties such as redshift, stellar mass, and others [AAB+24]. Using this approach on this data, we can study which physical property of the galaxy is encoded at which modality. To assess the potential benefits of this multimodal feature extraction for physical property inference, we compare the performance of the proposed extended DMVAE model with CLUB to that of the traditional DMVAE and VAE, as well as with single-modal data. In addition to our extended DMVAE approach, we investigate whether physical-model-based differentiable decoder models can guide the model to learn a semantically meaningful latent representation, thereby improving task performance. Specifically, we utilise a physical-model-based image decoder that aims to reconstruct galaxy images based on physical parameters, thereby enhancing the interpretability of the latent representation. At last, remote sensing hyperspectral data are investigated as a source of multimodal data.

This thesis guides you through related work, provides the necessary background, outlines the general framework that contains the methods used, and presents the experiments conducted to evaluate the framework and the datasets employed. It is structured in the following way:

In chapter 2, we present previous work related to processing/evaluating multimodal data and also reference previous work on the physics datasets used. Afterwards, chapter 3 introduces all necessary concepts and background from probability theory, including methods for estimating mutual information, such as CLUB [CHD+20], and an application of these methods, which is to estimate task contributions similar to our method. We then describe machine learning concepts, architecture, and generative models, such as VAEs. Here, concepts from disentangled representation learning (DRL) are also described. Afterwards, in chapter 4, the general problem setting is stated and the framework is introduced. Therefore, we look at the setup of the learning problem to be evaluated and the motivation behind it. Specifically, we examine the Disentangled Multimodal Variational Autoencoder (DMVAE) [LP21] proposed by Lee and Pavlovic, which serves as the primary architecture we study and base our framework on. Subsequently, the architecture is extended with CLUB loss to minimise the mutual information between representations. Based on that framework, two methods are described to investigate the contributions of unique and shared information between modalities to a downstream task. In chapter 5, we describe the experiments conducted along with the motivation behind them. For this, we first explain the evaluation tools, including several metrics and visualisation tools to evaluate the model's performance. We also outline the datasets used, which include the Multimodal Universe [AAB+24] dataset and, later on, also the HyPlant FLUO [SAC+19] remote sensing dataset. Also, corresponding preprocessing steps, model architectures and further training details are described. We then present several experiments and their results, including hyperparameter optimisation, several parameter studies, and general evaluation experiments that examine the combination of image and spectrum modalities and compare it to the use of single modalities. This is done using the VAE, DMVAE and DMVAE with CLUB. Additionally, we investigate the inclusion of a differentiable physical-model-based decoder for predicting a semantically meaningful latent space and its impact on downstream task performance. Furthermore, we experiment with hyperspectral data from HyPlant as a data source that combines images and spectra, by decomposing the hyperspectral data into these modalities to apply our framework and thereby studying which modality contains more information about the underlying hyperspectral data. At the end in chapter 6, we conclude our results.

In summary, the key contributions of this thesis are:

1. We propose an extension of DMVAE with CLUB loss capable of providing redundancy-free shared and private representations of multimodal data. We demonstrate how this approach can be utilised to assess the impact of both unique and shared features on downstream tasks, and we test how the additional CLUB loss affects the model's performance, latent space and mutual information between the representations.

2. We apply the framework to galaxy imaging and spectral data, evaluating the contributions of these modalities to the task of predicting physical properties for galaxies. Here, we conduct a thorough comparison of the physical property prediction task using VAEs, DMVAEs, DMVAEs with CLUB, and regression models on both single- and multimodal data. We also use our framework to analyse hyperspectral remote sensing data, aiming to identify the information that structural images and spectra convey about the underlying hyperspectral data.

3. We do several hyperparameter optimisations to study their impact on the model's performance. Additionally, we conduct a test on different latent sizes to investigate their effects. In doing so, we determine suitable representation sizes for image and spectral modalities, as well as for their unique and shared features.

4. We also study how to extract interpretable parameters from the data simultaneously. To address this, we specifically investigate whether incorporating a physical-model-based image decoder for galaxy images into the framework can help guide the learning of a semantically meaningful latent space and assess its impact on the downstream task.

Chapter 2

Related Work

Several studies have explored topics similar to those addressed in this thesis. We begin with a literature review, which is subdivided into work related to multimodal methods and models, as well as methods designed explicitly for (astro-) physics data.

2.1 Work related to Multimodal Methods

Variational Autoencoders [KW+13] were introduced by Kingma and Welling as a probabilistic method for mapping data into a lower-dimensional latent distribution, capturing the underlying generative factors of the data that can be used for generating novel data. Wang et al. provide an overview of various methods for extracting informative, disentangled representations of data using VAEs and other models [WCT+24]. An extension of the VAE that is more semantically meaningful is the Disentangled Multimodal Variational Autoencoder (DMVAE) [LP21] by Lee and Pavlovic. This model improves upon previous work, such as JMVAE [SNM16], JVAE [VFHM17], or MVAE [WG18], by utilising a product-of-experts approach to extract features that are unique to a modality or shared between modalities, and enables cross-reconstruction capabilities. An extension of this type of model is the SSDMM-VAE [MSSA23], which, in addition to shared and unique latent spaces, utilises both discrete and continuous latent spaces to enhance performance. Other current models include BridgedVAE [YSNH20]. Due to its meaningful latent structure, cross-reconstruction capabilities, and well-performing nature, this thesis focuses on the DMVAE.

Another method for estimating contributions of multimodal data to a downstream task is Partial Information Decomposition [WB10], introduced by Williams et al. This information-theoretic framework is designed to study the unique, synergistic, and redundant information that multiple modalities contain about some target variable and is used in neuroscience or machine learning due to its explainable nature.

2.2 Work related to (Astro-) Physics data analysis

Several previous works have applied autoencoder and VAE-like architectures to astronomical data. Schawinski et al. propose a method using a Fader network [LZU+17], a network that enables the controlled manipulation of specific attributes, to disentangle features and physical properties in the latent space and test which physical properties are responsible for transforming one data population into another [STZ18], thereby evaluating the plausibility of hypotheses. Later, Aragon-Calvo introduced an autoencoder framework that incorporates a physical model decoder of galaxy images to reconstruct the original data [ACC20]. Therby, the model automatically predicts the physically semantically meaningful input parameters of the physical model corresponding to the shape of the galaxy. Takeishi et al. propose a more sophisticated approach that incorporates a physical model into an autoencoder framework, enabling it to additionally model features in the data that are not captured by the physical model, using additional regularisation methods [TK21]. Later on, several VAE architectures have been explored on galaxy image data [XSdS+23], [Dia22]. Iwasaki et al. have applied the VAE to galaxy spectra from SDSS and examined low-dimensional latent representations on their information content about underlying physical properties [ICT23].

Recently, the Multimodal Universe dataset [AAB+24] was published, a large-scale multimodal scientific astronomical dataset designed for machine learning research. Based on this dataset, Parker et al. introduced AstroCLIP [PLG+24], a foundation model for embedding galaxy images and spectral data into a meaningful embedding space. For this, they utilise separate transformer encoders for images and spectra and minimise a contrastive loss in the embedding space to align corresponding image and spectrum embeddings. Based on these embeddings, they evaluate the performance of several downstream tasks. They predict physical properties including stellar mass, specific star formation rate, mass-weighted metallicity and mass-weighted stellar age from the PROVABGS catalogue [HKT+23], as well as the photometric redshift from DESI [AAA+16], and they use the Galaxy Zoo DECaLS [WLG+22] dataset to test galaxy morphology classification [PLG+24]. Compared to this approach, our approach can differentiate whether the information for the downstream task is encoded in the unique or shared features of the modalities. We can either obtain an embedding for a single modality or a more precise embedding when combining both modalities. Additionally, our model can directly cross-reconstruct the other modality instead of relying on similarity search. For the HyPlant dataset, we refer to [SAC+19] for more details.

Chapter 3

Essential Background

This chapter introduces the necessary machine learning concepts and methods, which are later utilised in the framework. We will first introduce concepts of probability theory, machine learning and architecture design, and then examine how these concepts can be applied to generate novel data. We then build upon the idea of autoencoders to a probabilistic generative extension called the Variational Autoencoder (VAE) [KW$^+$13]. We then examine the concept of disentangled representation learning (DRL), which aims to learn the underlying, separable generative factors of data. Several methods will be described that can achieve this.

3.1 Probabilistic Background

To understand the methods used later on, it is necessary to introduce some probabilistic background first. *Random variables* are functions $X : \Omega \to E$ that map events from a sample space Ω to a measurable space E. They are often written as capital letters, and the probability for some realisation $x \in E$ of the random variable is in the discrete case

$$P(X = x) = P(\{\omega \in \Omega | X(\omega) = x\}).$$

For easier notation, this will often be abbreviated by $p(x)$. For random variables, the distinction between the discrete and continuous cases is necessary, depending on whether the event space is discrete and countable or continuous. We focus on the continuous case here. However, for both cases, similar formulas and derivations hold, with the main difference being that for the continuous case, integrals are used, and we need to compute probabilities over intervals; for the discrete case, sums are used. Probability distributions p assign probabilities to events. In general, the properties $\int p(x)dx = 1$ and $p(x) \geq 0$ must hold.

In the case of multiple random variables, more concepts can be introduced. The joint distribution $p(x, y)$ describes the probability of observing x and y at the same time. Here we can also marginalise out random variables to obtain the probability for single variables $p(x) = \int p(x, y)dy$. The conditional likelihood $p(x|y)$ describes the probability of observing x if y is already known. It is defined as $p(x|y) = \frac{p(x,y)}{p(y)}$. A fundamental principle for updating prior beliefs based on new observations is Bayes' Theorem. It can be derived by applying the product rule twice:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)},$$

where

$$p(x) = \int p(x|y)p(y)dy.$$

After the basic probabilistic context is described, we briefly outline two important probabilistic measures that are used throughout the thesis. The first one is the so-called *Kullback-Leibler (KL) divergence*, defined as

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx.$$

This is a measure of how different two probability distributions P and Q are from each other, serving as a measure of the discrepancy between the distributions. In the context of machine learning, it is often used to estimate how closely a learned distribution approximates a target distribution. It has the important properties of being always non-negative $D_{KL}(P\|Q) \geq 0$ and that $D_{KL}(P\|Q) = 0$ if P and Q are the same distribution. Based on this, we can define the *mutual information* (MI) $I(X;Y)$, which is a symmetric measure for how much information one random variable X contains about the other Y. Formally,

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] = D_{KL}(p(x,y)\|p(x)p(y)).$$

If two variables X, Y are independent, then $I(X;Y) = 0$. Mutual information is a crucial measure, as it enables us to quantify how much one variable reveals about another and how dependent they are on each other. This can serve multiple purposes. By minimising this measure, we can reduce redundancies between the random variables X and Y. This can also serve as a measure of how much a variable X contributes to predicting some variable Y. In the general setting, when the closed form of the marginal distributions $p(x), p(y)$ or that of the joint distribution $p(x,y)$ is unknown, it is often infeasible to directly estimate it accurately. Here, several sampling methods, such as non-parametric binning, kernel density estimation, or K-NN entropy estimation, can be used [CHD+20]. However, these methods are often unreliable and struggle with high-dimensional data [CHD+20]. We now show a technique that is better suited for estimating MI and minimising it.

Upper and Lower Bounds on Mutual Information

The Contrastive Log-ratio Upper Bound (CLUB) [CHD+20] provides a tight differentiable upper bound on the mutual information between two random variables $\mathbf{X}$ and $\mathbf{Y}$. This bound relies on both negative and positive samples, as well as a conditional distribution $p(\mathbf{y}|\mathbf{x})$ for the considered variables. There exist several variations of this loss for several cases: 1. If the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is known; 2. If the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is unknown. In the first case, the upper bound can be estimated as

$$I_{CLUB}(\mathbf{X};\mathbf{Y}) := \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[\log p(\mathbf{y}|\mathbf{x})] - \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{y})}[\log p(\mathbf{y}|\mathbf{x})].$$

Here follows a short sketch of the proof that $I_{CLUB}(\mathbf{x};\mathbf{y})$ is an upper bound on mutual information as described in [CHD+20], where more details are explained. It needs to be shown that

$$I_{CLUB}(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}) \geq 0.$$

By inserting the upper bound and reordering the terms, we get [CHD+20]

$$I_{CLUB}(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}) = \mathbb{E}_{p(\mathbf{y})}[\log p(\mathbf{y}) - \mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{y}|\mathbf{x})]].$$

By Jensen's inequality, the following holds as log is concave: $\log p(\mathbf{y}) = \log(\mathbb{E}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]) \geq \mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{y}|\mathbf{x})]]$, see [CHD+20]. Hence, $\log(\mathbb{E}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]) - \mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{y}|\mathbf{x})]] \geq 0$ and thus

$$I_{CLUB}(\mathbf{X};\mathbf{Y}) - I(\mathbf{x};\mathbf{y}) \geq 0.$$

This upper bound can now be estimated with the samples (x_i, y_i) with $i \in \underline{N}$ as

$$\hat{I}_{CLUB} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\log p(\mathbf{y}_i|\mathbf{x}_i) - \log p(\mathbf{y}_j|\mathbf{x}_i)].$$

This upper bound takes the difference between the log-conditional distributions of positive and negative pairs and is therefore contrastive. For unknown conditional distributions, instead of using the true conditional distribution, we have to use a variational approach to approximate $p(\mathbf{y}|\mathbf{x})$ by $q_\theta(\mathbf{y}|\mathbf{x})$, which is represented by a neural network. This network learns to predict one variable from another and must be trained simultaneously. This is done by maximising the log-likelihood

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\log q_\theta(\mathbf{y}_i|\mathbf{x}_i).$$

Then the upper bound approximation with variational CLUB is defined as

$$\hat{I}_{vCLUB} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\log q_\theta(\mathbf{y}_i|\mathbf{x}_i) - \log q_\theta(\mathbf{y}_j|\mathbf{x}_i)],$$

which can no longer guarantee that it is always an upper bound on mutual information, but is a good approximation of it [CHD+20]. Since this bound is differentiable, we can use it to minimise the mutual information between two variables $\mathbf{x}$ and $\mathbf{y}$ sampled from some distribution $p_\sigma(\mathbf{x},\mathbf{y})$. The following needs to be done in each training step with the sampled pairs $(\mathbf{x}_i, \mathbf{y}_i)$ [CHD+20]:

1. Compute $\hat{I}_{CLUB}$ by computing $U_i = \log q_\theta(\mathbf{y}_i|\mathbf{x}_i) - \frac{1}{N}\sum_{j=1}^{N} q_\theta(\mathbf{y}_j|\mathbf{x}_i)$ for each $i \in \underline{N}$. Then update $p_\sigma(\mathbf{x},\mathbf{y})$ by minimizing $\hat{I}_{CLUB} = \frac{1}{N}\sum_{i=1}^{N} U_i$.
2. Maximize likelihood $\mathcal{L}(\theta)$.

We can compute the upper bound without the minimisation objective by not updating $p_\sigma(\mathbf{x},\mathbf{y})$. For minimising MI between both variables in the distribution $p_\sigma(\mathbf{x},\mathbf{y})$, the CLUB approximation delivers more stable gradients than other methods [CHD+20].

Additionally, there are several methods to derive a lower bound on MI. A standard method for estimating a lower bound on mutual information is Information Noise Contrastive Estimation (InfoNCE) [CHD+20]. It uses a scoring function $f(\mathbf{x},\mathbf{y})$ that compares the contrastive embeddings of $\mathbf{x}$ and $\mathbf{y}$ [OLV18]

$$I_{NCE} := \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{f(\mathbf{x}_i,\mathbf{y}_i)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f(\mathbf{x}_i,\mathbf{y}_j)}}\right].$$

With InfoNCE and CLUB, we now have a way to minimise mutual information between variables and to approximate the range in which MI lies.

Partial Information Decomposition

Mutual information can be used to investigate how much random variables reveal about each other. In the multimodal case, this can be extended to include two or more variables, which can be investigated in terms of their information content about a target variable, thereby translating to the influence of variables on downstream task performance. For this, we can refer to the Partial Information Decomposition (PID) framework [WB10], which is used, for example, in neuroscience. Assume we are given two modalities as random variables, M_1 and M_2, and want to predict a random variable T based on these. Then we can quantify how much useful information M_1 and M_2 contain about T as $I(T; M_1, M_2)$. Applying this partial information decomposition, we semantically get [WB10]

$$I(T; M_1, M_2) = \text{Unq}(T; M_1) + \text{Unq}(T; M_2) + \text{Rdn}(T; M_1, M_2) + \text{Syn}(T; M_1, M_2),$$

which means that the contributions of both modalities can be decomposed into unique, redundant and synergistic information contributions. The unique information is only present in the respective modalities, redundant information exists in both modalities, and synergistic information jointly complements each other. They can be computed in the following way. First, we need to compute estimates of $I(T; M_1)$, $I(T; M_2)$ and $I(T; M_1, M_2)$. Using the framework from [WB10], which defines redundancy as the expected minimum information that any modality provides [1], the contributions can be derived as:

$$\begin{aligned}
&\text{Rdn}(T; M_1, M_2) = \sum_t p(t) min_i(I(T = t; M_i)) \quad \textbf{Definition by [WB10]} \\
&\text{Unq}(T; M_1) = I(T; M_1) - \text{Rdn}(T; M_1, M_2) && (3.1) \\
&\text{Unq}(T; M_2) = I(T; M_2) - \text{Rdn}(T; M_1, M_2) && (3.2) \\
&\text{Syn}(T; M_1, M_2) = I(T; M_1, M_2) - \text{Unq}(T; M_1) - \text{Unq}(T; M_2) - \text{Rdn}(T; M_1, M_2)
\end{aligned}$$

However, for high-dimensional data M_1, M_2, T, this becomes more challenging as we cannot directly estimate the mutual information, but instead need to approximate it using, for example, trained neural network estimators. These can, however, also become inaccurate for complex, high-dimensional data and compressing the data to a low-dimensional embedding would be necessary, which could alter results. This makes the PID a good theoretical framework, but it is often unsuited for real-life tasks involving complex data. In the following sections, we lay the foundations for the framework, which we will propose in chapter 4, that aims to estimate both unique and shared contributions. Our framework utilises a more practical and heuristic deep learning approach, making it easier to apply and yielding more practically usable contribution estimations.

3.2 Machine Learning Fundamentals

In recent decades, machine learning has established itself as an important tool for data analysis in scientific applications, including physics data. From computer vision applications, including segmentation and scene understanding, to data clustering, time series

[1] There are other definitions for redundancy which can lead to different derivations for the appearing terms.

analysis, and deep generative models for generating data, it has many applications. Deep neural networks have gained traction, with the introduction of more powerful architectures, such as transformers with their attention mechanism, diffusion models, or variational autoencoders. With increasingly powerful GPUs, the large-scale training of large models has become feasible. The general idea in machine learning is to learn patterns from data. There are three main types of learning: *supervised*, *unsupervised*, and *reinforcement* learning.

In the supervised case we are given a dataset $\{(\mathbf{x}_1, \mathbf{t}_1), ..., (\mathbf{x}_N, \mathbf{t}_N)\}$ of input data $\mathbf{x}_i$ and output/target values $\mathbf{t}_i$ and the goal is to learn a function $f : \mathbf{X} \rightarrow \mathbf{T}$ which minimizes a loss function $L(f(\mathbf{x}_i), \mathbf{t}_i)$ that measures the similarity of the predicted output with the actual output for all data pairs. If the values in $\mathbf{t}_i$ are continuous, this is referred to as regression; if they are discrete, this is referred to as classification. In unsupervised learning, we are only given input data points $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, and several common tasks are associated with it, including clustering similar data, learning a compressed representation of the data, or estimating the probability density of the underlying data. In this case, we can also train models to generate data. A special case of unsupervised learning is self-supervised learning, where auxiliary labels are generated from the input data points, which can be used for specific tasks such as reconstructing masked images or learning a meaningful representation of the data with contrastive learning. The third case is reinforcement learning, where an agent interacts with its environment and learns a policy to maximise the rewards from the environment.

This thesis focuses primarily on the unsupervised case and, to a lesser extent, on the supervised case. For that, we have to recap some necessities. Assuming the supervised case, our data points are often assumed to be generated by some underlying function f with an input value x_i and a corresponding target value $\mathbf{t}_i$ where $\mathbf{t}_i = f(\mathbf{x}_i) + \epsilon$ where ϵ is noise. The goal is to learn this function as $f(\mathbf{x}, \hat{\theta})$ such that $\hat{\theta} = \arg\min_\theta \sum_i L(f(\mathbf{x}_i, \theta), t_i)$. Here θ stands for the parametrisation of f which depends on the function and is, in the case of neural networks, the weights. The loss function $L(\mathbf{x}_i, \mathbf{t}_i)$ measures the similarity or error between the predicted target value and the true target value. In regression, the main task is to minimise the expected loss

$$\mathbb{E}[L] = \int \int L(f(\mathbf{x}, \theta), \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{3.3}$$

A commonly used measure to compute the difference between the predicted vector $\mathbf{t}_i$ and true target vector $\mathbf{x}_i$ is the $L2$ loss: $L(\mathbf{x}_i, \mathbf{t}_i) = \|\mathbf{x}_i - \mathbf{t}_i\|_2^2$. Then the function can be optimised by minimising the mean squared error (MSE) over all N samples:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \|f(\mathbf{x}_i, \theta) - \mathbf{t}_i\|_2^2$$

For general regression problems, the function f has often the form $f(\mathbf{x}_i, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \Phi_j(\mathbf{x}_i)$ for one data sample $\mathbf{x}_i \in \mathbb{R}^M$ and with the weight $\mathbf{w} \in \mathbb{R}^M$ and basis functions Φ_j. The first entry is often set to 0, $\Phi_0(\mathbf{x}_i) = 1$, such that w_0 is a bias term. The loss function is then minimised by gradient descent, which reduces the loss value by iteratively following the negative gradient in the parameter space. In deep learning, the function $f(\mathbf{x}_i, \mathbf{w})$ is replaced by a more complex function consisting of several nested layers in a multilayer perceptron (MLP) structure. In an MLP, each node computes a value dependent on the entire previous layer. The general structure is visualised in fig. 3.1.

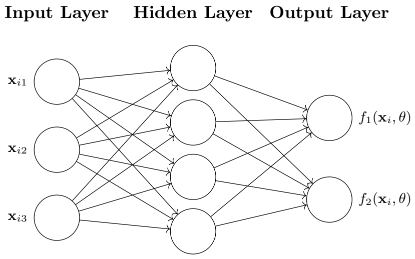


Figure 3.1: Visualisation of a Multilayer Perceptron (MLP) with one hidden layer.

The input vector $\mathbf{x}_i$ is first processed by the next (hidden) layer as a product of the input vector multiplied by a corresponding weight plus an additional bias. Then, a non-linear activation function Φ is applied to improve the expressivity of the network, as otherwise, the network could be reduced to a simple matrix-vector multiplication. The first layer is computed as

$$\mathbf{h}_{ik}(\mathbf{x}) = \Phi(\sum_{i=0}^{M-1} \mathbf{w}_{kj}\mathbf{x}_{ij}).$$

The next layer is then computed similarly, but with the output of the previous layer, $\mathbf{h}_i$, instead of $\mathbf{x}_i$. Multiple of these hidden layers can be concatenated before the last layer computes output values $f(\mathbf{x}_i, \theta)$. This general form of neural networks can be used to learn a function representing the relation from input $\mathbf{x}_i$ to target output $\mathbf{t}_i$. It can also be minimised by gradient descent. The backpropagation algorithm implements the gradient descent method for neural networks. In current frameworks, the gradient descent method is implemented efficiently as reverse mode automatic differentiation.

Convolutional Neural Networks

For structured data, such as images or spectra, using an inherent grid structure, a different architecture than MLPs is more effective. In computer vision and machine learning, it is desired to extract meaningful features from images and data. Early approaches applied hand-engineered filters/convolutions to images to extract features such as edges and corners. Convolutional neural networks perform the same function but are learned directly by the machine learning model.

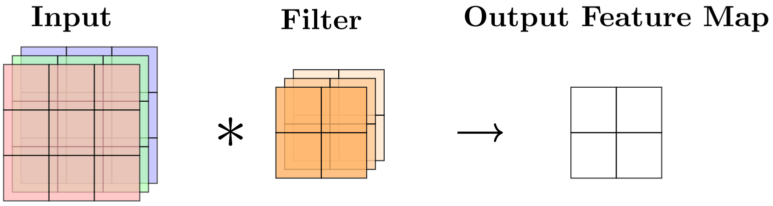


Figure 3.2: Visualisation of a convolution operation.

The convolution operation is visualised in fig. 3.2, where for an RGB input image with three channels, a filter is applied to each corresponding patch and channel in the input image, and the results are added to create an output feature map $I * F = I'$. As the filters are reused across the entire image, this drastically reduces the parameters the model has to learn. This operation can be utilised in a neural network and can be completed with batch norm layers, pooling and non-linear activation functions. To extract more complex patterns/features, several of these convolution layers are combined. Typical CNN models reduce the spatial size with every layer while increasing the number of channels (channel inflation). Another closely linked layer is the transposed convolution, which can increase spatial size and is often applied in generative models. Convolutional layers also exist for 1-D, 2-D, or 3-D shaped data.

