

Determining star formation histories and age-metallicity relations with convolutional neural networks

Enrique Galceran^{1,2} ^{*}, Patricia Sánchez-Blázquez^{1,2} , Artemi Camps-Fariña^{1,2} , Médéric Boquien³ , Ralf S. Klessen^{4,5} , Francesco Belfiore⁶ , Daniel A. Dale⁷ , Francesca Pinna^{8,9} , Ivan S. Gerasimov³ , Thomas G. Williams¹⁰ , and Hsi-An Pan¹¹ 

¹ Departamento de Física de la Tierra y Astrofísica, Universidad Complutense de Madrid, E-28040, Spain

² Instituto de Física de Partículas y del Cosmos, IPARCOS, E-28040, Spain

³ Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, France

⁴ Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Albert-Ueberle-Str. 2, 69120 Heidelberg, Germany

⁵ Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Im Neuenheimer Feld 225, 69120 Heidelberg, Germany

⁶ INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, I-50157, Firenze, Italy

⁷ Department of Physics and Astronomy, University of Wyoming, Laramie, WY 82071, USA

⁸ Instituto de Astrofísica de Canarias, calle Vía Láctea s/n, E-38205 La Laguna, Tenerife, Spain

⁹ Departamento de Astrofísica, Universidad de La Laguna, Avenida Astrofísico Francisco Sánchez s/n, E-38206 La Laguna, Spain

¹⁰ Sub-department of Astrophysics, Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH, UK

¹¹ Department of Physics, Tamkang University, No.151, Yingzhuan Road, Tamsui District, New Taipei City 251301, Taiwan

Received Decembre 4, 2025; accepted May 6, 2026

ABSTRACT

Context. Recovering the star formation and chemical enrichment histories of galaxies is essential for understanding the physical processes that govern their formation and evolution. Classical full spectral fitting techniques have enabled major advances in this field, but the inversion problem remains highly degenerate, particularly when the available data have limited wavelength coverage and moderate signal-to-noise ratios.

Aims. We aim to develop a state-of-the-art tool to infer detailed star formation histories (SFHs) and age-metallicity relations from realistic observational data, while mitigating classical degeneracies and substantially reducing computational cost. In particular, we seek to exploit the complementarity of spectroscopic and photometric data to improve constraints on the spatially resolved SFH and metallicity evolution of nearby galaxies in the PHANGS collaboration.

Methods. We constructed and trained a convolutional neural network (CNN) that combines convolutional layers, attention mechanisms, and a shared latent space to jointly predict SFHs and metallicities in 16 age bins. The network simultaneously processes integral-field spectroscopic data from PHANGS-MUSE and five-band photometric fluxes from PHANGS-HST. Training was performed on a dataset of 165 000 synthetic spectra and photometric measurements spanning a broad range of SFH shapes, metallicity evolution, dust attenuation, and signal-to-noise ratios representative of the observations.

Results. The CNN accurately recovers SFHs and age-metallicity relations over a wide range of evolutionary scenarios. The inferred mass- and luminosity-weighted mean ages and metallicities show a negligible bias, with dispersions of ~ 0.12 dex in age and ~ 0.03 dex in metallicity. When applied to real PHANGS-MUSE and PHANGS-HST data for NGC 3627, the network produces smooth, spatially coherent maps of stellar age and metallicity that recover physically meaningful structures, including younger populations tracing the spiral arms and star-forming regions. The CNN is approximately 5×10^3 – 2×10^4 times faster than traditional full spectral fitting codes, providing a powerful and efficient alternative for the analysis of large spectro-photometric surveys.

Key words. Methods: data analysis – Methods: numerical – Galaxies: stellar content

1. Introduction

The study of the star formation histories (SFHs) of galaxies offers crucial insights into their evolutionary pathways by enabling us to reconstruct how they assembled, while also highlighting the role of diverse physical processes in shaping the properties observed today (see e.g. Cid Fernandes et al. 2004; Hopkins et al. 2006; Asari et al. 2007; Panter et al. 2003; Sánchez-Blázquez et al. 2011; Carnall et al. 2018; Pessa et al. 2023). In the Milky Way and nearby Local Group galaxies, SFHs can be reconstructed in detail by fitting isochrones to colour–magnitude diagrams (CMDs) of resolved stars (see review by Tolstoy et al. 2009). For the vast majority of galaxies, however, the individual

stars cannot be resolved, and SFHs must instead be inferred by comparing their spectral energy distribution (SED) or the integrated spectral information with predictions from stellar population models.

Establishing a connection between models and observations requires a statistical inference framework. Full spectral fitting has become a widely used approach for deriving stellar population properties from integrated spectra, typically by comparing observational data with a combination of single stellar population (SSP) models (e.g. Ocvirk et al. 2006a,b; Tojeiro et al. 2007; Wilkinson et al. 2017; Cappellari 2023; Cid Fernandes et al. 2013). However, the inference of the SFH from integrated spectra is an ill-conditioned problem, where small uncertainties or noise in the input spectrum can lead to large, unstable, or non-unique variations in the derived SFH (e.g. Ocvirk et al. 2006a,b).

* e-mail: egalcera@ucm.es

Some codes alleviate this problem by establishing a regularisation of the inversion, favouring smoothly varying solutions over other equally probable solutions (e.g. [Ocvirk et al. 2006b](#); [Cappellari 2023](#)); however, their success depends heavily on the signal-to-noise ratio (S/N) of the input data. More recently, stellar population inference methods have increasingly adopted Bayesian formalisms (e.g. [Turner et al. 2021](#); [Johnson et al. 2021](#); [Maksymowicz-Maciata et al. 2024](#); [Iglesias-Navarro et al. 2024](#)), where Markov chain Monte Carlo (MCMC) sampling is commonly employed to characterise posterior distributions. Under the assumption of approximate Gaussianity, this allows for an efficient mapping of the degeneracies inherent in the high-dimensional parameter space, but it is very time-consuming.

Challenges persist, and different codes usually recover different solutions depending on the strategies used to fit the data, even when using the same model templates (see e.g. [Ruiz-Lara et al. 2015](#); [Sánchez et al. 2016](#); [Magris et al. 2015](#); [Ge et al. 2018](#)). Certainly, the problems arise primarily from the existing degeneracies between the parameters defining the solution ([Worthey et al. 1992](#)) and the limited ability of fitting strategies to distinguish subtle spectral differences or differences spread across correlated pixels.

Machine-learning (ML) techniques have seen significant advances, propelling them into everyday use. Scientific applications have also benefited from this improvement, which automates procedures that have traditionally required human supervision and improves the accuracy and speed of data analysis. Convolutional neural networks (CNNs), due to their capacity for feature extraction and pattern recognition, have proven to be versatile and extremely successful for several astrophysical tasks, such as the morphological classification of galaxies ([Dieleman et al. 2015](#); [Huertas-Company et al. 2015](#); [Cabayol et al. 2019](#); [Cheng et al. 2020](#); [Vega-Ferrero et al. 2021](#); [Walmsey et al. 2022](#); [Domínguez Sánchez et al. 2022](#); [Hausen & Robertson 2020](#)); the detection of low surface brightness features ([Domínguez Sánchez et al. 2023](#)); the stellar cluster classification ([Hannon et al. 2023](#)); the identification of deeply obscured clusters ([Graham et al. 2025](#)); determining stellar parameters ([Fabbro et al. 2018](#)); stellar cluster identification ([Maschmann et al. 2024](#)); among others.

The application of CNNs to stellar population inference has also shown promising results. Several studies have demonstrated that stellar ages and metallicities can be recovered with an accuracy comparable to, or even exceeding, that of traditional full spectral fitting methods when using photometric data as input (see e.g. [Liew-Cain et al. 2021](#); [Walter et al. 2026](#)). More recently, [Woo et al. \(2024\)](#) reported substantial improvements in both the bias and the scatter of mean age and metallicity estimates derived from spectra relative to conventional fitting techniques. The work of [Lovell et al. \(2019\)](#) pioneered the use of CNNs to recover not only average stellar ages and metallicities, but also the full star formation history. However, their model assumed a single, constant metallicity. To the best of our knowledge, the ability of CNNs to infer the same level of detail as full spectral fitting methods, i.e. both the full SFH and the age-metallicity relation, has so far been explored only in [Anwar et al. \(2025\)](#).

In this work, we present a new CNN to infer the SFH and the age-metallicity relations from a combination of spectra observed with the multi unit spectroscopic explorer (MUSE) integral field spectrograph at the European Southern Observatory very large telescope (ESO VLT) and photometric fluxes in five wide field camera 3 (WFC3) at the Hubble space telescope (HST) near ultra violet (NUV) and optical bands that mimic

the data taken as part of the physics at high angular resolution in nearby galaxies surveys¹ (PHANGS). The PHANGS survey aims to provide a unified picture of star formation across the different phases of the interstellar medium in a representative sample of nearby, massive, star-forming galaxies (see [Leroy et al. 2021](#)). Its main objectives are to constrain star-formation timescales and efficiencies, assess the role of stellar feedback processes, trace chemical enrichment and mixing across galaxy discs, and understand how young stellar clustering both emerges from and reshapes the surrounding medium. To this aim, the survey is obtaining multi-wavelength observations able to resolve key structures such as molecular clouds, H II regions, and star clusters on scales of a few to ~ 100 pc in a sample of star-forming galaxies (see survey papers [Leroy et al. 2021](#); [Lee et al. 2022, 2023](#); [Emsellem et al. 2022](#); [Chandar et al. 2025](#)). It also samples a wide range of galactic environments—bars, spiral arms, and centres—to evaluate environmental effects, while ensuring consistency with the broader population of star-forming galaxies along the main sequence. Our work directly supports several key goals of the PHANGS project, including constraining the timescales of star-formation processes, quantifying the role of different stellar feedback mechanisms, and resolving chemical enrichment and mixing across galaxy discs.

This article is structured as follows. Section 2 describes the observational datasets that motivate this work, together with the construction and encoding of the synthetic training set designed to reproduce their spectro-photometric characteristics. Section 3 details the architecture and implementation of the CNN and outlines the configuration of the pPXF full spectral fitting used for comparison. The performance of both approaches is evaluated in Sect. 4, where we also explore the robustness of the CNN to different input conditions. In Sect. 5 we compare the performance of our trained CNN using real observational data from the nearby spiral galaxy NGC 3627. Finally, Sect. 6 discusses the advantages and limitations of our method and its prospects for deriving realistic SFHs and age-metallicity relations. In this paper, we assume a flat Λ CDM cosmology with $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$, and a Hubble constant $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

2. Data

In this work, we simulated observations designed to reproduce the characteristics of the PHANGS-MUSE ([Emsellem et al. 2022](#), hereafter [EM22](#)) and PHANGS-HST ([Lee et al. 2022](#)) surveys. The PHANGS-MUSE dataset comprises 168 pointings obtained with the MUSE integral-field spectrograph at the ESO VLT, targeting 19 nearby galaxies in the redshift range $0.0012 \leq z \leq 0.0046$ and spanning stellar masses $9.4 < \log(M_\star/M_\odot) < 11.0$. These observations achieve a median spatial resolution of ~ 50 pc. The spectra cover the wavelength range 4750–9350 Å, with a spectral resolution varying from FWHM ~ 2.9 Å at 4800 Å to ~ 2.4 Å at $\lambda > 6500$ Å. In this work, we restrict our analysis to $\lambda < 7000$ Å to minimise contamination from sky-subtraction residuals (see [EM22](#)).

