Emergent Denoising of SDSS Galaxy Spectra Through Unsupervised Deep Learning

Oliver C. Camilleri[1]*, Zahra Sharbaf[2,3], Ignacio Ferreras[2,3,4]†

[1] *Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, Surrey, GU2 7XH, UK*
[2] *Department of Physics and Astronomy, University College London, London WC1E 6BT, UK*
[3] *Instituto de Astrofísica de Canarias, Calle Vía Láctea s/n, E38205, La Laguna, Tenerife, Spain*
[4] *Departamento de Astrofísica, Universidad de La Laguna, E38206 La Laguna, Tenerife, Spain*

ABSTRACT

Spectroscopy represents the ideal observational method to maximally extract information from galaxies regarding their star formation and chemical enrichment histories. However, absorption spectra of galaxies prove rather challenging at high redshift or in low mass galaxies, due to the need to spread the photons into a relatively large set of spectral bins. For this reason, the data from many state-of-the-art spectroscopic surveys suffer from low signal-to-noise (S/N) ratios, and prevent accurate estimates of the stellar population parameters. In this paper, we tackle the issue of denoising an ensemble by the use of unsupervised Deep Learning techniques trained on a homogeneous sample of spectra over a wide range of S/N. These methods reconstruct spectra at a higher S/N and allow us to investigate the potential for Deep Learning to faithfully reproduce spectra from incomplete data. Our methodology is tested on three key line strengths and is compared with synthetic data to assess retrieval biases. The results suggest a standard Autoencoder as a very powerful method that does not introduce systematics in the reconstruction. We also note in this work how careful the analysis needs to be, as other methods can – on a quick check – produce spectra that appear noiseless but are in fact strongly biased towards a simple overfitting of the noisy input. Denoising methods with minimal bias will maximise the quality of ongoing and future spectral surveys such as DESI, WEAVE, or WAVES.

Key words: techniques: spectroscopic – methods: data analysis – galaxies: statistics – galaxies: evolution – galaxies: stellar content – galaxies: fundamental parameters

arXiv:2510.08411v1 [astro-ph.IM] 9 Oct 2025

1 INTRODUCTION

Galaxy spectra in the wavelength interval from near ultraviolet to near infrared encode a vast amount of information regarding the properties of the underlying stellar populations. The continuum and absorption lines of the photospheres of constituent stars leave their imprint on the integrated spectra, and represent the workhorse of galaxy formation studies concerning the star formation and chemical enrichment histories. Spectroscopic surveys of galaxies, such as the Sloan Digital Sky Survey (York et al. 2000) have greatly helped deepen our understanding of galaxy formation. However, spectroscopy requires large integration times, as the faint light from galaxies is spread with respect to wavelength. Often times, spectra have been used – mainly by cosmologists – as a tool to derive redshift and thus determine the large-scale distribution of galaxies. However, exploring the galaxies themselves through their stellar populations requires deeper data, at a higher signal-to-noise (S/N) ratio, in order to compare the observations with detailed models of stellar population synthesis (see, e.g. Walcher et al. 2011; Conroy 2013). In this regard, it is not uncommon to need values of S/N per resolution element above $\gtrsim$20-30, if not higher, for detailed analyses of subtle differences in the populations, such as variations of chemical abundances or the initial mass function (e.g. La Barbera et al. 2013, 2017; Ferreras et al. 2019). In this regard, spectroscopic surveys tend to be optimised to produce large volumes of data, at the cost of a lower S/N, and so, any algorithm aimed at increasing the S/N of the data proves a very valuable tool, especially with the ongoing and upcoming surveys such as DESI (DESI Collaboration et al. 2025), WEAVE (Jin et al. 2024) or WAVES (Driver et al. 2019).

The actual S/N of an observation is borne out of the competition between the photons from the galaxy and spurious photons or counts coming from unwanted sources, such as the background sky, airglow, detectors and reduction artifacts, etc. Increasing the S/N of a single spectrum is typically not feasible unless models are used, therefore introducing large systematics. Our approach to this problem starts from a large ensemble of spectra taken by the same instrument and following the same reduction process. The ensemble should also include a large amount of high quality (i.e. high S/N data,

* E-mail: oc00149@surrey.ac.uk
† Corresponding author: ignacio.ferreras@iac.es

so that the method can somehow interpolate among the ensemble members to produce optimised versions of the data. Traditional data driven methods, such as Principal Component Analysis, have been applied to stellar and galaxy spectra, for instance to remove the emission lines from airglow (Wild & Hewett 2005). These methods are commonly used for classification purposes (Madgwick et al. 2003; McGurk et al. 2010) and can also help in the interpretability of the information content, for instance by exploring the resultant latent space (e.g., Sharbaf et al. 2023, 2025). However, as a linear method, PCA is less versatile to encode and model the many intricacies of the spectra in a large ensemble.

Deep Learning (DL) methods are seeing a rapid uptake in use throughout astronomy, helping to address problems that traditional data-driven approaches struggle with. For example, DL has been applied to galaxy and stellar spectra for classification purposes (e.g., Folkes et al. 1996; Fabbro et al. 2018; Wu et al. 2024), dimensionality reduction (e.g., Portillo et al. 2020), recovery of spectra with bad quality (e.g., Wang et al. 2017), general analysis (e.g., Lovell et al. 2019; Melchior et al. 2023) or in the search for anomalies (e.g., Baron & Poznanski 2017; Liang et al. 2023). In this paper, we adopt a set of DL algorithms, each trained on a large set of galaxy spectra from the Legacy part of SDSS, and then assess their ability to increase the S/N of input spectra. These models are unsupervised, and in contrast to works like Scourfield et al. (2023), do not rely on adding synthetic noise to *training* data. Instead, the denoising effects seen are emergent, ultimately stemming from information bottlenecks and aided by the formulation of the objective function. Furthermore, we experiment with the reconstruction of spectra from incomplete data and attempt to explain deep model decision making.

