Applications of Dimensional Reduction Methods to Galactic Images

A thesis presented in partial fulfillment of the requirements for the bachelor's degree at Graduate School of Science, Nagoya University

By

LOH Guan Hui Dillon

Student ID: 061801914

Laboratory of Galaxy Evolution

Division of Particle and Astrophysical Science,

Graduate School of Science

Supervisor: Dr. Tsutomu T. Takeuchi

June 2022

Nagoya University

Abstract

We introduce the use of unsupervised neural networks in extracting features from galactic images, and demonstrate how such latent image features could be used as a simple method for estimating the physical properties of a galaxy in the absence of other data. As an illustrative example, we ran a dataset of 12539 i band SDSS images through various types of unsupervised neural networks and showed that even with just 2 latent features, physical and morphological properties are distinctly clustered in latent space. We also compare the results of using unsupervised neural networks with methods explored by previous works, like Principal Component Analysis (Uzeirbegovic et al. 2020).

Introduction

Galaxies are, in essence, large systems of stars, stellar remnants, interstellar gas, dust, and dark matter held together by gravity (Phillipps 2005). The distribution of a galaxy's stellar matter - and in turn its light distribution - is called its 'morphology'. This morphological structure is what we observe visually when we look at galaxies, and comes in various shapes and sizes. However, while galactic morphology is largely unique for each galaxy, there are certain shared features commonly observed. which include bulges, spiral arms, and discs. Morphological classification systems, like the famous Hubble Tuning Fork, are based on such shared features. Galaxies which contain uncommon features are generally classified as 'Irregular' galaxies.

A galaxy's structure has been shown to be strongly related with various physical properties, such as stellar mass (Bundy et al. 2005) and star formation rate (SFR). For example, spiral galaxies tend to have higher associated SFRs when compared with elliptical galaxies. This is due to their spiral arms, which help trigger shocks in molecular gas clouds and cause them to undergo gravitational collapse when they pass through, hence accelerating star formation activity. In addition, a galaxy's morphology is also a reflection of its evolution - under the 'bottom-up' theories of galaxy formation, the earliest galaxies formed initially with spiral or disc structures, before eventually undergoing merger to become the larger elliptical galaxies. A galaxy's morphological properties could hence be viewed as a proxy for estimating its underlying or hidden properties.

To do so, we require an understanding of the relationship between a galaxy's morphol-

ogy and its physical properties. One way of doing so is by understanding how a galaxy's structure can be decomposed into more fundamental 'latent features' which exhibit correlations with different properties. Previous works by Uzeirbegovic et al. (2020) have shown that Principal Component Analysis (PCA) is one way of doing so, and has shown that most galaxies can be represented simply as unique sums of a small set of latent features. Some of these are immediately recognisable as well-understood features like spiral arms, center bulges, etc., while other latent features are more novel. We will attempt to apply this to a dataset of near-infrared spatially resolved galactic images, and examine the relationship between the principal components of a galaxy and its physical properties.

However, such methods are limited as they can only find linear manifolds of the higher dimensional space. If the underlying manifold is non-linear in nature, PCA will fail to find suitable latent vectors. As such, we are motivated to explore non-linear methods like neural networks too. In particular, we will make use of various autoencoder architectures (a deep unsupervised learning method) to study the latent features of the same dataset.

Data

Galaxy Catalogue

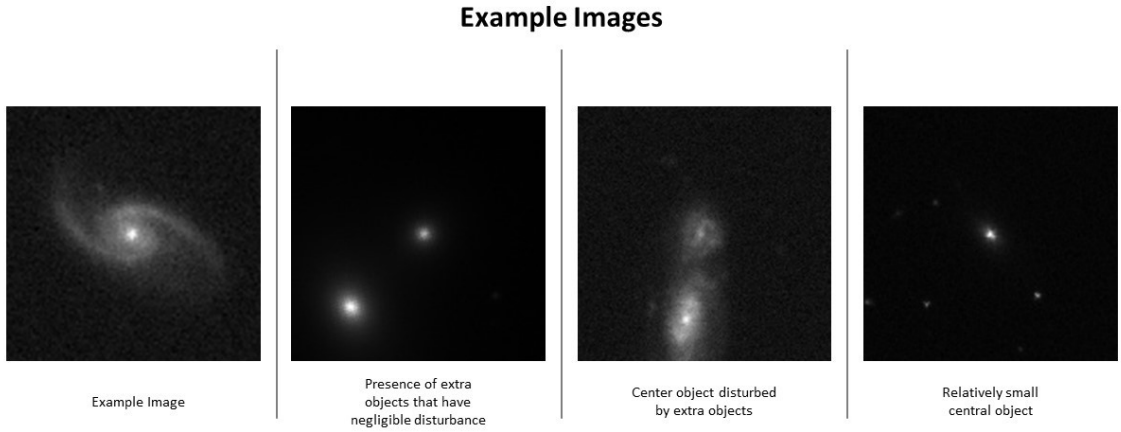
We utilise the Reference Catalog of galaxy SEDs (RCSED) dataset, which is a value-added join between the Sloan Digital Sky Survey (SDSS), Galaxy Evolution Explorer (GALEX), and UKIRT Infrared Deep Sky Survey (UKIDSS) datasets, as our main galaxy catalogue. In particular, this catalogue consists of low to intermediate redshift galaxies only ($0.007 < z < 0.6$).

We further used the Galaxy Zoo 2 dataset taken from the Galaxy Zoo project to filter and keep only face-on galaxies. The Galaxy Zoo 2 dataset was collected by showing human collaborators images of galaxies, and having them vote for whether a feature is applicable to the galaxy. If more than 80% of the votes for a classification is true, a 'clean' flag is raised. We used this clean flag to determine if a galaxy was face-on. This was done as the more interesting features of a edge-on galaxy are not visible to us due to our position, causing most of them to have similar sharp elliptical profiles.

Galaxies with 'oddities', such as those undergoing merging, containing dustlanes, or whose position overlaps with other galaxies behind it, were also removed in the same way, as they were considered to be outliers.

Galactic Images

We downloaded SDSS Data Release 17 i-Band images using NASA's SkyView API. i-Band images were chosen as these wavelengths, which are near-infrared, are sensitive to the emissions of long-living stars. As such, as the long-term structure of a galaxy is defined by the distribution of such long-living stars, this band would give us the best representation of our galaxy's morphology. The downloaded single-channel gray scale



images were set to be 150 x 150 pixels in dimension. As the SkyView application was not made to handle batch downloads, we were limited to 12539 galaxies downloaded within a reasonable period of time.

Image Preprocessing

Since the relative size of the galaxies in each image was different, we had to pre-process them to ensure that focal galaxy took up the same proportion of each image's dimensions. To do so, we first conducted Sersic Fitting via the statmorph python package on all the objects present in each image, before isolating the sersic function of the center-most object. Each fit generates a 2D sersic profile (i.e. 2 sersic fits along perpendicular directions), which allows us to create ellipticals with the longest and shortest edges being along each

perpendicular sersic fit.

We then rotated the images according to the angle between the y-axis and the major axis of this fitted elliptical, such that the major axis of all galaxies are vertically aligned. We then cropped down each image to the 1.5 times the galaxy's petrosian half-light radius, before finally scaling all the images to a size of 100 x 100 pixels. A visualisation of the steps can be seen in figure X. We then flattened each image to 10000 dimensional vectors, which is required for passing into our neural network. Finally, each image vector is than stacked on each other to form a 12539 rows x 10000 columns image data matrix.