3.3 Generative Models

In this section, we describe how generative models, such as the Variational Autoencoder, work, building up the ideas for it step by step. When we want to generate data items from a data distribution, it means that we want to draw samples from the underlying real data distribution. If the underlying data distribution is $p(\mathbf{x})$, then we want to draw $\mathbf{x} \sim p(\mathbf{x})$. This probability distribution, however, is often unknown or intractable in closed form. To still perform this inference task, there are two primary methods: Sampling methods and Variational Inference.

Sampling methods rely on samples to estimate the target distribution and to estimate statistical properties based on these samples. A popular sampling method is MCMC, which produces a sequence of dependent samples using an iterative scheme with a Markov chain whose stationary distribution is the target distribution. By drawing every n-th sample, the samples become approximately independent. Sampling becomes more accurate with a large number of samples, which can make the process computationally expensive.

In contrast, variational methods attempt to directly approximate the intractable posterior distributions with a simpler distribution that can be optimised to be close to the actual distribution. These simpler distributions, unlike the real distribution, are tractable to optimise to approximate the real underlying distribution, while they can also introduce a bias. Variational methods are often more efficient than sampling, but their effectiveness depends on the task and must be derived. This is the approach used for VAEs.

3.3.1 Autoencoder

Consider we are given a data item $\mathbf{x}$ from which we want to extract important features in a compressed format. This can be achieved through unsupervised learning, where an encoder $f(\mathbf{x})$ takes the input and outputs a lower-dimensional representation $\mathbf{z} = f(\mathbf{x})$. Then, a decoder $g(\mathbf{z})$ has the task of reconstructing the original data item x using the compressed representation as input, where $g(\mathbf{z}) = \hat{\mathbf{x}} \approx \mathbf{x}$. This model can be trained by minimising the mean square error with respect to the encoder and decoder parameters

$$L = \frac{1}{N} \sum_{i=1}^{N} \|x - g(f(\mathbf{x}))\|_2^2.$$

This task has a bottleneck due to the compressed representation in the so-called *latent space*. Since the latent space often has a far lower dimension than the original data item, the model must learn the most important features of the data, which are most informative for reconstruction. Suppose the model learns to extract such features into the latent space and is capable of reconstructing the original data. In that case, it is in principle possible to generate new data items if it were possible to sample meaningful latent vectors. However, as the latent space is not regularised, it has an irregular format from which we cannot just draw latent values. Instead, a more advanced probabilistic model is needed, for which latent values can be sampled to generate novel data.

3.3.2 Variational Autoencoder

The variational autoencoder [KW+13] utilises the structure of the autoencoder, which can generate data points based on a latent vector, but extends it probabilistically. Sampling from the latent manifold directly is not possible for autoencoders because the latent distribution is unknown. If we could sample from the true underlying latent distribution $p_{\theta^*}(\mathbf{z})$ and then sample from the true conditional distribution $p_{\theta^*}(\mathbf{x}|\mathbf{z})$, we could generate novel data points [Lei24]. For VAEs, we can choose $p_{\theta^*}(\mathbf{z})$ to be a standard zero-mean, unit-covariance Gaussian distribution, and $p_{\theta^*}(\mathbf{x}|\mathbf{z})$ can be represented as a neural network. Then, this scenario corresponds to the task of maximising the likelihood [LP21]

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z} = \mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z})}[p_\theta(\mathbf{x}|\mathbf{z})].$$

Despite $p_\theta(\mathbf{z})$ and $p_\theta(\mathbf{x}|\mathbf{z})$ being computable as a Gaussian and a neural network, computing $p_\theta(\mathbf{x})$ requires marginalising over all $\mathbf{z}$, which is analytically intractable [Lei24]. Variational autoencoders address this dilemma by employing variational inference, which involves deriving an evidence lower bound (ELBO). By maximising this ELBO, we get another way to maximise the likelihood in a tractable form. For this, another neural network $q_\Phi(\mathbf{z}|\mathbf{x})$ is introduced, which acts as an encoder and approximates the posterior $p_\theta(\mathbf{z}|\mathbf{x})$. The encoder and decoder are probabilistic. They predict a Gaussian distribution by predicting mean $\mu_{\mathbf{z}|\mathbf{x}}$ and (co-) variance $\sigma_{\mathbf{z}|\mathbf{x}}$ for the encoder and by predicting mean $\mu_{\mathbf{x}|\mathbf{z}}$ and (co-) variance $\sigma_{\mathbf{x}|\mathbf{z}}$ for the decoder, from which we can sample. Then the latent variable $\mathbf{z}$ can be sampled as $\mathbf{z}|\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\mu_{\mathbf{z}|\mathbf{x}}, \sigma^2_{\mathbf{z}|\mathbf{x}}I)$ and the reconstructed data point $\mathbf{x}$ can be sampled as $\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{x}|\mu_{\mathbf{x}|\mathbf{z}}, \sigma^2_{\mathbf{x}|\mathbf{z}}I)$. The goal is now to maximize $\mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|(\mathbf{x})}[p_\theta(\mathbf{x}|\mathbf{z})]$ instead of $\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z})}[p_\theta(\mathbf{x}|\mathbf{z})]$. We follow the steps shown in [Lei24]. To simplify calculations, the log-likelihood is used:

$$\begin{aligned}
\log p_\theta(\mathbf{x}_i) &= \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i)] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(\mathbf{x}_i|\mathbf{z})p_\theta(\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x}_i)}\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{p_\theta(\mathbf{x}_i|\mathbf{z})p_\theta(\mathbf{z})q_\Phi(\mathbf{z}|\mathbf{x}_i)}{p_\theta(\mathbf{z}|\mathbf{x}_i)q_\Phi(\mathbf{z}|\mathbf{x}_i)}\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{q_\Phi(\mathbf{z}|\mathbf{x}_i)}{p_\theta(\mathbf{z})}\right] + \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{q_\Phi(\mathbf{z}|\mathbf{x}_i)}{p_\theta(\mathbf{z}|\mathbf{x}_i)}\right]
\end{aligned}$$

Here, the KL-divergence naturally appears. Hence, we get

$$\log p_\theta(\mathbf{x}_i) = \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z})] - D_{KL}(q_\Phi(\mathbf{z}|\mathbf{x}_i)\|p_\theta(\mathbf{z})) + D_{KL}(q_\Phi(\mathbf{z}|\mathbf{x}_i)\|p_\theta(\mathbf{z}|\mathbf{x}_i)). \tag{3.4}$$

The first term is responsible for reconstructing the data. As the decoder predicts a Gaussian distribution, the expectation value can be predicted as

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}\left[\log p_\theta(\mathbf{x}_i|\mathbf{z})\right] &= \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}[\log \mathcal{N}(\mathbf{x}_i|\mu_{\mathbf{x}_i|\mathbf{z}}, \sigma^2_{\mathbf{x}_i|\mathbf{z}}I)] \\ &= -\frac{1}{2}\sum_{j=1}^{d}\mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}\left[\log(2\pi\sigma^2_{\mathbf{x}_i|\mathbf{z},j}) + \frac{(\mathbf{x}_{i,j}-\mu_{\mathbf{x}_i|\mathbf{z},j})^2}{\sigma^2_{\mathbf{x}_i|\mathbf{z},j}}\right].\end{aligned}$$

In practice the covariance $\sigma_{\mathbf{x}|\mathbf{z}}$ term is often dropped. Then the expectation value can be approximately minimised by minimising the mean squared error $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mu_{\mathbf{x}_i|\mathbf{z}}\|_2^2$. Since the $\mathbf{z}$ values are sampled, it normally wouldn't be possible to do gradient descent through the latent space to the encoder. However, we can utilise the reparameterization trick [KW$^+$13] to obtain differentiable samples. This is done by defining the random variable $\mathbf{z}$ as

$$z_i = \mu_{\mathbf{z}|\mathbf{x},i} + \sigma_{\mathbf{z}|\mathbf{x},i}\epsilon_i$$

using $\epsilon_i \sim \mathbb{N}(0,1)$. Then this becomes differentiable. The second term in eq. (3.4) can be solved in a closed form as the prior is a standard Gaussian, and the encoder gives the other distribution. The last term in eq. (3.4) is analytically intractable. Since $D_{KL}(P\|Q) \geq 0$, the term can be neglected and we get a lower bound on $\log p_\Theta(x)$. Hence:

$$\log p_\Theta(\mathbf{x}_i) \geq \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z})] - D_{KL}(q_\Phi(\mathbf{z}|\mathbf{x}_i)\|p_\theta(\mathbf{z})) =: ELBO(\mathbf{x}_i, \theta, \Phi)$$

The first term is for reconstruction, and the second term conditions the encoder to learn a latent space that is close to the prior, which is a standard Gaussian. In summary, the optimisation objective is to maximise the log-likelihood by maximising the lower bound [Lei24]

$$\begin{aligned}\log p_\theta(\mathbf{x}_i) &\geq \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x})}[ELBO(\mathbf{x}_i, \theta, \Phi)], \\ \theta, \Phi &= \arg\max_{\theta,\Phi}\sum_{i=1}^{N} ELBO(\mathbf{x}_i, \theta, \Phi).\end{aligned}$$

The optimisation thus faces a tradeoff of reconstruction performance and closeness of the latent space to a standard Gaussian. Then the VAE learns a latent representation of the data distribution mapping a data item $\mathbf{x}$ to a latent representation consisting of a mean $\mu_{\mathbf{z}|\mathbf{x}}$ and the inherent uncertainty represented by the variance $\sigma^2_{\mathbf{z}|\mathbf{x}}$. The encoder essentially learns an approximate posterior distribution $q_\Phi(\mathbf{z}|\mathbf{x})$. This latent representation captures the essential underlying factors of the data and can be seen as a compression of the data. The mean $\mu_{\mathbf{z}|\mathbf{x}}$ can here be seen as a maximum a posteriori (MAP) estimate of the approximate posterior distribution and is in practice often used as the deterministic, most likely encoding of the input $\mathbf{x}$ in latent space, that is usable for downstream tasks and as a representative embedding.

3.3.3 Disentangled Representation Learning

To investigate downstream task performance, it is necessary to extract a meaningful and robust representation of the data on which downstream models can be applied. *Disentangled representation learning* (DRL) is a paradigm for learning to separate the distinct, independent, and informative generative factors of variation in the data [WCT$^+$24]. There are several categories of models that can perform this task, including VAE-based methods. As previously described, VAEs can be used to extract essential features of data. To adjust the amount of disentanglement between the latent features, the $\beta - \mathrm{VAE}$ extends the normal VAE with a weighting factor β on the KL-divergence, which controls how much the variables shall be disentangled [WCT^{+}24]. This affects whether disentanglement is favoured over reconstruction capabilities.

$$ELBO(\mathbf{x}_i, \theta, \Phi) = \mathbb{E}_{\mathbf{z} \sim q_\Phi(\mathbf{z}|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z})] - \beta D_{KL}(q_\Phi(\mathbf{z}|\mathbf{x}_i) \| p_\theta(\mathbf{z}))$$

The $\beta - \mathrm{TCVAE}$ extends this even further by decomposing the KL-divergence into three terms [WCT^{+}24]:

$$\begin{aligned} D_{KL}(q_\Phi(\mathbf{z}|\mathbf{x}_i) \| p_\theta(\mathbf{z})) &= D_{KL}(q_\Phi(\mathbf{z}, \mathbf{x}_i) \| q_\Phi(\mathbf{z}) p_\theta(\mathbf{x}_i)) && \textbf{Mutual Information} \\ &+ D_{KL}(q_\Phi(\mathbf{z}) \| \prod_j q_\Phi(z_j)) && \textbf{Total Correlation} \\ &+ \sum_j D_{KL}(q_\Phi(z_j) \| p_\theta(z_j)) && \textbf{Dimension-wise KL-divergence} \end{aligned} \tag{3.5}$$

For each of the three terms, the model applies a penalty factor β_i for $i \in \{1, 2, 3\}$, giving more control over what should be penalised [WCT^{+}24]. This reveals that a higher penalty β_{KL} on the KL-divergence leads to an increased penalty on the mutual information between the data $\mathbf{x}$ and the corresponding latent representation $\mathbf{z}$, which harms reconstruction performance and feature extraction capabilities as it reduces the dependence of $\mathbf{z}$ on $\mathbf{x}$ while making $\mathbf{z}$ more independent. To retrieve an informative latent space, this is not necessarily desired, as we want to capture as much information as possible. Thus, we need to reduce β_{KL} accordingly. This is the model that we will later use for the VAE due to its increased control capabilities.

Chapter 4

General Framework and Methods

Here, we propose a framework that enables the investigation of unique and shared contributions of multimodal data to a prediction task. For this, one can examine the problem from an abstract perspective. Without loss of generality, two modalities are assumed. These modalities together contain information that is *shared* (redundant or synergistic) between both modalities, and they both contain *unique* information that is not captured by the other modality. When these modalities are combined and used together for a given prediction task, all three distinct information subsets are implicitly jointly used for the task. However, not all information captured by each of these subsets can be used for the given task. This is visualised in fig. 4.1.

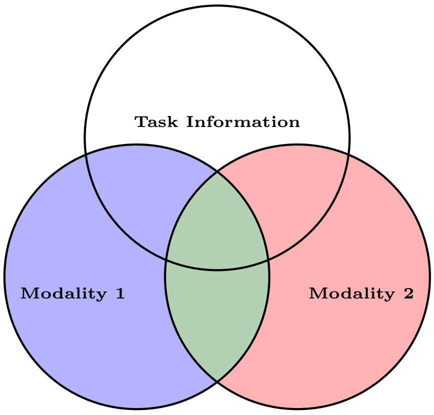


Figure 4.1: Visualisation of information overlap of private and shared information of two modalities and the information required for a task.

The diagram displays the information from both modalities, where blue represents the unique information from modality 1, red represents the unique information from modality 2, and green represents the shared information. At the top, you can see all the information that can be used for a given prediction task. The combined information from both modalities potentially only partially contains useful information for the task. It depends on the task and the modality how much overlap in information there is. The coloured regions only partially intersect the task information, as the modalities can also contain information that is not relevant to the task. Additionally, other information, not captured by the modalities, could also be used for the prediction task. When evaluating whether adding more data modalities increases task performance, one must consider the relative contributions of each modality, broken down into its private and shared components. Specifically, the unique contribution and the added synergistic information could influence the task performance. To extract these disentangled representations of shared and private information, we first need to explain the Disentangled Multimodal Variational Autoencoder, which serves as a base model.

4.1 Disentangled Multimodal Variational Auto-encoder

The Disentangled Multimodal Variational Autoencoder (DMVAE) [LP21] extends the Variational Autoencoder to a multimodal model. The goal is to obtain separate latent spaces that contain features shared between multiple modalities and private features that contain unique features of each modality. To do this, an unsupervised learning task is set up with one en-/decoder for each modality. Every Encoder i has the task to predict latent values $\mathbf{z}_i$ that contain modality-specific private features $\mathbf{z}_{p_i}$ and inter-modality shared features $\mathbf{z}_{s_i}$. In this thesis, only two modalities are assumed, and the inclusion of more modalities leads to more complicated shared features. This means that encoders 1 and 2 predict

$$q_{\Phi_1}(\mathbf{z}_1 \mid \mathbf{x}_1) \sim \mathbf{z}_1 = (\mathbf{z}_{p_1}, \mathbf{z}_{s_1}), \quad q_{\Phi_2}(\mathbf{z}_2 \mid \mathbf{x}_2) \sim \mathbf{z}_2 = (\mathbf{z}_{p_2}, \mathbf{z}_{s_2}),$$

where the models should learn to predict

$$\mathbf{z}_{s_1} = \mathbf{z}_{s_2} = \mathbf{z}_s.$$

The shared features are then combined using a product of experts (PoE). PoE is a method for combining probability distributions from multiple experts into a single joint probability distribution that captures more complex dependencies. It is constructed as a product of all expert distributions $p_i(\mathbf{z})$ normalised by some partition function. For N experts, it is defined as [LP21]

$$q(\mathbf{z}_s|\mathbf{x}_1, ..., \mathbf{x}_N) \propto p(\mathbf{z}_s) \prod_{i=1}^{N} q(\mathbf{z}_s|\mathbf{x}_i).$$

Each expert $q(\mathbf{z}_s|\mathbf{x}_i)$ may encode different constraints about the latent variable $\mathbf{z}_s$. By multiplying the experts, the resulting PoE distribution assigns high probability to those $\mathbf{z}_s$ values that all experts support. This is especially useful in multimodal learning as different views of multiple modalities on the latent variable $\mathbf{z}_s$ provide a sharper, more constrained posterior. For the prior, a Gaussian is assumed $p(\mathbf{z}_s) = \mathcal{N}(\mathbf{z}_s|\mathbf{0}, I)$ and the experts are also Gaussian $q(\mathbf{z}_s|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}_s|\mu_i, \mathbf{C}_i)$ with mean μ_i and covariance $\mathbf{C}_i$ [LP21]. Hence the resulting distribution of the PoE in closed form is $q(\mathbf{z}_s|\mathbf{x}_1, ..., \mathbf{x}_N) = \mathcal{N}(\mathbf{z}_s|\mu, \mathbf{C})$ [LP21] with

$$\mathbf{C}^{-1} = \sum_{i=1}^{N} \mathbf{C}_i^{-1}, \quad \mu = \mathbf{C} \sum_{i=1}^{N} \mathbf{C}_i^{-1} \mu_i.$$

The decoders have the task of reconstructing their respective modality based on samples from the shared latent space and their corresponding private latent space. This design has the advantage that the shared latent variable $\mathbf{z}_s$ can also be inferred when just one modality $\mathbf{x}_i$ is given $q(\mathbf{z}_s|\mathbf{x}_i)$, enabling cross-reconstruction capabilities together with samples from a normal distribution resembling the unknown private latent vector [LP21]. Assuming the two-modality case, the learning objective can now be defined similarly to that of VAEs, consisting of multiple reconstruction and KL-divergence terms:

$$\begin{aligned} ELBO_{DMVAE} := \sum_{i\in\{1,2\}} \mathbb{E}_{x_i\sim p(x_i)} \Bigg[& \lambda_i \mathbb{E}_{q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i), q_\Phi(\mathbf{z}_s|\mathbf{x}_1,\mathbf{x}_2)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}_{p_i},\mathbf{z}_s)] \\ & - \beta_i KL(q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i) q_\Phi(\mathbf{z}_s|\mathbf{x}_1,\mathbf{x}_2) \| p(\mathbf{z}_{p_i}) p(\mathbf{z}_s)) \\ & + \sum_{j\in\{1,2\}} \Bigg(\lambda_i \mathbb{E}_{q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i), q_\Phi(\mathbf{z}_s|\mathbf{x}_j)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}_{p_i},\mathbf{z}_s)] \\ & - \beta_i KL(q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i) q_\Phi(\mathbf{z}_s|\mathbf{x}_j) \| p(\mathbf{z}_{p_i}) p(\mathbf{z}_{s_i})) \Bigg) \Bigg] \end{aligned} \tag{4.1}$$

This objective contains six terms: For each combination of latent spaces that the decoder can receive, there is a reconstruction and a KL-divergence term, such that each combination resembles a normal distribution from which we can reconstruct respective modalities. For each modality, there are three combinations which are $(\mathbf{z}_{p_i}, \mathbf{z}_{s_i})$ for direct reconstruction, $(\mathbf{z}_{p_i}, \mathbf{z}_s)$ for joint reconstruction and $(\mathbf{z}_{p_i}, \mathbf{z}_{s_j})$ for cross-reconstruction for each combination of modalities $i, j \in \{1, 2\}$ and $i \neq j$. We now describe their corresponding reconstruction and KL-divergence terms:

1. The first terms are for the reconstruction using the combined shared representation from the PoE and the respective private representation from the encoder i. These are then used by the respective decoder:

   $$\mathbb{E}_{q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i), q_\Phi(\mathbf{z}_s|\mathbf{x}_1,\mathbf{x}_2)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}_{p_i},\mathbf{z}_s)] \text{ and } KL(q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i) q_\Phi(\mathbf{z}_s|\mathbf{x}_1,\mathbf{x}_2) \| p(\mathbf{z}_{p_i}) p(\mathbf{z}_s))$$

2. Then there are terms corresponding to direct reconstruction using the private and shared latent values inferred from the corresponding encoder i which are utilised by the corresponding decoder:

   $$\mathbb{E}_{q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i), q_\Phi(\mathbf{z}_s|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}_{p_i},\mathbf{z}_s)] \text{ and } KL(q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i) q_\Phi(\mathbf{z}_s|\mathbf{x}_i) \| p(\mathbf{z}_{p_i}) p(\mathbf{z}_{s_i}))$$

3. At last, there are terms to enable cross-reconstruction through the use of the shared representation inferred from modality j with the private representation inferred from modality i, which are used by the corresponding decoder i. In practice the cross-reconstruction can be achieved by sampling $\mathbf{z}_j \sim p(\mathbf{z}_{p_j})$ from a normal distribution:

   $$\mathbb{E}_{q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i), q_\Phi(\mathbf{z}_s|\mathbf{x}_j)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}_{p_i},\mathbf{z}_s)] \text{ and } KL(q_\Phi(\mathbf{z}_{p_i}|\mathbf{x}_i) q_\Phi(\mathbf{z}_s|\mathbf{x}_j) \| p(\mathbf{z}_{p_i}) p(\mathbf{z}_{s_i}))$$

By training the DMVAE on this objective, the model is encouraged to extract shared features, as it needs to utilise them for improved (cross-) reconstruction performance using either the shared representation inferred by the corresponding modality, the other modality, or the joint PoE representation, which by design only captures shared information. Everything that cannot be captured in the shared representation should be put into the private latent spaces. Then, the model learns unique features in the private latent spaces, redundant features in the shared latent space, and synergistic information is only indirectly considered through the PoE, which could extract synergistic information into the shared latent space; however, this is not enforced.

4.2 DMVAE with CLUB

Based on the DMVAE and CLUB, we propose a framework for evaluating multimodal data. The DMVAE serves as a basis, with corresponding encoders and decoders responsible for feature extraction and reconstruction for each modality. The DMVAE already encourages that encoders compute private and shared features that are disentangled from one another. However, it does not directly ensure that the private and shared latent spaces contain no mutual information and no semantic dependencies. Consequently, semantic dependencies, redundancies, or leakage between latent spaces can occur, which is undesirable for evaluating contributions. Our goal is to minimise the mutual information between the shared and private latent spaces $I(z_s; z_{p1})$ and $I(z_s; z_{p2})$ [1]. To address this, we use a differentiable approximation of mutual information, employing variational CLUB, because the mentioned conditional distributions from shared to private latent spaces are unknown. This CLUB loss is incorporated between the shared and private latent spaces. The architecture is visualised in fig. 4.2.

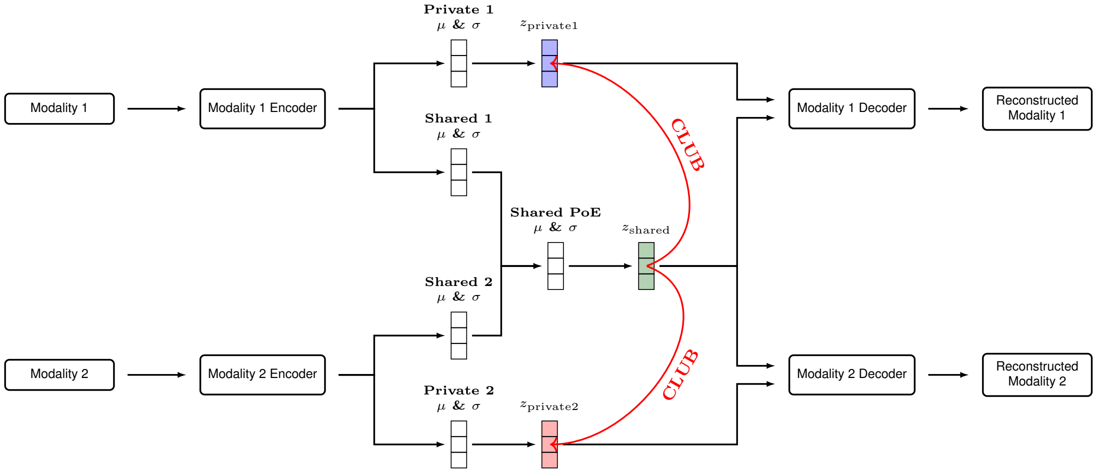


Figure 4.2: Architecture of the Disentangled Multimodal Variational Autoencoder (DMVAE) with added CLUB loss.