Please note it is important to ensure that the process does not alter the sample in a systematic way. We also emphasise that this method is not a smoothing process, where the S/N can be increased at the cost of a lower spectral resolution. A similar type of work has been recently presented in Scourfield et al. (2023), mainly focussed on retrieval of emission line data. The authors conclude that a Variational Autoencoder performs better than PCA, and study the effect of denoising DESI data from an SDSS-trained set regarding the relationship between stellar mass and (gas phase) metallicity. Melchior et al. (2023) also consider the use of autoencoders to analyse galaxy spectra, and look into the interpretability of latent space, with results mostly focused on emission lines, which is where the spectral variance fully dominates in star-forming and AGN systems. The authors indeed suggest that these methods can be adopted to denoise ensembles of spectra. This work complements these previous studies, turning to the more challenging case of the absorption line spectra, that is commonly used to constrain the stellar population content and the past star formation history. We also consider the issue of potential biasing of the denoised data by use of synthetic spectra that are adopted as ground truth to assess the fidelity of the recovered measurements.

The structure of the paper is as follows, we give a brief presentation of the working sample in § 2, along with a description of the various methods tested for denoising in § 3. The comparison of retrieved and original data is shown in § 4, including a comparison of synthetic data with added noise. A brief discussion of the explainability of the DL performance

is presented in § 5. Finally, our concluding remarks are given in § 6.

2 THE SAMPLE

We select our working sample from the Legacy set of galaxy spectra of the Sloan Digital Sky Survey (York et al. 2000). The data include a large number of spectra (of order 1 million), and most importantly covers a wide range of S/N, with a large number of high quality data at a S/N higher than 10-20, and many spectra at lower S/N. This represents an ideal training sample as the data processing is homogeneously performed, minimising biases, and covers all types of evolutionary stages of galaxies, mass, morphology, etc. Moreover, each observation includes, in addition to the actual spectrum, a best fit model that can be adopted as a synthetic case to which noise is added, as we will show below. The sample is the same as the one presented in Sharbaf et al. (2023) and the motivation behind the constraints regarding the quality of the data as an ensemble is identical. For instance, in order to set this exercise in the best defined scenario, we include a constraint in (stellar) velocity dispersion between 100 and $150 \, \mathrm{km \, s^{-1}}$ as a wider range will also introduce the expected bias in effective spectral resolution caused by the kinematic kernel. The data are taken from SDSS Data Release 16 (Ahumada et al. 2020), and correspond to single fibre measurements of the central parts of galaxies (3 arcsec diameter), at spectral resolution $\mathcal{R} \sim 2,000$ (Smee et al. 2013). The targets are selected as completely as possible down to a Petrosian flux level for the target galaxy in the SDSS-r band of $r < 17.77 \, \mathrm{AB}$ (Strauss et al. 2002). The data were further constrained in redshift, z∈[0.05,0.1] and in S/N measured as an average within the SDSS-r band higher than 15 per pixel ($\Delta \log(\lambda/\text{Å}) = 10^{-4}$). The total sample comprises 68,794 spectra, which were retrieved from the main SDSS database, de-redshifted and de-reddened regarding foreground dust absorption from the Milky Way, following a standard Fitzpatrick (1999) attenuation law.

To remove the variations caused by different stellar mass and redshift of the galaxies, all spectra are normalized to the same average flux in 6,000-6,500Å window, in the rest-frame. The spectral range is restricted to the rest-frame wavelength $\lambda \in [3700, 7500]$Å. Finally, for one of the tests, we explore the denoising procedure when training on data without continuum. For that purpose, we take the robust high percentile method of Rogers et al. (2010) to define the continuum that is removed from each spectra for this test case. For more details about the sample, please see Sharbaf et al. (2023).

3 METHODOLOGY: DENOISING THE SPECTRA

3.1 Butterworth Filtering

The Butterworth filter (BF) is a classical signal processing approach used to selectively attenuate unwanted frequencies within a general signal. Its maximally flat frequency response in the passband supresses signal distortion, making it a popular choice in a range of domains. The squared magnitude of the frequency response, $|H(\omega)|^2$, at angular frequency, ω, is
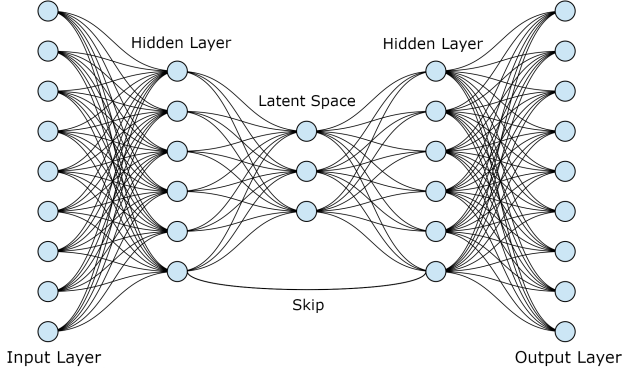
Figure 1. A 9-6-3-6-9 autoencoding network with an added skip connection between two hidden layer neurons. The input is embedded within a learned space of lower-dimensionality known as a latent space. From this representation, the input is reconstructed in the output layer.

Table 1. Network architectures for our full spectrum (FS), narrow windows (NW), narrow windows-skip (NW-S), and continuum subtracted (CS) models. ⊕ denotes layer concatenation and inputs/outputs are in bold. The red arrows indicate a skip connections between layers.