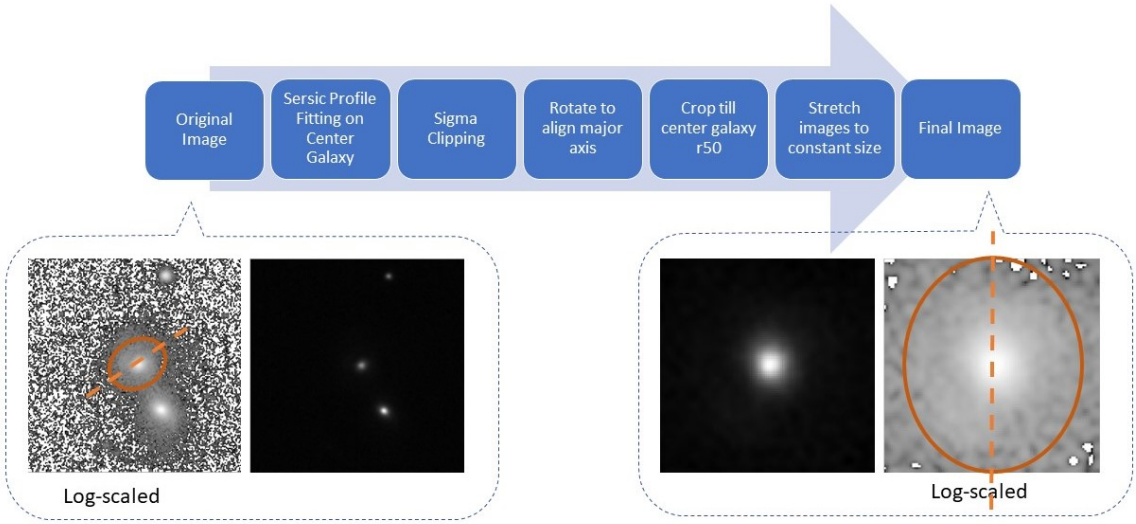


Figure 1:

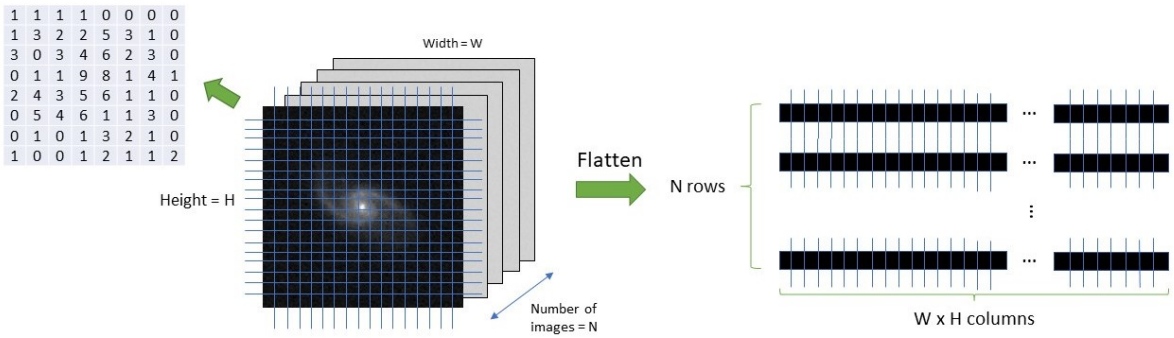


Figure 2:

Method

Principal Component Analysis

Consider a data set matrix, where each row is a data point and each column is one of the data point's dimensions. The 'Principal Components' (PC) of such a data set is the set of orthogonal unit vectors where average squared distance from data points vector line is minimised. Importantly, when data is expressed in coordinates defined by these principal components, each dimension of the data is linearly uncorrelated with one another. The information held by each of these dimensions does not overlap with each other.

Furthermore, the principal component vectors are defined such that when the data is scalar projected, the greatest possible variance is achieved precisely when the data is projected onto the first principal component. The next highest variance results from projecting onto the second principal component, and so on. Indeed, by looking at the example in figure X, we see that PC1 is precisely the direction along which the variance of the data is maximum, followed by PC2.

We also note that if we were to change our coordinate frame to that defined by PC1 and PC2, the data points are scatter with minimal variance along PC2. We interpret this as PC2 containing very little information that uniquely defines each data point, especially in comparison with PC1. As such, we can choose to ignore this dimension altogether, effectively reducing our data set's number of dimensions and making it simpler to work with.

In the case of our galactic images, we also want to find a lower dimensional representation that has minimal loss in this 'information'. We hence look for the principal components of our image data matrix, and choose a suitable number of principal components that still makes up a reasonable explained variance ratio.

To calculate the principal components, one method is by finding the eigenvectors of the data set's covariance matrix:

For a $n \times p$ dataset matrix B, its $p \times p$ covariance matrix C is given by

$$C = \frac{1}{n-1} B^* B$$

, where $\frac{1}{n-1}$ is used due to Bessel's correction.

One can then calculate the eigenvectors v_i of the covariance matrix to get the principal components. The eigenvalues λ_i of each eigenvector will be precisely the explained variance of that particular principal component. The explained variance ratio of PC i is given by $\frac{\lambda_i}{\sum_{n=1}}$.

Autoencoders

Autoencoders are a type of unsupervised neural network that:

- Learns the optimal way of encoding an image into a compressed, low-dimensional representation.
- Learns to decode the encoded image such that the final output is similar to the input.

Mathematically, this means that we are searching for two functions, ϕ and ψ, where

$$\phi : X -> F$$

$$\psi : F -> \hat{X}$$

such that the loss function $L(X, \hat{X})$ between input data set X and output data set $\hat{X}$ is minimised.

Each node in the input layer corresponds to the value of one dimension from our input data. In the case of our $n \times 1$ dimensional image data, each node would contain the value of a row in the flattened image vector (which in turn corresponds with one pixel from the original image array).

For each neuron j in the following second layer, we repeat the following steps. Firstly, the value from each node i of the input data x, x_i, is multiplied by a weight w_ij, before the weighted values from all the nodes are then summed up. This output is then passed into a non-linear activation function σ.

$$Output_j = \sigma(\sum_i (w_ij \times x_i))$$

The output from this activation function is than finally passed as the value for one of the destination nodes in the next hidden layer. This is then repeated for every layer in the network. Note that the output from the final layer will have the same dimensions as the input layer.

In particular, the dimensions of the hidden layers (number of nodes in the layer) are intentionally constrained in the middle. This forces the neural network to learn how to compress the N number of dimensions initially present at the input layer into a smaller number of dimensions with minimal loss in information.

The neural network trains itself to by first checking the error between the output and input values. Through a process called 'backpropagation', it then decides how the weights w need to be adjusted to reduce this error. In particular, we will make use of a loss function called the mean squared error (MSE), which for a single input vector y and

output vector $\hat{y}$ is defined as

$$MSE = (y - \hat{x})^2$$

For a given dataset Y of n vectors $y_1, y_2, ..., y_n$, the overall loss function is defined as

$$L(Y, \hat{Y}) = \frac{1}{n} \sum_{i}^{n} (y_i - \hat{y}_i)^2$$

We then check how this loss function changes when we make changes to the weights w (and biases) by finding the gradients with respect to them, and search for its local or global minimum.

Variational Autoencoders

Unlike a basic autoencoder, the main feature of a variational autoencoder (VAE) is that it finds a probabilistic distribution for the latent features of each input vector, rather then a fixed value. In practice, we assume that our latent distributions are Gaussian distributions, such that the encoded input data is a set of means and variances, $\mu_i and \sigma_i$, for each latent dimension i. Our decoder than takes a random sample from each distribution and attempts to reconstruct the original image.