In doing so, we ensure that the mutual information between the shared and private latent spaces is minimised, thereby maintaining a clear separation between the shared and private information. The private latent spaces are assumed to be independent by construction, as they do not interact with each other. Hence, it is expected that the mutual

[1] We assume that the shared and unique information can be disentangled with as little MI as possible, although this can sometimes be challenging in practice, when there are causal relations between them, which may lead shared features to contain unintuitive information.

information between the private latent spaces is by default very low, which is why we do not use an extra CLUB loss between the private latent spaces. Furthermore, suppose any mutual information exists between private latent spaces. In that case, the model should naturally incorporate it into the shared latent space, thereby directly reducing the mutual information between the private latent spaces. Our framework now allows us to study the shared information across multiple data modalities and the information that is unique to each modality. Note that this idea of removing redundant information from the feature representation of a modality to obtain a unique representation of the modality is similar to that of PID, as seen in eq. (3.1) and eq. (3.2), but not in relation to a specific target variable here. Based on eq. (4.1), the learning objective is then to minimise the following loss, with λ_{CLUB} for weighting the influence of CLUB:

$$-ELBO_{DMVAE} + \lambda_{CLUB}\hat{I}_{vCLUB}(z_s; p_1) + \lambda_{CLUB}\hat{I}_{vCLUB}(z_s; p_2)$$

This learning objective then has the three main components: total reconstruction loss ($R_{total,i}$) that contains all three reconstruction terms for modality i, total KL-divergence ($KL_{total,i}$) that contains all three KL terms for modality i and the corresponding CLUB loss in the latent space:

$$\sum_{i=1}^{2} \lambda_i R_{total,i} + \beta_{KL} KL_{total,i} + \lambda_{CLUB}\hat{I}_{vCLUB}(z_s; p_i)$$

4.3 Downstream task

This DMVAE with CLUB loss framework is now capable of strictly disentangling modality-specific and shared features, which can be utilised for downstream tasks. When applying a downstream model to all latent spaces (mean values), we can evaluate the performance of the multimodal data on some downstream task. We can now directly investigate the influence of each latent space by training a downstream model on all non-empty combinations of latent spaces $\mathcal{Z} = \{Z_{p1}, Z_{p2}, Z_s\}$, evaluating each element in the set $C = \mathcal{P}(\mathcal{Z})\backslash\emptyset$. This **subset-based method** offers multiple ways for evaluating downstream contributions. By examining the performance drop when excluding one latent space compared to the entire latent space, it can be observed how much this latent space contributed to the task and how well it complemented the other latent spaces. Additionally, observing the performance of single latent spaces reveals their direct performance, without synergistic effects between the different latent spaces on the downstream task.

In addition to this method, we now describe another **SHAP-based method** that also aims to estimate contributions while potentially saving compute, as we only need to train the downstream model on all latent spaces once [2].

Task Contribution Estimation using SHAP

SHapley Additive exPlanation (SHAP) [LL17] is a method targeting the explainability of machine learning model predictions. It quantifies the contribution of each input feature

[2] The subset based method is applicable to more general downstream tasks that can also include tasks like reconstruction, while the SHAP based method can only be applied to predictions about single parameters.

to the model's output prediction. Each feature value can be either positive, indicating a positive impact on a prediction, or negative, indicating a negative impact on a prediction. The SHAP value is a *local method* for explaining the feature contributions of a single input $\mathbf{x}$ to model predictions. For each feature i the SHAP value can be computed as [ICT23]

$$\Phi_i(\mathbf{x}) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{\mathbf{x}}(S \cup \{i\}) - f_{\mathbf{x}}(S)].$$

Here F is the set of all features and $f_{\mathbf{x}}(S)$ is the prediction of a regression model that is only evaluated on the input features S while $f_{\mathbf{x}}(S \cup \{i\})$ is the prediction of a regression model evaluated on the input features $S \cup \{i\}$ [LL17] [ICT23]. By computing the difference between these two regression models for all subsets of input features and weighting the terms in a way that each difference makes a fair contribution, we can calculate SHAP values for each feature that comprehensively represent its contribution to the output. In practice, the effect of the absence of features on the regression model is estimated, for example, over samples [LL17]. This can now be applied to a downstream task in our framework, where we can measure the contribution of each latent variable to the downstream prediction. This can now be used to estimate contributions of each latent space in the following way. SHAP has the property of being an *additive feature attribution method* to the original prediction model f with an explanation model g that approximates f

$$g(z') = \Phi_0 + \sum_{j=1}^{M} \Phi_j z'_j,$$

which essentially means that the sum of the appearing SHAP contributions plus a baseline can explain the model's predictions. Whether a simplified input feature j appears and has a contribution is indicated by the boolean variable z_j. For more details, we refer to [LL17]. This sum can be decomposed into three sums corresponding to the three latent spaces, private 1, private 2 and shared, where each sum then approximates the contribution of the corresponding latent space. By taking the absolute values of each SHAP value and then summing them, we can quantify the absolute contribution of each latent space to the prediction. We take the absolute values as we are concerned only with the magnitude of its influence, regardless of whether the features positively or negatively impact the prediction. Assuming three latent sizes of size P_1, P_2, S and the corresponding SHAP values $\Phi^{(i)}_{p1,j}, \Phi^{(i)}_{p2,j}, \Phi^{(i)}_{s,j}$ for sample i and feature j, we compute the latent spaces' absolute contributions over the entire dataset of size N by

$$\Phi_{p1} = \sum_{i=1}^{N} \sum_{j=1}^{P1} |\Phi^{(i)}_{p1,j}|, \quad \Phi_{p2} = \sum_{i=1}^{N} \sum_{j=1}^{P2} |\Phi^{(i)}_{p2,j}|, \quad \Phi_{s} = \sum_{i=1}^{N} \sum_{j=1}^{S} |\Phi^{(i)}_{s,j}|,$$

and then we can compute their relative contribution by

$$C_{p1} = \frac{\Phi_{p1}}{\Phi_{p1} + \Phi_{p2} + \Phi_s}, \quad C_{p2} = \frac{\Phi_{p2}}{\Phi_{p1} + \Phi_{p2} + \Phi_s}, \quad C_s = \frac{\Phi_s}{\Phi_{p1} + \Phi_{p2} + \Phi_s}.$$

This approach is valid because the private and shared representations have by construction minimal dependencies. Recall that SHAP is a local explanation method for single data inputs, and by computing the SHAP contribution over the entire dataset, we assume that this approximates the global contributions of each latent variable. This assumption may not always hold, as SHAP does not guarantee that it generalises over the entire dataset. However, since we use a simple downstream model, we assume that contributions can be computed on the whole dataset.

Note that the SHAP-based method estimates contributions based on performance drops when leaving out one feature for all combinations of latent values as input for a regression model. Meanwhile, our first approach can only demonstrate a performance drop when entire latent spaces are excluded. Hence, both methods could differ in their contribution estimation, which should be compensated for by the fact that the latent spaces have minimal MI, reducing synergistic effects between individual values from different latent spaces.

Chapter 5

Experiments and Evaluation

This chapter describes the experiments designed to study the impact of multimodal data on downstream tasks for physics data. We describe the experimental setup, the motivation behind it and evaluate the results of the experiments using several metrics.

These experiments study the proposed framework and investigate the impact of including CLUB loss on the model's performance, as well as investigate physics data. Using images and spectra, we aim to examine the amount of information these modalities contain about each other using their cross-reconstruction performance and the unique and shared information they provide about underlying physical properties. To achieve this, we first need to identify and fix suitable hyperparameters for the KL-divergence weight and the CLUB loss weight, and then determine an appropriate learning rate schedule for training. We also need to determine suitable latent sizes for private and shared features that align with the modalities. This also helps to study the impact of the CLUB loss. We also directly compare DMVAEs with CLUB to DMVAEs without CLUB to see the effect of the CLUB loss, and compare them to VAEs on single-modal data. We also evaluate the structure of the latent space as well as the mutual information between latent spaces by computing lower and upper bounds. We examine where exactly relevant information for the downstream tasks is stored.

The experiments are structured as follows: We first conduct five experiments on galaxy image and spectral data from the Multimodal Universe dataset [AAB$^+$24] to investigate their usability for predicting physical properties. Here, we also investigate how the inclusion of a physical model can predict interpretable parameters and how this impacts the performance of downstream tasks. In the sixth experiment, we investigate hyperspectral data from the HyPlant FLUO [SAC$^+$19] dataset, which is remote sensing data recorded by airborne sensors. We decompose this data into image and spectrum to apply our framework to with hyperspectral reconstruction as a downstream task. This helps to determine whether the structural RGB image or spectrum contains more underlying information about the hyperspectral data. We first introduce the evaluation metrics used and present the primary dataset we aim to investigate, along with some background and details on the training and architectures used, before proceeding to the experiments.

5.1 Evaluation Tools and Metrics

Several evaluation tools are required to evaluate the model's performance. To study the reconstruction performance of the modalities, assess downstream task performance, and investigate latent space properties, several evaluation metrics are necessary to test the performance of different parameter configurations. When minimising an optimisation function that consists of several components, such as MSE for reconstruction performance, KL-divergence, and CLUB, we can directly use these components to evaluate the model's performance and latent properties. Still, more metrics are needed to assess the model's performance under various aspects. We use three metrics for evaluating the reconstruction performance compared to the original data item: Mean Squared Error (MSE), Structural Similarity Index (SSIM) and the Fréchet Inception Distance (FID). Note that there are also metrics specifically designed to evaluate the generative performance of galaxy images. These can be found in [FIL+13] but are not further studied here. To evaluate the structure and downstream-task performance of the latent space, we use t-SNE and R^2.

Mean Squared Error (MSE)

The MSE metric is a classic method that compares the reconstructed data point by point with the original data item. For an image x and its reconstruction x' this is defined as [TCERP+23]

$$MSE = \frac{1}{HWC} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} (x_{c,i,j} - x'_{c,i,j})^2,$$

averaged across all samples. In the optimal case where the model perfectly reconstructs the data item, this becomes 0.

Structural Similarity Index (SSIM)

Instead of relying on per-pixel accuracy, SSIM utilises local perceptual features to compare the similarity between two images. This metric is specifically designed for images and better reflects how humans perceive them. Specifically, it compares for each image region luminance, corresponding to the mean intensity (μ_x, μ_y), contrast, corresponding to the standard deviation (σ_x, σ_y) of the intensity, and structure, corresponding to the normalized covariance σ_{xy} between the two images x and y. It is defined as [TCERP+23]

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

Here, $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$ are just stabilising constants where L is the dynamic value range and K is some small constant. This metric is computed for all patches of the image and then averaged across them. It returns values between -1 and 1, but typically only between 0 and 1, where 1 indicates perfect similarity and 0 indicates no similarity. This can help estimate whether the structure of the images is similar. This metric is helpful in generative tasks, as small changes in brightness or global intensity are not penalised too much by the metric, better aligning with VAEs, as it is expected that they return non-pixel-perfect reconstructions but a distribution of similar images.

Fréchet Inception Distance (FID)

The FID metric is used for reviewing the reconstruction performance of the entire reconstructed image distribution compared to the original image distribution. It compares deep perceptual feature distributions. For the original and the reconstructed image, inception networks are applied, and features at some specific layer are extracted [TCERP+23]

$$\mathbf{x}' = Inception(\mathbf{x}), \quad \hat{\mathbf{x}}' = Inception(\hat{\mathbf{x}}).$$

Here $\mathbf{x}$ is the original image and $\hat{\mathbf{x}}$ is the reconstructed image. Assuming the features follow a multivariate Gaussian distribution with the means $\mu_{\mathbf{x}}$, $\mu_{\hat{\mathbf{x}}}$ and covariances $\mathbf{\Sigma}_{\mathbf{x}}$, $\mathbf{\Sigma}_{\hat{\mathbf{x}}}$ of the respective features from the original and reconstructed data, then, FID is defined as [TCERP+23]

$$\text{FID}(\mathbf{x}', \hat{\mathbf{x}}') = ||\mu_{\mathbf{x}} - \mu_{\hat{\mathbf{x}}}||^2 + Tr(\Sigma_{\mathbf{x}} + \Sigma_{\hat{\mathbf{x}}} - 2(\Sigma_{\mathbf{x}}\Sigma_{\hat{\mathbf{x}}})^{1/2}),$$

providing a metric that evaluates the similarity of the entire distribution of reconstructed images to the original ones globally, while also capturing better how humans perceive them [TCERP+23].

Coefficient of Determination (R^2)

A standard metric for evaluating the performance of a regression model is MSE. Additionally, another popular metric is the coefficient of determination, R^2. This metric can describe how well the model captures the variation of the target variable [TCERP+23]. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

with the mean $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ which is used to comupte the total variance of y and $\hat{y}_i$ representing the reconstructed data y_i. If the value is 1, then the model is perfect, and if it is 0, then it is as good as a mean predictor [TCERP+23].

t-Distributed Stochastic Neighbour Embedding (t-SNE)

The t-distributed stochastic neighbour embedding (t-SNE) can be used to reduce high-dimensional datapoints to a two or three-dimensional datapoint representation for illustration purposes [MH08]. It serves as a probabilistic, non-linear dimension reduction method that maps similar points close together and dissimilar points far apart from each other. It does this by computing the local similarity for all pairs of data points. Then it assigns points to the low-dimensional space while measuring their similarity using a specific measure. Then, the low-dimensional points are iteratively updated to minimise the difference between the similarity distributions in the low- and high-dimensional spaces, pushing similar points together and dissimilar points away from one another [MH08].

5.2 Datasets: Multimodal Universe

To evaluate this model, suitable physics datasets are necessary that provide multimodal data combined with an appropriate property prediction downstream task. Astronomical data can provide vast amounts of imaging, spectral, hyperspectral, time-series and tabular multimodal data. Specifically, imaging and spectral data can describe galaxies from multiple perspectives, containing information about their physical properties. The physical properties can be used as a downstream prediction task as they reveal how much information modalities contain about such underlying properties. Additionally, we want to investigate how much information is encoded in one modality about the other modality and which modality contains more usable information for the downstream task. For that, we can apply the proposed framework to study where exactly the useful information is encoded.

For these reasons, we chose the Multimodal Universe dataset (MMU) [AAB+24]. This is a large-scale, multimodal astronomical dataset that is specifically designed for machine learning tasks that can leverage such data. MMU includes the previously described modalities. The data measurements are taken using ground-based and space-based telescopes in various surveys, which can be cross-matched using built-in functionalities. We will now discuss each modality in detail:

1. MMU contains galaxy image data from multiple surveys, including Legacy Surveys DR10, Legacy Surveys North, HSC or JWST. We use images from the Hyper Suprime-Cam Subaru Strategic Program (HSC) because of its high-quality 160×160 images [AAB+24]. Images are captured in multiple channels, where each channel corresponds to a broad wavelength range. These channels are defined by the photometric system, which assigns a letter to each passband of wavelengths corresponding to its respective filters. The HSC program captured images in the g, r, i, z, and y channels, which correspond to the optical range visible to the human eye and the infrared range. The respective transmission rates of the filters used, which correspond to different wavelength intervals, are shown in more detail in Appendix fig. A.2. Unlike normal images, astronomical images have a very high dynamic range, meaning that there is a large span of multiple magnitudes between the brightest and dimmest signal sources in the image [AAB+24]. Additionally, the images are often noisy due to the sensors' high sensitivity, and depending on the telescope, there are multiple potential sources of noise [AAB+24]. These can include light pollution from Earth's atmosphere, noise introduced by the sensor due to imperfections or readout noise, sudden high-energy photons striking the sensor, and statistical variations in brightness and noise from long exposures, which occur due to the low number of photons reaching the sensor. The resulting image from the sensor can be described in the following form when relating it to the true underlying data [AAB+24]:

$$I = S * \Pi + n \tag{5.1}$$

   I: Captured signal

   S: Intrinsic source emission (true original signal)

   $*$: Convolution operator

   Π: Instrumental response/Point Spread function (PSF)

n: Measurement noise

Due to the high dynamic range, the data cannot be visualised well in its original form. Instead, we need to compress the data in a way that both bright and dim objects are visible. We need a way to map the data to RGB images. For this, we use an algorithm by Lupton which compresses the range by applying an *arcsinh* function such that outliers do not disturb the image [LBF+04]. As the HSC study contains g, r, i, z, and y channels, we need to drop some channels and retain only three channels. For representing these channels in a plausible way to RGB, we use g, r and z and map the longest wavelength to the longest one $z \to r$, the intermediate wavelength to the intermediate one $r \to g$ and the shortest wavelength to the shortest one $g \to b$ to make it look as plausible as possible since some of the wavelengths are not visible to the human eye. The algorithm also depends on two parameters: *stretch* determines the dynamic range, and Q determines the smoothness of the transition. For this algorithm, we inspect different combinations visually fig. 5.1.

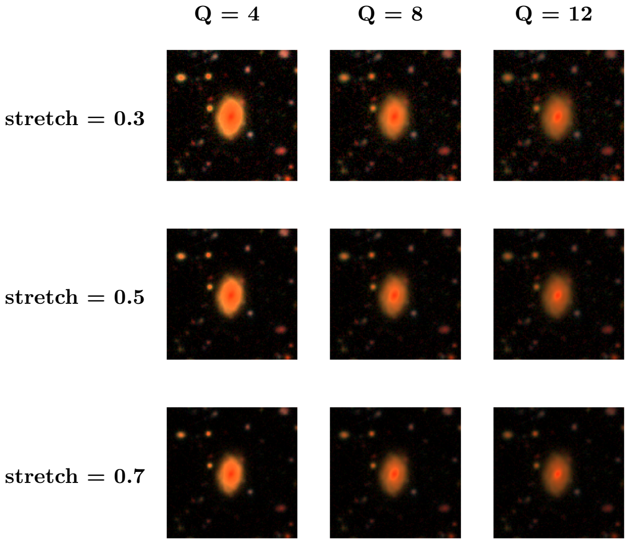


Figure 5.1: Galaxy images processed using the Lupton algorithm on g, r, z channels.

To minimise the number of parameters to optimise, we fix these values directly to $stretch = 0.5$ and $Q = 8$, which are close to the default values and provide a compromise between low noise, high dynamic range, and visible structures.

2. Spectra represent the distribution of wavelengths in the light emitted by galaxies that are captured by the telescope. It is captured similarly to an image in a telescope, but instead of using a camera, it uses a spectrograph to analyse the signal. This results in a 1-D signal for which a similar relation of true signal to observed signal holds as in eq. (5.1)(see [AAB+24] for more details). We use the spectra from the Dark Energy Spectroscopic Instrument (DESI), as they can be directly cross-matched with the physical property dataset (PROVABGS) [AAB+24].

3. Hyperspectral data can be represented as a three-dimensional data cube, combining spatial and spectral information: Two dimensions represent the spatial structure, and one dimension encodes the intensity for specific wavelengths. Compared to (RGB) images, it contains more spectral bands. Images can therefore be seen as a special case

obtained by averaging over spectral intervals, while spectra can be seen as a special case obtained by averaging over the spatial domain. One such dataset for galaxies is MaNGA [AAB+24].

4. Time-series data captures the brightness of objects over time. This is particularly relevant for objects like supernovae, which vary in brightness over time, providing valuable insights [AAB+24].

5. The last main modality captured in the dataset is tabular data. This can include various types of data, such as galaxy morphology and physical property data about galaxies. We use the PRObabilistic Value-Added Bright Galaxy Survey (PROVABGS) dataset, which inferred the physical property data from DESI spectra and contains galaxy properties such as [AAB+24]

   (a) Stellar mass $\log(M_*)$: Total mass of all stars in solar units, where one stellar unit represents the sun for comparison.

   (b) Average star formation rate $avgSFR$: Average rate at which stars are formed in a galaxy (we always apply log to compress the range).

   (c) Specific star formation rate $sSFR$: Rate at which stars are formed in a galaxy, normalised by stellar mass ($\log avgSFR - \log Z_{MW}$).

   (d) Metallicity Z_{MW} Abundance of elements heavier than helium, weighted by galaxy mass (we always apply log to compress the range).

   (e) Redshift Z_{HP}: How much the light spectrum has been shifted towards longer wavelengths, helpful for analysing the distance to a galaxy.

   (f) Age $t_{age,MW}$: Average stellar age, weighted by galaxy stellar mass.

We will study the image and spectrum modality combined with the physical properties. To use this as a single dataset with multiple modalities, we need to cross-match corresponding data items from multiple observational studies with one another. This is handled directly by the MMU dataset. Each data item contains the coordinates of right ascension and declination, which identify a point on the celestial sphere. This celestial sphere, as shown in fig. A.1 for reference, is an idealised infinitely large sphere centred on Earth, commonly used to project astronomical objects onto it for mapping their positions. We cross-match all elements that appear within a distance of 1 arcsecond of each other.

We utilise a multimodal dataset comprising the HSC, DESI, and PROVABGS datasets. After cross-matching, a dataset of size 8,503 remains, which we split into train, validation, and test datasets at ratios of 70%, 15%, and 15%, respectively. Due to the relatively small dataset size, it is reasonable to use CNNs instead of transformer-based models, which would require more data for training. For all subsequent experiments, training is conducted for 50 epochs, with a batch size of 256.

5.2.1 Physics background

The following is a brief chapter on the physics background of galaxies. Galaxies are collections of stars, gas, dust, dark matter, and often supermassive black holes, held together by their gravity. They are a complex system with complex dynamics. Galaxies originally formed due to matter density fluctuations, which collapsed, clustered and merged into galaxies under the pull of dark matter halos [MVdBW10]. Before galaxies fully assembled, the very first generation of massive stars formed. Within galaxies, new stars form in gas clouds that collapse under the influence of gravity. Then, depending on their mass, they undergo multiple phases of nuclear fusion, where lighter element nuclei are fused into heavier atoms (up to iron), releasing energy radiated as photons. Depending on its evolutionary phase and the dominant fusion processes, a star produces different energy outputs and emits photons with different spectral distributions. Typical bright galaxies have around 10^{10} stars [MVdBW10]. These are statistically distributed throughout the galaxy, depending on the type of galaxy, which can appear in various forms and shapes (morphologies) when observed from a distance. Many other processes in galaxies also produce or affect photons. Gas clouds can be ionised by high-energy photons or collisions of particles and can then re-emit this energy, when combining with free electrons, in characteristic wavelengths. During this, electrons are excited to a higher energy level by an incoming photon and afterwards return to a lower state while emitting energy in the form of a photon [MVdBW10]. For example, for hydrogen gas, a characteristic wavelength is that of the $H\alpha$ emission line. Atoms in gas clouds or stars also absorb specific wavelengths, which are then not visible in the spectrum; this can be used to analyse the apparent elements. Dust can also scatter ultraviolet and visible light, which re-radiate in infrared wavelengths. Other processes that can produce radiation are supermassive black holes in the centre of galaxies and supernovae, which emit a characteristic spectrum [MVdBW10].

Since the objects inside the galaxy move in different directions, one must also consider the Doppler effect, which describes the change in wavelength that occurs when the source of a signal is moving relative to the observer. If the source moves away from the observer, the wavelength becomes longer, and if the source moves towards the observer, the wavelength becomes shorter. This means that this effect widens the galaxy's spectrum due to the different velocities of objects inside the galaxy. Additionally, the galaxies are moving away from Earth, introducing a redshift z of their spectra, which is generally defined as

$$z = \frac{\lambda_{observed} - \lambda_{emitted}}{\lambda_{emitted}},$$

where λ resembles the wavelengths. If the redshift is given, it can be used to reconstruct the original spectrum using

$$\lambda_{emitted} = \frac{\lambda_{observed}}{z+1}. \tag{5.2}$$

This is also referred to as the rest-frame [MVdBW10]. In general, the relativistic Doppler effect is given by

$$\frac{\lambda_{observed}}{\lambda_{emitted}} = \sqrt{\frac{1+v/c}{1-v/c}},$$

where v is the relative speed difference between source and observer and c is the speed of light. For small $v \ll c$, z can be approximated through a first-order Taylor expansion by $z \approx v/c$. The redshift captures how much the spectrum of a galaxy is shifted to longer/redder wavelengths when the galaxy is moving away from us. It is essential to note that it is not the galaxy itself that is moving away, but rather the space between galaxies that is expanding. Through the expansion of the universe, galaxies that are further away from Earth are observed with a higher redshift. This connection is captured by Hubble's law for galaxies ($v \ll c$) [MVdBW10]

$$v = H_0 d \quad \text{with} \quad H(t) = \frac{\dot{a(t)}}{a(t)},$$

where v is the speed of the galaxy moving away, H_0 is the current time Hubble constant, and d is the distance to the galaxy. For more distant galaxies, one must take into account the scale factor of the universe $a(t)$, depending on the time when the signal was emitted compared to the scale factor at the time the signal was received. The scale factors $a(t)$ can be computed using the Friedmann–Lemaître equations.

$$1 + z = \frac{a(t_{observed})}{a(t_{emitted})}$$

All these factors influence the light emitted by a galaxy. They can therefore affect the image and spectrum captured by the telescope, which can be used to analyse its inner processes and properties.

5.2.2 Models' Architectures

For images and spectra from this dataset, we require suitable encoder/decoder architectures that can extract useful information and reconstruct each modality. These can then be used in VAEs, DMVAEs, and regression models as well. In the Appendix, we present the detailed architectures of the encoder/decoder pairs for images (see section A.1) and for spectra (see section A.1).

5.2.3 Downstream task: Physical property prediction

Based on the latent representations of the VAEs and DMVAEs, we utilise physical property prediction of galaxies as a downstream task. These include $log(M_*)$, $avgSFR$, $sSFR$, Z_{MW}, Z_{HP} and $t_{age,MW}$ which were previously described. For this, we use a simple MLP, as described in [PLG$^+$24], to assess the informativeness of the latent spaces. The exact architecture is shown in section A.1.

5.2.4 Data Preprocessing

Before the data is used for the experiments, several preprocessing steps are done to improve training performance and reduce overfitting. These steps differ slightly for images and spectra. For images:

1. The high dynamic range multichannel (g, r, i, z, y) images are reduced to three channels and mapped to RGB (z, r, g) and compressed to a more human-readable image using Lupton's algorithm [LBF$^+$04] and rescaled to 128×128. The input range is then reduced from $[0, 255]$ to $[0, 1]$.