FS	NW	NW-S	CS
3800	**400 400**	**400 400**	**3800**
3000	400 400	400 400	2800
2100	800 800	800 800	1200
1500	1100 1100	1100 1100	800
850	⊕	⊕	450
350	1800	1800	400
110	2100	2100	450
300	2400	2400	800
750	2800	2800	1100
1400	3000	3000	1600
2700	3400	3400	2800
3800	**3800**	**3800**	**3800**

given by

$$|H(\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}} \qquad (1)$$

where ω_c is the cutoff frequency and n is the order of the filter. For the majority of our tests, ω_c was set to 0.8, although we did weaken the filter for some experiments, increasing ω_c up to 1.32. In all cases, n was held at 5.

While alternative techniques such as the Savitzky-Golay filter are known for their ability to preserve sharp features (see Nikonov et al. 2017), we found that achieving adequate emission/absorption line preservation often required high polynomial orders, which in turn limited the denoising capability of the filter. In contrast, the Butterworth filter provided a better trade-off between smoothing and signal fidelity, with intuitive control over attenuation in the frequency-domain via ω_c. For these reasons, in addition to its frequency response characteristics, we use the Butterworth filter as a classical baseline to better contextualise our deep methods.

3.2 Deep Learning

We developed four different deep models, each tasked with reconstructing spectra through some form of information restriction. The full spectrum (FS) model reproduces spectra using the entire spectrum as input. In contrast, the narrow windows (NW) and continuum subtracted (CS) models perform this reconstruction using just two specific wavelength windows and continuum-removed spectra, respectively. The windows in question, defined in Sharbaf et al. (2023), span 3800-4200Å and 5000-5400Å. Going forward, the former will be referred to as the "blue" region while the latter will be referred to as the "red" region. FS and CS possess autoencoder-like bottleneck architectures. Autoencoders are a type of unsupervised feed-forward neural network that reduce the dimensionality of input data by learning to embed it within a compressed, abstract space known as a latent space. This kind of data compression network is illustrated in Fig. 1, while our exact network configurations are provided in Tab. 1 .

We probe the influence of skip connections using the narrow windows-skip (NW-S) model. Skip connections, illus-

trated in Fig. 1, are shortcuts that allow unmodified information to flow from earlier to later layers, ensuring crucial information is not lost and potentially helping networks learn more useful representations. A theoretical exploration of skip connections and their benefits can be found in Veit et al. (2016).

We use either ReLU or PReLU activation functions for all layers other than those corresponding to outputs, which are linear. The ReLU function is defined as

$$\mathrm{ReLU}(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \qquad (2)$$

PReLU was used only in the CS case, as unlike ReLU, it can meaningfully represent negative values. It is defined as

$$\mathrm{PReLU}(x) = \begin{cases} \alpha x, & x < 0 \\ x, & x \geq 0 \end{cases} \qquad (3)$$

where α is a learnable parameter. The objective function for all three models was a mean absolute error (MAE) loss, defined by

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{S}_i - S_i \right| \qquad (4)$$

where $\hat{S}_{1...N}$ are predicted spectra and $S_{1...N}$ are the originals. Crucially, the MAE reconstruction objective exhibits robustness to noise, reducing the influence of outliers in the data. This benefit is demonstrated in Appendix A, which compares the performance of MAE to that of the more commonly used mean squared error (MSE) objective.

We optimise all models using the Adam method (as defined in Kingma & Ba 2014) and an initial learning rate of 10^{-4}. This is reduced during training via adaptive learning rate decay with a reduction factor of 0.78.

4 COMPARISONS OF SPECTRAL DENOISING

A set of 50,000 spectra are used to train and validate the models described in the previous section and the recovered (denoised) data are compared with a set of 8,000–10,000 spectra not seen by the models. We consider two cases for the reconstruction: 1) actual SDSS spectra from the same dataset;
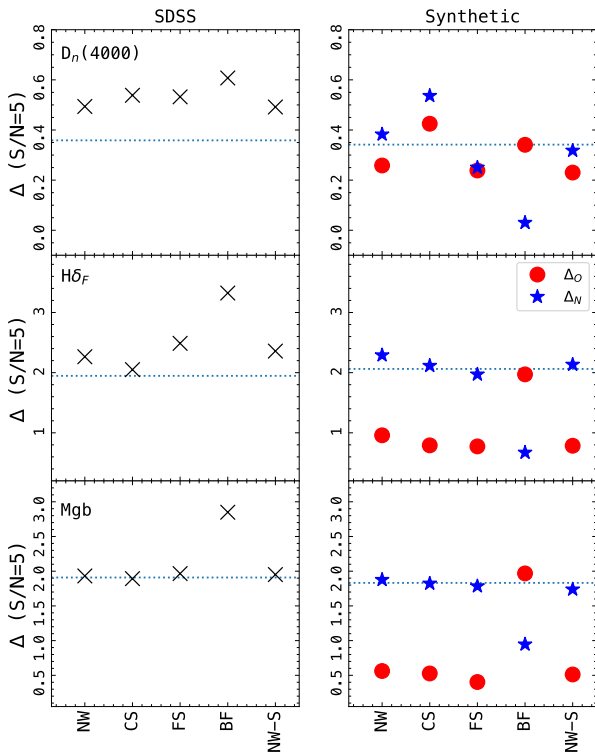
Figure 2. Residual statistic (Δ) estimated at S/N=5 for the reconstruction of SDSS galaxy spectra (left) and synthetic data (right). The horizontal dashed line represents the residual Δ for the comparison between the observed data and the best fit spectra. See text for details.

and 2) best fit spectra of unseen SDSS data (i.e. "noiseless" model fits) with added noise. Comparing the recovery in both of these cases allows us to assess whether any of the adopted models overfit to the training data. We want to avoid the use of synthetic data for the training to ensure the denoising procedure does not add unwanted systematics. Note that the synthetic data is produced from fits to the original SDSS spectra, thus providing an ideal case where the same ensemble of galaxies is used. The more traditional approach using a grid of models can fail in this respect as the parameter space explored need not map the true sample. The best fit spectra of case 2 are taken from the official spSpec data model from SDSS-Legacy (see, e.g., Aihara et al. 2011). These correspond to the official SDSS pipeline best fit models for redshift and classification purposes, and can be considered for our purposes the ground truth.