However, it is important to note that sampling directly from the $N_i(\mu_i, \sigma_i)$ will prevent us from finding the partial derivative of the loss function with respect to the weights attached to the latent layer. This means that we will be unable to train our neural network as we are unable to perform backpropagation. To get around this, we implement the 'reparameterisation trick', where we instead randomly sample from a unit Gaussian $N(0, 1)$, and shift it by the mean μ_i before scaling it by the variance σ_i.

Since we are now taking fixed values from the latent features μ_i and σ_i, we are able to take the partial derivatives and perform backpropagation normally.

The main advantage of a VAE is that we are able to find smoother latent distributions

when compared to a basic autoencoder. The loss function of a VAE contains an additional term on top of the MSE, called the Kullback-Leibler (KL) divergence term.

$$L(Y, \hat{Y}) = \frac{1}{n} \sum_{i}^{n} (y_i - \hat{y_i})^2 + D_{KL}(N(\mu_Y, \sigma_Y) || N(\mathbf{0}, \mathbf{I}))$$

The KL Divergence $D_{KL}(P||Q)$ measures how different the probability distribution P is from Q. By imposing that the distribution of our data points in latent space should be close to a unit Gaussian $N(0, 1)$ via the KL divergence term, we are able to 'smoothen' our distribution. This allows for increased interpretability of our results, since there are no empty gaps in latent space between clusters of points. An extension of this idea, the β-VAE, adds a hyperparameter coefficient in front of the KL divergence term, which allows one to adjust the balance between reconstruction accuracy and distribution smoothness.

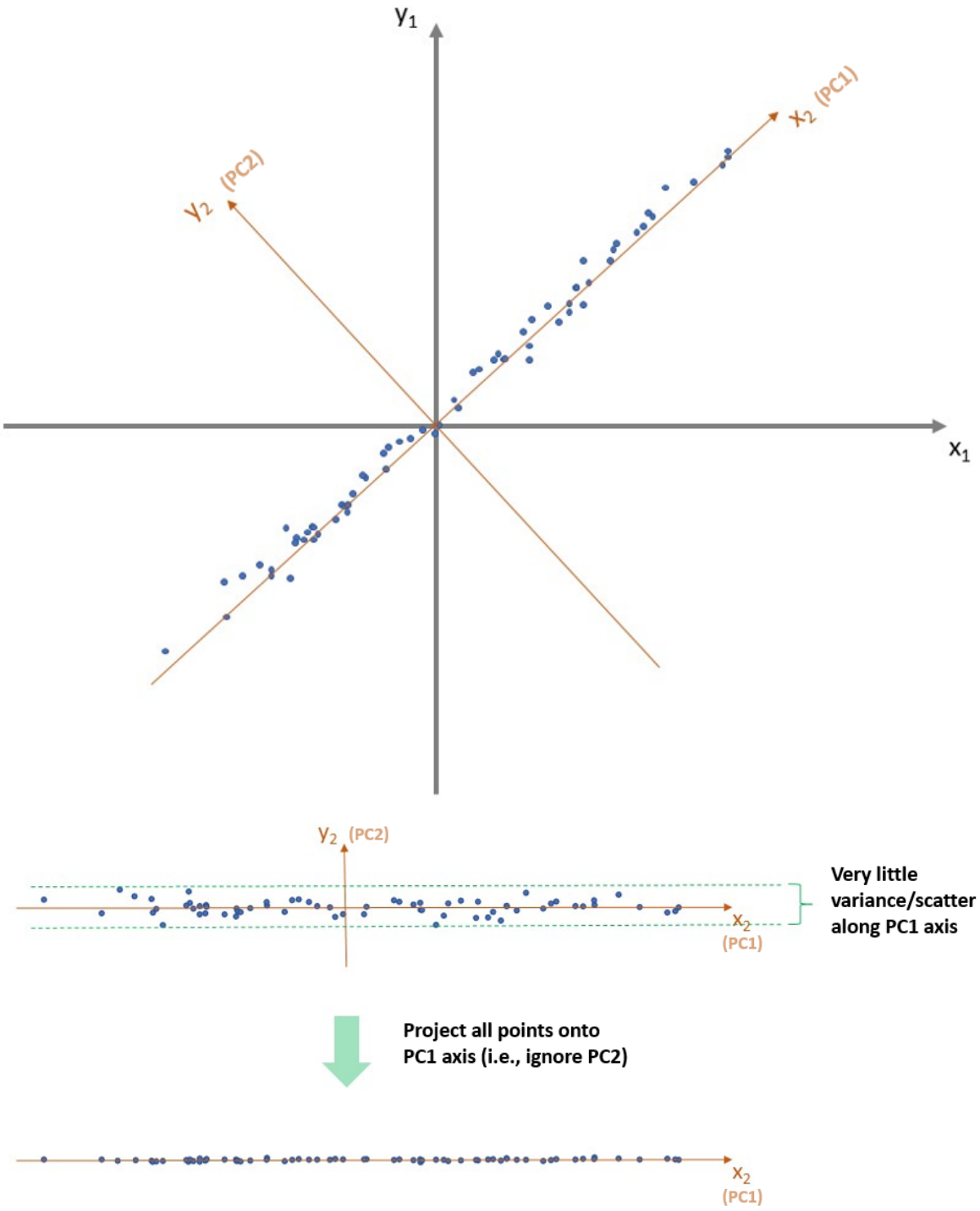

y1
x1
x2 (PC1)
y2 (PC2)
y2 (PC2)
x2
(PC1)
Very little
variance/scatter
along PC1 axis
Project all points onto
PC1 axis (i.e., ignore PC2)
x2
(PC1)

Figure 3:

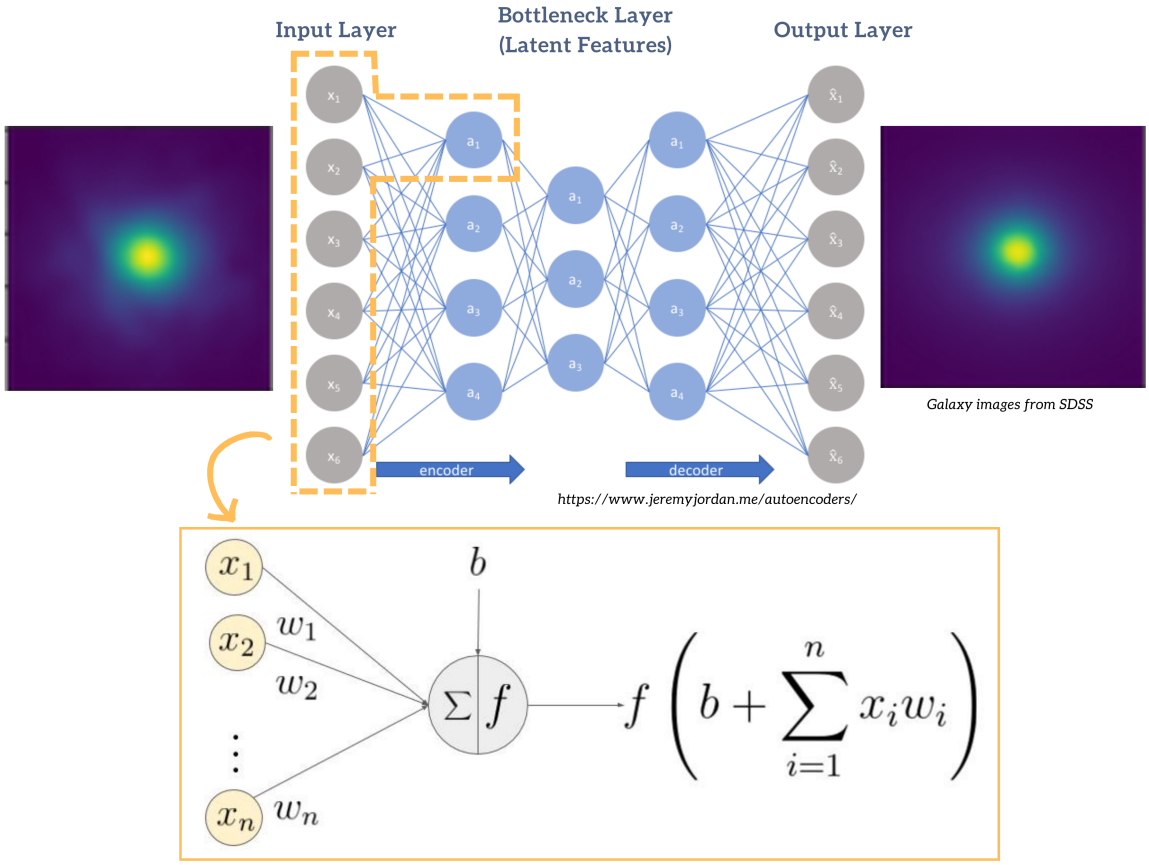


Figure 4:

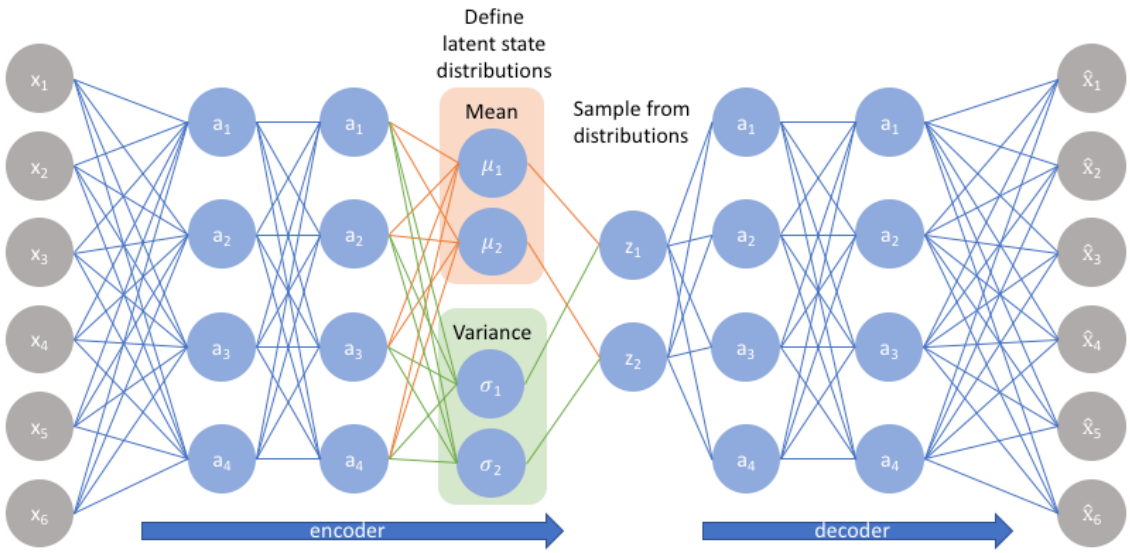


Figure 5:

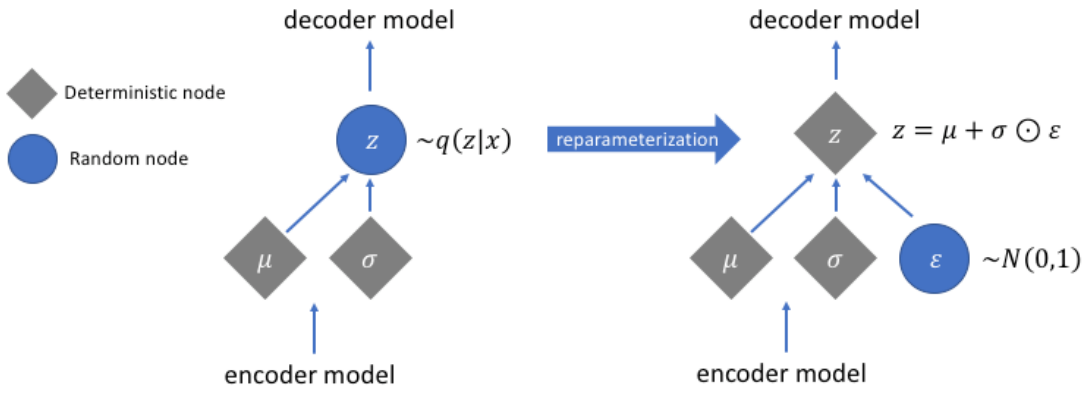

decoder model
Deterministic node
Random node
z
$\sim q(z|x)$
μ
σ
encoder model
reparameterization
decoder model
z
$z = \mu + \sigma \odot \varepsilon$
μ
σ
ε
$\sim N(0,1)$
encoder model

Figure 6:

Results/Discussion

Principal Component Analysis

With reference to figures X, the first 5 principal components were sufficient for explaining 85% of the explained variance ratio in our data set, while 12 would explain 90% of the explained variance ratio. This represents a significant compression from 10000-dimensional image data (one dimension for each pixel) to just 12 dimensions. Among these principal components, PC1 is visually similar to the active galactic nuclei (AGN) - along with its surrounding accretion disc) that we would find at the core of many galaxies, which explains its high explained variance ratio (40%).

We note that the nature of our principal components are very different from those of Uzeirbegovic et al. (2020). In particular, while the principal components in Uzeirbegovic et al. (2020) were radially symmetrical and decomposed galactic images into superposed 'eigengalaxies', our application yielded individual structural features instead.

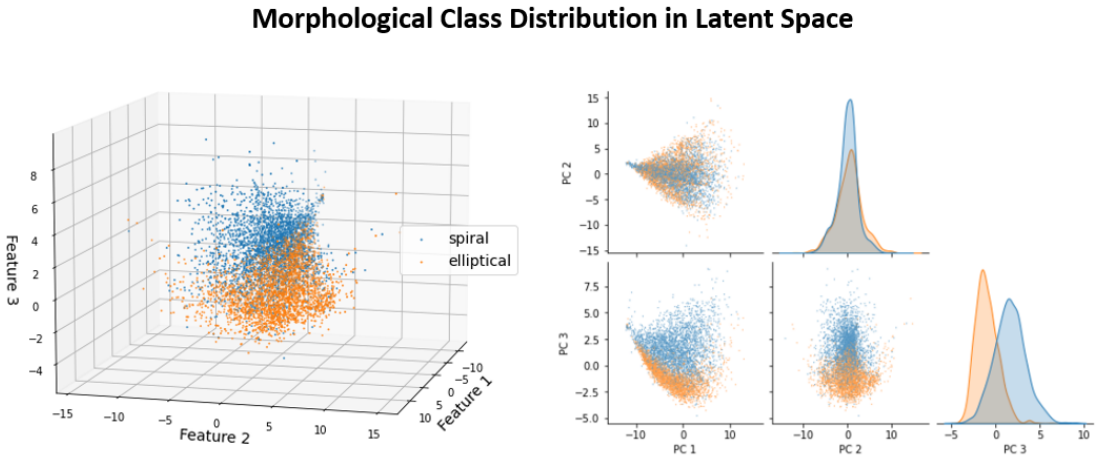
Reconstruction of images from varying numbers n of principal components gave the results in figure X. Reconstructions from the first 2, 4, and 8 PCs yielded largely elliptical-like galaxies. However, at n ¿ 16, as more spiral-like PCs (like PC10, 11 and 16), the reconstructed images began showing clearer spiral-like features. The mean-squared error loss of the output and input images fell from 0.007 with 2 PCs, to ¡0.0015 at 64 PC.