2. The standard deviation is computed channel-wise, and the image is normalised to $x' = \frac{x}{\sigma}$. The mean is ignored, allowing the ReLU output function to fit the normalised data.

3. The images are augmented by random horizontal and vertical flips.

For spectra

1. Padding values are removed.
2. The flux is resampled on 4000 logarithmically placed wavelength positions between 3800 and 8400 as in [ICT23].
3. The standard deviation and mean are computed channel-wise, and the spectrum is normalised to $x' = \frac{x-\mu}{\sigma}$.

The preprocessing of the spectra is done similarly to [ICT23]. The reason for the logarithmic resampling of the wavelength is that, due to the Doppler shift, the velocity behaves additively in log-space. When observing some wavelength in log-space, we get

$$\log(\lambda_{observed}) = \log(\lambda_{emitted}(1+z)) = \log(\lambda_{emitted}) + \log(1+z).$$

Hence, we get the difference between the emitted and observed wavelength

$$\Delta \log \lambda = \log \lambda_{observed} - \log \lambda_{emitted} = \log(1+z) \approx v/c.$$

This means that the Doppler effect introduces an additive term in log-space, which depends linearly on the velocity of the galaxy. This should help simplify the learning task, as the features maintain their relative positions in log wavelength space, with only an additional velocity-dependent term. Additionally, one could use eq. (5.2) to normalise the spectrum to the rest-frame such that wavelengths are directly comparable. However, this would presuppose that the redshift is given. We do not assume this here, as we want to predict the redshift and leave this step out.

5.2.5 Training details

The loss function consists primarily of three components: Reconstruction, KL-divergence, and CLUB losses. These components need to be weighted appropriately for a good compromise. Here, we first describe and adjust the weighting factors of the reconstruction terms λ_1 and λ_2, while the KL and CLUB weights are determined through experiments.

The simplest method would be to set the weights for the reconstruction terms to both 1. Here, the weights for both modalities, λ_1 and λ_2, are determined dynamically using the validation dataset. The idea behind this is the following: We aim for the reconstruction of both modalities to have a comparable influence on the latent space. This could help best extract the shared and private features, as otherwise, one modality might have a disproportionately large impact on the learned latent space. This could make it harder for the model to learn shared features between both modalities. Assume that the model should learn two modalities, but one modality inherently has a much lower training loss. Then the gradient of the loss is dominated by one modality making it potentially more difficult to to learn shared features [1]. By reweighing the loss of both modalities to be of similar magnitude, the model can potentially learn latent features better. The weights λ_1, λ_2 are determined as follows:

$$\lambda_1 = 2\frac{a}{a+1}, \quad \lambda_2 = 2\frac{1}{a+1},$$

[1] In our case, the model quickly learned that it could achieve a low training loss for the galaxy images by simply reconstructing black images, although it had not yet learned anything about the galaxies' appearances at that point, whereas the loss for the spectra was still quite high.

where

$$a = \frac{R_{total,2}}{R_{total,1} + R_{total,2} + \epsilon}$$

and $R_{total,i}$ represents the total reconstruction loss including all three reconstruction losses regarding the corresponding modality and a is clipped between $(10^{-3}, 10^{3})$, resulting in $\lambda_1 + \lambda_2 \approx 2$. This process is repeated every two epochs, automatically balancing both modalities.

For the CLUB loss, we only train the neural network to predict the conditional distribution between latent spaces in the first five epochs, and only afterwards include CLUB in the loss. This ensures that only a useful, relatively accurate CLUB loss is used during training, thereby reducing the noise introduced by the CLUB loss.

5.3 Experiments

We now give an overview of the experiments. In the first five experiments, we conduct experiments on galaxy image, referred to as modality 1, and spectral, referred to as modality 2, modalities, performing them on both multimodal data using the DMVAE (with CLUB loss) and on single-modal data using VAEs, and in the sixth experiment, we use hyperspectral data. The chapters are structured as follows:

1. We study and determine the weighting factors of the CLUB loss λ_{CLUB} and the KL-divergence β_{KL} for the DMVAE with CLUB. The reconstruction weights are determined dynamically as previously mentioned. Here, we can already see the impact of these weights on the model.

2. Once we have fixed the weights for CLUB and KL, an optimal learning rate and a learning rate schedule are determined for both DMVAE with CLUB and the VAEs. We use the exponential learning rate scheduler, which depends on a starting learning rate η_0 and learning rate decay γ. These are the parameters we optimise here and fix for further experiments [2].

3. After having determined training-relevant hyperparameters, we systematically test different latent sizes. For the DMVAE (with CLUB), we examine different sizes of both private latent spaces and the shared latent space, and for the VAEs, we only have to test various sizes for one latent size together with β_{KL}. This helps to determine suitable representation sizes for the modalities and to assess the impact of CLUB. We consider these different configurations using our pre-determined metrics and on downstream task performance.

4. Upon completing the test on different latent sizes, we can study single configurations in more detail. To do so, we compare the DMVAE with and without CLUB loss and evaluate lower and upper bounds on mutual information to assess how effectively the CLUB loss performs. We also consider the structure of the latent space and where different physical parameter values are placed within it. We then assess the

[2] Although it is generally possible that the different parameter configurations in the later experiments have different optimal learning rates, by evaluating this in advance, we can remove this hyperparameter from later, save compute and make the experiments easier to compare.

downstream task performance on all combinations of latent spaces to test where the underlying corresponding information is encoded. We also compare this to VAEs and evaluate all models on a similar configuration.

5. After having investigated the default framework, we aim to extend it by incorporating a physical model of galaxy images in the image decoder, which should help the latent space learn semantically meaningful geometric attributes of the galaxy and possibly increase downstream performance. We investigate it and evaluate the performance of this model using several metrics.

6. Once these experiments are completed, we turn to a different dataset containing hyperspectral data from the HyPlant dataset. We investigate it as a source of multimodal image and spectral data to see how much information these modalities contain about hyperspectral data.

5.3.1 Experiment 1: Finding suitable weights for KL and CLUB

Our loss function for the DMVAE with CLUB consists of three terms: Reconstruction, KL-divergence and CLUB loss, whose weights we need to balance. In this experiment, we investigate different values of the weights β_{CLUB} and λ_{CLUB}.

Outcome: We end up with the values $\lambda_{CLUB} = 0.1$ and $\beta_{KL} = 0.01$, which we fix for further experiments.

Details: We now describe in detail how we come to these weight values. Since we determine the reconstruction weight dynamically, to balance both modalities according to the previously described formula, we only need to balance the KL divergence and CLUB. Balancing different learning terms dynamically is also subject to the paradigm of Multitask learning (MTL). However, here, we conduct a test on various combinations of constant weights β_{KL} and λ_{CLUB} to determine a suitable combination that offers the best tradeoff between reconstruction, KL-divergence, and CLUB performance. It is expected that β_{KL} can have a large impact on the latent space's information content because it implicitly controls how much the mutual information between input x and latent z is penalised. This dependency is evident in eq. (3.5). It is important to note that the optimal KL and CLUB weights might depend on the latent space size. However, due to the large size of parameter combinations, we cannot test all parameters jointly. Instead, we first test for a suitable weight combination and later check for a suitable latent size. Here, we first use a latent size of 8 for all latent spaces, as a small latent size saves space and naturally has less correlation between latent spaces. The results are shown in fig. 5.2, where the test losses are visualised, decomposed into their components (with all components weighted equally in the test loss) [3].

[3] Note that the loss components were evaluated on the test set, and the CLUB loss was here just trained on the training set, such that the evaluation on the test set is only an approximation of the mutual information estimate. In later experiments, we evaluate CLUB and mutual information more accurately.

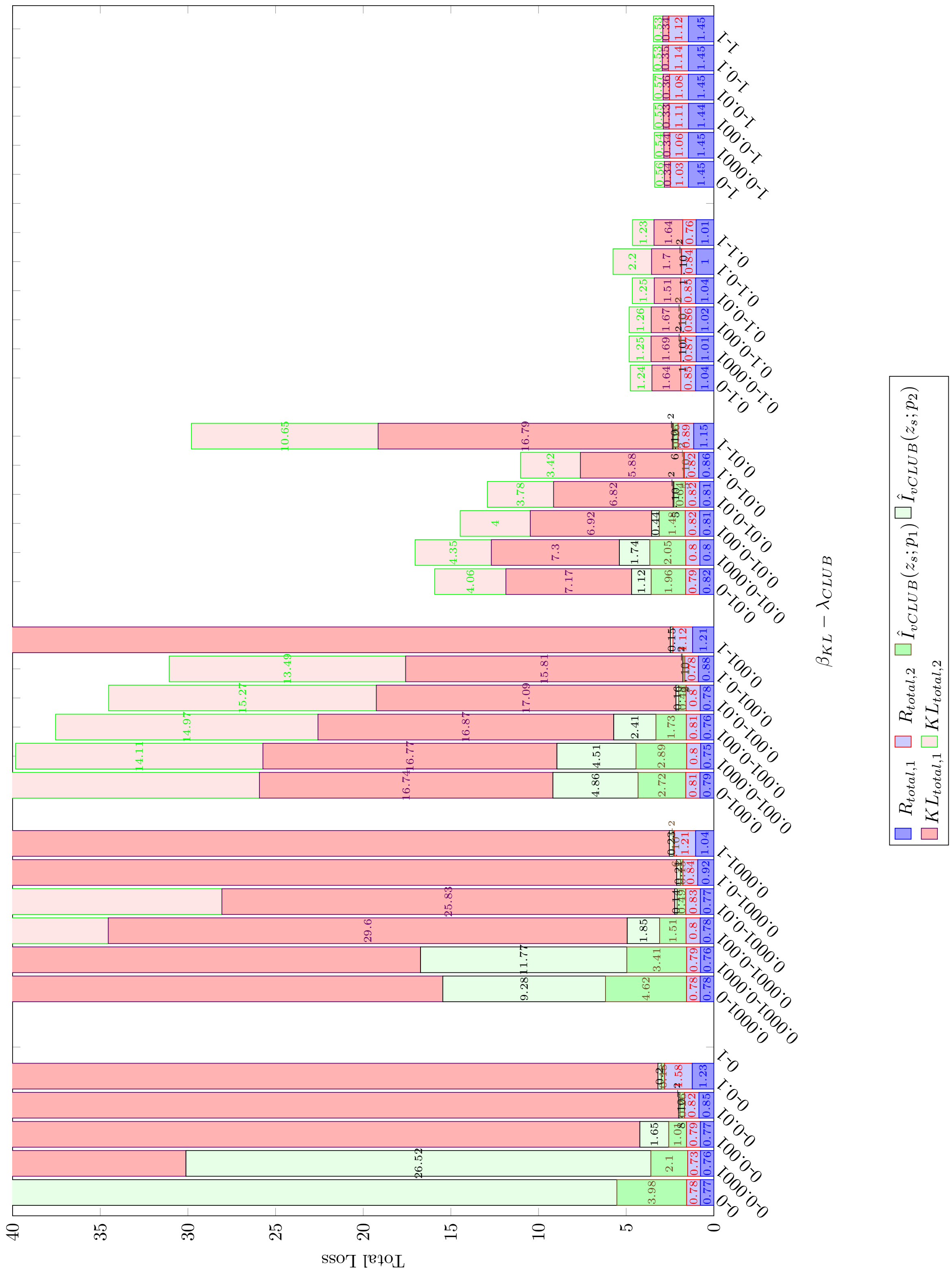


Figure 5.2: Test loss components for different $\beta_{KL} - \lambda_{CLUB}$ configurations, decomposed for modality 1 and 2: $R_{total,1}$, $R_{total,2}$ (reconstruction performance), $KL_{total,1}$, $KL_{total,2}$ (KL-divergence), and $\hat{I}_{vCLUB}(z_s; p_1)$, $\hat{I}_{vCLUB}(z_s; p_2)$ (MI estimation).

The configurations are grouped from left to right by their KL weight, and within each block, they are sorted by the CLUB weight. This plot then shows, as expected, that assigning lower weights leads to higher losses in the corresponding components. Since the overall loss function reflects a trade-off between reconstruction quality and latent regularisation, increasing the weights of KL and CLUB inevitably reduces reconstruction performance. It is desired that the mutual information between the latent spaces is minimal while having the best possible reconstruction performance. A suitable low CLUB loss is generally reached for $\lambda_{CLUB} = 0.1$. A good reconstruction performance with a low test loss and low KL-divergence is achieved with $\beta_{KL} = 0.01$. A higher β_{KL} can further decrease the test loss; however, it comes at the cost of significantly worse reconstruction performance.

5.3.2 Experiment 2: Learning Rate Schedule

In this experiment, we aim to determine a suitable learning rate schedule for VAEs on images, VAEs on spectra, and the DMVAE with CLUB loss. We differentiate between these models as they differ substantially and can have different learning rates. We use the exponential learning rate scheduler as it is a standard scheduler for decaying the learning rate over time, which can help achieve a lower loss.

$$\eta_t = \eta_0 \cdot \gamma^t$$

Outcome: The best found configurations are shown in table 5.1 and fixed for later experiments.

Model	Image VAE	Spectrum VAE	DMVAE with CLUB
(η_0, γ)	$(10^{-3}, 0.99)$	$(10^{-4}, 0.96)$	$(10^{-4}, 0.99)$

Table 5.1: Best hyperparameter configurations for each model.

Details: We now describe how we arrive at these parameter values by evaluating the performance for different decays $\gamma \in \{0.99, 0.975, 0.96\}$ and start learning rates $\eta_0 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ on image/spectrum VAEs and on DMVAE with CLUB. We evaluate solely based on the test loss.

VAE for images: In this small experiment, we utilise the image modality with a latent space of size 16 (resembling the same latent size that the decoders of the DMVAE had in the previous experiment of $8+8$), $\beta_{KL} = 0.01$, which serves as a baseline here, as previous testing has demonstrated that these parameters already yield satisfactory results. It is essential to note that the total test loss is evaluated with $\beta_{KL} = 1$ for easier comparability in subsequent experiments. The results are shown in table 5.3

γ	$\eta_0 = 10^{-3}$				$\eta_0 = 10^{-4}$				$\eta_0 = 10^{-5}$				$\eta_0 = 10^{-6}$			
	rec	KL	scaled	total	rec	KL	scaled	total	rec	KL	scaled	total	rec	KL	scaled	total
0.99	0.22	1.88	**0.24**	2.11	0.23	1.88	0.25	2.11	0.48	1.62	0.5	2.10	0.86	1.47	0.87	2.34
0.975	0.22	1.98	0.24	2.20	0.24	1.85	0.26	2.08	0.57	1.46	0.58	2.03	0.92	1.25	0.93	2.17
0.96	0.22	2.03	0.24	2.25	0.25	1.88	0.27	2.12	0.69	1.31	0.7	2.00	0.97	0.93	0.98	1.89

Table 5.2: Test loss components (reconstruction, KL-divergence, scaled test loss, total test loss with $\beta = 1$) for different η_0 and γ.

Considering the down-weighting factor of the KL-divergence by 0.01, we get that a starting learning rate of 10^{-3} and $\gamma = 0.99$ results in the best test loss with the down-weighting factor included, since the reconstruction loss here is the best with a relatively low KL loss. Although other configurations have a lower total test loss, we choose the configuration based on the fact that reconstruction is more important for our goals. The training plots are included in fig. A.6. Additionally, when evaluating the reconstruction performance visually, we can see that the model achieves satisfying results, as shown in fig. 5.3.

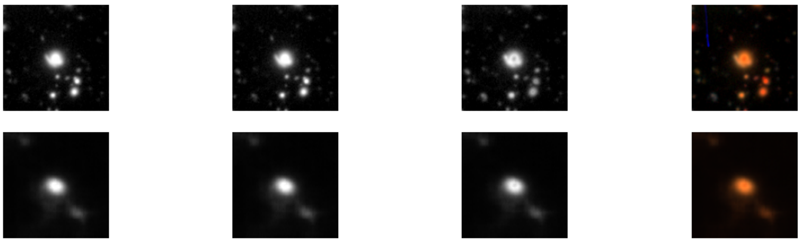


Figure 5.3: Visualisation of test dataset sample. Top row: original image after preprocessing (unnormalized); bottom row: reconstructed image (unnormalized). Each column shows, from left to right: red (z), green (r), and blue (g) channels, followed by the combined RGB image.

VAE for spectra: We also test for the best start learning rate η_0 and decay γ with the same configuration as for images.

γ	$\eta_0 = 10^{-3}$				$\eta_0 = 10^{-4}$				$\eta_0 = 10^{-5}$				$\eta_0 = 10^{-6}$			
	rec	KL	scaled	total	rec	KL	scaled	total	rec	KL	scaled	total	rec	KL	scaled	total
0.99	0.28	3.78	0.32	4.06	0.28	0.69	0.29	0.97	0.28	1.02	0.29	1.30	0.39	1.91	0.41	2.30
0.975	2.96	10752.07	110.48	10755.03	0.27	0.63	0.28	0.90	0.27	1.07	0.28	1.34	0.51	1.58	0.53	2.09
0.96	0.27	2.88	0.30	3.15	0.27	0.75	**0.28**	1.02	0.30	1.14	0.31	1.44	0.54	1.60	0.56	2.15

Table 5.3: Test loss components for different η_0 and γ.

We conclude that a starting learning rate of 10^{-4} with $\gamma = 0.96$ yields the best scaled test performance, along with the best reconstruction and a reasonable KL performance. The training plots are attached in fig. A.7. There, it is evident that the reconstruction loss during training is quite noisy, and the loss only slowly decreases, indicating that the reconstruction task is challenging for the model to learn. In fig. 5.4, you can see the reconstruction performance visually.

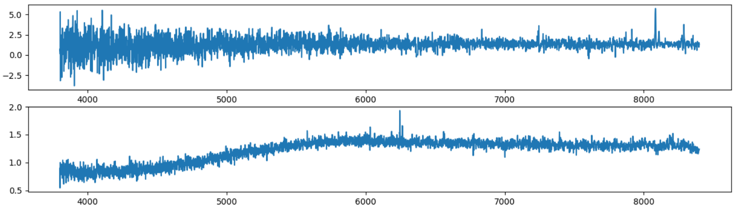


Figure 5.4: Visualisation of test dataset sample. Top row: original spectrum after preprocessing (unnormalized); bottom row: reconstructed spectrum (unnormalized). On the x-axis is the wavelength in $\mathring{A}$ and the y-axis shows the flux in $10^{-17}\frac{erg}{scm^2\mathring{A}}$.

DMVAE with CLUB: We repeat this now for the DMVAE with CLUB, with again a latent size of 8 for each latent space, and we set $\beta_{KL} = 0.01$ and $\lambda_{CLUB} = 0.1$. Similarly to the VAE case, we need to consider that the learning objective down-weights specific loss terms. Thus, we examine the components of the total test loss separately and reweigh each to one, as visualised in fig. 5.5.

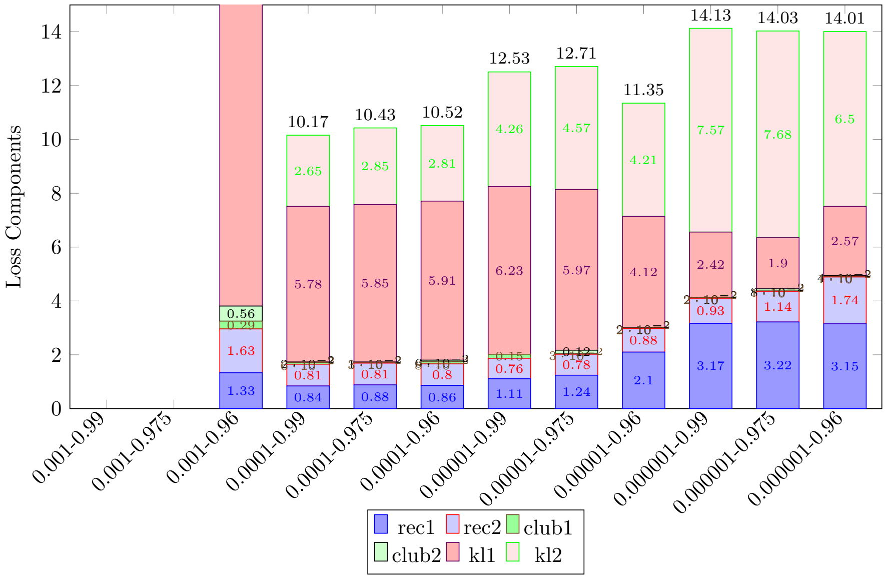
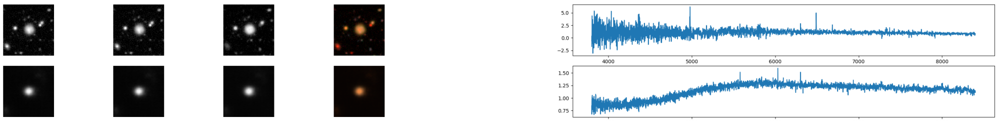


Figure 5.5: Test loss components $R_{total,1}$, $R_{total,2}$, $KL_{total,1}$, $KL_{total,2}$, $\hat{I}_{vCLUB}(z_s; p_1)$, and $\hat{I}_{vCLUB}(z_s; p_2)$ for different configurations of η_0 and γ. Failed configurations are empty.

It is visible that a starting learning rate of 10^{-4} combined with a γ of 0.99 has the best test loss. It also has a reasonably low KL-divergence, low CLUB loss, as well as good reconstruction performance on both modalities. Here, reconstruction is the primary objective, as we aim for the latent space to learn as much as possible about the modalities. It is visible that higher learning rates are too large to reach good reconstruction abilities. The smaller the learning rate gets, the worse the reconstruction ability, as it is too low to learn suitable parameters for the model. Below a suitable starting learning rate of 10^{-4}, the trend is visible that lower γ reduces the reconstruction performance, implying that keeping a higher learning rate longer is better than trying to reduce it too much. The reconstruction performance is visualised in fig. 5.6 and its training plot is shown in fig. A.8.



(a) Top row: original image after preprocessing (unnormalized); bottom row: reconstructed image (unnormalized). Each group shows, from left to right: red (z), green (r), and blue (g) channels, followed by the combined RGB image.

(b) Top row: original spectrum after preprocessing (unnormalized); bottom row: reconstructed spectrum (unnormalized). On the x-axis is the wavelength in $\mathring{A}$ and y-axis in $10^{-17}\frac{erg}{scm^2\mathring{A}}$.

Figure 5.6: Visualisation of test dataset samples.

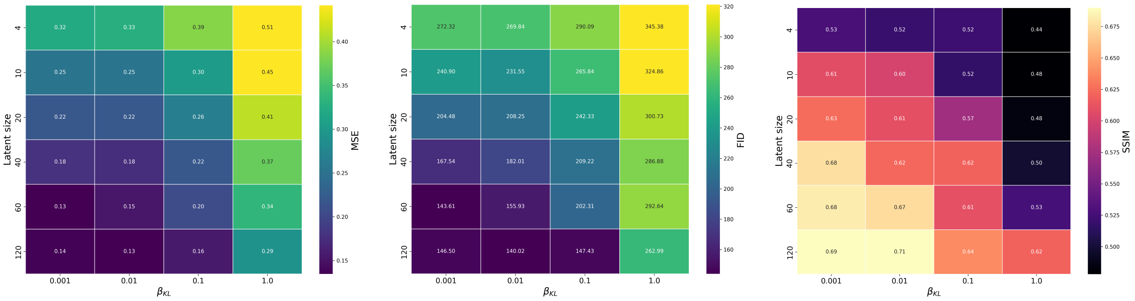
5.3.3 Experiment 3: Evaluation of different latent space sizes

Here, different latent sizes are compared to study their impact on the metrics and downstream task performance and to find a suitable representation size in the latent spaces for each modality. This step is critical because we need to implicitly optimise a tradeoff of latent sizes to be large enough to capture all essential features, leading to a good reconstruction performance, while also not making the latent spaces too large, which could increase correlation and therefore increase mutual information, leading to higher CLUB loss, which could decrease the reconstruction loss.

Outcome: We find that for VAEs, smaller β_{KL} values generally lead to improved reconstruction and downstream performance. Depending on the modality, either a lower (for spectra) or a larger (for images) latent size is beneficial for the downstream task. For reasonable reconstruction performance, images require a larger latent size of 40 or 60, while spectra perform better with smaller latent sizes, such as 4. For DMVAE with CLUB, we find that larger private latent space increase their respective reconstruction abilities while larger shared latent spaces generally harm them. When excluding the CLUB loss, a larger shared latent size also improves reconstruction performance. The reconstruction performance of the spectrum modality does not benefit from larger latent sizes. For both DMVAE variants, the downstream performance does not consistently depend on the private latent space, but rather on the shared latent size, which decreases/increases for larger latent sizes when CLUB is used/not used, respectively. The cross-reconstruction performance is significantly worse than for normal reconstruction.