In order to quantify the performance of the models, we opt for a figure of merit that targets the residual of three key line strengths. While a full spectral residual is a valid description, our approach focuses on the use of DL methods in optimising the analysis of stellar population parameters. Previous work on real and synthetic galaxy spectra reveals a rather small set of spectral windows where information is encoded Ferreras et al. (2023). These include the "blue" and "red" regions defined earlier. Unsurprisingly, these windows happen to be the intervals where traditional line strengths were defined. In this paper we focus on three of the most prominent line strengths: the 4000Å break (as defined by Balogh et al. 1999), the fine (i.e. narrow) definition of Hδ Balmer absorption of

Worthey & Ottaviani (1997), and the traditional Mgb index of the Lick system (Trager et al. 1998).

The figure of merit (Δ) is defined for each line strength measurement as follows: we produce a vector with the residuals of the output and the reference for each spectrum: $\delta_s(\mathcal{I}_i) \equiv [\mathcal{I}_{i,s} - \mathcal{I}_{i,r}]$, where s represents the output spectrum and r the reference spectrum. The mean or median of the ensemble $\{\delta_s(\mathcal{I})\}$ are expected to be close to zero, and the standard deviation represents how well the data are recovered. Therefore, we adopt the standard deviation of the residuals as our figure of merit:

$$\Delta(\mathcal{I}) \equiv \sqrt{\langle \delta^2(\mathcal{I}) \rangle - \langle \delta(\mathcal{I}) \rangle^2}. \qquad (5)$$

Also note that the reference case, $\mathcal{I}_{i,r}]$, is the comparison spectra, that can be defined in two ways, it is either the ("noiseless") best fit data (Δ_O), or the noisy original SDSS data (Δ_N). The former allows us to quantify the denoising process, whereas the latter is used to test overfitting.

Fig. 2 shows the residual statistic measured at a S/N=5 (as an average over all spectra with S/N∈[5,6]) for the reconstruction of an additional, unseen set of real SDSS spectra (left) or a set of synthetic spectra (right). In the case with real data, we only show the data points corresponding to Δ_O, as Δ_N would trivially compare the noisy data with itself. For the synthetic case, we can compare the residual statistic for the "ground truth" case (Δ_O), and for a noisy realization of this ideal case, with Gaussian noise that shows the same S/N as the original data (Δ_N). In this framework, the case $\Delta_N < \Delta_O$ would be indicative of a kind of overfitting. The opposite would be suggestive of true denoising. For reference, the value of Δ_O for the comparison of noiseless and noisy synthetic data – i.e. the variance expected by the presence of noise in the spectra – is shown in each case as a horizontal dashed blue line. The performance of the different methods is comparable with the SDSS data reconstruction (left panels), although the BF method appears to fare worse.

The more interesting results are found for the synthetic data (right panels), where we can discriminate between the recovery of the input, noisy data (Δ_N, blue stars), or a more desirable reconstruction of the original, noiseless, spectra. (Δ_O, red circles). One thing that stands out quite clearly is that the 4000Å break strength is poorly determined by the CS method. This may be quite expected, as the D$_n$(4000) is wider than the other two and relies on the continuum. However, we performed this test as there is a well-known degeneracy between parameters, so that, for instance, D$_n$(4000) indices are correlated with, e.g, Mgb. The strong covariance between line strengths found in Ferreras et al. (2023) indicates that the absorption line spectrum would encode similar information as the continuum. This experiment suggests that discovering and leveraging this trend is challenging for DL methods.