Pearson Correlations between PCs and Galaxy Zoo Votes			
	PC 1	**PC 2**	**PC 3**
Has Smooth Profile	0.392331	-0.010117	-0.422918
Has Features or Disk	-0.394702	0.008477	0.421957
Has Spiral Arms	-0.211708	-0.009526	0.454300
Has Obvious Bulge	-0.152348	0.042601	-0.447374
Is Completely Round	-0.092058	0.049670	-0.411865

Table 1:

Principal Components relationship with Morphological Features

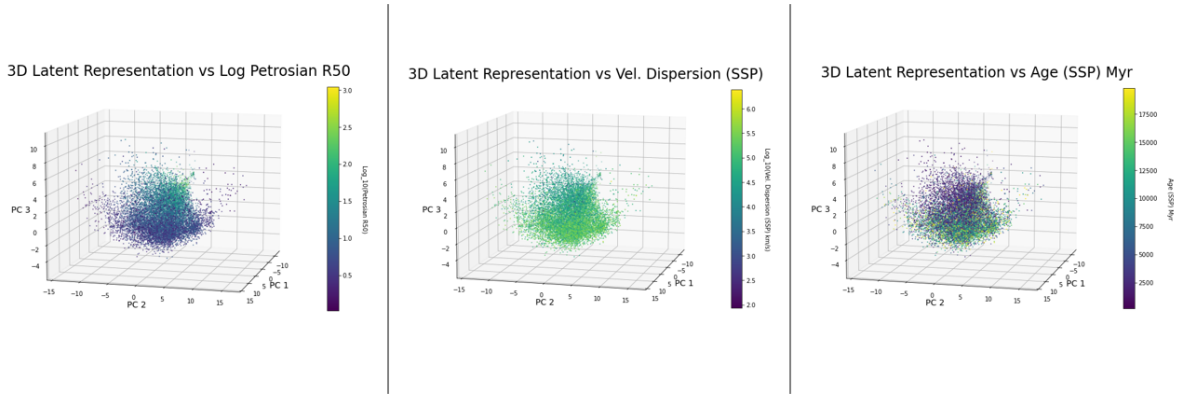
From figure X, we note that if we only consider the first 3 PCs (explained variance ratio 77%), the distribution of galaxies in PC-space is such that different morphological classes are distinctly clustered. Spiral galaxies tend to be found in regions of higher PC3, while the opposite is true for elliptical galaxies. However, there is little clustering along PC1 and PC2, indicating that these features are not particularly unique to either of these two broad classes.



The principal components are also linearly correlated with the Galaxy Zoo 2 vote data. For example, as shown in figure X, the Pearson correlation matrix indicates that higher amounts of PC1 is moderately correlated with a greater probability of a galaxy

having a smooth profile, and a lower probability of having spiral or disc features. This is indicative of elliptical class galaxies. Similarly, PC3 is moderately positively correlated with the probability of the galaxy having spiral/disc features, which is characteristic of spiral classes.

Principal Components relationship with Physical Features



Pearson Correlations between PCs and Galaxy Zoo Votes			
	PC 1	**PC 2**	**PC 3**
Vel. Dispersion (SSP) km/s	-0.131828	0.053716	-0.463102
Age (SSP) Myr	-0.195082	0.056888	-0.408934
Metallicity (SSP)	-0.198687	0.050650	-0.435227
Vel. Dispersion (exp SFH) km/s	-0.128208	0.054953	-0.470396
Age (exp SFH) Myr	0.239812	-0.058860	0.442931
Metallicity (exp SFH)	-0.232993	0.053867	-0.406294
Petrosian 50 Radius arcsec	-0.645232	0.035007	0.413253
LOGSFRSED	0.063654	-0.008977	0.012052
LOGMSTAR	0.052606	-0.002355	-0.047605
LOGSSFR	0.032332	-0.038885	0.396317

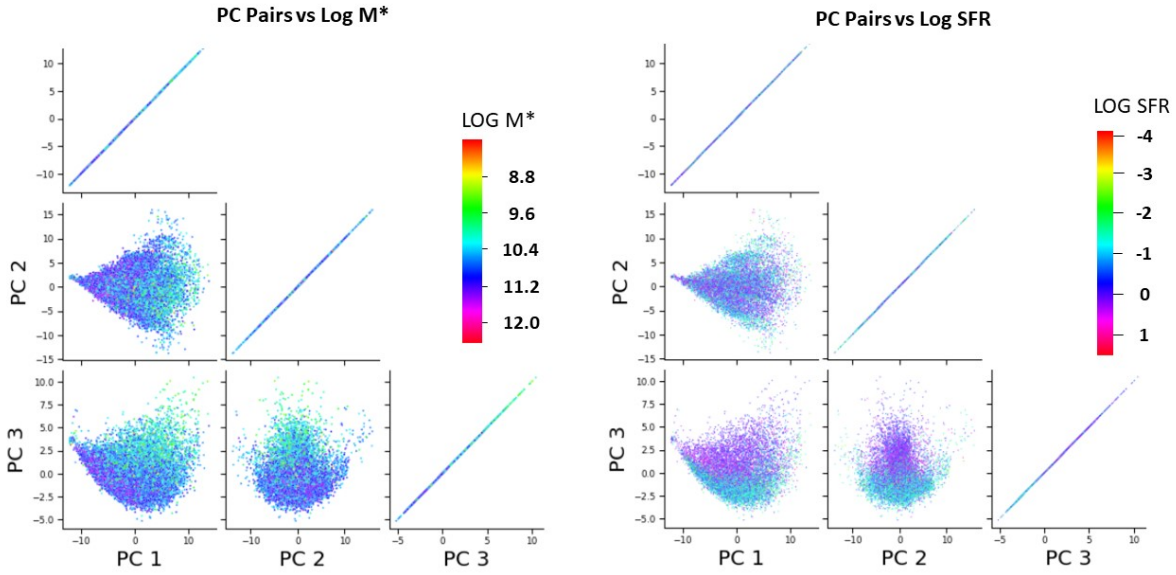
Table 2:

The principal components also show moderate-strong correlations with many of the galaxy's physical properties. In particular, PC1 exhibits a strongly negative correlation with the Petrosian 50 Radius of a galaxy (corr = -0.645). As discussed before, PC1

represents the presence of AGNs, which are typically the brightest sources of EM radiation. As such, a large AGN presence would lead to most of its emissions being concentrated towards the center, which explains the negative correlation with r50.