The PHANGS-HST survey provides imaging in the *NUV–U–B–V–I* bands for 38 nearby spiral galaxies, including the full PHANGS-MUSE sample, with a median spatial resolution of ~ 5 pc. The inclusion of the NUV band is particularly valuable, as emission in this wavelength range is highly sensitive to small fractions of young stellar populations, making it a powerful tracer of recent, low-level star formation that may remain

¹ <http://www.phangs.org>

undetected in optical spectra (e.g. Werle et al. 2019; Salvador-Rusiñol et al. 2020). Moreover, while different combinations of age and metallicity can produce nearly indistinguishable optical spectra, they give rise to significantly different spectral energy distributions in the NUV. Consequently, the inclusion of NUV photometry helps to break the age–metallicity degeneracy and improves the robustness of the inferred stellar population properties (e.g. Kaviraj et al. 2007; Anghopo et al. 2024).

2.1. Generation of synthetic SFH

Supervised ML requires a training set with known inputs (the observable quantities) and corresponding outputs (labels). In our case, the labels are the stellar mass fractions and mean metallicities in 16 age bins. To construct this dataset, we generated a diverse set of synthetic SFHs and computed the associated spectra and HST fluxes under different assumptions of chemical enrichment and dust attenuation.

We employed the GP-SFH module of DenseBasis (Iyer et al. 2019), which parametrises SFHs using five quantities: the total stellar mass, the specific star formation rate (sSFR) at $z=0$, and three look-back times at which the cumulative stellar mass reaches fixed fractions of the final mass. We fixed these fractions to 1/4, 1/2, and 3/4.

Since our objective is not to infer the total stellar mass², we set this parameter to $15 \cdot 10^6 M_\odot$. The remaining parameters were chosen randomly from the following distributions.

The sSFR at $z = 0$ was sampled from a uniform distribution in the range $10^{-12} \text{ yr}^{-1} \leq \text{sSFR} \leq 10^{-9} \text{ yr}^{-1}$. Although galaxies in the PHANGS sample are selected to lie on the star-forming main sequence (with global sSFR typically above 10^{-11} yr^{-1}), our analysis focusses on spatially resolved regions within galaxies, where local sSFR values can vary above and below the global average.

The three look-back times were drawn from a Dirichlet distribution (Multivariate beta; Kotz et al. 2000) with concentration parameter $\alpha = 1$. Despite Iyer et al. (2019) recommending $\alpha = 5$ for SED fitting applications, we found that $\alpha = 1$ provides a more balanced mixture of narrow and extended SFH profiles for the purpose of generating a wide variety of training examples.

Using GP-SFH, we generated 150 000 synthetic galaxies. However, the resulting colour distribution did not sufficiently populate the reddest and bluest regions of the PHANGS colour space. To better sample these extremes, we generated an additional 15 000 SFHs using skew-normal functions, designed to represent regions with very old populations and red colours or regions where young stellar clusters dominate the light.

Skew-normal functions are characterised by a peak age, a width, and a skewness parameter, allowing for flexible control of both the timing and asymmetry of the SFH. The peak age was drawn from a Gaussian distribution with mean $\mu = 12 \text{ Gyr}$ and standard deviation $\sigma = 3 \text{ Gyr}$, ensuring that the bulk of the star formation occurs at early cosmic times. We sampled the width of the distribution from a Gaussian distribution with $\mu = 1 \text{ Gyr}$ and $\sigma = 0.4 \text{ Gyr}$, favouring relatively short formation timescales. The skewness parameter was drawn from a uniform distribution in the range $-0.25 \leq \alpha \leq 0.75$. This asymmetric range was intentionally chosen to preferentially generate SFHs skewed towards earlier times, thereby enhancing the number of systems dominated by old stellar populations.

² The stellar mass can be determined a posteriori from the inferred mass-to-light ratio and the observed flux.

In 40% of these SFHs, we superimposed a secondary, short-duration burst of recent star formation (40–50 Myr) occurring within the last 500 Myr. The burst age was sampled from a log-uniform distribution, while its width and mass fraction (ranging from 1–10% of the total stellar mass) were drawn from uniform distributions within the specified limits. All parameter distributions are summarised in Table 1.

In total, we generated 165,000 SFHs:

- 150 000 with GP-SFH,
- 9 000 with the skewed-normal parametrisation, and
- 6 000 with a skewed-normal SFH plus an added burst.

All SFHs were normalised to the same total mass as DenseBasis. The resulting training set spans a wide variety of evolutionary shapes—from smooth, gradually declining histories to rapidly varying cases with young bursts at different epochs. We did not constrain the SFHs to follow physically motivated evolutionary tracks, as our primary goal is to train the network to recognise how variations in the SFH at different times imprint themselves on the spectrum and photometry. Furthermore, the network is designed to analyse spatially resolved SFHs and age-metallicity relations, whose shapes can differ significantly from galaxy-integrated averages.

Table 1: Distributions of parameters for skewed-normal SFHs.

Parameter	Distribution	μ / min	σ / max
Skewed-normal SFH parameters			
Peak age	normal	12 Gyr	3 Gyr
Width	normal	1 Gyr	0.4 Gyr
Skewness	uniform	−0.25	0.75
Burst parameters			
Age	log-uniform	10 Myr	500 Myr
Width	uniform	40 Myr	50 Myr
Mass fraction	uniform	1%	10%

Notes. For Gaussian distributions, μ and σ denote the mean and standard deviation; for uniform and log-uniform distributions, the listed values indicate the minimum and maximum limits.

2.2. Generation of synthetic observations

For each SFH, we used ProSpect (Robotham et al. 2020) to generate the synthetic spectra and photometric fluxes used as inputs to our CNN. ProSpect computes synthetic spectra from specified star formation and metallicity histories, incorporating both dust attenuation and emission through the Charlot & Fall (2000) two-component attenuation model and the Dale et al. (2014) templates, respectively, as well as ionised gas emission lines using the MAPPINGS-III (Allen et al. 2008) photoionisation tables. Other widely used codes, such as MAGPHYS and CIGALE (da Cunha et al. 2008; Noll et al. 2009; Boquien et al. 2019), employ similar strategies to produce synthetic observables. However, ProSpect provides great flexibility to customise and extend the available templates for stellar and dust emission, and to implement a wide range of prescriptions for SFH and age-metallicity relations.

We used the E-MILES stellar population models (Vazdekis et al. 2010, 2016), with the extension to young ages (6.3 Myr to 63 Myr) as described in Asa'd et al. (2017). The E-MILES

templates cover the wavelength range $\lambda = 1680 - 50\,000\text{ \AA}$ with a spectral resolution of $\text{FWHM} = 2.5\text{ \AA}$, combining four empirical stellar libraries: NGSL (Gregg et al. 2006), MILES (Sánchez-Blázquez et al. 2006), Indo-US (Valdes et al. 2004), and CaT (Cenarro et al. 2001). We adopted the version based on the Padova2000 isochrones (Girardi et al. 2000), which spans ages from 63 Myr to 17.8 Gyr and includes six metallicities. The reason is that this version can be reliably joined at 63 Myr with the young extension (ages between 6.3 and 63 Myr; Asa'd et al. (2017)), which uses scaled-solar isochrones from Bertelli et al. (1994), extended to low-mass stars ($M < 0.5M_{\odot}$) with the stellar tracks from Pols et al. (1995). We use the versions computed with a Chabrier (2001) initial mass function (IMF).

Dust extinction was incorporated by multiplying each model spectrum by the transmission function

$$T(\lambda) = \exp \left[-\tau_{\lambda_{\text{piv}}} \left(\frac{\lambda}{\lambda_{\text{piv}}} \right)^n \right], \quad (1)$$

where $\tau_{\lambda_{\text{piv}}}$ is the optical depth at $\lambda_{\text{piv}} = 5500\text{ \AA}$ and n is the attenuation slope. We adopt the two-component attenuation prescription of Charlot & Fall (2000), where the optical depth includes a diffuse component, τ_{ISM} , and, for stellar populations younger than 10 Myr, an additional birth-cloud contribution, τ_{BC} . We fixed $n = 0.7$, consistent with the empirical correlation between the far-infrared-to-ultraviolet luminosity ratio and the UV spectral slope in star-forming galaxies (Charlot & Fall 2000).

The fiducial model of Charlot & Fall (2000) adopts $\tau_{\text{ISM}} = 0.3$ and $\tau_{\text{BC}} = 1$, values that reproduce the typical attenuation of O-type and supergiant stars in the Milky Way. However, the effective optical depth is expected to vary from galaxy to galaxy, depending on global properties such as metallicity and gas content, as well as geometric factors such as inclination. To account for this diversity, we introduced object-to-object variations in the synthetic dataset by sampling these parameters from physically motivated distributions. Specifically, τ_{BC} was drawn from a log-uniform distribution in the range 0.5–4.0, while τ_{ISM} was sampled from a Gaussian distribution centred at 0.3 with a standard deviation of 0.08. These ranges were chosen to reproduce the attenuation properties observed in the PHANGS–MUSE sample (see EM22), inferred from the Balmer decrement for τ_{BC} and from the stellar $E(B - V)^3$ derived through full spectral fitting for τ_{ISM} . The resulting distributions are shown in Fig. 1.

Although the wavelength coverage of our current dataset does not reach the far-infrared (FIR), we account for dust re-emission of absorbed light by incorporating the FIR template library of Dale et al. (2014). This inclusion enables future extensions of the model to incorporate PHANGS–JWST (Lee et al. 2023) observations. At each stage of the attenuation process, we enforce energy conservation by re-emitting the total absorbed luminosity of the stellar radiation field, which determines the temperature of the emitting dust. This allows for different values of τ to be applied to the diffuse ISM and the birth cloud.

ProSpect incorporates a simple energy balance scheme to produce nebular emission features for a range of gas-phase metallicities. The key default assumption is that flux short of the Lyman limit is absorbed with an efficiency determined by the UV photon escape fraction. The integrated intrinsic stellar flux is then re-emitted using line energies determined by MAPPINGS–III as per the tables provided by Levesque et al. (2010).

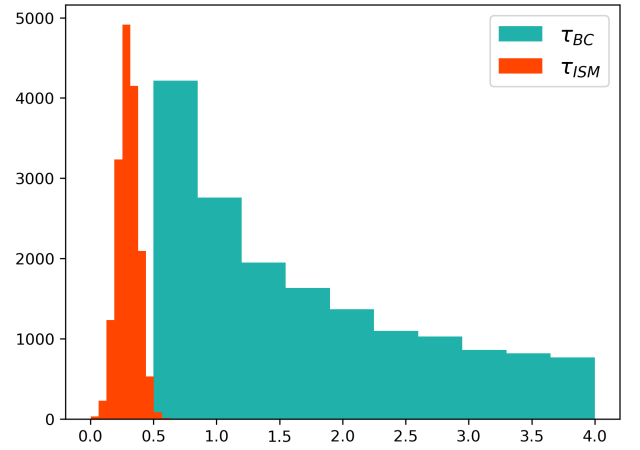


Fig. 1: Distribution of optical depths for the diffuse and birth-cloud components of the Charlot & Fall (2000) attenuation curve used to simulate internal extinction in the synthetic data.

Rather than generating models with constant metallicities, we implemented a simple prescription in which metallicity increases in tandem with star formation. To avoid enforcing a specific (and potentially unrealistic) enrichment law, we allowed the metallicity yield and the final metallicity to vary among models.