The figure also shows an uncanny inversion of the star - circle order (i.e. Δ_N vs Δ_O) in the BF method with respect to the other algorithms. We emphasize that this method does not use information from the ensemble, and only relies on a careful filtering of high frequencies, many of which would be ascribed to noise. This would be equivalent to a truncation in a Fourier series. The results presented here reveal that the BF method tends to overfit, so that it produces an optimal reconstruction when using noisy data as input, but underperforms when the noiseless synthetic data are considered.
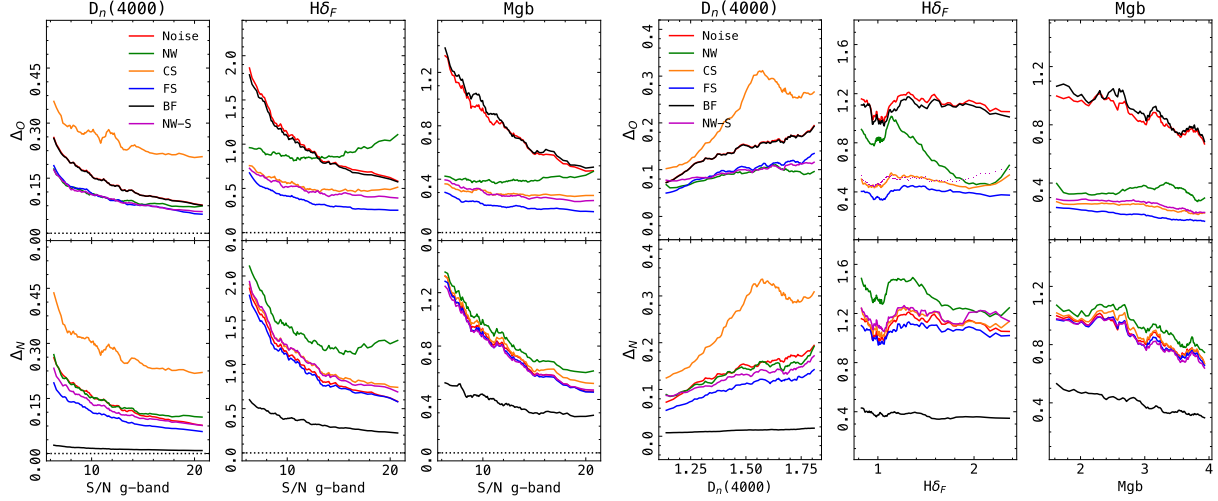
Figure 3. Standard deviation of the residuals of three line strengths, as labelled, showed with respect to the S/N in the SDSS-*g* band (left) and the actual line measurement (right). They correspond to synthetic data with added noise (see text for details). The comparisons are made between the recovered spectra and the original, noiseless data (Δ_O, top) or the noisy input (Δ_N, bottom).
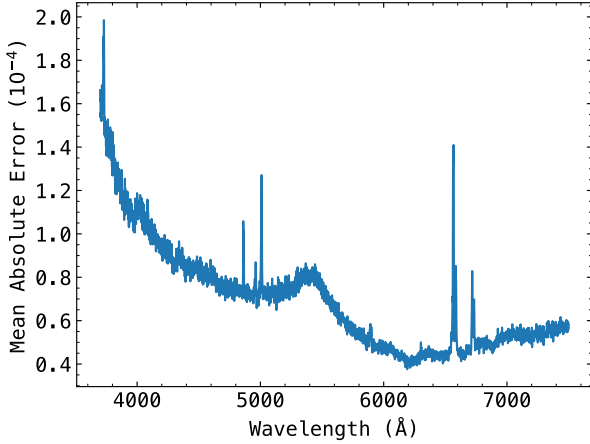

Figure 4. The limitations of the CS model are illustrated using the mean errors for each wavelength. To produce this plot, the absolute differences between model reconstructions and corresponding ground truth SDSS test set spectra were computed and then averaged.

Figure 5. Example of the recovery of a spectrum with original S/N∼5. Note how the Butterworth Filter (BF) optimises the residuals with respect to the input (noisy) data, i.e. overfits, whereas FS improves the residuals with respect to the original (noiseless) spectra.

The other methods perform equally well, although it appears that the FS model is favoured, as will be seen in more detail within the next figure.

The results for the reconstruction of the full batch of 10,000 synthetic data is shown in Fig. 3 as a running median when sorting the test sample with respect to the overall S/N (measured in the *g* band, left) or with respect to the measurement of the respective line strength (right). The latter gives an indication of the performance of the denoising with respect to the type of galaxy: for instance, younger stellar populations roughly correspond to a lower value of the 4000Å break strength. Please note that in order to focus more on the bluer part of the spectrum, where the signal tends to be weaker in most galaxies, we opt for the S/N averaged in the SDSS-*g* band (instead of the *r* band adopted in the threshold imposed for the sample, as discussed in § 2.
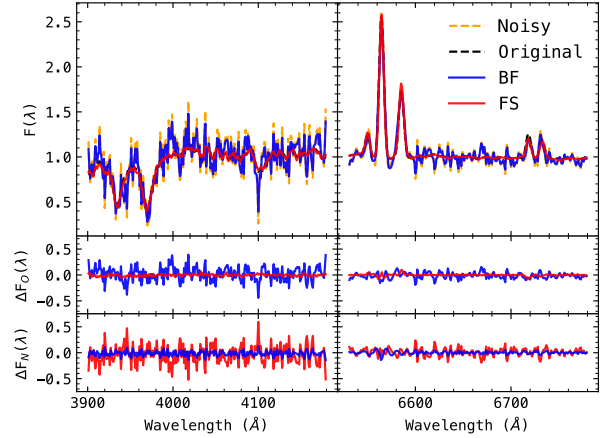
Both Δ_N (bottom panels) and Δ_O (top panels) are shown, for the same three line strengths. The characteristic decrease of the residual statistic is evident with respect to increasing S/N. The strong trend of Δ with $D_n(4000)$ is also expected, as stronger breaks imply fainter flux in the blue at around $\lambda \sim$3,800Å where the signal is weaker. The overfitting of the BF method (black lines) produces the lowest Δ_N (lower panels), but a residual statistic Δ_O that is comparable with the noisy data, i.e. the Butterworth-reconstructed spectra closely resemble the input noisy SDSS data, but not the original, noiseless data. Out of the other DL methods that appeared to perform similarly in Fig. 2, the FS (blue) appear to be optimal. The subpar performance of the CS model is also clear as in the previous plot, and so this model is discarded going forward. To better understand where it fails, Fig. 4 is included as a visualisation of the average error as a function of wavelength. Emission lines and blue wavelengths appear par-
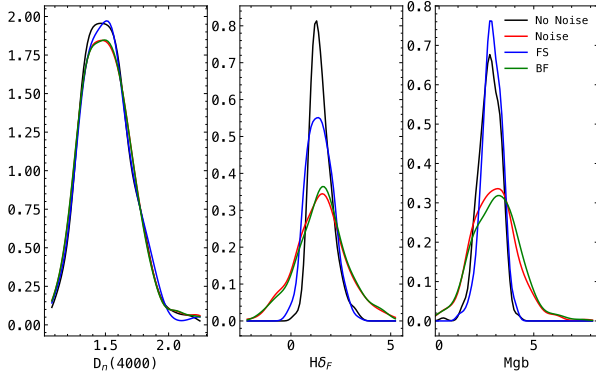
Figure 6. Distribution of the three targeted line strengths, for the original SDSS spectra, the noiseless fits and two of the methods explored, as labelled.

ticularly troublesome. This is expected, especially at the blue end of the spectrum where the dominant feature, $D_n(4000)$, is associated with particularly high variance. This figure will be explored further in Section 5.