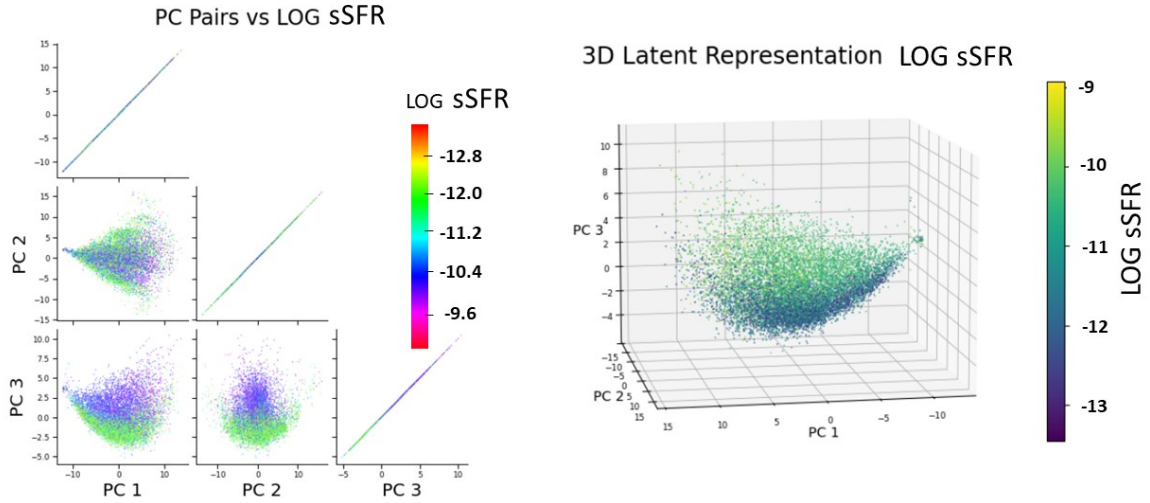
PC3 also exhibits moderate correlations with many of the physical properties. In particular, PC3 is positively correlated with the logarithm of the Specific Star Formation Rate (LOGSSFR). Since this structure is an accelerant of star formation activity and surrounds the core, it could be posited that this is an underlying substructure found within the spiral arm structures.

Despite this, the 3 principal components do not exhibit linear correlations with either the logarithm of star formation rate (SFR) or stellar mass (MSTAR).



Autoencoders

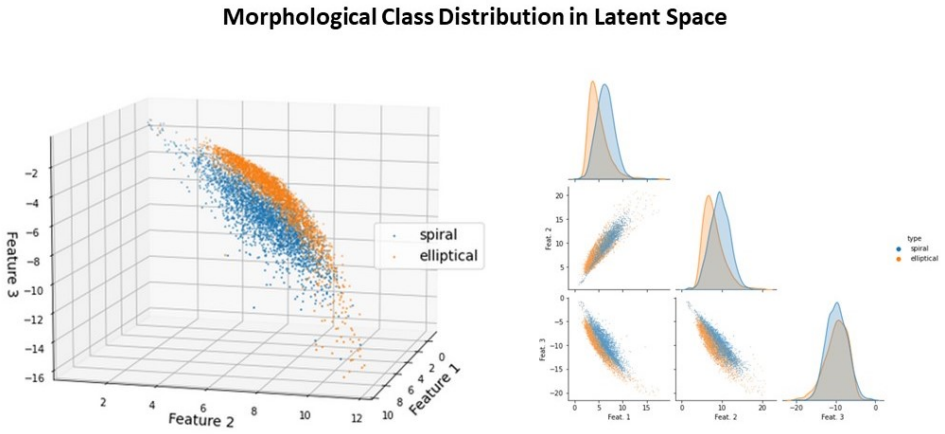
The image data was scaled logarithmically prior to input into the neural network, in order to dynamically scale the pixel values and ensure better representation of less bright features like spiral arms.



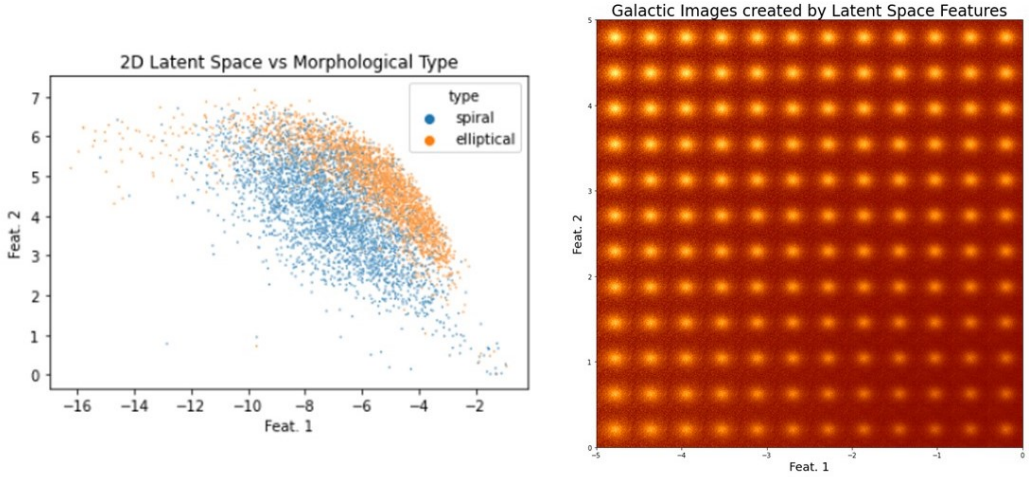
Latent Features Relationship with Morphological Properties

For a basic autoencoder with a 3-dimensional bottleneck, we were able to train it to learn a latent space where the galaxies are again distributed in distinct morphological clusters. From figure X, however, we note that there is likely a degeneracy between latent feature 1 and feature 2, as they exhibit close to extremely high linear correlations. This implies that the actual underlying latent space is likely to be 2 dimensional - i.e. we can compress our image data into just 2 dimensions while retaining most of the important defining information.

The 2-dimensional latent space, along with galactic images generated by linear interpolation of the features, can be seen in figure X below. We note that again, there is distinct clustering of classes, but no more degeneracy. Lower amounts of feature 1 tends to lead to the galaxies having more diffused disc-like structures, while higher amounts of feature 2 is related to brighter and larger center bulges. However, the boundary between the galaxy classes along each axis is not straightforward. Rather, it is observed that the boundary value for each feature is linearly dependent on the other.

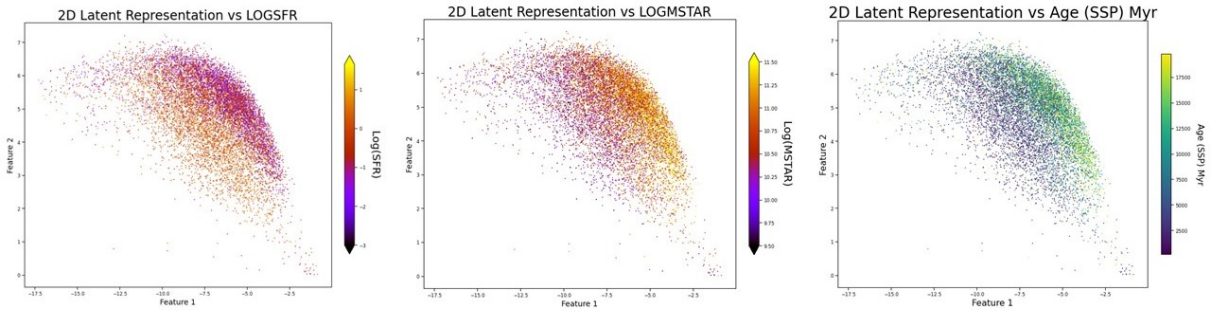


We also note that Feat.2 is moderately positively correlated with the likeliness of a galaxy to have a smooth profile, and negatively correlated with the likeliness of having spiral/disc features. Furthermore, the likelihood of a galaxy being spiral is negatively correlated with Feat.2, as evidenced in figure X.