Details: We investigate the latent spaces of $\{4, 10, 20, 40, 60\}$. We choose these values because [XSdS$^+$23] also studied suitable latent sizes, including sizes of $10, 20, 40, 60$ and determined that a latent size of 40 was able to reproduce most morphological features in a galaxy image. For spectra, a lower latent size seems suitable, as [ICT23] concluded that a latent size of 4 is suitable to represent galaxy spectra from SDSS spectral data. For the DMVAEs, we test each combination of latent sizes $(\text{private1}, \text{private2}, \text{shared}) \in \{(s, p1, p2) | s, p1, p2 \in \{4, 10, 20, 40, 60\}\}$. Due to the sheer size of the combinations ($5^3 = 125$), we cannot test β_{KL} and λ_{CLUB} values jointly, which is why we conducted the first experiment to fix these values. For VAEs, it is suitable to test different β_{KL} and latent sizes jointly as computing all combinations in $(\beta_{KL}, \text{latent size}) \in \{(\beta_{KL}, l) | \beta_{KL} \in \{0.001, 0.01, 0.1, 1\}, l \in \{4, 10, 20, 40, 60, 120\}\}$ is computationally feasible. Here, we also include the latent size of 120 for the VAEs, as in the DMVAE, each modality can be represented by a latent size of up to $60 + 60$, which ensures better comparability. For the following plots, to improve readability, the colour spectrum only visualises the $[0.05, 0.95]$ percentile of all values, and the shown metric values are rounded to two decimal places.

VAEs for Images: Here, traditional VAEs are used to evaluate the performance that can be expected if only one modality is given. We compare different combinations of β_{KL} and latent size, as both parameters can significantly impact both reconstruction performance and downstream task performance.



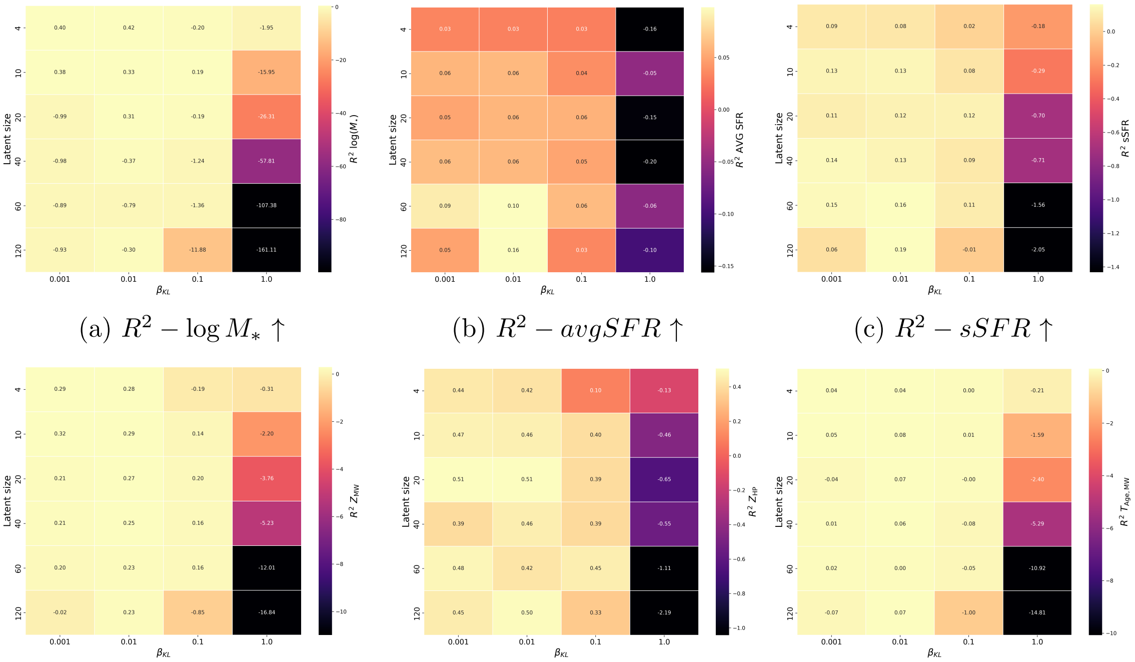
(a) MSE ↓

(b) FID ↓

(c) SSIM ↑

Figure 5.7: Reconstruction performance of image VAE: MSE, FID and SSIM metric visualisation dependent on latent size and β_{KL} (↓: Lower is better, ↑: Higher is better).

First, we evaluate the reconstruction quality of the image in fig. 5.7. It is visible that both a larger latent space and a lower β_{KL} consistently improve the reconstruction performance across all metrics. The pixel-wise metric MSE, as well as the perceptual metrics FID and SSIM, all show this trend. The best reconstruction performance is consistently achieved for the largest latent sizes and the lowest β_{KL}. It can be observed that a latent size of 40 or 60 can represent images well with $\beta_{KL} \leq 0.01$.



(a) $R^2 - \log M_* \uparrow$

(b) $R^2 - avgSFR \uparrow$

(c) $R^2 - sSFR \uparrow$

(d) $R^2 - Z_{MW} \uparrow$

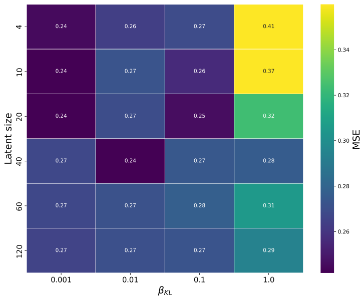
(e) $R^2 - Z_{HP} \uparrow$

(f) $R^2 - T_{AGE,MW} \uparrow$

Figure 5.8: Downstream regression performance of physical properties, measured in R^2 dependent on β_{KL} and latent size (↑: Higher is better).

In fig. 5.8, the downstream task performance is shown for every output separately. The performance varies between different physical properties, but generally increases if β_{KL} is sufficiently low. Additionally, it performs well for small latent sizes, with only slight improvements for larger ones, indicating that while a larger latent size significantly enhances reconstruction performance due to increased expressivity, it has only a subtle impact on downstream performance. One reason could be that the image also contains other non-related galaxies, which are captured in larger latent sizes, but do not always provide a benefit to property prediction. Meanwhile, lower β_{KL} perform better on both tasks, because the mutual information between x and z is penalised less, preserving more information.

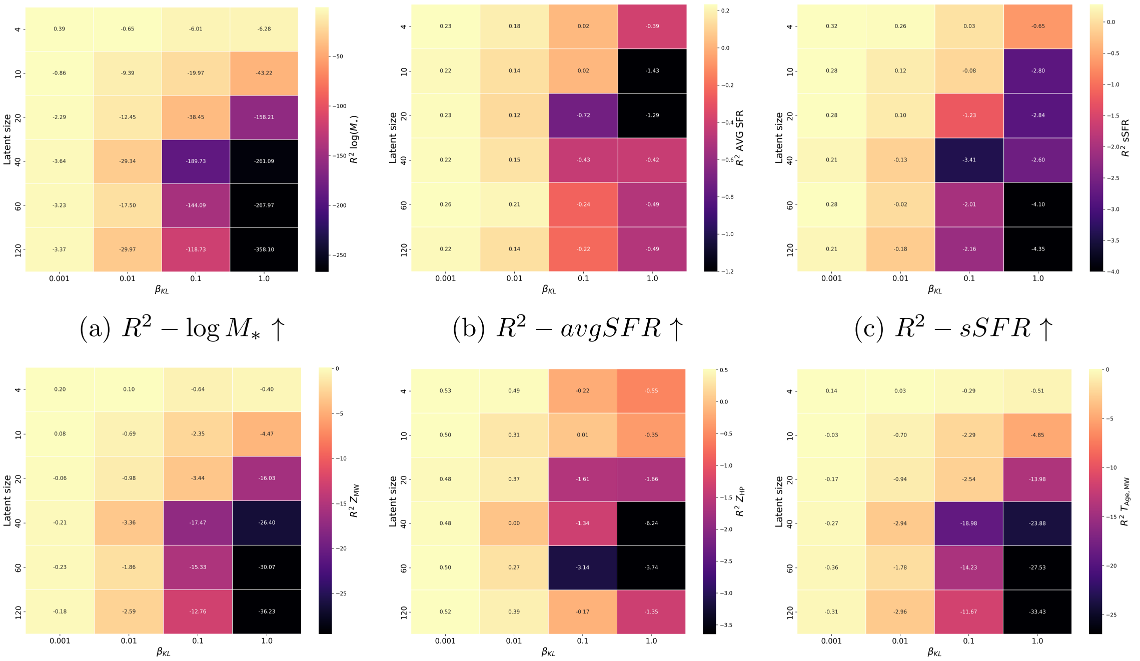
VAEs for Spectra: Here, we apply VAEs to spectral data and compare different combinations of β_{KL} and latent size.



(a) MSE ↓

Figure 5.9: Reconstruction performance of spectrum VAE dependent on latent size and β_{KL} (↓: Lower is better).

First, the reconstruction quality of the image is evaluated in fig. 5.9. It is evident that mainly a lower β_{KL} improves the reconstruction performance up to a certain point, while low latent sizes seem to reconstruct the data better than larger ones.



(a) $R^2 - \log M_* \uparrow$

(b) $R^2 - avgSFR \uparrow$

(c) $R^2 - sSFR \uparrow$

(d) $R^2 - Z_{MW} \uparrow$

(e) $R^2 - Z_{HP} \uparrow$

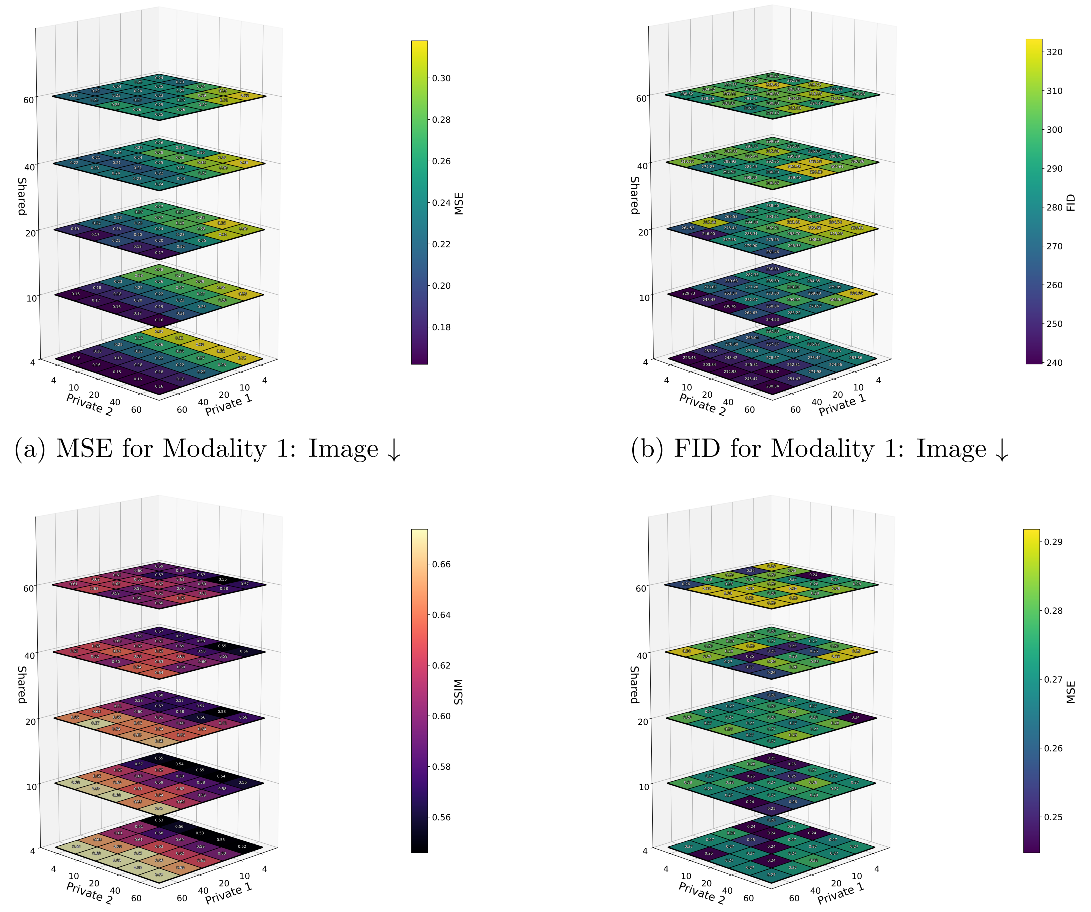
(f) $R^2 - T_{AGE,MW} \uparrow$

Figure 5.10: Downstream regression performance of physical properties, measured in R^2 dependent on β_{KL} and latent size (↑: Higher is better).

In the fig. 5.10, the downstream task performance is evaluated for every physical property.

Here, the performance varies between different physical properties, but generally increases for lower β_{KL} and smaller latent sizes. This shows that a small latent size can capture the essential features of a spectrum while being more robust in doing so than for larger latent sizes. This result aligns with findings from [ICT23] that a latent size of 4 is already suitable for spectra. Now, the multimodal DMVAEs (with CLUB) are investigated.

DMVAE with CLUB: We investigate various combinations of latent spaces to assess their impact on different metrics and determine a suitable representation size for both modalities. We seek a tradeoff between a reasonable size that captures all details while maintaining a low inherent mutual information between the representations. This also gives indications on how well the model can learn the modalities and where information is stored. Due to the size of the parameters to optimise, the $\beta_{KL} = 0.01$ and $\lambda_{CLUB} = 0.1$ values from the first experiment, along with an initial learning rate of 10^{-4} and $\gamma = 0.99$, are used. We study all the described combinations of latent sizes. First, we examine the reconstruction performance for each modality when both modalities are provided as input, using the MSE, FID, and SSIM metrics.



(a) MSE for Modality 1: Image ↓

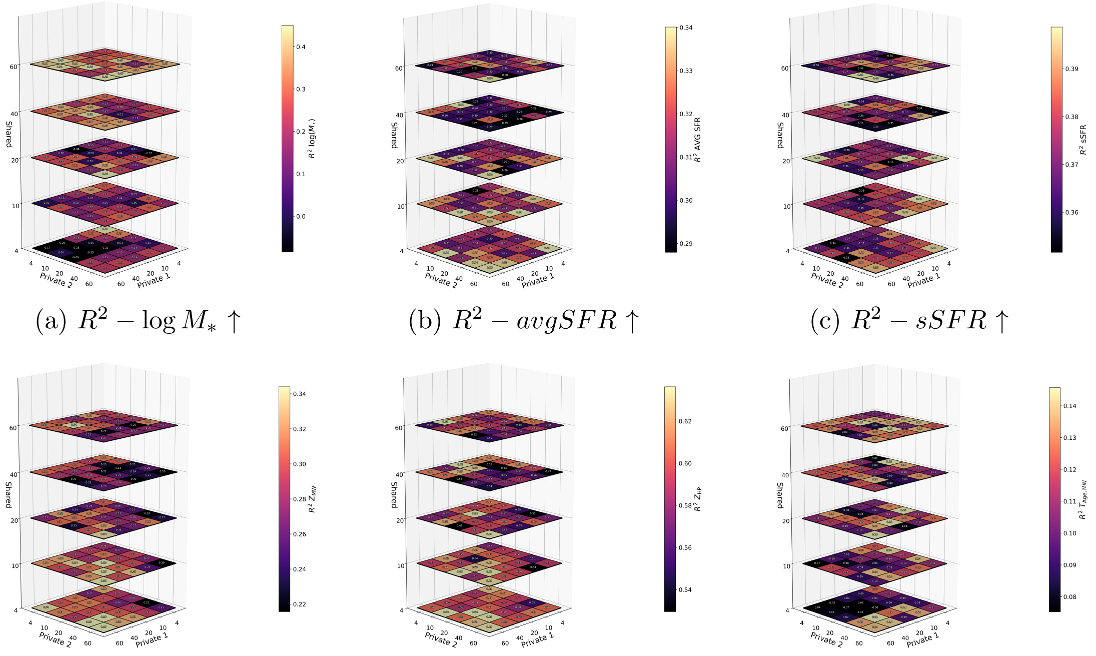
(b) FID for Modality 1: Image ↓

(c) SSIM for Modality 1: Image ↑

(d) MSE for Modality 2: Spectrum ↓

Figure 5.11: Reconstruction performance of DMVAE with CLUB of modality 1 and modality 2: MSE, FID and SSIM metric visualisation dependent on latent size and private 1, private 2 and shared latent size (↓: Lower is better, ↑: Higher is better).

When looking at the pixel-wise reconstruction performance as in fig. 5.11, multiple trends are visible: For reconstructing images (modality 1), a larger private 1 latent size increases the reconstruction performance significantly, as it enables the model to capture more features of the images, resulting in more detailed reconstructions. It is also evident that increasing the private 2 (spectra) latent size has a minimal impact on the reconstruction capabilities of modality 1. It is also visible that a larger shared latent size seems to decrease the reconstruction performance. When we evaluate the FID and SSIM metrics, which assess perceived reconstruction performance rather than pixel-wise performance, we observe similar trends. For the reconstruction capabilities of spectra in fig. 5.11d, it is also visible that a higher shared latent space seems to decrease the reconstruction performance. However, it is evident here that the private latent space of modality 2 and 1 has no consistent impact on its reconstruction performance. Overall, either a private latent size of 40 or 60 for images and minimal latent sizes for private 2 and shared seem to represent the data well. The reason for this trend, that a larger shared latent size decreases reconstruction performance, is not immediately apparent, because one would typically expect that a larger latent space should always improve reconstruction performance. The CLUB loss could be responsible for this, which we will study in more detail later. However, it may not be necessary to achieve perfect reconstruction performance, as not all features in the image are important for the downstream task. When comparing the reconstruction performance to that of the VAE (see fig. 5.7), it can be observed that the image VAE achieves better reconstruction performance according to MSE, FID, and SSIM, particularly for the largest latent space configurations. In contrast, for spectra, there is no significant difference (see fig. 5.9). This could be because the VAE offers more latent space that is only used for the image modality and is not restricted due to the inclusion of a PoE. Still, the DMVAE can achieve better downstream task performance despite its often worse reconstruction performance, due to its multimodality, as we now examine.



(a) $R^2 - \log M_* \uparrow$

(b) $R^2 - avgSFR \uparrow$

(c) $R^2 - sSFR \uparrow$

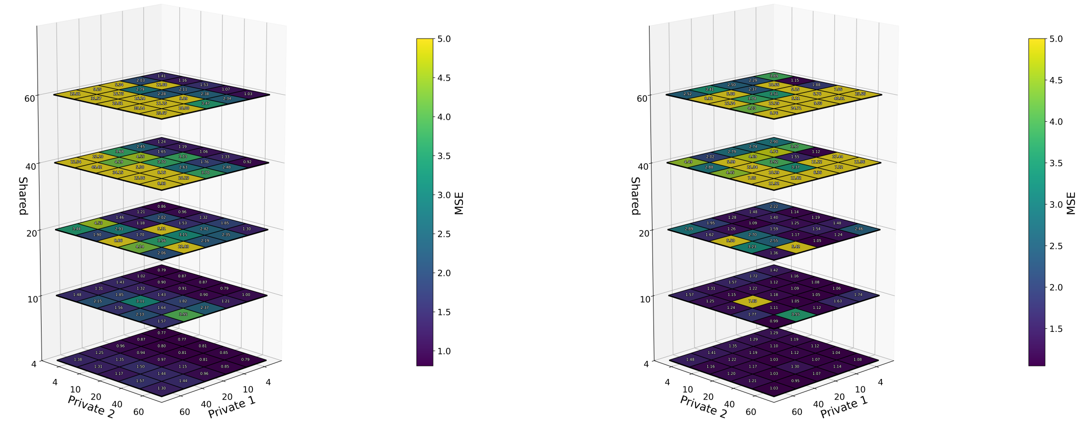
(d) $R^2 - Z_{MW} \uparrow$

(e) $R^2 - Z_{HP} \uparrow$

(f) $R^2 - T_{AGE,MW} \uparrow$

Figure 5.12: Downstream regression performance of physical properties, measured in R^2 dependent on private 1, private 2 and shared latent size for the DMVAE with CLUB (↑: Higher is better).

Specifically, we examine the downstream task performances of physical property prediction to assess how the latent sizes impact this. For every predicted physical property, its R^2 value is visualised in fig. 5.12. It is visible that the performance varies significantly. A slight trend is visible, indicating that lower shared latent sizes may lead to better performance on specific properties, such as the redshift Z_{HP}. However, on others, a larger shared latent size appears to be beneficial for properties like stellar mass log M_*. This suggests that the large shared latent size can potentially contain relevant information for some properties, while negatively impacting prediction performance for others due to the overall poorer reconstruction performance. We can also evaluate how much information each modality contains about the other by examining its cross-reconstruction capabilities.



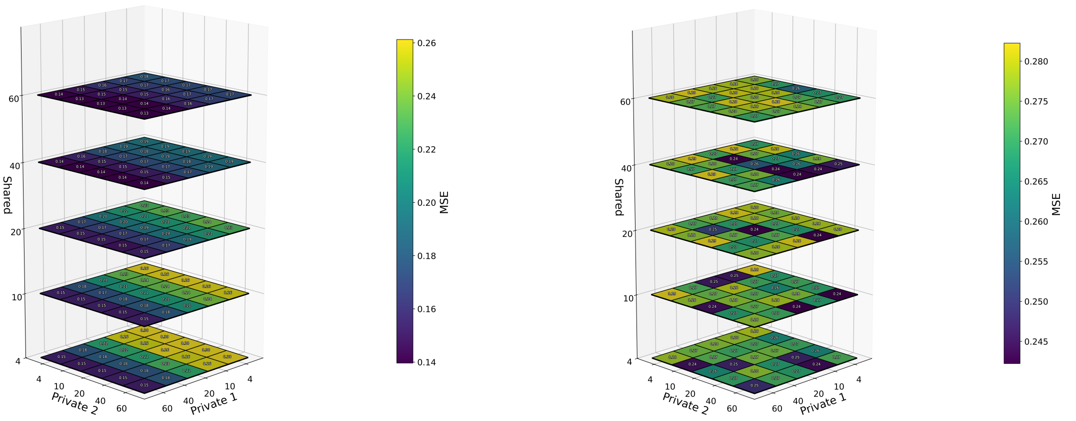
(a) MSE cross-reconstruction of Modality 1 ↓

(b) MSE cross-reconstruction of Modality 2 ↓

Figure 5.13: Cross-reconstruction performance when only other modality is given, dependent on shared, private 1 and private 2 latent size for the DMVAE with CLUB (↓: Lower is better). The upper colour was differently rescaled for a clearer impression.

This also shows the not directly obvious behaviour that larger shared latent sizes lead to significantly worse cross-reconstruction for both modalities. The cross-reconstruction performance remains significantly below that of joint reconstruction, as seen in fig. 5.11. This could be a consequence of CLUB or because they do not share enough information for accurate cross-reconstruction, insufficient training or because the KL weight is too low to approximate a Gaussian in latent space accurately, leading to unrealistic sampled latent values, which is especially relevant for cross-reconstruction. We now investigate whether this effect of lower-performing large shared latent sizes is caused by CLUB loss.

DMVAE without CLUB: We repeat the experiment on the same combinations of latent spaces on the DMVAE without CLUB loss.

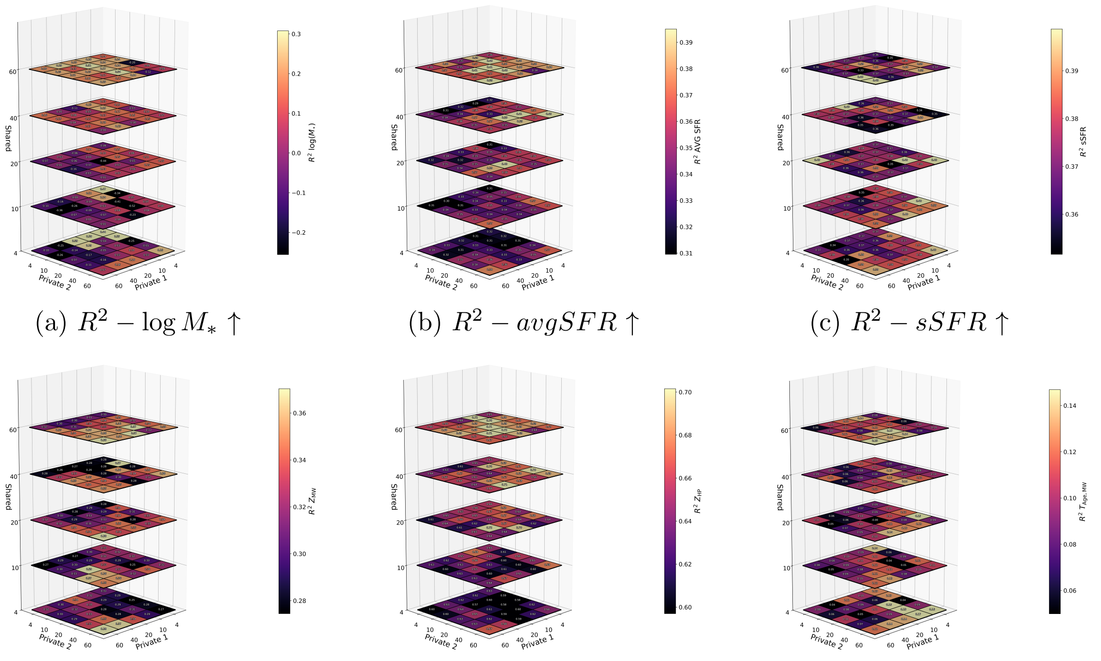


(a) MSE for Modality 1: Image ↓

(b) MSE for Modality 2: Spectrum ↓

Figure 5.14: Reconstruction performance of DMVAE without CLUB of modalities 1 and 2 dependent on private 1, private 2 and shared latent size (↓: Lower is better).