Fig. 3 reveals that the NW and NW-S exhibit inferior denoising, producing higher residuals than FS in most cases. They are, however, attempting to fulfil a significantly more challenging reconstruction objective. Additionally, their bottleneck differs in that it does not arise from an autoencoder-style constriction, but a lack of information inherent to the training data. The skip connection within NW-S appears to assist in utilising this limited information; it is interesting to see skip connections providing a small benefit in this context, and the limited advantage of NW-S highlights the fundamental difficulty in this kind of recovery. As in the case of CS, we discard the NW models going forward.

As an illustration of the difference in the reconstruction, we show in Fig. 5 details of a noisy spectrum (S/N∼5) in two characteristic wavelength intervals, around the 4000Å break (left) and in the red region where prominent emission lines are present: the Hα and [NII] complex and the [SII] doublet. We show the noiseless best fits, the noisy synthetic data, and two of the methods: BF and FS. To ease the comparison, in addition to the standard representation with flux (top), we include the residual with respect to the noiseless spectrum (middle panels) and the noisy one (lower panels). As anticipated, BF produces smaller ΔF_N residuals, but higher ones in ΔF_O. The FS method shows an ideal behaviour and appears to be able to use information from the ensemble to minimise the differences with respect to the ground truth. Using the alternative, gentler set of BF parameters defined in Section 3, the fundamental issue of over-fitting to ΔF_N remained.

A final test to assess a possible systematic in the reconstruction is a comparison of the actual distribution of line strengths in the noiseless and noisy cases and in the reconstructions. This is shown in Fig. 6, where the histograms of the three line strengths are shown for the same cases as in the previous figure. The small difference in the distribution of $D_n(4000)$ is a consequence of the robustness of this index with respect to S/N. Indeed, this is one of the better indices to consider even if the S/N is not high, as the width of the index and the large contrast with respect to the stellar population parameters ensures a meaningful estimate even at S/N∼5.

The other two indices show a more characteristic behaviour between the BF (overfitting) algorithm and the DL method. Note that both $H\delta_F$ and Mgb produce distributions closer to the ground truth with FS, whereas BF closely resembles the wider shape of the histogram of noisy data. From these tests, we conclude that a DL method is successful at improving the quality of galaxy spectra if a representative, large ensemble with a wide range of S/N, including high quality data with the same instrumental / data characteristics are available.

5 EXPLAINING MODEL BEHAVIOUR

Deep models are often considered "black boxes" as their mappings and decision-making are notoriously difficult to interpret. To better understand how spectral features are being leveraged, we employ SHAP (SHapley Additive exPlanations, Lundberg et al. 2020). SHAP is a game-theoretic method that explains model outputs by assigning each feature an importance value, based on its contribution to the gap between the model's actual prediction and its mean prediction, averaged over all possible feature subsets. Fig. 7 shows the mean SHAP scores corresponding to input flux for the trained FS model. In this case, the "prediction" is the compressed, latent space encoding. While emission lines like Hα and [NII] clearly dominate as individual features, it is interesting to note that the importance and therefore predictive power of the continuum is biased considerably towards bluer wavelengths.

These findings are in broad agreement with the negentropy-based analysis of Ferreras et al. (2023), although the observed variation in SHAP scores suggests that "useful" information is much more broadly distributed across the continuum. SHAP has its limitations, and care must be taken when interpreting scores. Despite this, the observed discrepancy highlights the value of considering alternative proxies for information content beyond variance. Entropy-based techniques and classical data-driven methods such as PCA scale the variance in a way that may overlook subtle yet important dependencies within spectral data. Given the SHAP dominance of emission lines, such subtly may be encouraged by masking these features in future studies and therefore forcing models to leverage more obscure patterns.

The figure also highlights the scores associated with the red and blue regions. While previous research has suggested the strong information content of the former in terms of variance, it does not contain a particularly noteworthy share of the SHAP importance; the Fe and Mg features within this window do not appear to play a notable role. Nonetheless, the plot makes it clear that their combined input acts as a useful signal to the FS model, providing insight into how the NW models achieve a limited ability to predict some unseen spectral regions, as would be expected by the high level of "information" entanglement in the absorption spectra across a wide range of wavelengths.

Finally, an important consideration is the overall shape of the plot, which appears to closely mirror the error distribution shown in Fig. 4. In other words, the wavelengths that the CS model fails to represent accurately tend to hold strong predictive power in the FS model. This reinforces the idea that the continuum holds crucial information that pose a challenge for deep reconstruction methods, despite known
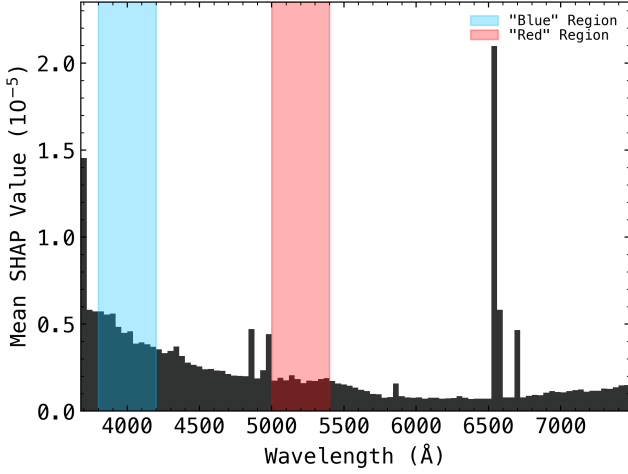
Figure 7. Using the FS model, mean SHAP values for flux are plotted as a function of binned wavelength (bin size = 40). The SHAP scores contained within the "blue" and "red" wavelength windows of interest are highlighted.

correlations between the continuum and absorption/emission lines.