Latent Features Relationship with Physical Properties

We note that the various physical properties gradate smoothly across the latent space. In particular, there are strong negative linear correlations between Feat. 2 and r50, which agrees with the idea discussed above that feat.2 represents the brightness and size of the center bulge. It is observed that there is a general trend of negative and positive gradation



when moving towards higher feature 1 and feature 2. This area of latent space tends to contain older galaxies with lower sSFR , which is characteristic of elliptical galaxies. Note that unlike with PCA, Figure X shows that the latent features here actually exhibit bivariate linear correlations with SFR and MSTAR.

Pearson Correlations between Feats and Galaxy Zoo Votes		
	Feat. 1	**Feat. 2**
Has Smooth Profile	-0.123796	0.563262
Has Features or Disk	0.126552	-0.561337
Has Spiral Arms	-0.022620	-0.425911
Has Obvious Bulge	0.292862	0.088417
Is Completely Round	0.172631	0.014948

Table 3:

Pearson Correlations between PCs and Galaxy Zoo Morphological Classification		
	Feat. 1	**Feat. 2**
spiral	0.018326	-0.390460
elliptical	0.282897	-0.045579
uncertain	-0.249939	0.376698

Table 4:

Variational Autoencoders

We set our variational autoencoder's latent space to be 2-dimensional. We immediately note that the distribution of our galaxies is smoother than when we used a basic autoen-

Pearson Correlations between Latent Feats and Physical Properties		
	Feat. 1	**Feat. 2**
Vel. Dispersion (SSP) km/s	0.321578	0.135615
Age (SSP) Myr	0.326011	0.025508
Metallicity (SSP)	0.321414	0.049316
Vel. Dispersion (exp SFH) km/s	0.322730	0.142525
Age (exp SFH) Myr	-0.369470	0.013757
Metallicity (exp SFH)	0.339767	0.001796
Petrosian 50 Radius arcsec	0.361540	-0.746581
LOGSFRSED	-0.058724	0.055595
LOGMSTAR	-0.024831	0.075587
LOGSSFR	-0.190473	-0.172685

Table 5:

coder, and is close to a 2D Gaussian.

Latent Features Relationship with Morphological Properties

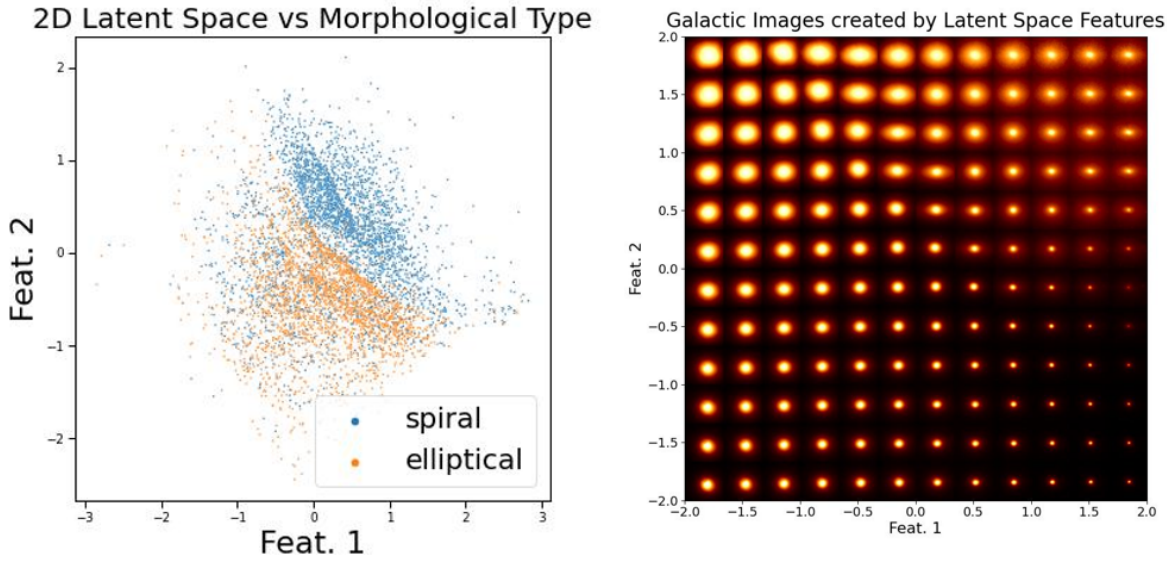


Figure 7:

We note that higher amounts of latent feature 1 is representative of a smaller center bulge, while feature 2 is proportional to the relative size of a diffused disc structure. However, unlike the AE's latent features, there is less overlap between the physical appearance

Pearson Correlations between PCs and Morphological Class		
	Feat. 1	**Feat. 2**
spiral	0.322450	0.198545
elliptical	0.137961	-0.344143
uncertain	-0.394076	0.112321

Table 6:

Pearson Correlations between PCs and Galaxy Zoo Morphological Classification		
	Feat. 1	**Feat. 2**
Has Smooth Profile	-0.483305	-0.118766
Has Features or Disk	0.484643	0.116122
Has Spiral Arms	0.335572	0.247038
Has Obvious Bulge	-0.017369	-0.427548
Is Completely Round	-0.016771	-0.217173

Table 7:

of feature 1 and feature 2, which indicates that the VAE's latent features are more linearly independent.

we note that while the morphological classes of our galaxies exhibit distinct clustering, there is some overlap around the origin. This is likely because, in this region of low Feat.1 and Feat. 2, neither the bulge nor disc features are pronounced enough to conclusively define the morphological class of a galaxy.

Latent Features Relationship with Physical Properties

We note that the physical properties of the galaxies gradates in varying ways across latent space. The r50 of a galaxy increases smoothly and proportionately across Feat. 1, but exhibits less variance along Feat. 2. This can be explained as the r50 of a galaxy being greatly dependent on the size of its center bulge, given that it is typically much brighter than its diffused disc. On the other hand, the metallicity of a galaxy proportionately increases along Feat. 2, but exhibits less variance along Feat. 1.

sSFR and Age, however, exhibit a clearer bivariate relationship with the latent features. In particular, galaxies of high age and low sSFR are found in regions of higher Feat.