The metallicity evolves with stellar age according to

$$Z(\text{age}) = Z_{\text{start}} + (Z_{\text{final}} - Z_{\text{start}}) \times \frac{m_{\text{age}}}{M_{\text{tot}}}, \quad (2)$$

where Z_{start} and Z_{final} represent the metallicities of the oldest and youngest stellar populations, respectively, and $\frac{m_{\text{age}}}{M_{\text{tot}}}$ is the cumulative mass fraction formed up to a given age bin. We fixed $Z_{\text{start}} = 0.008$, while Z_{final} was randomly sampled between 0.010 and 0.030, effectively allowing for a range of enrichment slopes.

To obtain the HST fluxes, we convolved the synthetic spectra with the transmission curves of the filters available for the PHANGS–HST sample (WFC3: F275W, F336W, F438W, F555W, and F814W). The values of the filters were kept in the same units as the spectra and not converted into magnitudes.

2.3. Preprocessing of synthetic data

To mimic MUSE observations, we resampled the synthetic spectra to match the wavelength-dependent resolution and sampling of the MUSE observations (4800Å to 7000Å with a step of 1.25Å). Furthermore, we applied an additional broadening to all synthetic spectra up to a common velocity dispersion of 150 km s^{−1}. We did not add any velocity shift to the spectra.

These same preprocessing steps must be applied to the observed spectra before using the CNN to infer their properties. Specifically, the spectra need to be shifted to the rest frame, broadened to a common velocity dispersion of $\sigma_f = 150\text{ km s}^{-1}$, using $\sigma_{\text{broadening}} = \sqrt{150^2 - \sigma_{\text{gal}}^2}$ (km s^{−1}), where σ_{gal} represent the stellar velocity dispersion. Additionally, each spectrum was normalised by its mean flux in the 5300–5500 Å window, chosen because it aligns with the F555W pivot wavelength and is free of strong emission lines.

As our goal is to infer stellar population properties from real galaxies, we incorporated noise into the synthetic data to better mimic observational conditions. Although the PHANGS–MUSE spectra are Voronoi-binned to achieve a minimum S/N of ~ 35

³ A selective extinction $R_V = 3.1$ is assumed.

per Å, internal tests showed that training on a range of S/N values leads to improved model performance compared to training at a single value. We therefore added Poisson noise to the synthetic spectra, scaling its amplitude to produce S/N values of 15, 20, 35, and 50. In Sect. 4.2, we further assess the performance of the model in relation to the noise level of the input data, and to more realistic noise patterns derived from the residuals of fits to real data.

3. Neural network model

Our CNN architecture combines convolutional, pooling, attention, and fully connected (dense) layers. A brief overview of their functions is provided below, but we refer to Dumoulin & Visin (2016); Khan et al. (2020); Li et al. (2022); and Zhao et al. (2024) for a more detailed explanation. For an in-depth analysis of the peculiarities of deep learning and applications of CNNs, we refer to the excellent review LeCun et al. (2015).

A convolutional layer is the fundamental building block of a CNN. It applies a set of learnable filters (or kernels) that convolve across the input data to extract local features such as edges, gradients, or spatial patterns (Dumoulin & Visin 2016). Each filter produces a feature map that highlights specific structures in the input, enabling the network to learn hierarchical representations: from low-level features in the first layers to increasingly complex patterns in deeper layers (Lecun et al. 1998). This local connectivity and weight sharing make convolutional layers particularly effective for analysing structured data such as images, spectra, or spatial maps. We use one-dimensional (1D) convolutions, which are well-suited to spectral data.

A max-pooling layer is a downsampling operation commonly used in CNNs to reduce the spatial dimensionality of feature maps while retaining the most relevant information. It divides the input into small regions and outputs the maximum value within each region, effectively preserving the strongest responses of the convolutional filters. This operation allows our network to focus on spectral or spatial features that exhibit the maximum response to variations in stellar population properties, while enhancing translational invariance, reducing computational cost, and mitigating overfitting (Dumoulin & Visin 2016).

A fully connected (dense) layer connects every neuron in one layer to every neuron in the next, allowing the network to combine and interpret the features extracted by previous convolutional or pooling layers (Rumelhart et al. 1986). Each neuron computes a weighted sum of its inputs, followed by a non-linear activation, enabling it to learn complex relationships among features. In this work, fully connected layers integrate spectral patterns identified in earlier stages of the network to produce global predictions, such as SFH and age-metallicity relations.

An attention mechanism (originally developed for sequence processing in natural language tasks for large language models; Vaswani et al. 2017) is a computational module designed to enhance the performance of NNs by allowing them to focus selectively on the most informative parts of the input data. Instead of processing all features with equal weight, the network learns to assign higher importance to specific regions or features that are most relevant for the task at hand. In astrophysical applications, attention mechanisms enable the model to identify and emphasise key spectral or spatial signatures while suppressing less informative regions. Furthermore, attention enhances the network’s ability to capture long-range dependencies—such as correlations between spectral features at widely separated wavelengths—that may be inaccessible to local convolution

operations alone (Melchior et al. 2023). These dependencies change dynamically; that is, the network focusses on the relevant features on a case-by-case basis.

3.1. Architecture

Our network takes as input a spectrum and five photometric bands (see Sect. 2) and gives as output the stellar mass contribution and their average metallicity in 16 different age bins with central values: 0.007, 0.021, 0.0425, 0.076, 0.1145, 0.1615, 0.23, 0.345, 0.545, 0.865, 1.37, 2.17, 3.44, 5.455, 8.645, and 14.2 Gyr. The central ages of the bins were chosen to sample, in approximately uniform intervals, the age range covered by the models on a logarithmic scale. As a result, the bin widths are narrower at younger ages, where the observational characteristics exhibit stronger variations. The total number of bins was set to the maximum value at which the network’s results showed no improvement.

We note that the ages of some SSP templates in the model grid formally exceed the age of the Universe. These very old ages should not be interpreted literally, but rather as reflecting the upper limit of the ages permitted by the models. Current uncertainties in the model ingredients can result in offsets of a few Gyr in the age scale; for example, a modest 50–100 K offset in the colour– T_{eff} transformation, well within typical uncertainties, can translate into differences of ~ 2 Gyr in the inferred ages (see e.g. Vazdekis et al. 2010). Although some authors choose to exclude the oldest templates from their analyses, doing so would impose an artificial truncation on the parameter space and introduce a saturation effect, biasing age estimates towards the imposed maximum rather than reflecting the intrinsic uncertainties of the models.

In Fig. 2 we present a schematic representation of the architecture. The model is split into three main sections:

- The first section consists of two parallel branches that encode the spectral and photometric information independently. The branch that processes the spectra has three convolutional and max-pooling layers, followed by an attention layer and two fully connected layers. The second branch, dedicated to the photometric data, consists of three fully connected layers. The photometric branch is smaller and shallower, as the number of input data is significantly lower (five photometric inputs vs 1761 spectral points), albeit each element is more information-rich.
- The middle section of the network contains that latent space, which combines the outputs from the two encoding branches. The features of these outputs (64 features from the spectral branch and 32 from the photometric branch) are concatenated and passed to a single layer referred to as the latent space. In our model, the latent space is a 40-dimensional compressed representation that integrates the most informative features extracted from both inputs. The latent space acts as a bottleneck, constraining the information transmitted to the decoding block to the most relevant features for accurately inferring the target outputs (i.e. the SFH and metallicity evolution).
- Finally, the decoding block transforms latent-space information into the target outputs using a single branch composed of three fully connected layers. Combining both outputs within a single branch, rather than defining separate decoders, reduces the number of trainable parameters and mitigates the risk of overfitting. From a physical perspective, this structure also reflects the intrinsic connection between stellar mass

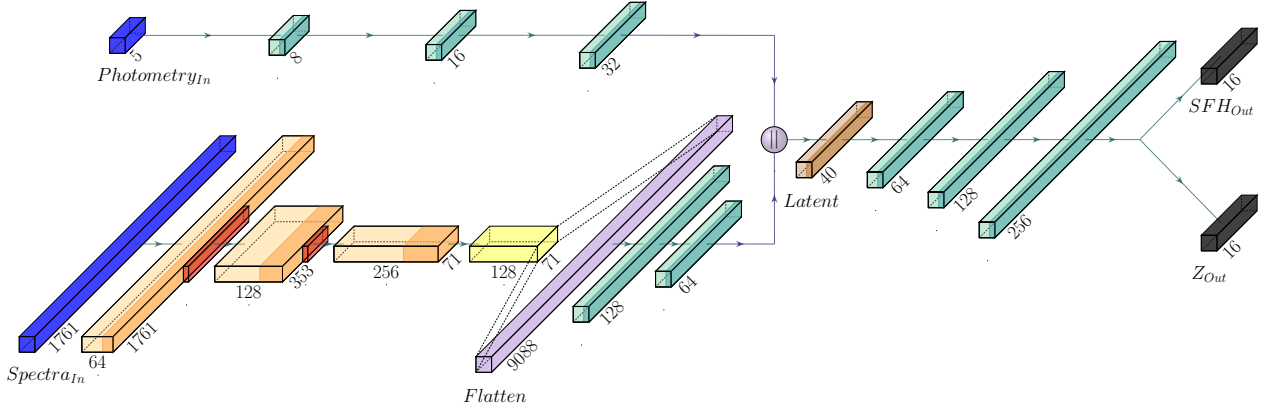


Fig. 2: Schematic representation of our CNN architecture as described in Section 3.1. The colour code identifies the different types of layers. Orange: convolutional layers, red: max pooling layers, yellow: attention layer, purple: flattening layer, cyan: dense layers. Darker shades, when present, represent the activation and Instance Normalisation. The latent, input and output layers are represented in brown, blue, and black, respectively, while the sphere indicates the position of the concatenation between the spectrum and photometry branches.

growth and chemical enrichment: both the SFH and metallicity evolution are strongly interdependent, and much of the relevant information linking the latent space to the outputs can therefore be shared. Previous studies (e.g. Krizhevsky et al. 2017; Hannon et al. 2023) have demonstrated that, with appropriate training, multiple correlated quantities can be effectively represented within a shared intermediate layer, without requiring independent processing for each output.

Normalising values across layers helps stabilise the learning process, improve training efficiency, and reduce internal covariate shift. batch normalisation (Ioffe & Szegedy 2015), which adjusts the distribution of values within each batch, is one of the most widely used methods for this task. However, we adopted instance normalisation (IN, Wu & He 2018), which normalises each sample independently rather than across a batch (i.e. all samples in a single training step). Note that this normalisation is applied to the features extracted by each layer within the CNN, and it is not related to the normalisation of input data (see Sect. 2.2). The IN works better than batch normalisation when batch statistics vary widely, for example, if the training data includes heterogeneous galaxy spectra. We found that IN provided greater training stability and slightly better predictive accuracy, likely because it reduced inter-sample dependencies and improved the consistency of feature scaling across diverse inputs. Several works have shown that the inclusion of IN often reduces the need for dropout layers (see e.g. Ioffe & Szegedy 2015). We found that once IN is added, the dropout layers (which set random neurons to zero to reduce overfitting and help with regularisation) did not improve the solution and instead slowed training, leading to their removal.

The number of layers and neurons per layer (i.e. depth and width of the hidden layers) of each branch were chosen according to the lowest number of each that achieved the best fit. Increasing the number of layers or neurons did not appreciably improve the network’s accuracy, so we kept the network as small as possible to avoid overfitting.