6 CONCLUSIONS

This study highlights both opportunities and challenges when applying DL methods to the denoising and reconstruction of galaxy spectra. A key insight is that reproducing spectra through some form of information bottleneck with good generalisation limits the capacity to memorise random noise. Importantly, an MAE loss seems to enhance this denoising effect. Among the tested architectures, the FS model performed best, demonstrating that unsupervised autoencoders can serve as a practical tool to increase the S/N of both an SDSS training set as well as generalising to unseen spectra. We show in Fig. 8 a comparison of the recovered (in colours) and the original (grey) noisy SDSS spectra of six typical cases. The colour coding corresponds to their evolutionary classification as quiescent (red), star forming (blue) and AGN (Green), following the classification of (Sharbaf et al. 2023) based on the nebular emission properties. The figure zooms in three interesting spectral regions commonly used for the analysis of the stellar and gaseous phase of galaxies. We emphasize that the DL method has the advantage of extracting the information from the ensemble to "predict" the spectra at higher S/N, minimising the biases, limited, of course, to the parent sample.

Compared to traditional denoising algorithms, which require careful parameter tuning to balance smoothing with the preservation of sharp features, our approach leverages statistics learned from the ensemble in a fully emergent way. Tests using synthetic spectra suggest that classical filtering techniques tend to produce smoothed outputs which do not always represent the underlying spectrum. Our DL framework appears stronger at recovering this true signal, provided that the noisy input is processed to be consistent with training data. Furthermore, it stands distinct from other deep denoising models: unlike supervised approaches that rely on the
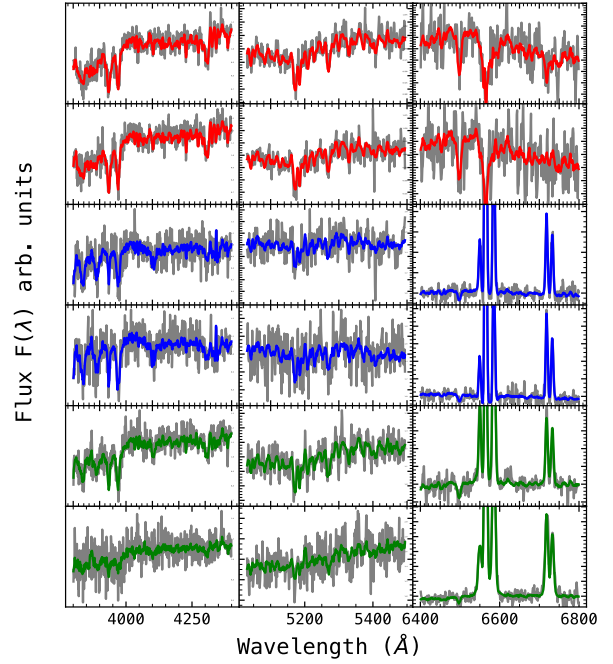

Figure 8. Illustration of the recovery with the FS architecture, in three characteristic spectral windows, as labelled. The shaded regions represent real galaxies at low S/N from SDSS and the lines show the recovery. A few cases are shown for galaxies classified as quiescent (red), star-forming (blue) or AGN (green), following the same criteria as in Sharbaf et al. (2023), based on the classification of the nebular emission lines.

addition of synthetic noise, our results emerge directly from real spectral data alone.

Beyond denoising capability, our SHAP analysis provides insight into how the models utilise spectral features, highlighting differences from traditional information-theoretic techniques and reinforcing the importance of the continuum. More generally, this line of investigation demonstrates the utility in employing explainability methods within astronomy. As DL is increasingly adopted within the field, it is crucial to maintain the capacity for human understanding.

Looking ahead, several avenues for improvement remain. The standard MAE loss used in this study treats all wavelengths as equally important. Instead, biasing the loss to focus on physically important regions could result in increased performance. In addition, neural networks are known to be biased towards learning low-frequency signals, as shown by Rahaman et al. (2018). Strategies, such as multi-stage neural networks, may assist in representing and utilising sharp features like emission lines. Finally, the adoption of more sophisticated DL architectures, like attention mechanisms (see Bahdanau et al. 2014), could better leverage the subtle correlations found within galaxy spectra.

ing for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

DATA AVAILABILITY

This work has been fully based on publicly available data: galaxy spectra were retrieved from the SDSS DR17 archive and stellar population synthesis models can be obtained from the respective authors. Code for each model presented in this paper will be made openly available on GitHub.