This plot illustrates what one would expect to see for the reconstruction performance of images, which increases for larger shared and private image latent sizes due to the larger representation size that the model can learn to utilise for the image. For the spectrum modality, it can be observed again that the latent size has a limited impact on its reconstruction performance, as spectra do not require a large size for representation. A latent size combination of 60 × 60 × 60 appears to be suitable for representing the data. This model comes much closer to the reconstruction performance of the VAEs, as seen in fig. 5.7 and fig. 5.9, and generally performs better for most configurations on reconstruction compared to DMVAE with CLUB, which reflects in its downstream task performance.



(a) $R^2 - \log M_* \uparrow$

(b) $R^2 - avgSFR \uparrow$

(c) $R^2 - sSFR \uparrow$

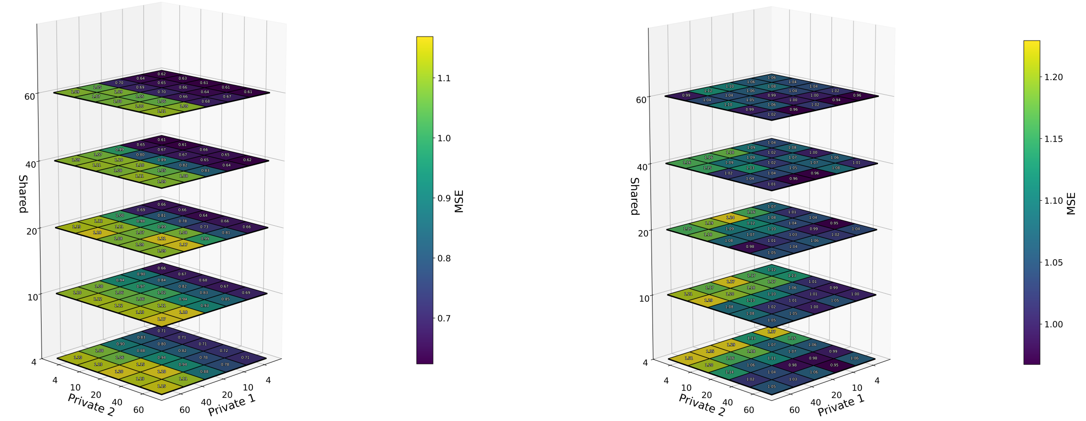
(d) $R^2 - Z_{MW} \uparrow$

(e) $R^2 - Z_{HP} \uparrow$

(f) $R^2 - T_{AGE,MW} \uparrow$

Figure 5.15: Downstream regression performance of physical properties, dependent on private 1/2 and shared latent size for the DMVAE without CLUB (↑: Higher is better).

This is visualised in fig. 5.15. Here, the downstream task performance again varies significantly, for the same possible reasons as before; however, a slight trend emerges that larger shared latent sizes may increase task performance. This is, for example, visible for the redshift Z_{HP}. We can also evaluate the cross-reconstruction performance in fig. 5.16.



(a) MSE for Modality 1: Image ↓

(b) MSE for Modality 2: Spectrum ↓

Figure 5.16: Cross-reconstruction performance when only other modality is given, dependent on shared, private 1 and private 2 latent size for the DMVAE without CLUB (↓: Lower is better).

This performs better than in fig. 5.13 and shows trends that one would expect. The first trend is that a larger shared latent space improves cross-reconstruction performance, as more information can be utilised for cross-reconstruction when the shared latent space is larger. This is visible for the cross-reconstruction of both modalities. For the cross-reconstruction of images, a second trend is visible: a smaller private 1 latent size increases the corresponding cross-reconstruction performance. This occurs because the ratio of shared to private latent size becomes larger, allowing more information from the other modality to be used for reconstruction. However, for the spectrum modality, the opposite trend is visible, meaning that a larger private latent space for spectra increases its reconstruction performance. This is counterintuitive and may be due to the minimal dependency of the spectrum on the latent size.

By removing the CLUB loss, the behaviour is as one expects, which shows that the CLUB loss causes the unexpected behaviour. This behaviour can now be explained in the following way. As the shared latent space becomes larger, it can capture more shared information, introducing additional correlations between latent spaces. As the CLUB loss relies on predicting the private latent spaces from the shared latent spaces, it becomes easier to predict them more accurately, leading to higher loss values. This trend is evident in fig. A.9, which illustrates that the approximated CLUB loss values increase with a larger shared latent size. These larger loss values reduce the impact of the reconstruction loss and the KL-divergence as their relative loss reduces when compared to the total optimisation objective. Therefore, CLUB can also harm cross-reconstruction performance for large latent sizes. If good cross-reconstruction performance is desired, one would have to potentially increase β_{KL} and re-estimate the other hyperparameters on the objective of cross-reconstruction. As this is not the primary target here, we proceed to the next experiment.

5.3.4 Experiment 4: Latent space and downstream task evaluation

In this section, we compare the downstream task performance of VAEs and DMVAEs with regression models, and evaluate the latent spaces of VAEs and DMVAEs (with/without CLUB). For this, we study the downstream task performances and use t-SNE [4] to visualise how well the physical properties are separated in latent space. If the values are well separated in latent spaces, it is easier for the downstream model to predict the correct value. This helps to comprehend why the performance differs on different physical properties. Furthermore, we study the contributions of unique and shared features to the downstream task using the methods described in chapter 4. To ensure that the mutual information is minimised between the private and shared latent spaces, we compute lower and upper bounds on it.

Outcome: We observe that VAEs exhibit poor performance on downstream tasks for both modalities, whereas the DMVAEs perform significantly better on these tasks. Here, the DMVAE without CLUB loss outperforms the one with CLUB and even surpasses the multimodal regression model in specific properties, which is otherwise the best. While studying the contributions of each latent space to the downstream task, we observe that CLUB loss enhances the model's ability to capture shared information that contributes to the downstream task compared to when it is left out, while reducing helpful information in the private representations. We find that the DMVAEs better separate the properties in their latent space than VAEs. We also see that the DMVAE with CLUB loss can minimise mutual information between latent spaces.

Details: This section describes how these observations are obtained. For a fair comparison, we choose a latent size of 120 for the VAEs and latent sizes of 60 for the DMVAES latent spaces and $\beta_{KL} = 0.01$ and $\lambda_{CLUB} = 0.1$.

Regression models: Before examining the performance of the VAEs and DMVAEs, first, we directly train regression models to predict physical properties from the modalities. This should serve as a baseline and an upper bound on the maximum performance expected from the VAE/DMVAE models' downstream task performance. The regression models use the same architecture as for the VAEs, meaning that the same image/spectrum encoders are used for the corresponding modalities, with the downstream model directly attached to them. For the multimodal case, the outputs of the encoders are concatenated and fed into the downstream regression model. These models are treated as a single model and trained directly on the regression task. The performance of these regression models is limited by the model's complexity, the dataset size, the inherent noise in the data, and the extent to which the modalities inherently reveal information about the physical properties. The models were trained for 50 epochs. The results are visualised in table 5.4, which serve as reference values. It can also be observed that the multimodal regression model performs the best overall.

Model	$R^2 - \log M_*$	R^2 – avgSFR	R^2 – sSFR	$R^2 - Z_{\mathrm{MW}}$	$R^2 - Z_{\mathrm{HP}}$	$R^2 - T_{\mathrm{AGE,MW}}$
Image Regressor	-0.576	0.393	0.443	0.316	0.424	0.170
Spectral Regressor	**0.033**	0.432	0.429	0.312	0.497	0.059
Image + Spectral Regressor	-0.054	**0.452**	**0.470**	**0.380**	**0.553**	**0.244**

Table 5.4: Regression model R^2 performance on different modalities.

[4] We run t-SNE on default hyperparameters for every experiment.

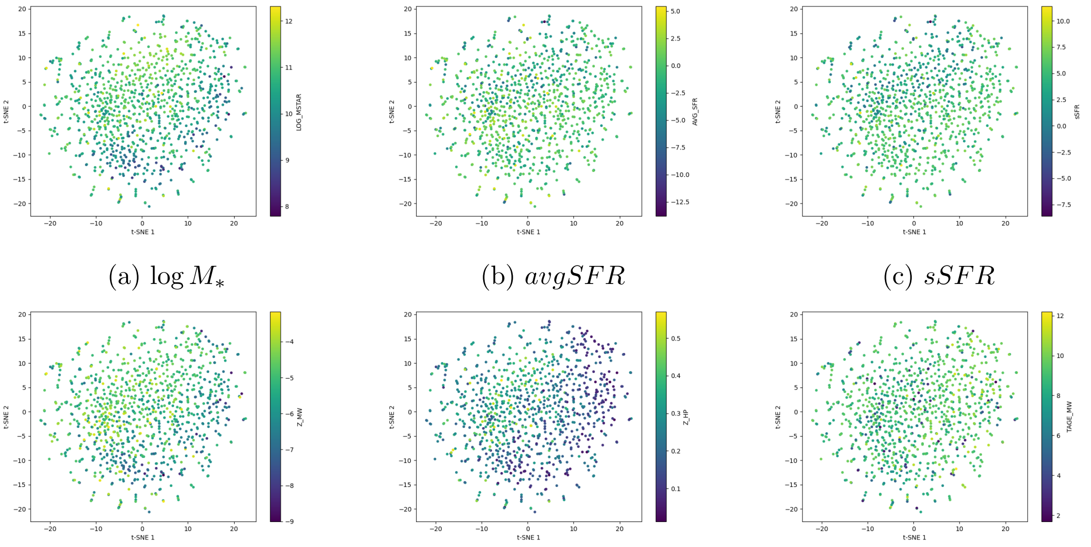
VAEs: Here, the VAEs applied to image and spectral modalities are examined in terms of their downstream task performance in predicting physical properties based on their latent values, serving as a reference for the DMVAVE, whose downstream performance is evaluated later.

Model	$R^2 - \log M_*$	$R^2 - \text{avgSFR}$	$R^2 - \text{sSFR}$	$R^2 - Z_{\text{MW}}$	$R^2 - Z_{\text{HP}}$	$R^2 - T_{\text{AGE,MW}}$
Image VAE	$-\mathbf{0.437}$	0.105	**0.150**	**0.247**	**0.470**	**0.047**
Spectrum VAE	-17.824	**0.130**	-0.160	-1.568	0.276	-1.575

Table 5.5: VAE downstream performance R^2 for image and spectrum modalities.

The results are visualised in table 5.5. It is evident that the image VAE performs best on the downstream task of predicting redshift Z_{HP}. The other property predictions perform worse, and $\log M_*$ achieves the lowest performance. Compared to that, the VAE performs worse on spectral data for the physical property prediction almost everywhere, as seen in table 5.5. This suggests that the model has greater difficulty in extracting relevant features from the spectra. As already observed in fig. 5.8 and fig. 5.10, the performance depends on the latent size. Especially for spectra, a larger latent size harms performance, which could be due to non-robust embeddings in unnecessarily large latent spaces.

These performance differences between properties, as well as the performance drop, can also be studied by visualising the latent space and colouring the respective latent points by their physical property values, as shown in fig. 5.17. The reason for this is that the latent values are distributed in a high-dimensional space, and when they are well separated in this space, it is easier for the downstream model to make a prediction.

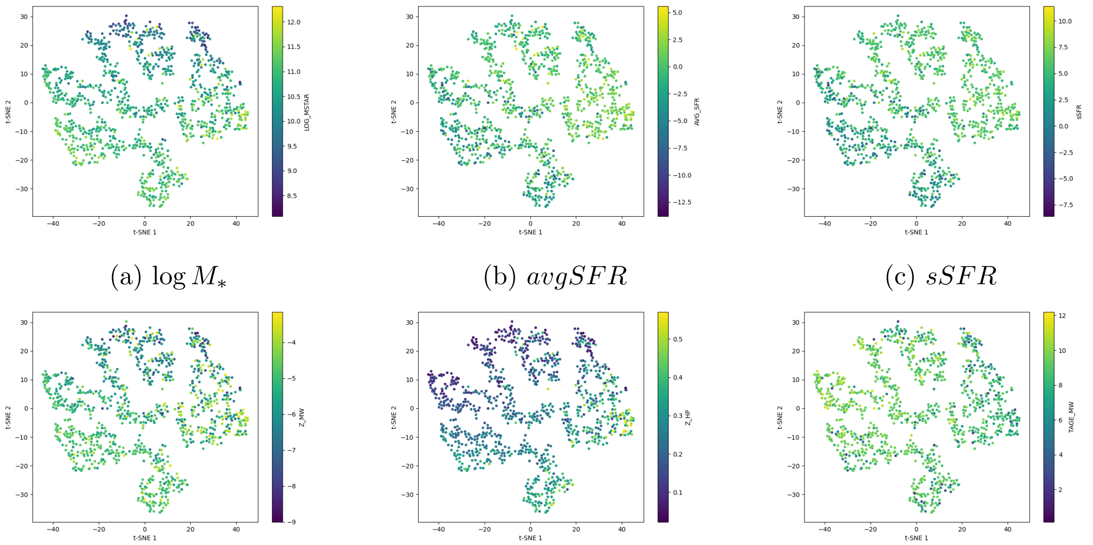


(a) $\log M_*$ (b) $avgSFR$ (c) $sSFR$

(d) Z_{MW} (e) Z_{HP} (f) $T_{AGE,MW}$

Figure 5.17: Visualisation of the latent space of the VAE on image data using t-SNE, where datapoints are coloured by their corresponding physical property.

It is visible that the latent points are not clustered and that the physical properties are not always well-separated. For the redshift Z_{HP}, a clear separation between high and low values is visible, while this is less visible for the other properties. In contrast to this, we also look at the latent space structure of the VAE with spectral data fig. 5.18.



(a) $\log M_*$

(b) $avgSFR$

(c) $sSFR$

(d) Z_{MW}

(e) Z_{HP}

(f) $T_{AGE,MW}$

Figure 5.18: Visualisation of the latent space of the VAE on spectral data using t-SNE, where datapoints are coloured by their corresponding physical property.

It is clearly visible that the latent space is significantly more clustered. Here, we can also observe a similar trend, where the redshift values are more distinct compared to the other properties. After evaluating the direct regression performance and downstream performance of the VAEs on single modalities, we proceed to the DMVAEs.

DMVAE with CLUB: Here, we evaluate how the information for different property prediction tasks is distributed among latent spaces. To achieve this, we first evaluate the downstream task on all subsets of the latent space.

R^2 Latents	$\log M_*$	avgSFR	sSFR	Z_{MW}	Z_{HP}	$T_{\mathrm{AGE,MW}}$
$\mathbf{Z_{p1}} \cup \mathbf{Z_{p2}} \cup \mathbf{Z_s}$	0.357	0.312	0.392	0.238	0.553	0.113
$\mathbf{Z_{p2}} \cup \mathbf{Z_s}$	0.252	0.303	0.372	0.244	0.553	0.128
$\mathbf{Z_{p1}} \cup \mathbf{Z_s}$	0.341	0.246	0.329	0.233	0.501	0.077
$\mathbf{Z_{p1}} \cup \mathbf{Z_{p2}}$	0.292	0.281	0.362	0.206	0.519	0.113
$\mathbf{Z_{p1}}$	-0.009	0.038	0.111	0.160	0.318	0.013
$\mathbf{Z_{p2}}$	0.078	0.230	0.315	0.133	0.419	0.076
$\mathbf{Z_s}$	0.028	0.255	0.349	0.174	0.478	0.024

Table 5.6: DMVAE with CLUB downstream performance of all latent subspaces shown in R^2 values.

The results are visualised in table 5.6. This can now be used to evaluate which unique/shared components from the modalities contain how much information about the corresponding property. The best performance is always expected for the case containing all subspaces, because this contains all unique, redundant and synergistic information that can complement each other for a downstream task. When comparing this performance to the performance of leaving one latent space out, we can study its impact by how much the performance drops when compared to the case when all latents are apparent. By removing one latent space, the impact of its information on the property prediction can be evaluated. When examining a specific property, such as $\log M_*$, we can observe that the performance drops most when omitting the private 1 latent space, compared to when all latent spaces are utilised. This shows the significant impact of the unique features from modality 1 on the downstream task. When examining the performance of $\log M_*$ for the corresponding single latent spaces, it is apparent that they are significantly lower than their joint performance of 0.357, indicating that both contain complementary and synergistic information, which collectively contain most of the information about the property. Compared to the VAEs, the downstream performance on all joint latent spaces is better in terms of all properties. Also, we can compare the results more fairly with the VAEs in cases where only private and shared latent spaces are used, $(p_1 + s)$ for the image VAE and $(p_2 + S)$ for the spectral VAE, which is information that could also be extracted from single modalities (except the synergistic ones). When comparing these respective performances in table 5.6 with table 5.5, we observe that the DMVAE consistently outperforms the VAE in this case as well, which could be attributed to the shared features capturing more complementary information usable for the downstream task. Compared to the performance of the regression models, the DMVAE with CLUB comes close but does not fully reach the capabilities of the multimodal regression model for most properties, see table 5.4. We now apply the second SHAP-based method, described in chapter 4, to evaluate the impact of each latent space on the downstream task. The results are shown in fig. 5.19.

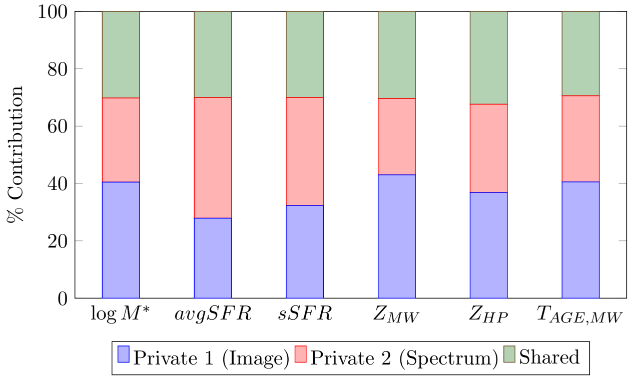
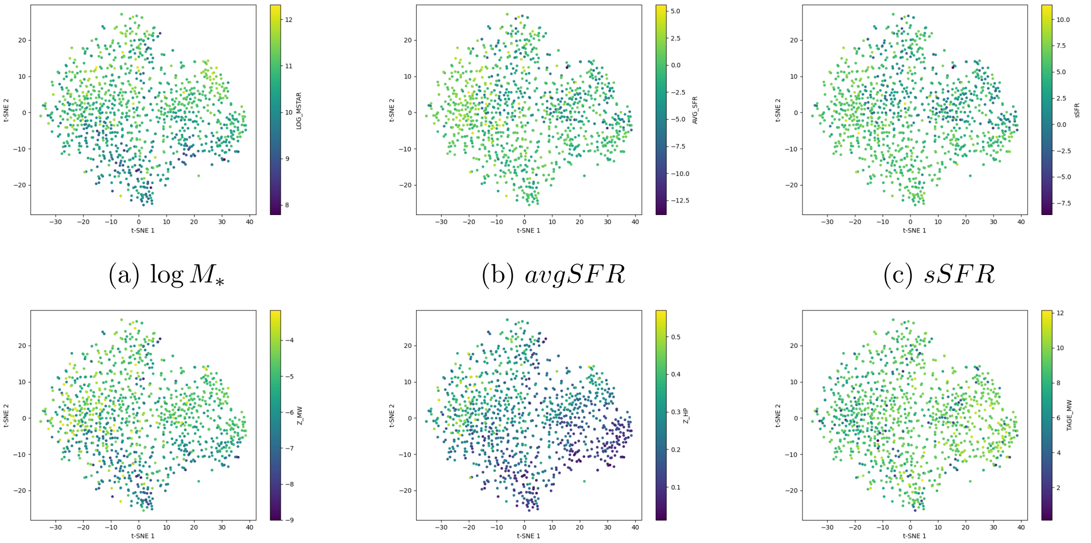


Figure 5.19: Contributions of each latent space to regression performance for DMVAE with CLUB computed using the method introduced in chapter 4 based on SHAP.

This method should directly approximate the contributions of each latent space and exhibit similar trends to those shown in table 5.6, which it largely does. For the $\log M_*$ property, this method also predicts that most information comes from the private 1 latent spaces. However, sometimes this method also disagrees with the subset-based method. Here, for example, the contribution of the unique spectrum information to the redshift (Z_{HP}) is underestimated when compared to table 5.6, which shows the most significant performance drop when private 2 is left out. Here, we also evaluate the structure of the joint latent space, which separates all properties reasonably well fig. 5.20. Again, the redshift Z_{HP} has the most apparent separation.



(a) $\log M_*$ (b) $avgSFR$ (c) $sSFR$

(d) Z_{MW} (e) Z_{HP} (f) $T_{AGE,MW}$

Figure 5.20: Visualisation of the latent space of DMVAE with CLUB using t-SNE, where datapoints are coloured by their corresponding physical property.

DMVAE without CLUB: To investigate how much of the previous results are attributed to the CLUB loss, we also examine these properties without the CLUB loss. Again, the downstream task performance is computed on all latent space combinations.

R^2 / Latents	$\log M_*$	avgSFR	sSFR	Z_{MW}	Z_{HP}	$T_{\mathrm{AGE,MW}}$
$\mathbf{Z_{p1}} \cup \mathbf{Z_{p2}} \cup \mathbf{Z_s}$	0.253	0.388	0.406	0.380	0.694	0.135
$\mathbf{Z_{p2}} \cup \mathbf{Z_s}$	0.138	0.293	0.321	0.324	0.656	0.110
$\mathbf{Z_{p1}} \cup \mathbf{Z_s}$	-0.088	0.128	0.176	0.293	0.511	0.106
$\mathbf{Z_{p1}} \cup \mathbf{Z_{p2}}$	0.125	0.395	0.427	0.345	0.685	0.081
$\mathbf{Z_{p1}}$	-0.709	0.070	0.137	0.224	0.447	0.001
$\mathbf{Z_{p2}}$	0.274	0.287	0.324	0.239	0.618	0.135
$\mathbf{Z_s}$	0.477	0.061	0.140	0.301	0.309	0.063

Table 5.7: DMVAE without CLUB downstream performance of latent subspaces shown in R^2 values.

In table 5.7, it is visible that the private representations have the most significant influence on the downstream task performance. This is especially apparent when comparing the performance to table 5.6. It is also visible on almost all properties that the downstream performance changes only marginally when the shared latent space is excluded. Similar trends are visible when visualising the contributions using SHAP in fig. 5.21.

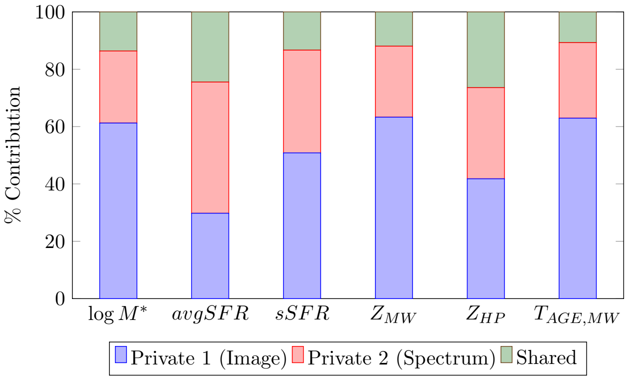
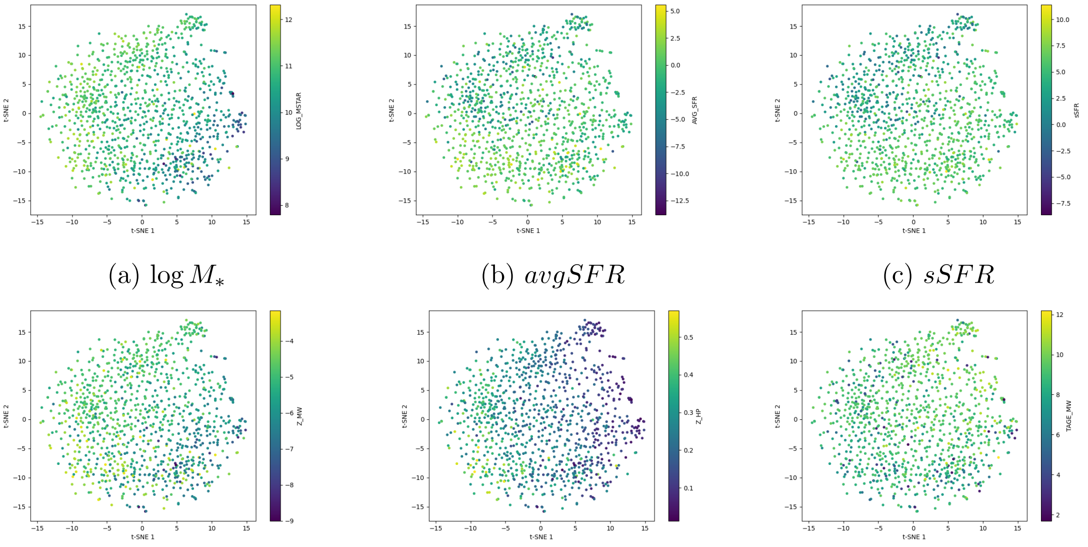


Figure 5.21: Contributions of each latent space to regression performance for DMVAE without CLUB computed using the method introduced in chapter 4 using SHAP.

It is, however, important to note that here the contributions are less meaningful, since the MI between latent spaces is not minimised. However, the trend is visible that the shared latent space has only a small contribution to the downstream task, which aligns with the subset-based method.