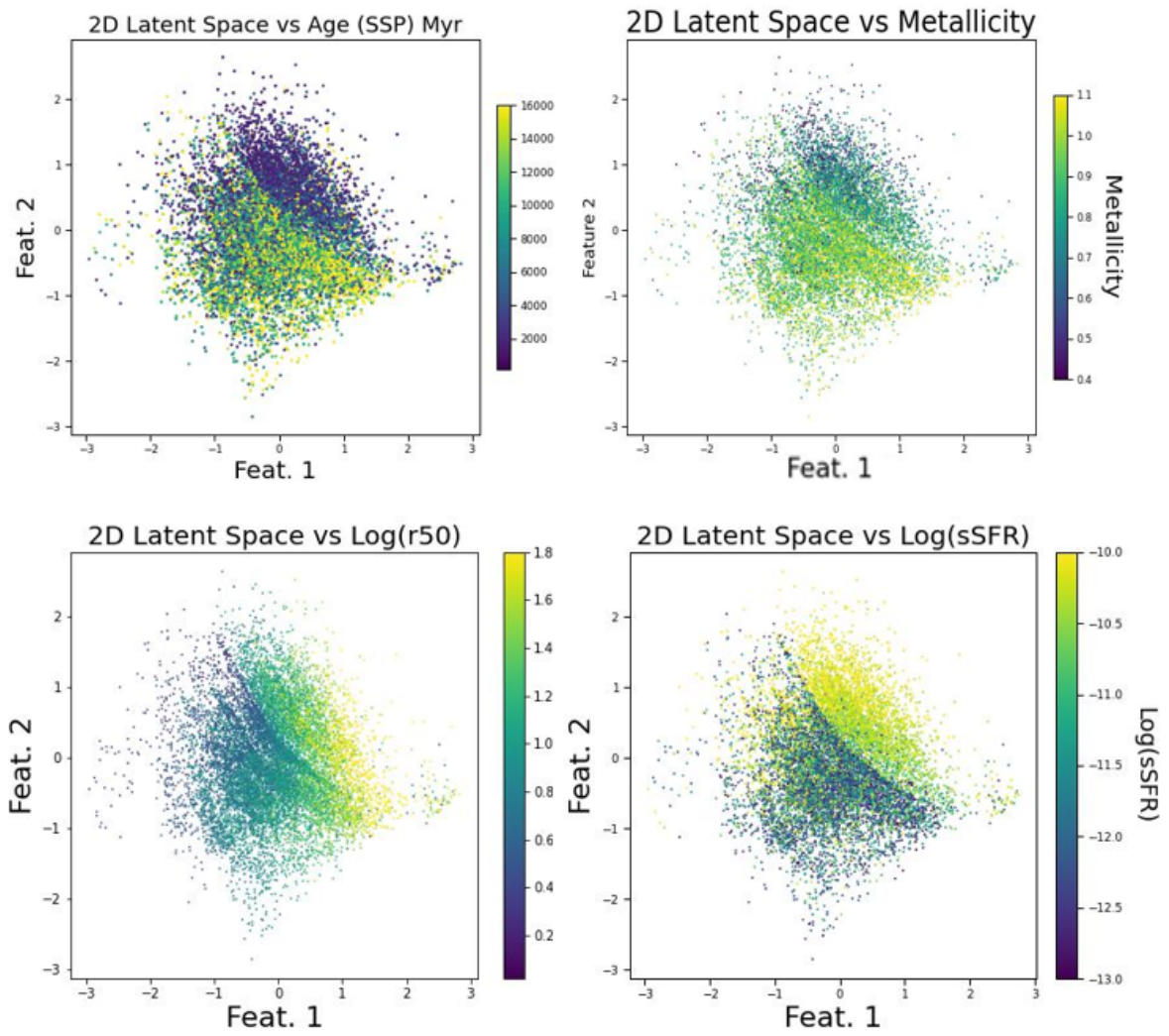


Figure 8:

1 and Feat. 2. This makes intuitive sense, as this regions are representative of galaxies with big diffused discs (characteristic of older galaxies) and relatively small bulges (which correlates with low star formation activity).

We also note that the sSFR and r50 of galaxies are inversely related with each other in the spiral clusters, which agrees with the theory that star formation rate is linearly related to the surface density for spiral galaxies.

Pearson Correlations between PCs and Physical Properties		
	Feat. 1	**Feat. 2**
Vel. Dispersion (SSP) km/s	-0.051947	-0.465508
Age (SSP) Myr	0.034974	-0.417873
Metallicity (SSP)	0.019536	-0.455701
Vel. Dispersion (exp SFH) km/s	-0.058639	-0.472199
Age (exp SFH) Myr	-0.049727	0.466863
Metallicity (exp SFH)	0.057441	-0.445452
Petrosian 50 Radius arcsec	0.717488	-0.033900
LOGSFRSED	-0.058900	0.045080
LOGMSTAR	-0.069458	-0.008703
LOGSSFR	0.110948	0.334421

Table 8:

Merging and Irregular Galaxies in VAE Latent Space

We also explored how merging and irregular galaxies, which were excluded in the training dataset, would be distributed in this latent space.

Merging galaxies were largely found along the 'border' between the elliptical and spiral galaxies. When coupled with the distribution of galaxy ages, we get a picture of how a galaxy's morphology evolves with time. This agrees with existing theory, where the younger spiral galaxies merge eventually to form ellipticals. Comparing this distribution with the sSFR along the border also indicates that sSFR drops rapidly upon merging, which supports the idea that galaxy mergers can initiate quenching (Davies et al. 2022).

Irregular galaxies, on the other hand, are found largely in the high Feat. 2 and positive Feat. 1 regions where spirals are found. As irregular galaxies tend to born from gravitational interactions that disrupt regular shapes (like ellipses), this could explain the similarities in their morphologies. More interestingly, the irregular galaxies are largely found in the regions of very high sSFR, which agrees with existing theory that irregular galaxies can have comparable sSFR with spirals, though the exact mechanism requires further investigation.

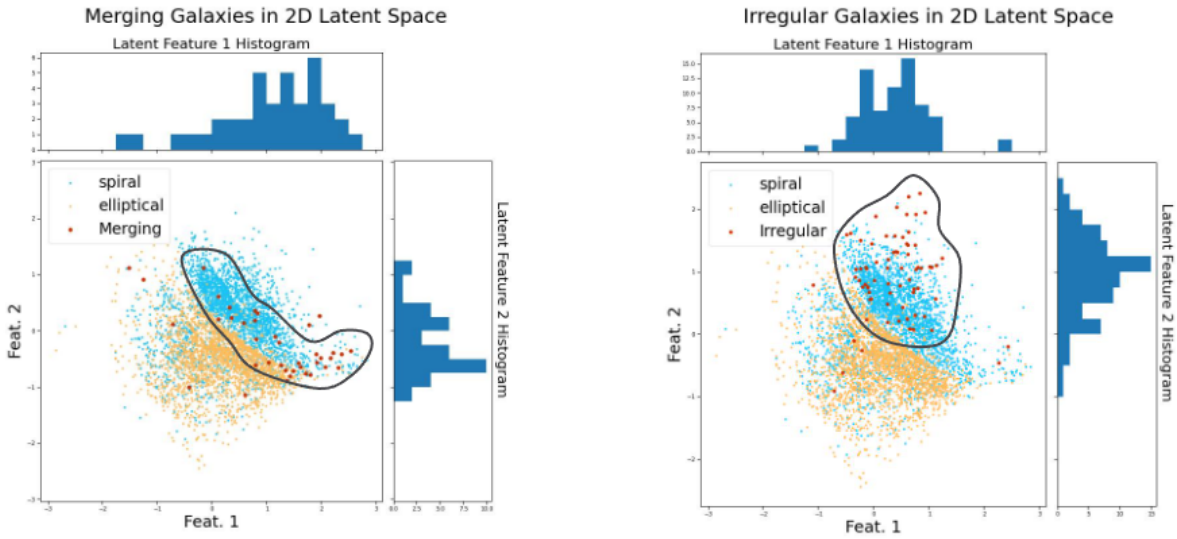

Merging Galaxies in 2D Latent Space
Latent Feature 1 Histogram
spiral
elliptical
Merging
Feat. 2
Feat. 1
Latent Feature 2 Histogram
Irregular Galaxies in 2D Latent Space
Latent Feature 1 Histogram
spiral
elliptical
Irregular
Feat. 2
Feat. 1
Latent Feature 2 Histogram

Figure 9:

Conclusions/Future Work

TBC

Bibliography

Bundy, K., Ellis, R. S., & Conselice, C. J. 2005

Davies, J. J., Pontzen, A., & Crain, R. A. 2022, Monthly Notices of the Royal Astronomical Society, 515, 1430, doi: 10.1093/mnras/stac1742

Phillipps, S. 2005, The Structure and Evolution of Galaxies

Uzeirbegovic, E., Geach, J. E., & Kaviraj, S. 2020