3.2. Activation function

Activation functions introduce non-linearity into NNs, allowing them to model complex relationships between inputs and out-

puts. In our model, we employ parametric rectified linear unit (PReLU; He et al. 2015), a variant of the standard rectified linear unit (ReLU; $\text{ReLU} \equiv \max(0, x)$) that introduces a learnable parameter controlling the slope for negative input values. While the traditional ReLU sets all negative activations to zero, potentially leading to inactive (‘dead’) neurons (which leads to the dying ReLU problem; Lu et al. 2020), PReLU allows a small, trainable negative response, enabling the network to retain gradient flow even for negative inputs. The function is defined as

$$\text{PReLU}(\alpha, x) = \max(0, x) + \alpha \cdot \min(0, x), \quad (3)$$

where α is a trainable parameter that controls the slope of the negative inputs. There is no clear consensus on what the optimal placement of the activation function is within the network architecture with respect to the intermediate regularising and normalising step (in our case, the IN layer). We explored all possible permutations of the order of the IN, activation and dropout layers and found differences in loss value to be $\lesssim 5\%$ (see Sect. 3.3). Ultimately, our final configuration only applies the activation function after the IN layer to avoid potential artefacts in the normalisation caused by the activation function.

3.3. Loss function

The loss function quantifies the discrepancy between the model’s predictions and the true target values, providing the basis for optimisation during training. We tested several alternatives and found that the best performance is achieved using the mean squared error (MSE), the symmetric mean absolute percentage error (SMAPE), and the logarithm of the hyperbolic cosine (Log-Cosh) loss functions.

The MSE loss function is often used as the default loss function for regression tasks because it is simple, differentiable, and easy to optimise with gradient-based algorithms such as gradient descent. However, MSE assumes that the target outputs follow a normal distribution, and as the errors are squared, large deviations have a disproportionately strong impact on the loss. This property can bias the predicted SFH, since larger errors are expected for older stellar populations that are less constrained by the observations. To mitigate this effect, we additionally tested

SMAPE and Log-Cosh. The former expresses the deviation between predicted and true values as a percentage, independent of whether the prediction is an over- or underestimation (see e.g. Lovell et al. 2019), while the latter provides a smooth approximation to the mean absolute error (MAE) and is less sensitive to outliers than MSE. Both SMAPE and Log-Cosh are therefore considered more robust to outliers and sparsity in the data (Saleh & Ehsanes Saleh 2022).

No significant differences were found between the predictions of the model trained with SMAPE and Log-Cosh. Still, Log-Cosh had the fastest convergence during the training stage, and we therefore adopted this loss function. Log-Cosh introduces a smoother transition around zero (i.e. perfect prediction), allowing the network to fine-tune its weights more efficiently, thereby reducing the risk of overfitting and improving overall training stability. The Log-Cosh function is defined as

$$\text{LogCosh}(y, y^p) = \sum_{i=1}^n \log \left[\cosh(y_i^p - y_i) \right], \quad (4)$$

where y and y^p are the target and predicted vectors, respectively. We compute two independent loss values corresponding to the two output vectors—one for the SFH and the other for the age-metallicity relation—and combine them into a single weighted loss function. Since the accurate recovery of the SFH is our primary objective, its contribution to the total loss is increased as much as possible without degrading the metallicity predictions. We found an optimal weighting ratio of 80:20 between the SFH and the age-metallicity relation, yielding a final loss function of the form

$$\mathcal{L}(y_{\text{SFH}}, y_Z, y_{\text{SFH}}^p, y_Z^p) = 0.8 \cdot \text{LogCosh}(y_{\text{SFH}}, y_{\text{SFH}}^p) + 0.2 \cdot \text{LogCosh}(y_Z, y_Z^p). \quad (5)$$

3.4. Training

The network weights were initialised using the uniform scheme of Glorot & Bengio (2010) and optimised via Adam (Kingma & Ba 2014) with an initial learning rate of 10^{-4} and a linear decay schedule⁴. From the 165 000 synthetic observations, 70% were used for training, 20% for validation, and 10% for testing. Training proceeded in epochs, with mini-batches of 128 samples per iteration. After each epoch, model performance was evaluated on the validation set, and the resulting validation score was used to track convergence and mitigate overfitting.

To improve training stability and efficiency, we employed three callback strategies. First, the learning rate was automatically reduced by a factor of 0.8 when the validation loss did not improve for five consecutive epochs. Secondly, we saved the model weights corresponding to the lowest validation loss to prevent overfitting. Thirdly, we stopped training pre-emptively when the validation loss did not improve for 20 consecutive epochs. The full training run required approximately four hours on a GPU cluster with an NVIDIA GeForce RTX 2070 Ti and about 12 hours on an Apple Silicon M2 processor.

Given the dimensionality of the output space, we evaluated the network performance as a function of the size of the training-set. We constructed randomly selected subsets of the full training set with sizes ranging from 20 000 to 165 000, and trained identical networks on each subset. We evaluated the final loss reached by the CNN, finding that it steadily decreases with increasing

training size until $\sim 125\,000$, beyond which increasing the size of the training sample does not significantly improve the predictions (see Fig. A.1 in Appendix A). This behaviour indicates that the adopted training-set size is sufficient to constrain the model and mitigates concerns about overfitting driven by an undersampled training distribution. Once training is complete, we evaluate the performance of the CNN model using the independent test dataset (see Sect. 4).

3.5. pPXF analysis

We run penalized pixel-fitting (pPXF; v9.4.2; Cappellari 2012, 2023) on the same test dataset used for evaluating the CNN model. pPXF is a widely adopted method for deriving SFHs and age-metallicity relations by minimising the residuals between an observed spectrum and a linear combination of SSP templates through a non-linear least-squares optimisation. Its most recent implementation allows the simultaneous fitting of spectra and photometry (Cappellari 2023), enabling a direct comparison with our network under identical input conditions and using the same SSP template set (see Sect. 2.2). The fitting procedure followed the default configuration recommended for datasets of this type: a regularisation error of 1%, a multiplicative polynomial of order 8, no additive components, and a Calzetti et al. (2000) extinction law with no constraints imposed on A_V .

It is important to note that pPXF does not incorporate theoretical nebular emission predictions when determining stellar population properties. Instead, emission lines are fitted beforehand to constrain the kinematics and are removed from the spectra before the stellar population fit. Emission lines, however, are retained in the photometric fitting.

The output of pPXF consists of V-band luminosity weights associated with the grid of SSP templates used in the fit, each characterised by a specific age and metallicity. To enable a direct comparison with the CNN model, we computed the mean metallicity at each age bin, weighted by the corresponding mass fraction.

4. Results

In this section, we evaluate the predictive performance of the CNN by comparing its outputs with the ground-truth labels in the test sample and with the results obtained using the widely adopted spectral-fitting code pPXF. Figure 3 shows four representative test cases, illustrating the SFHs and age-metallicity relations recovered by both methods. The figure also displays the input and reconstructed data generated from the solutions, enabling a direct visual assessment of the quality of the fit.

The examples shown in Fig. 3 demonstrate that the CNN is able to recover both the mass distribution of stellar populations across age bins and their corresponding metallicities with only minor deviations from the ground truth. The pPXF results also exhibit good overall agreement; however, in some cases, they introduce spurious contributions from very young stellar populations, often associated with metallicities that differ significantly from those of the previous components. We note that these low-level components and metallicity fluctuations are likely sensitive to the adopted regularisation scheme, extinction law, and multiplicative polynomials used in the fitting process, and they may be reduced with appropriate tuning of these parameters.

The figure also compares the input spectra with those reconstructed from the inferred solutions. Both the CNN and pPXF produce model spectra that closely reproduce the input data, although the χ^2 of the residuals is lower for the pPXF solutions

⁴ The linear decay was set to 10^{-4} (Initial learning rate) / 250 (maximum allowed number of epochs).

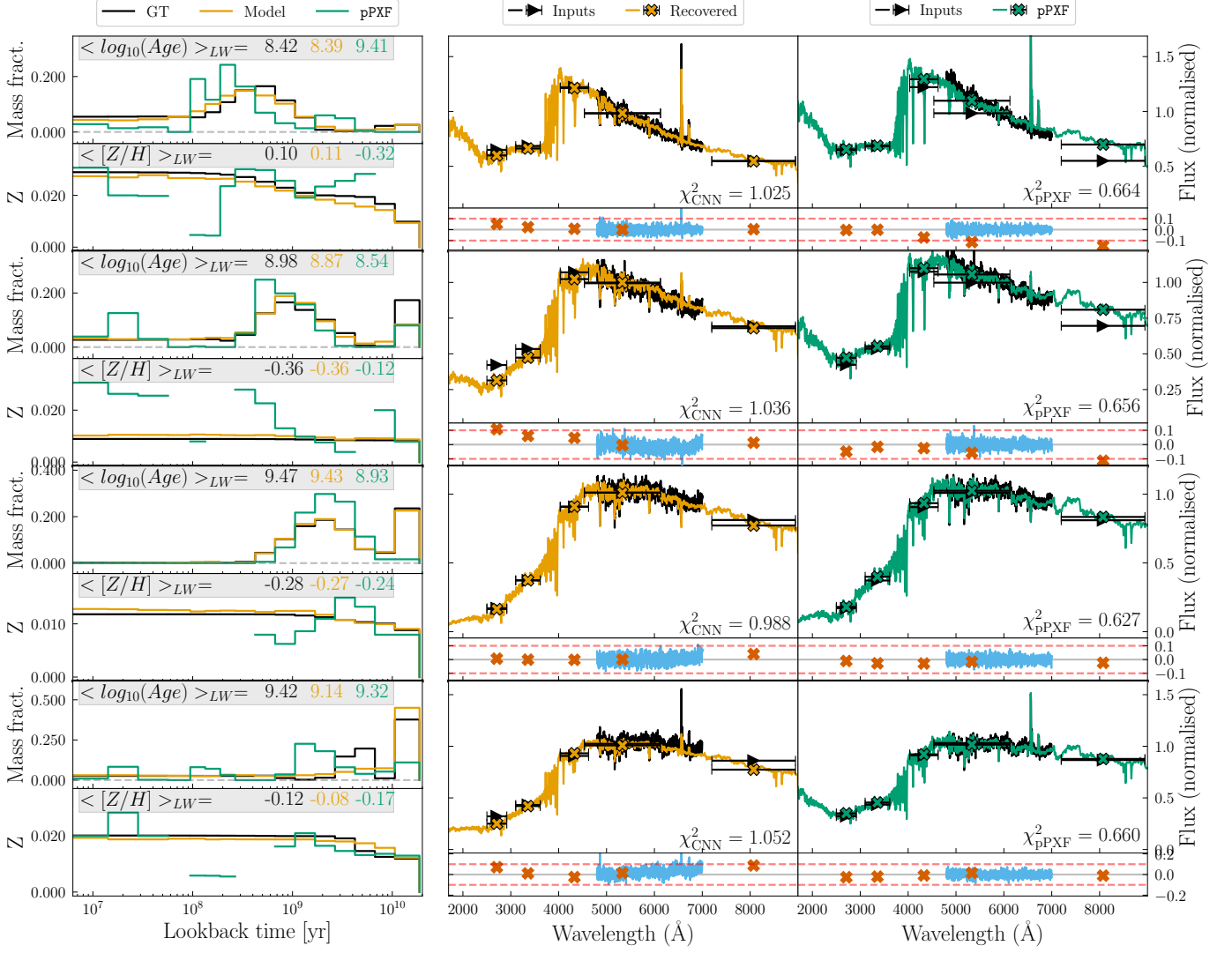


Fig. 3: Four representative examples of the recovered SFH and age-metallicity relation compared to the ground truth, using our CNN model and pPXF. Left column: For each example, we show the mass fraction of living stars (top panels) and the mean metallicity (bottom panels) across 16 age bins. The ground truth is shown in black, while the CNN and pPXF predictions are shown in orange and green, respectively. Gaps in the pPXF metallicity curves correspond to age bins with zero luminosity weight. Central column: Normalised input spectrum (black) and CNN reconstruction (orange). Photometric fluxes in the HST/WFC3 filters F275W, F336W, F438W, F555W, and F814W are shown as symbols (ordered from shorter to longer wavelengths), with horizontal error bars indicating the effective widths of the filters. Residuals are shown in the lower panels. Right column: Same as the central column, but for the pPXF reconstruction (green). The values reported in the left panels correspond to the luminosity-weighted $\log(\text{Age})$ and $[Z/H]$ for the ground truth (black), CNN predictions (orange), and pPXF results (green). The numbers in the central and right columns indicate the χ^2_{ν} values of the spectral fits.

than for the CNN, despite the fact that the CNN more accurately recovers the ground-truth stellar population properties. We believe this behaviour arises from several methodological differences. First, pPXF does not include the emission lines in the fit of the stellar continuum, and it employs multiplicative polynomials to minimise differences in the continuum shape, both of which strongly affect the variance across the spectrum. More fundamentally, full spectral-fitting methods are explicitly designed to minimise the difference between the observed and fitted spectra in flux space, whereas the CNN is directly optimised to recover the underlying physical parameters. As a consequence, the network is not necessarily driven by the spectral regions with the largest flux variations, but rather by those carrying the greatest information content for the inference of stellar population prop-

erties. In the context of stellar population analysis, where strong degeneracies allow different combinations of parameters to reproduce nearly identical continuum shapes, this represents a significant advantage over traditional fitting methods.