(a) $\log M_*$

(b) $avgSFR$

(c) $sSFR$

(d) Z_{MW}

(e) Z_{HP}

(f) $T_{AGE,MW}$

Figure 5.22: Visualisation of the latent space of DMVAE without CLUB using t-SNE, where datapoints are coloured by their corresponding physical property.

When evaluating the joint latent space, it is again evident that the model learns a latent space that can separate physical properties reasonably well, as shown in fig. 5.22. To directly study how well the mutual information minimisation works, we study the mutual information between all latent spaces by evaluating lower and upper bounds using InfoNCE and variational CLUB to approximate these. Note that these bounds only indicate trends and cannot accurately compute the exact mutual information value. For this, we train the neural networks needed for these methods for 50 epochs on the sampled latent values on the test dataset.

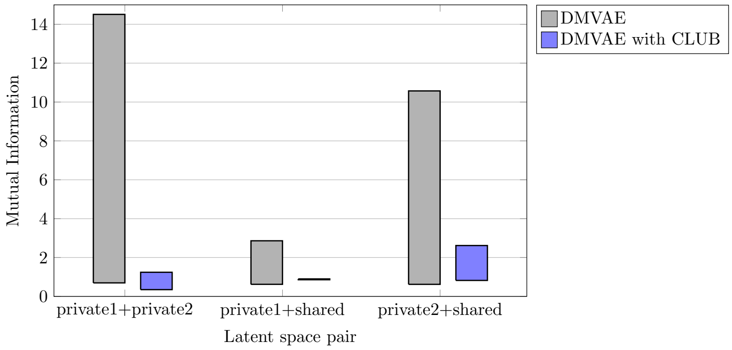


Figure 5.23: Comparison of InfoNCE lower bounds and CLUB upper bounds on mutual information for each latent space pair across models.

The results are visualised in fig. 5.23. The CLUB loss succeeds in minimising MI. It can significantly reduce the MI between the private latent spaces, which the model is only implicitly trained to do, and between the private and shared latent spaces. Without CLUB, there is considerable leakage and redundancy between the private information spaces. Between private 1 and shared information, the MI is already low without CLUB loss, possibly because images naturally contain many data points for which a large latent space is necessary, as we have previously seen. Here, the model needs to encode more information into the latent space, resulting in fewer redundancies and possibly more independent representative features, which could lead to reduced mutual information between the latent spaces. Between private 2 and shared information, the latent size is larger $(60 + 60)$ than necessary for representing spectra, leading to more redundancies. CLUB loss can minimise all of these redundancies.

5.3.5 Experiment 5: Physical-model-based decoder

Now, we test whether the inclusion of a physical model for galaxy images in the image decoder can help guide the latent space to learn a more semantically meaningful geometric representation, and if so, whether it influences downstream task performance. By using the differentiable model, the image encoder should learn to predict the correct underlying physically meaningful parameters, which are then input to the physical model to predict the shape of the galaxy.

Outcome: When incorporating a physical model in the decoder, the framework can predict meaningful geometric parameters approximately, while decreasing reconstruction performance for images. The model relies more heavily on private information from spectra and delivers downstream performance, as shown in table 5.8, which is inferior to the DMVAE with CLUB.

Model	$R^2 - \log M_*$	$R^2 - \text{avgSFR}$	$R^2 - \text{sSFR}$	$R^2 - Z_{\text{MW}}$	$R^2 - Z_{\text{HP}}$	$R^2 - T_{\text{AGE,MW}}$
DMVAE with CLUB+phy. decoder	0.135	0.245	0.326	0.179	0.552	0.099

Table 5.8: DMVAE with CLUB and physical-model based decoder downstream performance R^2.

Details: We now describe how we arrive at this downstream task performance in detail, by first describing the differentiable physical model and the architecture for incorporating it into a **physical-model-based image decoder**. This architecture uses a model G that approximates a physical process Ψ which depends on physical, meaningful parameters p to reconstruct the measurement x as described by [ACC20]:

$$G(p) \approx x \leftarrow \Psi(p)$$

In an autoencoder framework, part of the encoder predicts these physical parameters with an encoder function $f(x) = \hat{p}$. Then an autoencoder decodes $\hat{p}$ back into the reconstructed data item $\hat{x} = g(\hat{p})$. By replacing the decoder g with a differentiable physics model G, the encoder learns to predict the true underlying parameters of the model $f(x) = \hat{p} \approx p$ such that $G(\hat{p}) = x$ [ACC20]. Since the model is trained in an unsupervised way, we do not need the corresponding true model parameters p. Using this approach, the framework learns to predict the physical parameters, which helps guide the latent space to be more semantically meaningful. An image of a galaxy can be modelled by an exponential profile that is oriented, and it depends on the four parameters $p = \begin{bmatrix} I_0 & A & e & \theta \end{bmatrix}^T$. Here, I_0 sets the maximum intensity, A scales the length of the major axis, e controls how elliptic the profile is, and θ sets the galaxy's orientation. Based on these parameters, the intensity I can be formulated as follows [ACC20], [TK21]:

$$I(r) = I_0 exp(-r'),$$
$$r' = \sqrt{\left(\frac{x'}{A}\right)^2 + \left(\frac{y'}{B}\right)^2},$$

with the ellipticity e

$$B = A(1 - e),$$

where

$$x' = X_i cos\theta - Y_j sin\theta \quad \text{and} \quad y' = X_i sin\theta + Y_j cos\theta$$

and (X_i, Y_j) are uniformly sampled coordinates from a 128×128 grid from $[-1, 1] \times [-1, 1]$. We now build a decoder model based on this intensity model, which can then be used in conjunction with an encoder model that predicts physical parameters, as well as a classical VAE-like latent space for capturing residual features. This model is visualised in fig. 5.24.

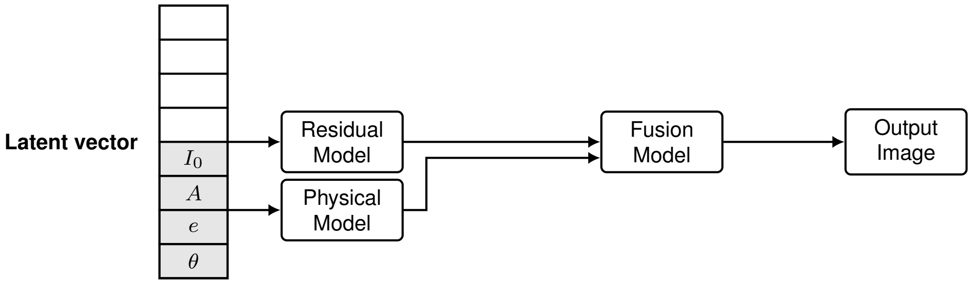
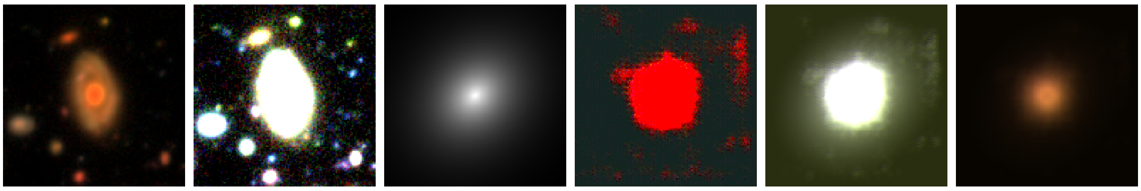


Figure 5.24: Physical-model-based decoder model architecture.

This decoder predicts images based on a latent space that is divided into physical parameters and a normal VAE-like latent space. The physical model returns a 1-dimensional intensity map of size 128 × 128. Since the physical model only includes a simplified representation of galaxy profiles, we need to use an additional network, based on the previously used decoder architecture, to predict the residual features that the model does not contain, to reconstruct images more accurately. This is done in conjunction with a fusion network, which is a simple CNN that utilises the model's intensity map, combined with the residual image, to reconstruct an RGB image. Then, the encoder learns to predict such physical parameters as well as latent values representing the residual data.

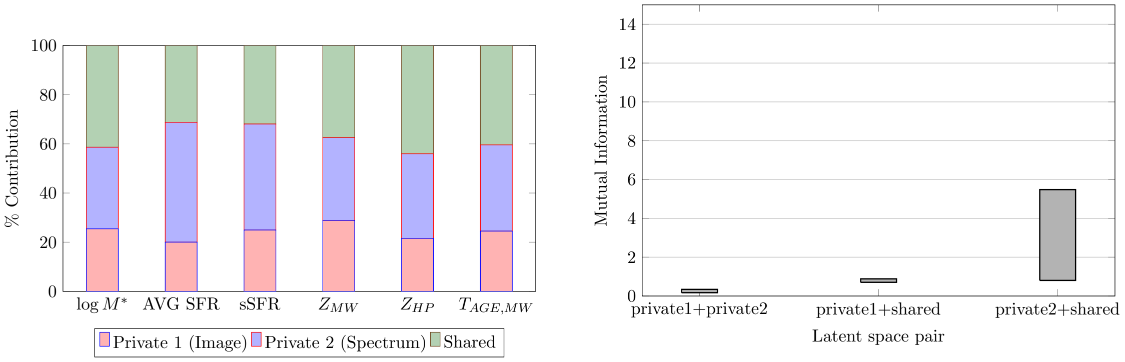
DMVAE with CLUB: Here, we investigate the inclusion of this physical-model-based decoder in the DMVAE with CLUB. The four parameters are predicted as part of the private latent space of the images. To guide the four latent parameters in a meaningful way, they only receive gradients from the physical model and the CLUB loss for disentanglement, while being excluded from the KL-divergence penalty and gradients from the residual neural network. To ensure that the model actively uses the physical model, it is necessary to use a residual model that is less powerful than the original decoder model. This is achieved by reducing the channel count of the original decoder to one-quarter of its original value. During testing, it showed that this was necessary for the physical model to learn meaningful parameters that predict an intensity map closely mirroring the original image, which can then be incorporated into the resulting generated image. This model was used to test downstream task performance.

In table 5.8, it is demonstrated that we achieve slightly worse downstream performance than the DMVAE with CLUB and without the physical decoder. The prediction performance is also visualised in more detail in the Appendix fig. A.10. As we mainly examine downstream task performance here, it is relevant that the physical model guides the latent space to contain physically meaningful values. These values, however, do not appear to improve downstream task performance, as the DMVAE with CLUB and without the physical model also captures the galaxy's shape in its latent variables, with better image reconstruction performance. The functionality of the decoder after training is illustrated in fig. 5.25, which shows how the physical model is applied to a test sample. The encoder appears to roughly estimate the parameters of the galaxy's shape, although with some inaccuracies. It is important to note that the physical model tries to predict the shape of the normalised galaxy images. Because some galaxies are irregularly shaped or contain unusual objects that overlap with the central galaxy, the model sometimes struggles to predict a fitting intensity map. Meanwhile, the neural network for residuals predicts the shape of galaxies and other objects that are not centred. The combination of both results in an image that primarily contains features from the neural network, with only minor contributions from the physical model. During training, it could be observed that for some runs, the model completely disregarded the physical models.



(a) Original (b) Original (normalised) (c) Physical model (d) Neural residual (e) Fused image (normalised) (f) Reconstructed

Figure 5.25: Relevant intermediate images produced by physical-model-based decoder.



Contributions to downstream predictions.

Mutual information lower and upper bound estimates.

Figure 5.26: DMVAE with CLUB and physics-based decoder plots.

Additionally, the contributions of all latent spaces to the property prediction task, as well as the mutual information between latent spaces, were evaluated in fig. 5.26. It shows that the private spectral information has a larger impact on the task than private image information and that MI was well minimised. A reason for the increased influence of private spectral features could be that image reconstruction and corresponding feature extraction perform worse due to the physical model, which is complicated for the model to use constructively. Then, the downstream model could learn to rely more on the spectrum modality instead. A solution could potentially involve longer training, different hyperparameters, or additional regularisation objectives, as in [TK21].

5.3.6 Summary

In the first five experiments, image and spectral galaxy data were used to compare single- and multimodal data on reconstruction, cross-reconstruction, and downstream tasks for VAEs, DMVAEs, and regression models. The results are shown in table 5.9.

	Regression models			**VAEs**		**DMVAEs**		
Metric	**Image**	**Spectrum**	**Image + Spectrum**	**Image**	**Spectrum**	**Without CLUB**	**With CLUB**	**With CLUB+phy. decoder**
MSE Image	-	-	-	**0.135**	-	**0.129**	0.229	0.42
MSE Spectrum	-	-	-	-	**0.263**	**0.27**	0.287	0.286
Cr-recon MSE Image	-	-	-	-	-	**1.089**	20.399	1.337
Cr-recon MSE Spectrum	-	-	-	-	-	**1.031**	15.573	1.333
$R^2 - \log M_*$	-0.576	**0.033**	-0.054	$\mathbf{-0.437}$	-17.824	0.253	**0.357**	0.135
R^2 – avgSFR	0.393	0.432	**0.452**	0.105	**0.130**	**0.388**	0.312	0.245
R^2 – sSFR	0.443	0.429	**0.470**	**0.150**	-0.160	**0.406**	0.392	0.326
$R^2 - Z_{\mathrm{MW}}$	0.316	0.312	**0.380**	**0.247**	-1.568	**0.38**	0.238	0.179
$R^2 - Z_{\mathrm{HP}}$	0.424	0.497	**0.553**	**0.470**	0.276	**0.694**	0.553	0.552
$R^2 - T_{\mathrm{AGE,MW}}$	0.170	0.059	**0.244**	**0.047**	-1.575	**0.135**	0.113	0.099

Table 5.9: Comparison of all metrics (reconstruction, cross-reconstruction, downstream performance) across all models investigated.

The regression performance should serve as an upper bound on the performance that could be expected from the VAE/DMVAE models in the optimal case, when either images or spectra are given, and for the multimodal case. Here, it is visible that for some properties, such as stellar mass, spectra perform better, while for others, like age, images perform better. However, in most properties, the multimodal regression model performs best.

The VAEs are often unable to achieve the same level of property prediction performance as the regression models. The image VAE achieves better downstream task performance on most metrics compared to the spectrum VAE.

The traditional DMVAE outperforms the VAEs on every property prediction. For redshift, it even exceeds the performance of the multimodal regression model, and only underperforms for stellar mass prediction. This indicates that the multimodal case contains more information that can improve accuracy on property predictions. The DMVAE with the additional CLUB loss exhibits lower performance on almost all metrics compared to the traditional DMVAE. This is also expected as the further loss terms result in a tradeoff with a lower impact of the reconstruction terms. The only exception is the stellar mass, which is predicted more accurately. Reaching the best performance, however, is not the primary goal of using the CLUB loss, as this term should primarily ensure clear feature separation between private and shared features to investigate where the information is stored. Adding a physical-model-based decoder to guide the latent space slightly harms the model's performance, compared to that of the DMVAE with CLUB. It can also be observed that the cross-reconstruction performance of DMVAE with CLUB is poor for this parameter configuration.

5.3.7 Experiment 6: Hyperspectral remote sensing data

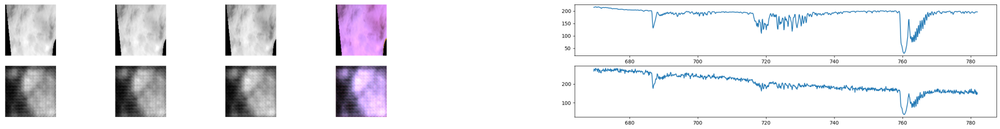
We now move on from images and spectra and look at hyperspectral data, which combines both data types in data cubes containing spatial and spectral components. Here, we aim to utilise hyperspectral data as a source for both image and spectral data due to their inherent relations, and investigate the amount of information these modalities contain about the underlying hyperspectral data. Images can be viewed as a special case of this data, obtained by selecting specific spectral bands, while spectra can be seen as a special case, obtained by averaging over the spatial domain. Using both modalities obtained from this, we apply our previously developed framework to the extracted RGB and spectral data, with the downstream task of reconstructing the hyperspectral data source. It is investigated where most of the information for this task is encoded. This is useful for studying whether expensive hyperspectral sensors can be approximated/replaced by a cheaper combination of an RGB/multispectral sensor and a spectrometer, which could contain similar information about the underlying data due to correlations.

Outcome: It can be observed that unique image features contain the most information about the hyperspectral data, while shared features contain less information, and the unique spectra information has the least amount of information. However, the hyperspectral reconstruction performance is limited when using images combined with spectra in this experiment.

Details: Here, the experiment is described in more detail. For the dataset, we investigate the remote sensing dataset Hyplant FLUO [5]. The Hyplant dataset contains airborne observations, taken by two sensors: DUAL and FLUO [SAC$^+$19]. DUAL captures a broad range of wavelengths (400-2500 nm) while FLUO captures data in the range of (670-780

[5] We also implemented the method on the MaNGA dataset; however, the data loading was too slow to complete.

nm) in a finer resolution [SAC+19]. We specifically use the radiance data. These datasets can be used to derive plant properties and retrieve sun-induced chlorophyll fluorescence (SIF), a signal emitted by plants that provides information about their photosynthetic properties [SAC+19]. We use the same setup as in previous experiments, so we do not have to repeat them. However, we increase all latent sizes to 200 as it is expected that the images can contain more information than for the galaxies. We also reduce the dataset size to $5,000$, the batch size to 32 and the number of epochs for the downstream task to 35, due to the large amount of data and processing required for the experiment.



(a) Top row: original extracted RGB image (lowest, middle and highest channel in hyperspectral cube); bottom row: reconstructed image. Each group shows, from left to right: red, green, and blue channels, followed by the combined RGB image.

(b) Top row: original mean spectrum (averaged across spatial domain in hyperspectral cube); bottom row: reconstructed mean spectrum. On the x-axis is the wavelength in nm and the y-axis shows the radiance in $\frac{mW}{m^2 srnm}$.

Figure 5.27: Visualisation of test dataset sample.

The reconstruction performance is visualised in fig. 5.27. As a downstream task, the hyperspectral cube is reconstructed to see how much information the image and spectrum capture about it. The results are shown in table 5.10.

Metric	$\mathbf{Z_{p1} \cup Z_{p2} \cup Z_s}$	$\mathbf{Z_{p2} \cup Z_s}$	$\mathbf{Z_{p1} \cup Z_s}$	$\mathbf{Z_{p1} \cup Z_{p2}}$	$\mathbf{Z_{p1}}$	$\mathbf{Z_{p2}}$	$\mathbf{Z_s}$
MSE Hyperspectral cube	364.92	403.17	369.47	345.74	351.06	730.72	408.49

Table 5.10: DMVAE with CLUB performance of all latent subspaces for HyPlant FLUO data.

The unique information from the RGB images contains most of the information about the hyperspectral cube. In contrast, the shared and unique spectral information contribute less to the task. For the individual representations, it is visible that unique image features perform best, while shared features also perform reasonably well, and unique spectral features on their own perform worst. This is the case because the spectrum only varies slightly, and the more essential features for reconstruction are the structure. When evaluating the mutual information fig. A.11, it can be seen that there are still large amounts of MI between private 1 and shared features, which explains that the shared information still captures a lot of information about the images. The reason for this is, as previously described, that minimising mutual information becomes more difficult for larger latent sizes, which we use here. We can conclude that further experiments are needed to evaluate whether the hyperspectral cube can be reconstructed, which may involve testing other hyperparameter configurations and possibly more complex models with larger dataset sizes and more training epochs. Due to time constraints, this isn't done here. We can, however, already observe that structural images seem to contain most of the information, and possibly evenly spaced multispectral images could further enhance the reconstruction performance. Whether it can, however, reach the precision necessary to be useful for methods based on hyperspectral data, such as SiF retrieval, remains to be evaluated.

5.4 Discussion

Previous experiments aimed to investigate the performance of the DMVAE with CLUB and the influence of the CLUB loss on the model, as well as to study two different multimodal physics datasets on their unique and shared features and their contributions to different downstream tasks. Additionally, the performance was compared between single and multimodal data.

The experiments show that the CLUB loss can minimise the mutual information between latent spaces while also slightly decreasing reconstruction and downstream task performance and strongly decreasing cross-reconstruction performance (for some parameter configurations). The experiments have also demonstrated that CLUB penalises the reconstruction performance for larger shared latent sizes, due to inherently more correlations between larger latent spaces leading to larger CLUB losses. This emphasises the need to find suitable latent sizes for a tradeoff of reconstruction, cross-reconstruction and low inherent MI between latent spaces. Meanwhile, the CLUB loss helps to increase the impact of the shared latent space on the downstream task performance, showing that it helps extract more shared features while reducing the information content in the private latent spaces. However, since the model is encouraged to learn shared features but penalised for extracting dependent representations, some unique or shared features that have causal relations to other latent representations can be lost. The SHAP- and the subset-based method can often reveal similar trends when evaluating contributions to a task. For the prediction performance, it was evident that it varies for different properties and that there is no clear trend for the latent sizes, indicating that large sizes, which capture more details of the modalities, do not always lead to increased prediction accuracy. This shows that even small latent sizes, which do not capture all relevant modality details, can often still capture a similar amount of meaningful information for downstream tasks.

Using the MMU dataset, we found suitable hyperparameters for the CLUB and KL-divergence weights as well as a reasonable learning schedule. When evaluating the MMU dataset, it was evident that the multimodal data consistently outperformed the single-modal data in downstream performance. The DMVAEs performed significantly better than the VAEs and achieved similar or even better performance than the regression models, which should serve as an upper bound on possible performance. The DMVAEs were able to separate physical properties well in the latent space. This demonstrates that multimodal data can increase accuracy in this case, highlighting the potential of combining image and spectral data to enhance task performance in other contexts. When introducing a physical-model-based differentiable decoder, a decrease in downstream performance can be observed. Other models, such as AstroCLIP [PLG^{+}24], can often reach better downstream performance for physical property predictions for galaxies, although on different datasets. Possible reasons for the low performance on certain metrics can include the preprocessing of the image and spectrum, which could remove essential information. Also, the dataset size may be too small and the model's complexity too low for it to reliably predict the physical properties. We additionally evaluated remote sensing hyperspectral data as a source of multimodal data. Here, it was apparent that the unique information of the structural image contains most information about the hyperspectral cube, while the unique spectral and shared information contribute less. The performance could be limited by the small dataset size, the inherent missing information for reasonable reconstruction of the hyperspectral data, the limited number of training epochs, and a model that is not complex enough to extract essential features and reconstruct the complex nature of the images.

Chapter 6

Conclusion and Outlook

This thesis aimed to investigate multimodal disentangled representation learning and utilise it to evaluate the influence of unique and shared information encoded across modalities on task performance, studying whether multimodal data can complement each other for increased accuracy in prediction tasks. For this, a multimodal framework was introduced and implemented based on the DMVAE, which included an additional CLUB loss term to minimise the mutual information between shared and private representations. This framework was applied to image and spectral physics data from the multimodal universe dataset, as well as to remote sensing hyperspectral data from the HyPlant FLUO dataset.

It was found that although the CLUB loss harms reconstruction and downstream performance, it can minimise redundancies and help evaluate more precisely the influence of shared and unique information on the task. CLUB was able to extract more shared information usable for the downstream task while also harming the reconstruction performance on larger shared latent sizes and reducing the information content in the private latent spaces. The addition of a physical-model-based decoder did not appear to have a positive impact on the performance metrics; however, it was able to approximate meaningful parameters.

This framework was able to evaluate the influence of the different modalities on downstream tasks such as physical property prediction and hyperspectral data reconstruction. It demonstrated a varying impact of shared and unique features on physical property prediction, as well as a significant impact of the unique feature of RGB images on hyperspectral reconstruction. Overall, the experiments conducted showed significantly increased accuracy when using multimodal data when compared to single-modal data. This highlights the general potential of complementary multimodal (physics) data for improving prediction accuracy in scientific tasks.

However, this approach also has some problems. The model is only able to extract unique and shared information and thereby does not differentiate between redundant and synergistic information. The framework has a large number of interdependent hyperparameters, which may depend on the data used; therefore, these parameters must be determined in advance. It is desired that the latent space is large enough to represent all essential features of the data, which also needs to be determined in advance, while keeping it low enough to minimise inherent correlations between the latent spaces and thereby keep CLUB losses small. The results can vary due to noisy training caused by CLUB loss and noisy data.

Also, it could potentially remove relevant shared or unique information with causal relations to other latent spaces. Furthermore, the cross-reconstruction performance between spectra and images is unsatisfactory for too large latent spaces, and the hyperparameters need to be fine-tuned for this when desired. It is also important to note that the model utilises shared and unique information with respect to both modalities, rather than with respect to the target variable from the downstream task, as is done for PID, which must be taken into account during the interpretation of the results.

Future work could apply this framework to other physics datasets and compare it to other similar methods, such as PID. Additionally, the influence of the three KL-divergence components — mutual information, total correlation, and dimension-wise KL-divergence — may be further investigated [1]. Especially the first one could have a large impact. Possible extensions of this model include separate latent spaces for unique, redundant, and synergistic information, or an improved latent structure that is closer to a Gaussian distribution through the use of latent diffusion methods. Additionally, the causal relationships between shared and unique features may be further explored. Also, it could be evaluated if this model can be extended to three or more modalities.

Ultimately, this thesis has demonstrated the potential of complementary multimodal data for improving the accuracy of prediction tasks. The implemented framework provides functionality to study the unique and shared contributions of multimodal data to prediction tasks. Thereby, it offers the potential to optimise multimodal sensor setups in a given scientific experiment for a prediction task by providing a way to study how well the multimodal data complements each other for the task.