REFERENCES

Ahumada R., et al., 2020, ApJS, 249, 3
Aihara H., et al., 2011, ApJS, 193, 29
Bahdanau D., Cho K., Bengio Y., 2014, arXiv e-prints, p. arXiv:1409.0473
Balogh M. L., Morris S. L., Yee H. K. C., Carlberg R. G., Ellingson E., 1999, ApJ, 527, 54
Baron D., Poznanski D., 2017, MNRAS, 465, 4530
Conroy C., 2013, ARA&A, 51, 393
DESI Collaboration Abdul-Karim M., et al., 2025, arXiv e-prints, p. arXiv:2503.14745
Driver S. P., et al., 2019, The Messenger, 175, 46
Fabbro S., Venn K. A., O'Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, MNRAS, 475, 2978
Ferreras I., Hopkins A. M., Lagos C., Sansom A. E., Scott N., Croom S., Brough S., 2019, MNRAS, 487, 435
Ferreras I., Lahav O., Somerville R. S., Silk J., 2023, RAS Techniques and Instruments, 2, 78
Fitzpatrick E. L., 1999, PASP, 111, 63
Folkes S. R., Lahav O., Maddox S. J., 1996, MNRAS, 283, 651
Jin S., Trager S. C., et al., 2024, MNRAS, 530, 2688
Kingma D. P., Ba J., 2014, arXiv e-prints, p. arXiv:1412.6980
La Barbera F., Ferreras I., Vazdekis A., de la Rosa I. G., de Carvalho R. R., Trevisan M., Falcón-Barroso J., Ricciardelli E., 2013, MNRAS, 433, 3017
La Barbera F., Vazdekis A., Ferreras I., Pasquali A., Allende Prieto C., Röck B., Aguado D. S., Peletier R. F., 2017, MNRAS, 464, 3597
Liang Y., Melchior P., Hahn C., Shen J., Goulding A., Ward C., 2023, ApJ, 956, L6
Lovell C. C., Acquaviva V., Thomas P. A., Iyer K. G., Gawiser E., Wilkins S. M., 2019, MNRAS, 490, 5503
Lundberg S. M., et al., 2020, Nature Machine Intelligence, 2, 56
Madgwick D. S., et al., 2003, ApJ, 599, 997
McGurk R. C., Kimball A. E., Ivezić Ž., 2010, AJ, 139, 1261
Melchior P., Liang Y., Hahn C., Goulding A., 2023, AJ, 166, 74
Nikonov A. V., Davletshin R. V., Iakovleva N. I., Lazarev P. S., 2017, J. Commun. Technol. Electron., 62, 1048
Portillo S. K. N., Parejko J. K., Vergara J. R., Connolly A. J., 2020, AJ, 160, 45
Rahaman N., Baratin A., Arpit D., Draxler F., Lin M., Hamprecht F. A., Bengio Y., Courville A., 2018, arXiv e-prints, p. arXiv:1806.08734
Rogers B., Ferreras I., Peletier R., Silk J., 2010, MNRAS, 402, 447
Scourfield M., Saintonge A., de Mijolla D., Viti S., 2023, MNRAS, 526, 3037
Sharbaf Z., Ferreras I., Lahav O., 2023, MNRAS, 526, 585
Sharbaf Z., Ferreras I., Negri A., Angthopo J., Vecchia C. D., Lahav O., Somerville R. S., 2025, MNRAS, 539, 1480
Smee S. A., et al., 2013, AJ, 146, 32
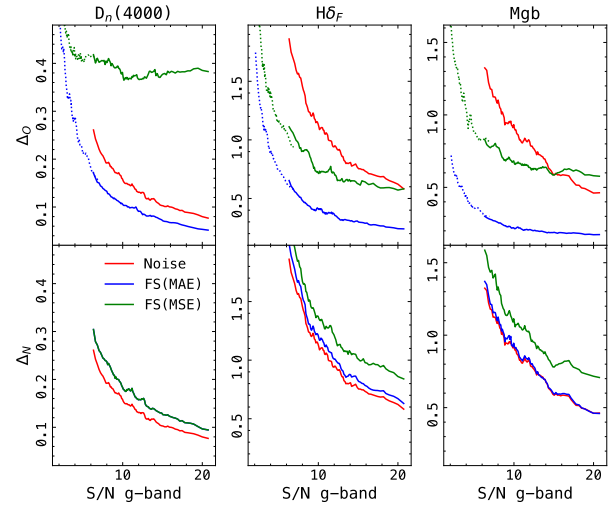Strauss M. A., et al., 2002, AJ, 124, 1810

Figure A1. Comparison of performance with two models that have the same architecture (FS) but different loss function, as labelled. The synthetic data in this case has a noise model corresponding to a Laplacian distribution.

Trager S. C., Worthey G., Faber S. M., Burstein D., González J. J., 1998, ApJS, 116, 1
Veit A., Wilber M., Belongie S., 2016, in Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16. Curran Associates Inc., Red Hook, NY, USA, p. 550–558
Walcher J., Groves B., Budavári T., Dale D., 2011, Ap&SS, 331, 1
Wang K., Guo P., Luo A. L., 2017, MNRAS, 465, 4311
Wild V., Hewett P. C., 2005, MNRAS, 358, 1083
Worthey G., Ottaviani D. L., 1997, ApJS, 111, 377
Wu Y., Tao Y., Fan D., Cui C., Zhang Y., 2024, MNRAS, 527, 1163
York D. G., et al., 2000, AJ, 120, 1579

APPENDIX A: A NOTE ON THE EFFECT OF THE LOSS FUNCTION

Among the many details in a DL model is the choice of the loss function. In this paper, we show that a standard architecture with a MAE loss function optimally performs in producing spectra at a higher S/N. However, the comparisons were made with synthetic data with noise being simplified to a Gaussian distribution with mean and standard deviation given by the observed flux and uncertainty in each spectral bin. Here we explore an additional case where the noise is modelled by a Laplacian distribution with scale given by the standard rule $\lambda = \sigma/\sqrt{2}$. Fig. A1 compares the results when Laplacian noise is added. We also show the same procedure for an identical DL architecture (FS), where the loss function is changed to MSE, which has a higher dependence on outliers. Furthermore, we included an additional batch of noisy data, where the original S/N was decreased by a factor of 3, presented as a dotted extension in the top panels. Note that even in the case of Laplacian noise (representing a higher number of "outliers" with respect to a Gaussian distribution) the FS(MAE) method performs well, whereas FS(MSE) produces noticeably higher residual statistic.

This paper has been typeset from a TeX/LaTeX file prepared by the author.