Figure 4 shows the mean difference between the input and CNN-predicted SFR, normalised by the mean SFR in each age bin. The relative errors are typically of the order of 2–4% across most age bins, increasing to $\sim 13\%$ in the 7–10 Gyr range. Overall, the CNN tends to overestimate the SFR at ages $\lesssim 4$ Gyr and underestimate it at older ages.

To enable a more quantitative comparison, we compute the logarithmic weighted mean ages and metallicities for both the

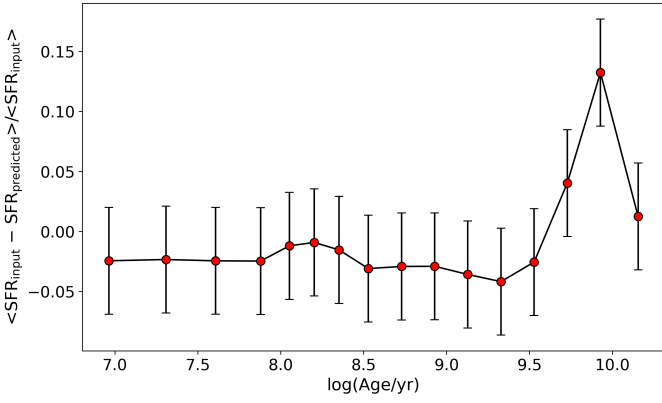


Fig. 4: Mean difference between the input and the predicted SFR as a function of age, normalised by the mean SFR in each bin. The error bars represent the standard deviation against the mean.

recovered and input SFHs as follows

$$\langle \log_{10}(\text{age/yr}) \rangle_{\text{XW}} = \frac{\sum_i \log_{10}(\text{age/yr})_i w_i}{\sum_i w_i}, \quad (6)$$

$$\langle [Z/H] \rangle_{\text{XW}} = \frac{\sum_i [Z/H]_i w_i}{\sum_i w_i}, \quad (7)$$

where w_i represents either the mass fraction (for mass-weighted quantities, $\text{XW} = \text{MW}$) or the V -band light fraction (for light-weighted quantities, $\text{XW} = \text{LW}$) contributed by the stellar population in age bin i .

Although the mass-weighted age is generally a more representative indicator of the overall stellar population properties, it is less sensitive to recent star formation when most of the stellar mass resides in old populations, as is typically the case in real galaxies. By contrast, light-weighted quantities are considerably more sensitive to small contributions from young stellar populations, owing to their much lower mass-to-light ratios.

Figure 5 compares the input mean ages and metallicities with those predicted by our CNN and by pPXF. Our CNN accurately recovers the mean values, with negligible biases in metallicity and in the mass-weighted ages, while the median bias in the light-weighted ages remains below 2%. The scatter in the age predictions is 30% and 37% for the mass- and luminosity-weighted values, respectively, and 10% and 6% for the corresponding mass- and luminosity-weighted metallicities.

In contrast, the ages recovered by pPXF exhibit significantly larger biases, particularly for the mass-weighted quantities, which are underestimated by approximately 65% on average. The metallicity biases obtained with pPXF are also larger, reaching ~10–14% for the mass- and luminosity-weighted values, respectively. The scatter in the predicted quantities is likewise substantially smaller for the CNN than for pPXF, especially in the case of the metallicity estimates.

4.1. Age-Z degeneracy

To investigate the impact of the age-metallicity degeneracy on the predictions of both methods, we show in Fig. 6 the correlation between the differences in the logarithm of the input and recovered mean ages for both mass- and luminosity-weighted estimates, as predicted by the CNN model and pPXF. There is an anti-correlation in the errors of both parameters, which are more readily present in the mass-weighted values.

However, the model shows a significantly shallower slope of the correlation, much less scatter overall, and the distribution is better centred around $\Delta = 0$ in both axes. pPXF shows a systematic shift towards underestimating ages regardless of how the age is weighed. The metallicities, on the other hand, are overestimated when weighting by mass and the opposite when weighting by light (Fig. 5). This indicates that pPXF systematically warps the shape of the metallicity evolution and overestimates recent SFR, on top of providing less accurate values in general. Another interesting difference between the model and pPXF is how the value of the metallicity correlates with the scatter; no significant gradient is observed for pPXF, but it is clearly present for the model, such that lower values of the metallicity are better determined, especially when the age is overestimated. Older stellar populations are expected to have deeper metal absorption features, so it makes sense that very low metallicities are better predicted when the model overestimates the age, since in this case, the absorption features are faint in the first place.

These results suggest that our CNN is able to capture the subtle differences in the spectra and flux values resulting from combinations of ages and metallicities that are difficult to capture using full spectral fitting methods such as pPXF. This advantage of CNNs has also been demonstrated in Woo et al. (2024), where they compare several fitting codes with the deep-learning CNN STARNET (Fabbro et al. 2018).

Although the synthetic spectra are attenuated using a transfer function to mimic dust extinction (see Sect. 2.2), the CNN is not trained to predict the corresponding optical depths. It is therefore important to evaluate whether the amount of extinction affects the recovered parameters. Figure 7 presents the differences between the ground-truth and predicted mean mass- and luminosity-weighted ages and metallicities as a function of the optical depth of the birth cloud (τ_{BC}) and the diffuse interstellar medium (τ_{ISM}). The biases in the predicted mean parameters do not exhibit any significant correlation with the level of extinction in either dust component included in the synthetic data. We also check that the network preserves the age-metallicity correlations present in the input data, indicating that the network does not wash out the underlying correlations, despite the introduction of a broad range of extinction and noise (see Fig. B.1 in Appendix B).

4.2. Effect of the S/N in the predictions

To assess the robustness of the predictions depending on the level of spectral noise, we evaluated the test sample after adding Poisson noise corresponding to six different S/N ($\text{S/N} = 5\text{--}100$) with the model. Figure 8 presents the median differences and associated scatter between the input and predicted values as a function of S/N.

As expected, the scatter in the recovered parameters increases towards lower S/N, approximately doubling between $\text{S/N} = 100$ and $\text{S/N} = 5$. In contrast, the impact on the mean bias is more modest: although the bias increases by a few per cent at low S/N, it remains small overall. The clear dependence of the scatter on S/N provides an important indication that the CNN is not overfitting the training data, but instead is extracting information from the underlying signal.

Real spectra may include additional sources of systematic noise arising, for example, from residuals of imperfect sky-line subtraction or bad pixels. To investigate the impact of such effects on our results, we also generated synthetic data using error spectra derived from the observations. These were constructed from the root mean squared of the residuals of a fit to the stel-

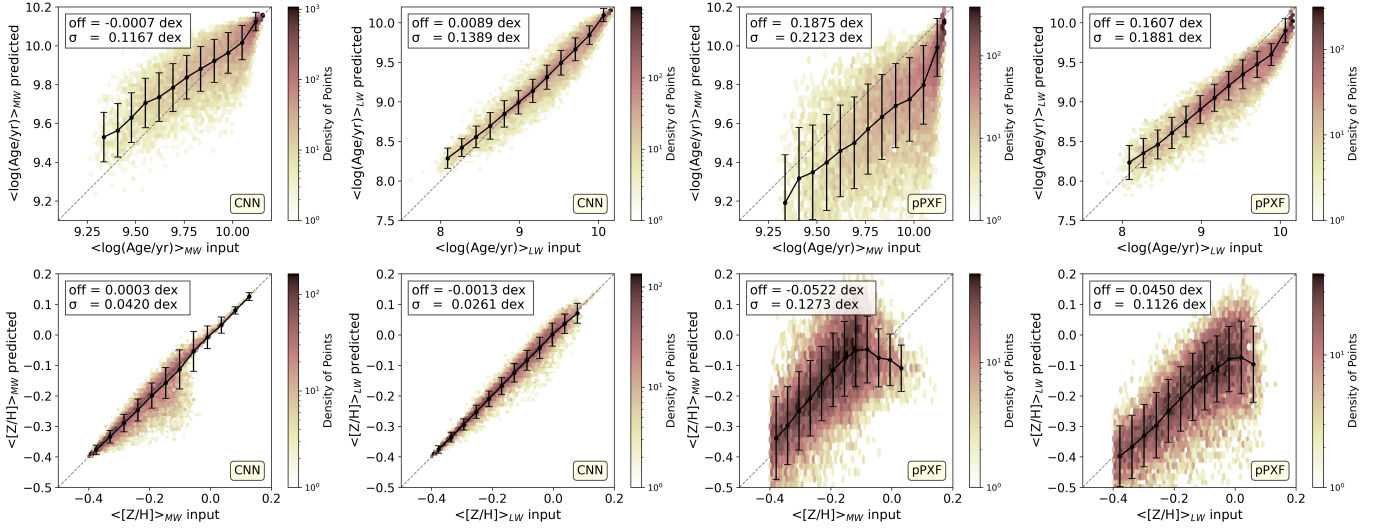


Fig. 5: Comparison between the ground truth mean mass- and luminosity-weighted ages (top row) and metallicities (bottom row) obtained with the CNN model (left panels) and with pPXF (right panels). The insets display the mean logarithmic differences between the input and predicted values ($\langle \log(\text{Age/yr})_{\text{input}} - \log(\text{Age/yr})_{\text{predicted}} \rangle$ and $\langle [Z/H]_{\text{input}} - [Z/H]_{\text{output}} \rangle$), together with their standard deviations. Each hexbin is coloured according to its point density, as indicated by the colour bar. Black symbols denote the median predicted values within bins of the corresponding input quantities.