[1] The framework has the functionality for penalising the different KL terms differently already implemented.

Bibliography

[AAA+16] Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balland, Otger Ballester, et al. The desi experiment part i: science, targeting, and survey design. *arXiv preprint arXiv:1611.00036*, 2016.

[AAB+24] Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, Benjamin M Boyd, David Chemaly, Brian Cherinka, Ioana Ciucă, Miles Cranmer, Aaron Do, Matthew Grayling, et al. The multimodal universe: enabling large-scale machine learning with 100 tb of astronomical scientific data. *Advances in Neural Information Processing Systems*, 37:57841–57913, 2024.

[ACC20] Miguel A Aragon-Calvo and JC Carvajal. Self-supervised learning with physics-aware neural networks–i. galaxy model fitting. *Monthly Notices of the Royal Astronomical Society*, 498(3):3713–3719, 2020.

[AM] Hussein Aljlailaty and Mohammad M. Mansour. Earth-centered inertial (eci) frame. `https://www.researchgate.net/figure/Earth-centered-inertial-ECI-frame_fig8_346655136`. [Accessed 2025-09-02].

[CHD+20] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[Dia22] Noé Dia. vae-galaxy. `https://github.com/Ciela-Institute/vae-galaxy`, 2022. GitHub repository, accessed 2025-07-07.

[FIL+13] PE Freeman, R Izbicki, AB Lee, JA Newman, CJ Conselice, AM Koekemoer, JM Lotz, and M Mozena. New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434(1):282–295, 2013.

[HKT+23] ChangHoon Hahn, KJ Kwon, Rita Tojeiro, Malgorzata Siudek, Rebecca EA Canning, Mar Mezcua, Jeremy L Tinker, David Brooks, Peter Doel, Kevin Fanning, et al. The desi probabilistic value-added bright galaxy survey (provabgs) mock challenge. *The Astrophysical Journal*, 945(1):16, 2023.

[ICT23] Daiki Iwasaki, Suchetha Cooray, and Tsutomu T Takeuchi. Extracting an informative latent representation of high-dimensional galaxy spectra. *arXiv preprint arXiv:2311.17414*, 2023.

[KW$^+$13] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[LBF$^+$04] Robert Lupton, Michael R Blanton, George Fekete, David W Hogg, Wil O'Mullane, Alex Szalay, and Nicholas Wherry. Preparing red-green-blue images from ccd data. *Publications of the Astronomical Society of the Pacific*, 116(816):133, 2004.

[Lei24] Bastian Leibe. Advanced machine learning lecture. Lecture at RWTH, Computer Vision Institute, 2024. SS 2024.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[LP21] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 1692–1700, 2021.

[LZU$^+$17] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *Advances in neural information processing systems*, 30, 2017.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[MSSA23] Arnab Kumar Mondal, Ajay Sailopal, Parag Singla, and Prathosh Ap. Ssdmm-vae: variational multi-modal disentangled representation learning. *Applied intelligence*, 53(7):8467–8481, 2023.

[MVdBW10] Houjun Mo, Frank Van den Bosch, and Simon White. *Galaxy formation and evolution*. Cambridge University Press, 2010.

[oJ] National Astronomical Observatory of Japan. Hyper suprime-cam. `https://subarutelescope.org/Projects/HSC/forobservers.html`. [Accessed 2025-09-02].

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[PLG$^+$24] Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, et al. Astroclip: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, 2024.

[SAC$^+$19] Bastian Siegmann, Luis Alonso, Marco Celesti, Sergio Cogliati, Roberto Colombo, Alexander Damm, Sarah Douglas, Luis Guanter, Jan Hanuš, Kari Kataja, Thorsten Kraska, Maria Matveeva, Jóse Moreno, Onno Muller, Miroslav Pikl, Francisco Pinto, Juan Quirós Vargas, Patrick Rademske, Fernando Rodriguez-Morene, Neus Sabater, Anke Schickling, Dirk Schüttemeyer, František Zemek, and Uwe Rascher. The high-performance airborne imaging spectrometer hyplant—from raw images to top-of-canopy reflectance and fluorescence products: Introduction of an automatized processing chain. *Remote Sensing*, 11(23), 2019.

[SNM16] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

[STZ18] Kevin Schawinski, M Dennis Turp, and Ce Zhang. Exploring galaxy evolution with generative models. *Astronomy & Astrophysics*, 616:L16, 2018.

[TCERP$^+$23] Juan Terven, Diana M Cordova-Esparza, Alfonso Ramirez-Pedraza, Edgar A Chavez-Urbiola, and Julio A Romero-Gonzalez. Loss functions and metrics in deep learning. *arXiv preprint arXiv:2307.02694*, 2023.

[TK21] Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *Advances in Neural Information Processing Systems*, 34:14809–14821, 2021.

[VFHM17] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

[WB10] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[WCT$^+$24] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[WG18] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

[WLG$^+$22] Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin Gal, et al. Galaxy zoo decals: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 2022.

[XSdS$^+$23] Quanfeng Xu, Shiyin Shen, Rafael S de Souza, Mi Chen, Renhao Ye, Yumei She, Zhu Chen, Emille EO Ishida, Alberto Krone-Martins, and Rupesh Durgesh. From images to features: unbiased morphology classification via variational auto-encoders and domain adaptation. *Monthly Notices of the Royal Astronomical Society*, 526(4):6391–6400, 2023.

[YSNH20] Ravindra Yadav, Ashish Sardana, Vinay Namboodiri, and Rajesh M Hegde. Bridged variational autoencoders for joint modeling of images and attributes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1479–1487, 2020.

Appendix A

Additional illustrations and resources

Code

The code for the framework implementation, experiments and evaluation methods can be found under `https://jugit.fz-juelich.de/ias-8/thesis-moritz-effen`. The code is designed to repeat all experiments that have been done directly. Additionally, the resulting data from the experiments is contained in this repository. For more details on the code and experiments, refer to the README and the corresponding code documentation.

Plots

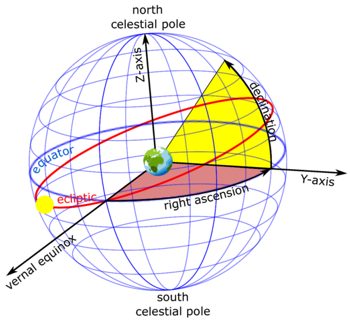


Figure A.1: Celestial sphere [AM].

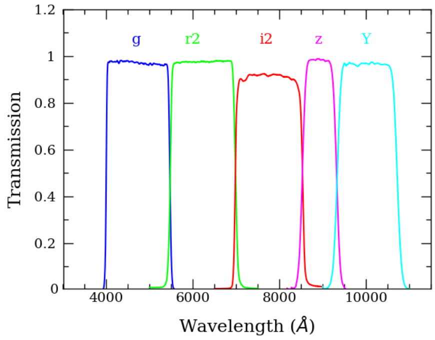


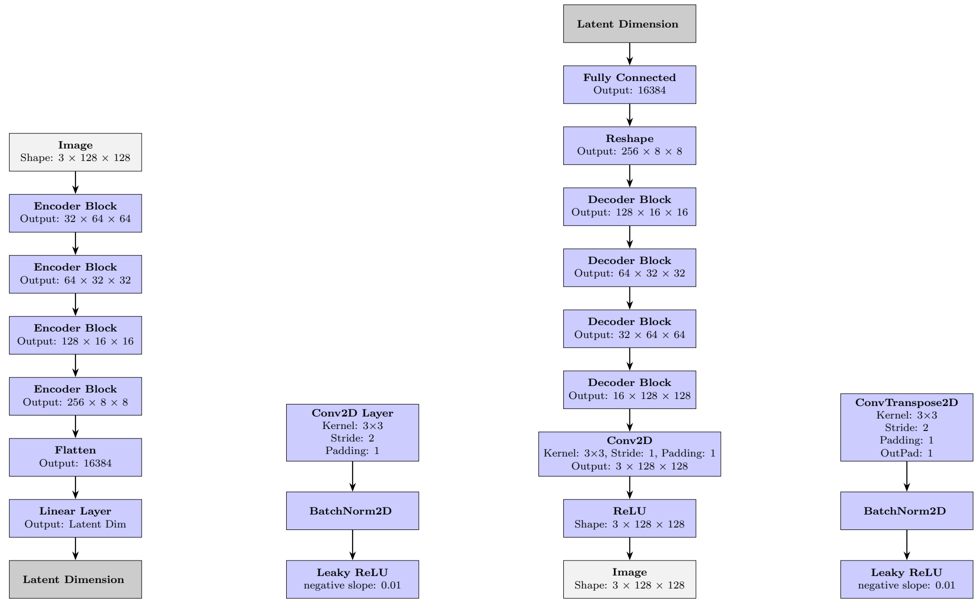
Figure A.2: HSC transmission filters [oJ].

A.1 Architectures

Here, the details of the models are described.

Image models

For the image encoder/decoder, we use a classical CNN structure. Here we use a similar structure to [XSdS+23], [Dia22]. Additionally, [ICT23] employed a similar encoder-decoder structure for images. The encoder consists of four encoder blocks, each containing a convolution, followed by a batchnorm and a ReLU. After these four blocks, the output is flattened and fed through a linear layer that outputs the latent space. Note that for the VAE, the linear layer predicts mean μ and log variance $\log(\sigma)$ and in the DMVAE, the linear layer predicts the μ_p and $\log(\sigma_p)$ for the private latent space and μ_s and $\log(\sigma_s)$ for the shared latent space. The dimension of the latent spaces can be modified. This is visualised in fig. A.3a. For the decoder, we mirror the encoder's structure. First, the latent variable is fed through a linear layer and then reshaped so that it can be fed into the transpose convolution layers. Here we have four decoder blocks, which consist of a transpose convolution followed by a batch norm layer and a leaky ReLU activation function. Then we have a final layer consisting of a convolution and a ReLU activation function for reconstructing the original image [XSdS+23]. This final layer enhances the decoder's power and should reduce checkerboard artefacts. This is visualised in fig. A.3c.



(a) Encoder architecture.

(b) Encoder block

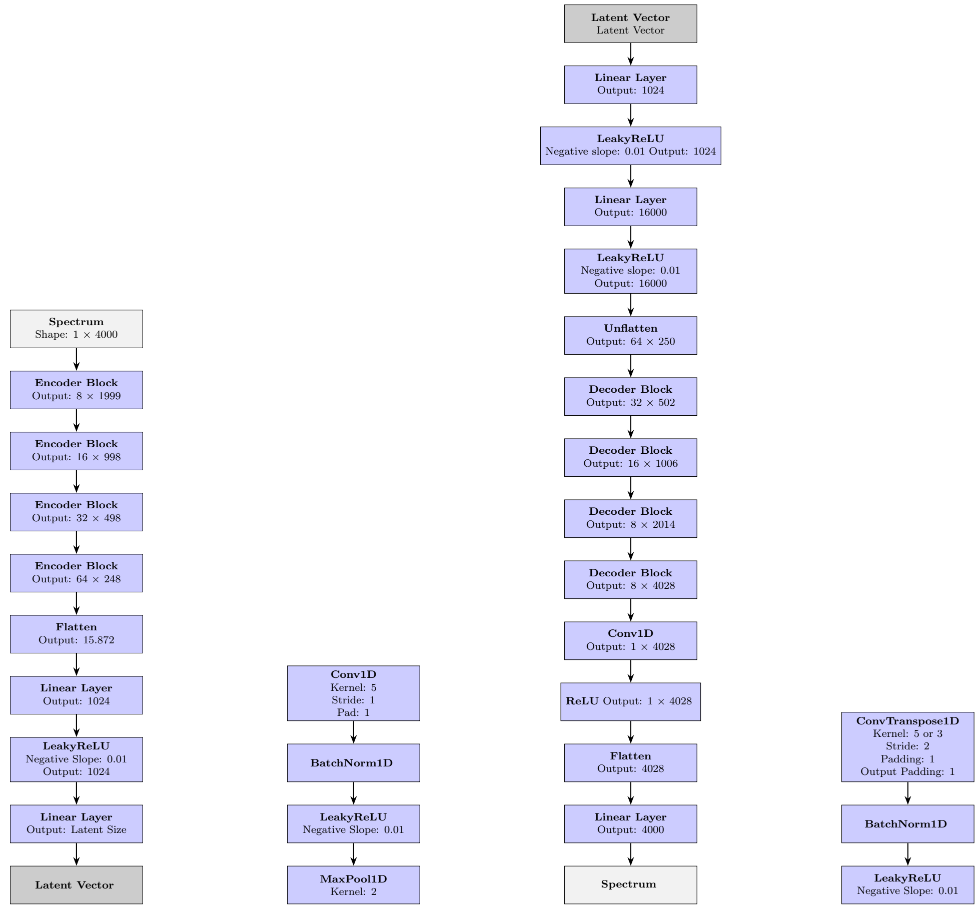
(c) Decoder architecture.

(d) Decoder Block

Figure A.3: Image encoder/decoder architecture, similar to [Dia22].

Spectum models

For the spectrum encoder and decoder, we use the architecture from [ICT23], which also consists of four decoder blocks, each containing a convolution, batch normalisation, leaky ReLU and downsampling with maxpooling. Afterwards, the output is fed through two linear layers, as shown in fig. A.4a. The decoder for the spectrum is again a version of the mirrored spectrum encoder, consisting of two linear layers and an unflatten layer that transforms the latent input into a spatial input for the subsequent steps. Here, four decoder blocks contain a transposed convolution, a batch normalisation and the leaky ReLU activation function. At the end, the final layers consist of a convolution, a flatten layer, and a linear layer, which outputs the spectrum. Similarly to the image decoder, this makes the decoder more powerful. Unlike [ICT23], we remove the activation function, allowing the model to fit the signal range. See the structure in fig. A.4c.



(a) Encoder architecture. (b) Encoder Block (c) Decoder architecture. (d) Decoder Block

Figure A.4: Encoder architecture for spectra.

Downstream model

The downstream model is a simple MLP for predicting physical properties. The architecture is from [PLG+24].

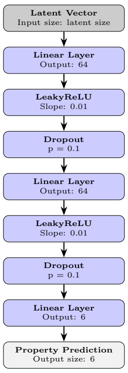
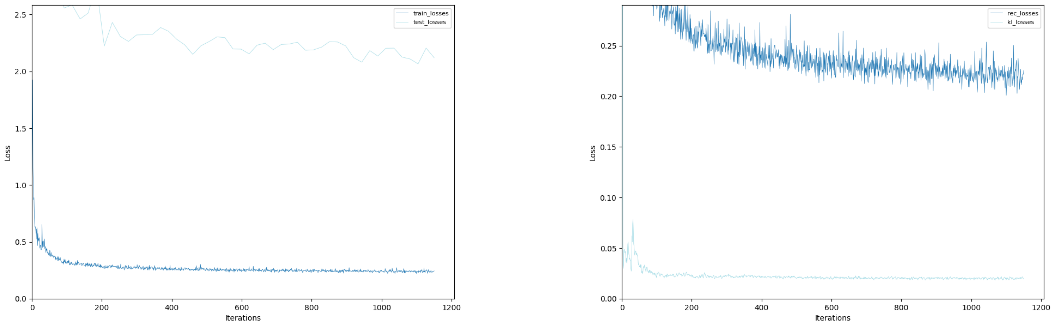


Figure A.5: MLP for physical property downstream regression.

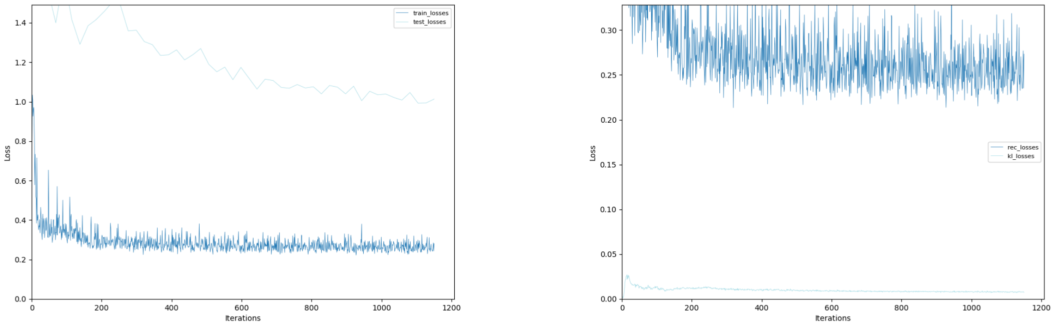
A.2 Experiment plots



(a) Visualization of train and test losses ($\beta_{KL} = 1$ in test loss).

(b) Decomposition of training loss according to training objective (scaling included).

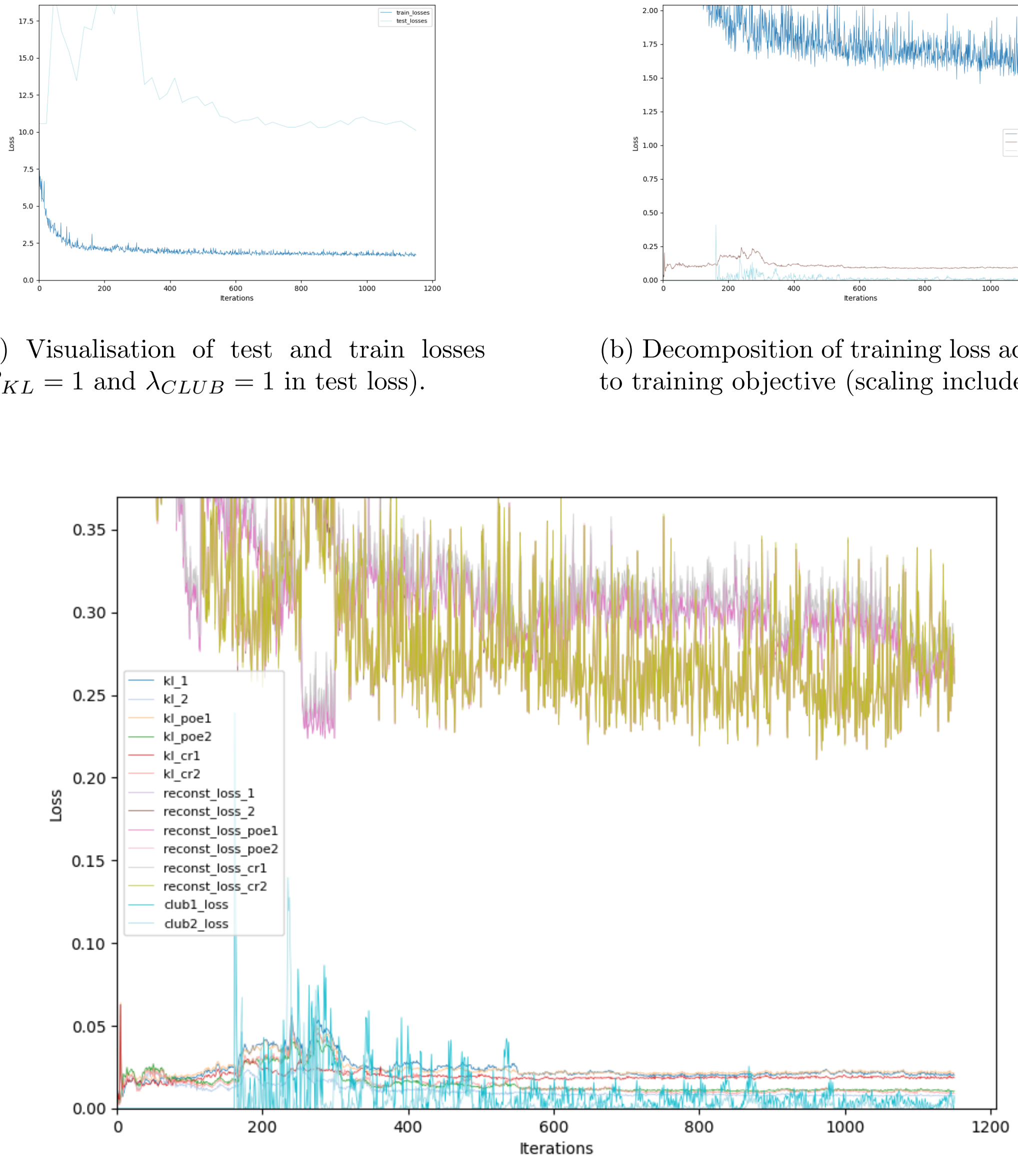
Figure A.6: Visualisation of test and training performance (VAE for images).



(a) Visualisation of test and train losses ($\beta_{KL} = 1$ in test loss).

(b) Decomposition of training loss according to training objective (scaling included).

Figure A.7: Visualisation of test and training performance (VAE for spectra).

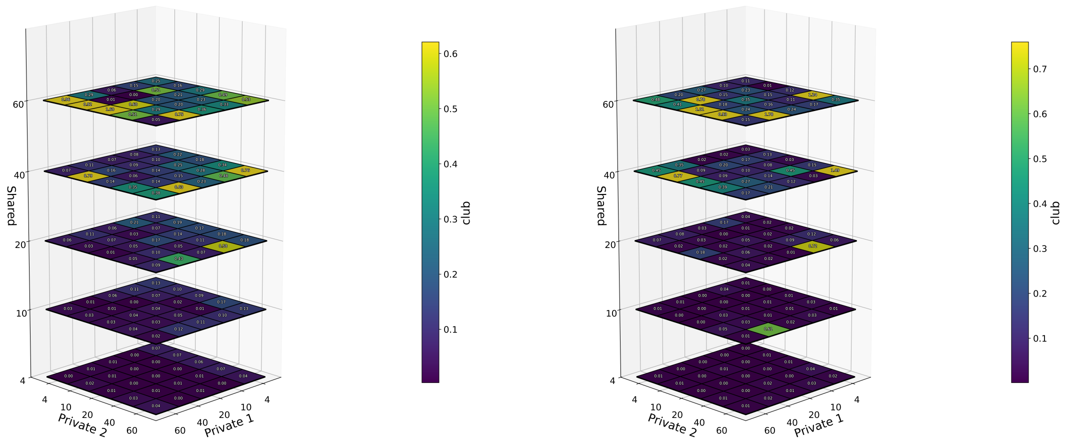


(a) Visualisation of test and train losses ($\beta_{KL} = 1$ and $\lambda_{CLUB} = 1$ in test loss).

(b) Decomposition of training loss according to training objective (scaling included).

(c) Decomposition of training loss according to detailed training objective (scaling included).

Figure A.8: Visualisation of test and training performance of DMVAE with CLUB loss and a latent size of 8 (Note that the losses can become significantly noisier for larger latent space sizes. For this, we refer to the folder containing all experimental data in the repository, where each training run has corresponding plots for showing the loss).



(a) CLUB 1 (b) CLUB 2

Figure A.9: Approximation of the CLUB loss, with mutual estimator trained on the train dataset and evaluation on the test dataset.

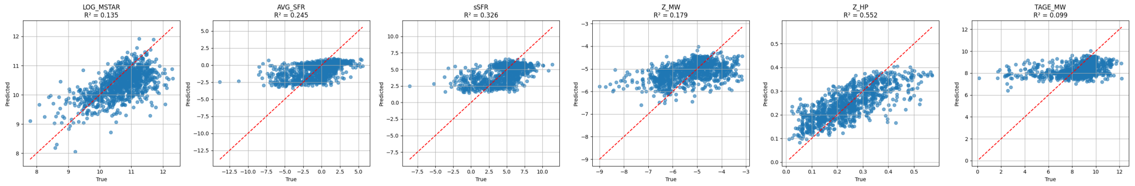


Figure A.10: Downstream performance of DMVAE with CLUB and physical decoder showing R^2 values.

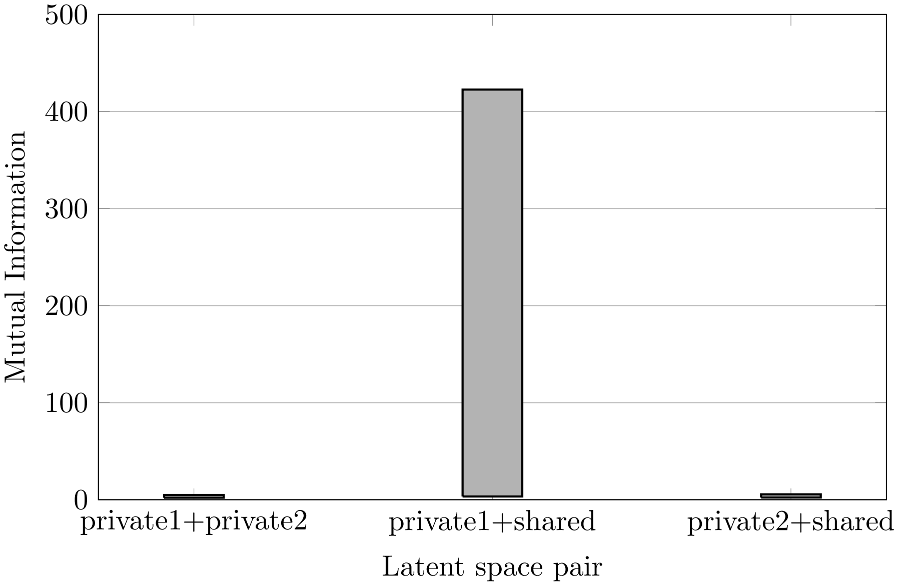


Figure A.11: Mutual information bounds for experiment 6.