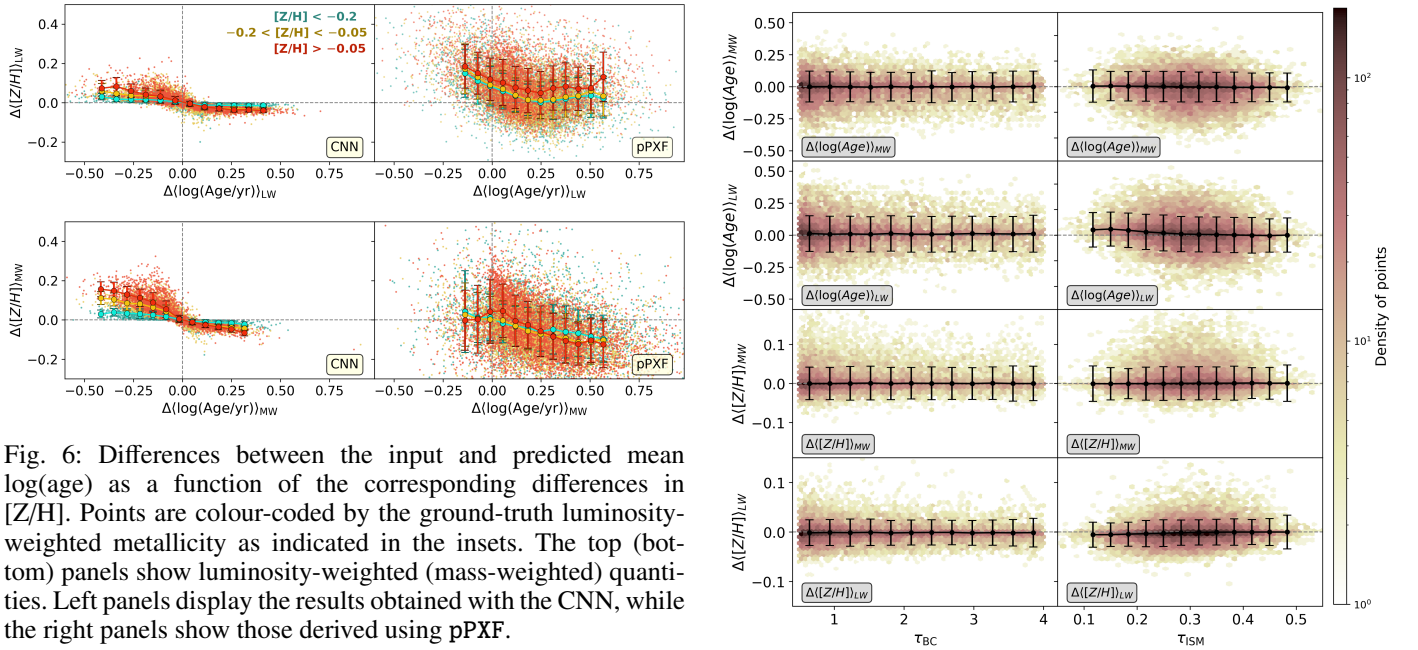


Fig. 6: Differences between the input and predicted mean log(age) as a function of the corresponding differences in $[Z/H]$. Points are colour-coded by the ground-truth luminosity-weighted metallicity as indicated in the insets. The top (bottom) panels show luminosity-weighted (mass-weighted) quantities. Left panels display the results obtained with the CNN, while the right panels show those derived using pPXF.

lar and gas kinematics obtained with pPXF. We run the model trained with Poisson noise on the test set, where we added noise using this ‘realistic’ error spectra, altering each pixel individually according to this error spectra. The results, which have been included in Appendix D, are consistent with those obtained on the test sample with Poisson noise. These tests demonstrate that our noise-injection prescription does not bias the model’s predictions, likely because the sky residuals do not correlate with any of the desired outputs, and the network ignores the inputs.

4.3. Ablation analysis of the input data

Identifying which input features carry the greatest predictive power provides insight into the spectral regions and photomet-

Fig. 7: Density plots of the differences between the ground-truth and predicted mean mass- and luminosity-weighted ages and metallicities as a function of the optical depth of the birth cloud (τ_{BC} ; left panels) and the diffuse interstellar medium (τ_{ISM} ; right panels). The black symbols indicate the median differences computed in bins of τ , while the error bars represent the corresponding standard deviations.

ric bands that most strongly constrain the SFH and the age-metallicity relation. To investigate this, we perform an ablation analysis in which the fluxes in selected spectral regions are replaced, pixel by pixel, with values randomly drawn from the empirical distribution of the same region across the full dataset. This procedure follows the permutation feature-importance ap-

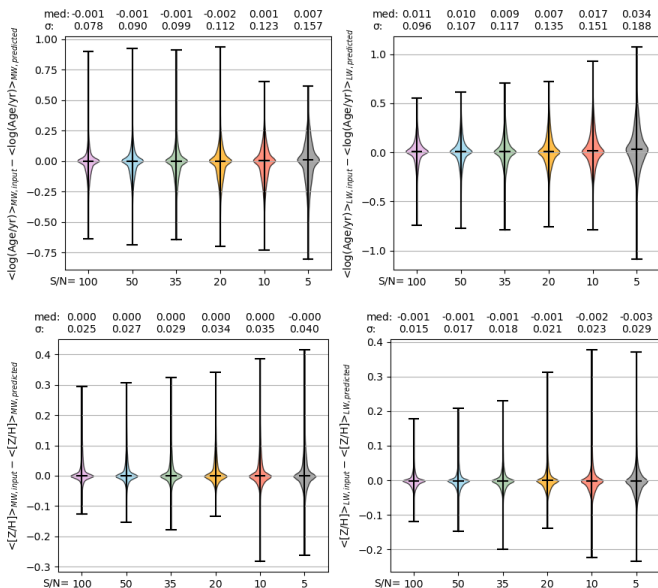


Fig. 8: Violin plots showing the distributions of the differences between the input and predicted mean values of the MW and LW age and metallicities for different S/N levels. The black horizontal line marks the median, and the error bars indicate the 16th and 84th percentiles of each distribution. The median values and standard deviations are annotated at the top of each panel.

proach introduced by Breiman (2001) for random-forest models and later formalised and generalised by Fisher et al. (2019), and has been successfully applied in related contexts (e.g. Kang et al. 2023). By construction, this method removes any association between the perturbed feature and the target variables while preserving the global statistical properties of the data.

We quantify the effect of removing individual HST fluxes on the bias and scatter of the predicted mean ages and metallicities, and we also consider the limiting case in which no photometric information is provided. We further assess the sensitivity of the CNN predictions to different spectral regions by replacing selected wavelength intervals in the input spectra. We test contiguous equal-length intervals of width $\Delta\lambda \approx 250$ Å covering the full wavelength range, as well as the regions containing H α (6500–6750 Å) and H β (4800–5050 Å). To evaluate the importance of the continuum shape, we also consider ‘flattened’ spectra, obtained by fitting and subtracting a second-order polynomial.

4.3.1. Ablation tests on the age predictions

Figure 9 quantifies the impact of removing individual photometric bands on the recovery of the mean light-weighted (LW) age as a function of the input mean age. The fiducial case, in which all photometric fluxes are included, is shown for comparison. Overall, removing individual HST bands degrades the predictions, introducing systematic offsets that strongly depend on the mean age of the system.

For populations with $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}} \leq 8.6$, the largest degradation is produced by removing F336W, followed by F814W and F275W. In the first case, the bias increases from 15% in the fiducial case to 100%, while removing F814W and F275W increases it to 80% and 50%, respectively. By contrast, removing F438W or F555W does not produce a significant increase in the bias for these young systems.

The impact on intermediate-age populations ($8.6 < \langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}} \leq 9.5$) is smaller, but still significant. In this regime, the bias in the mean LW age increases from 3% in the fiducial case to 50% and 35% when the F336W and F814W bands are removed, respectively. The age predictions of older systems ($\langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}} > 9.5$) depend less strongly on the photometric fluxes, with the notable exception of F275W. The omission of this band produces an average bias of 0.2 dex, together with an increase in the scatter of more than 100% relative to the fiducial case.

Removing all photometric fluxes also degrades the predictions substantially. In this case, the network systematically over-predicts the ages of young systems and under-predicts those of old ones, with biases reaching up to 75%. This demonstrates that the photometric information provides an essential anchor for the age scale across the full parameter space.

Figure 10 illustrates the impact of perturbing different spectral regions on the mean LW-age predictions. Overall, the inferred values show a strong dependence on the integrity of the input spectrum, with the replacement of individual wavelength intervals systematically increasing both the residual scatter and the bias. As in the case of the photometric ablation tests, the magnitude of the bias also depends strongly on the age of the system.

For the youngest systems, the strongest degradation occurs when removing the 6300–6500 Å and 6500–6750 Å intervals, where the median residuals can shift by up to ~ 0.5 dex. These wavelength ranges contain several strong emission lines, including H α , [N II], and [S II], which are associated with young ionising stellar populations. However, removing regions without strong emission lines, such as 5300–5550 Å and 5550–5800 Å, also degrades the age predictions for young systems by up to 0.4 dex. These intervals contain weak metal-blend absorption features, including the Lick indices Fe5335, Fe5406, Fe5709, and Fe5782, which become significantly stronger in populations aged 0.2–1 Gyr (e.g. Vazdekis et al. 2010). Furthermore, the continuum slope in the 5300–5550 Å segment, which corresponds to the normalisation range adopted for the input spectra, shows very little variation even among spectra with substantially different stellar populations. This indicates that the network does not rely exclusively on large-scale spectral variations or strong emission lines, but also exploits weaker absorption features in the determination of young stellar ages.

Intermediate-age and old populations show a similar trend: the strongest degradation again occurs when the 6300–6500 Å and 6500–6750 Å regions are removed, with the offset increasing by up to 90% relative to the fiducial case. In addition to H α absorption, these wavelength ranges include molecular bands and Fe lines whose strengths depend on both age and metallicity. Similarly, removing the 5050–5300 Å interval also introduces strong biases of up to $\sim 80\%$ in the predicted LW ages. This spectral range contains some of the strongest metallicity-sensitive absorption features, such as Mgb and Fe5270, and, together with the 6300–6500 Å segment, has one of the largest impacts on the metallicity predictions (see Sect. 4.3.2). The degradation caused by removing these regions likely reflects a poorer recovery of the metallicity, which in turn propagates into the age estimates for old systems, where the age-metallicity degeneracy is particularly strong.

Although the photometric fluxes (only photometry) have a strong influence on the predicted values, removing all the spectral information produces the largest degradation in the age predictions across all age ranges. The photometric fluxes are par-

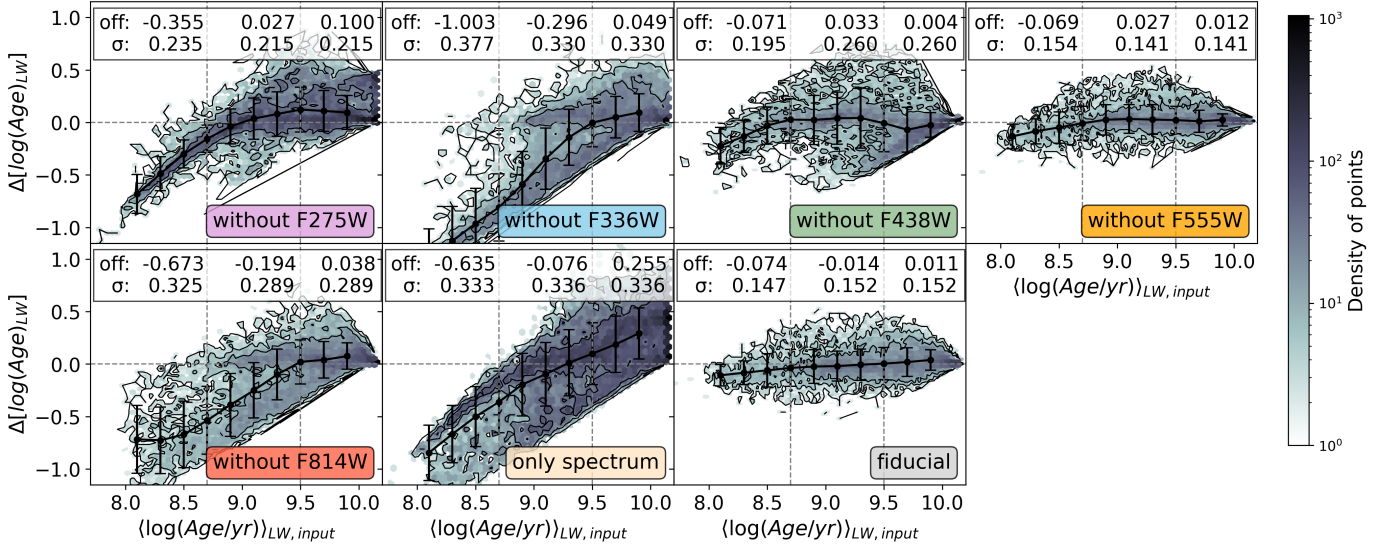


Fig. 9: Impact of removing photometric fluxes from the input data on the recovery of the LW age. Each panel shows the difference between the predicted and input LW age as a function of the input value, when individual photometric fluxes are replaced by values randomly drawn from the empirical distribution of the whole sample. Additionally, the full photometric input can be replaced with random values, leaving only the spectrum unchanged. For reference, the fiducial case in which all photometric fluxes are included is also shown. The running median and standard deviation are plotted as black circles with error bars. The insets report the mean biases and standard deviations in three age intervals, whose boundaries are marked by vertical lines: (1) $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}} \leq 8.6$; (2) $8.6 < \langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}} \leq 9.5$; and (3) $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}} > 9.5$.

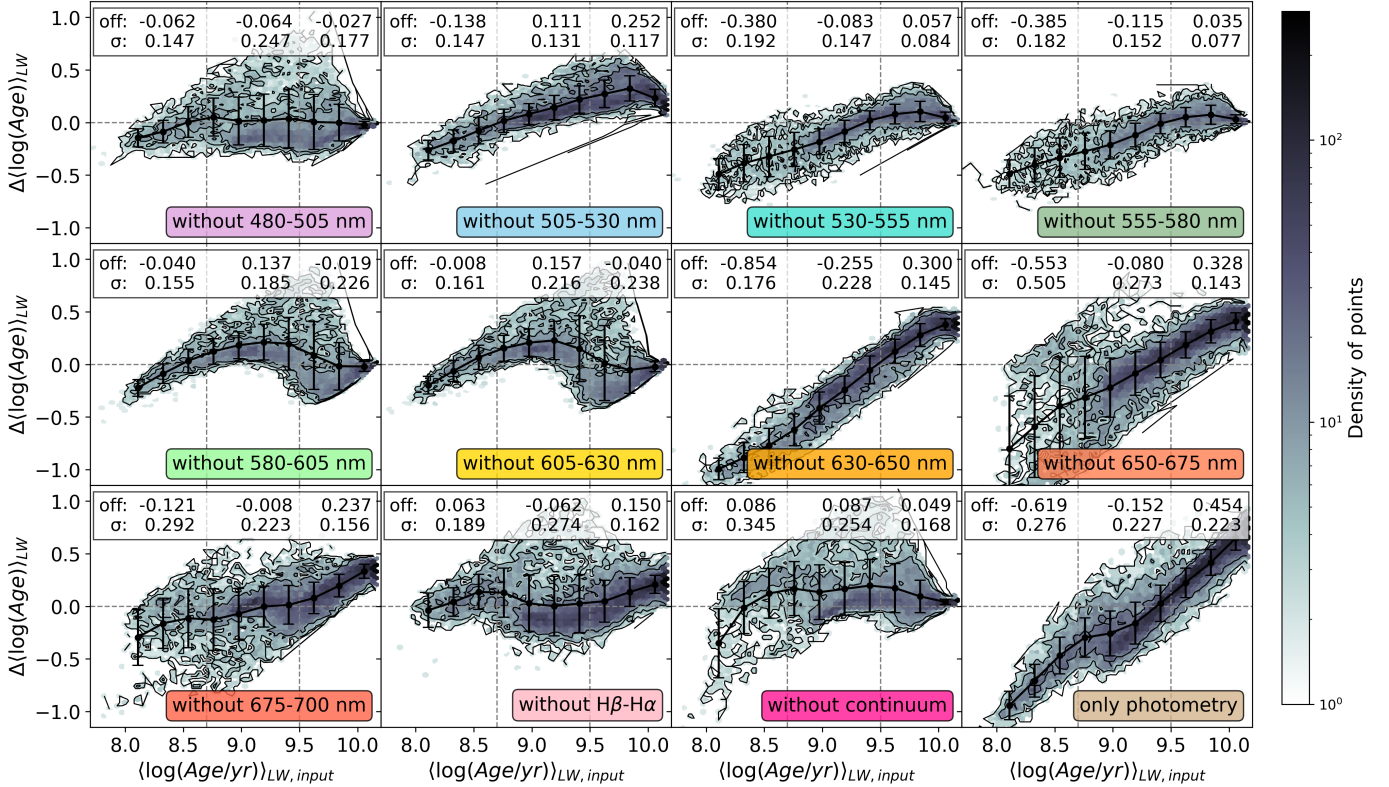


Fig. 10: Impact of different spectral regions on the LW-age predictions. Each panel shows the difference between the ground-truth and predicted $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}}$ for different perturbations of the input spectrum. The first nine panels (from left to right and top to bottom) show the results obtained when individual wavelength segments are replaced by random values drawn from the empirical distribution of the same segment across the full dataset. The panel labelled ‘without continuum’ shows the results when the continuum shape is removed from the spectra. The final panel shows the case in which the full spectrum is replaced by random values. The running median and standard deviation are plotted as black circles with error bars. The insets report the mean biases and standard deviations in three age intervals (see Fig. 9).

ticularly important for young systems since the MUSE spectral range does not extend sufficiently far into the blue, where sensitivity to young stellar populations is greatest. The spectrum, however, remains essential for accurate age determination across the full range of stellar populations. In particular, spectral regions containing strong absorption features, such as Mgb, TiO molecular bands, and prominent Fe lines, play a major role in determining the ages of old systems.

Flattening the input spectra (without continuum) results in a systematic over-prediction of the mean LW age, although the offset is not strongly correlated with the mean age itself. This suggests that the model is not only sensitive to age-dependent variations in the continuum shape of the spectra, but that detailed absorption features also remain essential for obtaining physically meaningful results.

The results of the ablation tests for the mass-weighted (MW) ages are shown in Figs. F.1 and F.2 of Appendix F. Overall, the conclusions are similar to those for the LW predictions, although the degradation is smaller. The most influential bands for the age predictions of young and intermediate-age populations are again F336W and F814W. However, the bias resulting from removing F275W is much lower for MW ages, as expected, since the NUV band primarily constrains very young populations, which usually contribute little to the total stellar mass. More generally, the MW ages depend less strongly on the global shape of the SED than the LW values, which suggests that the MW ages rely more on spectral features and less on the overall shape of the spectrum than the LW-ages.

4.3.2. Ablation tests on metallicity predictions

Figure 11 shows the impact of the photometric fluxes on the mean LW metallicity predictions. Unlike in the age case, the accuracy of the predictions does not show a clear dependence on the input photometry and appears largely insensitive to the removal of individual filters.

However, removing individual bands does increase the scatter, particularly for metal-rich systems. The filters with the largest impact are, in decreasing order, F336W, F814W, F438W, and F275W, whereas removing F555W does not produce any significant increase in either the bias or the scatter of the predictions. The effect is strongest for systems with intermediate or high metallicity ($\langle [Z/H] \rangle > -0.2$), where removing F336W and F814W increases the scatter in the mean LW metallicity by 15% and 12%, respectively, relative to the fiducial case. The scatter in low-metallicity systems is much less affected by photometry, with a maximum increase of only 4%. Since the synthetic data include metallicity evolution, metal-poor systems are typically older than metal-rich ones. We therefore interpret this behaviour as a consequence of the lower sensitivity of the photometric fluxes, relative to the spectral features, in constraining the properties of these older stellar populations (see Sect. 4.3.1).

Figure 12 shows the impact of the spectral ablation tests on the mean LW metallicity predictions. As in the photometric tests, the metallicity predictions do not show strong systematic biases when individual spectral regions are removed, regardless of the mean metallicity of the system. However, the scatter increases clearly with increasing metallicity. The largest degradation occurs when the 6300–6500 Å and 6500–6750 Å intervals are removed, producing a 12% increase in the scatter relative to the fiducial case for systems with $\langle [Z/H] \rangle_{LW} > -0.2$. This behaviour reflects the importance of strong emission lines, such as [N II], [S II], and H α , in constraining the metallicity of young,

metal-rich systems. The next most influential intervals are 6050–6300 Å and 5800–6050 Å, whose removal increases the scatter by about 9% in the same metallicity regime. The remaining spectral segments produce smaller increases, typically between 3% and 5%.

The equivalent figures for MW metallicities are presented in Figs. F.3 and F.4 of Appendix F. The results are very similar to those for the LW values, with only small numerical differences. The main discrepancy is the reduced scatter in the high-metallicity regime, likely due to the smaller number of synthetic spectra with high mean metallicity.

5. Application to PHANGS-MUSE data: NGC 3627

In this section, we test the performance of our trained CNN using real observational data of the nearby spiral galaxy NGC 3627. The aim is to assess the model’s performance on real data and to investigate the presence of possible domain biases.

We compare the results from our network with those obtained using the PHANGS-MUSE Data Analysis Pipeline (DAP⁵; EM22), which performs full spectral fitting with pPXF on the MUSE spectroscopic data without incorporating HST photometry. The spectra are Voronoi-binned to ensure a minimum signal-to-noise ratio of $S/N \sim 35$ per Å.

The public data release includes mass- and luminosity-weighted ages and metallicities derived using the E-MILES simple stellar population models (Vazdekis et al. 2016), generated with the BaSTI isochrones (Pietrinferni et al. 2004), with an age coverage from 0.15 to 14 Gyr. However, for a more consistent comparison with our results, we use values from an internal version of the PHANGS-MUSE analysis that adopts the same set of stellar population templates used in this work, based on E-MILES extended to very young ages (6.3–63 Myr). Nevertheless, the PHANGS-MUSE implementation does not include templates older than 14.12 Gyr or with metallicities below $[Z/H] < -0.7$, which introduces differences in the accessible parameter space relative to our model.

We use the same Voronoi-binned spectra as input for our CNN, preprocessed in the same way as the synthetic spectra used for training. In particular, all spectra are shifted to the rest frame, broadened to a final velocity dispersion of $\sigma = 150$ km s⁻¹, and normalised by their mean flux in the wavelength range 5300–5500 Å. For the first two steps, we use the radial velocities and velocity dispersions derived for each bin by the PHANGS-MUSE DAP.

To use the HST fluxes as input to our CNN, we first match the imaging data to the same spatial binning scheme as the spectra. This is achieved by first converting the images to surface brightness, convolving them to the MUSE point-spread function, and then averaging the fluxes within each Voronoi bin to obtain the corresponding photometric measurements. We only retain bins for which all spatially averaged fluxes are positive.

The preprocessed spectra and photometry are then fed into the trained CNN to infer the SFHs and age-metallicity relations in the predefined age bins. From these outputs, we compute the mean stellar ages and metallicities, weighted by both stellar mass and luminosity, following Eqs. 6 and 7.

Figure 13 shows the resulting spatially resolved maps of mass- and luminosity-weighted mean ages and metallicities across NGC 3627. The CNN-derived maps display clear spatial

⁵ The version of the pipeline used to analyse the PHANGS-MUSE galaxies is publicly available at <https://gitlab.com/francbelf/ifu-pipeline>.

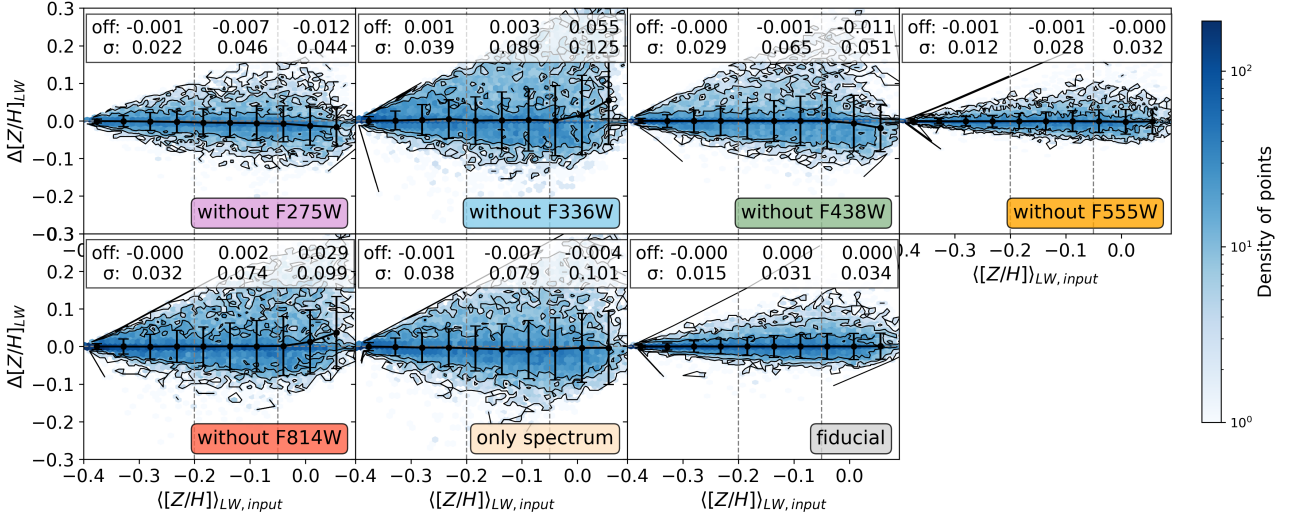


Fig. 11: Same as Fig. 9, but for the light-weighted metallicity. The insets report the mean biases and standard deviations in three metallicity intervals, whose boundaries are marked by vertical lines: (1) $\langle [Z/H] \rangle_{LW} \leq -0.2$; (2) $-0.2 < \langle [Z/H] \rangle_{LW} \leq 0.05$; and (3) $\langle [Z/H] \rangle_{LW} > 0.05$.

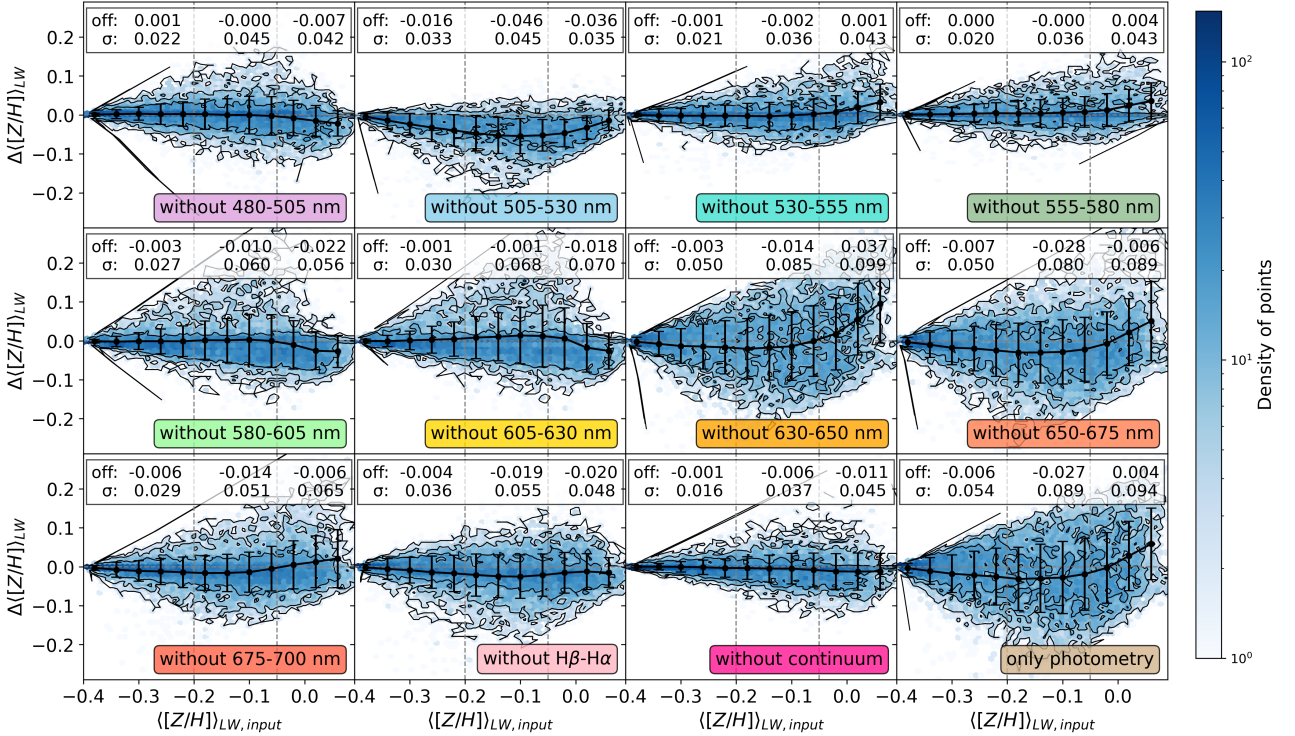


Fig. 12: Impact of different spectral regions on the predictions of the mean LW metallicity. The different panels are equivalent to those in Fig. 10, but for $\langle [Z/H] \rangle_{LW}$. Masking most wavelength intervals produces only minor changes in accuracy, while removing broad continuum information or relying on photometry alone leads to larger scatter, demonstrating the complementary role of spectral and photometric data in constraining metallicity. The insets report the mean biases and standard deviations in three metallicity intervals (see Fig. 11).

structure, with younger stellar populations tracing the spiral arms and star-forming regions, and older populations dominating the central regions. The maps are smooth and largely free of noise, which would not be expected if the CNN were not able to extract physically meaningful information from real data.

For comparison, we also show the maps obtained with pPXF, as well as maps of the integrated flux and key emission lines ($H\alpha$

and $[O III] \lambda 5007$). In addition, we include a map of the ionised gas metallicity derived using the O3N2 strong-line indicator, defined as $\log \left(\frac{[O III] \lambda 5007 / H\beta}{[N II] \lambda 6584 / H\alpha} \right)$, adopting the calibration of [Marino et al. \(2013\)](#). The emission-line fluxes are corrected for dust attenuation using the Balmer decrement, assuming case-B recombination and a [Cardelli et al. \(1989\)](#) extinction law.

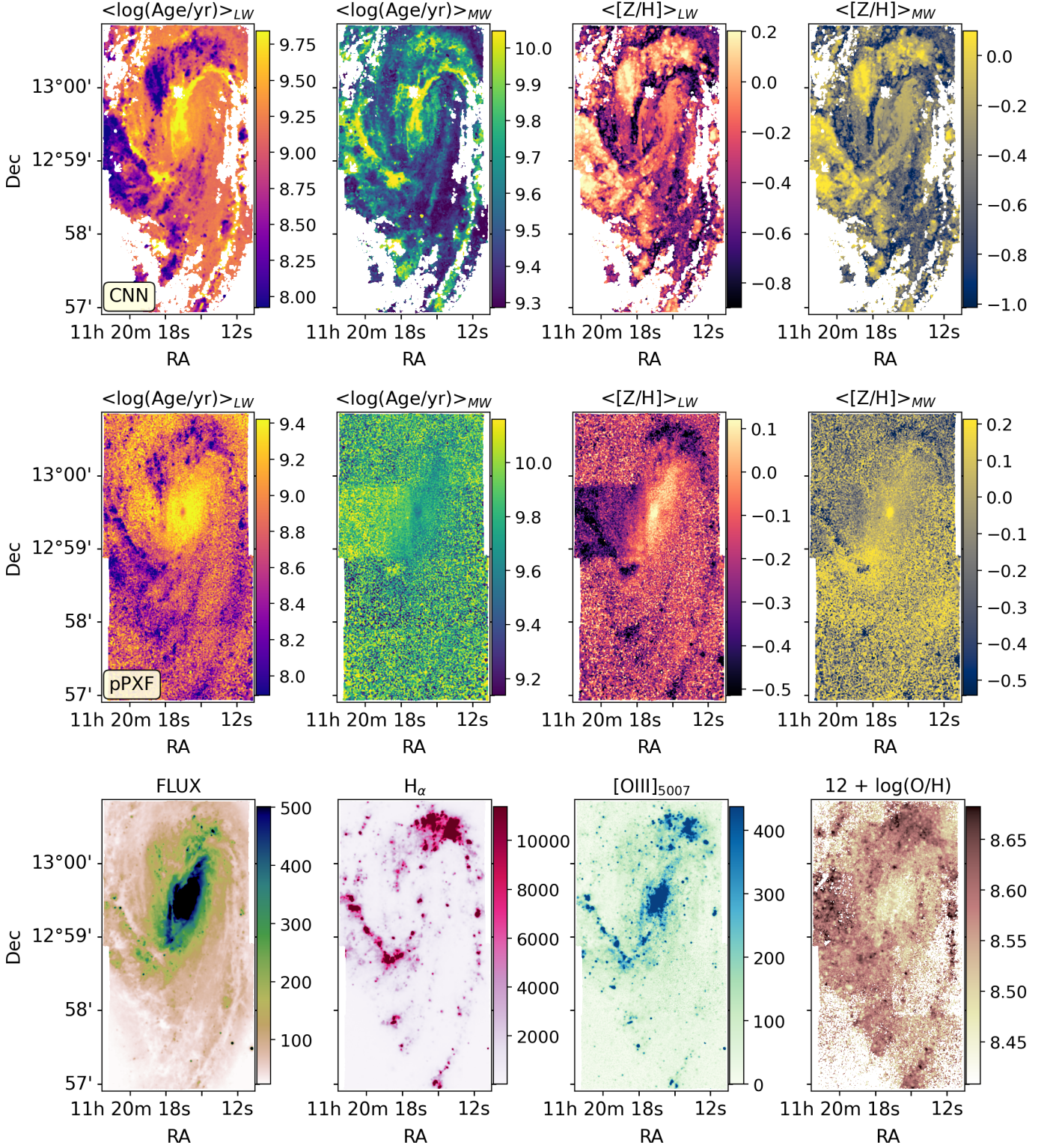


Fig. 13: Top row: Maps of the mass- and luminosity-weighted stellar ages and metallicities predicted by the CNN for NGC 3627. Middle row: Equivalent maps obtained with pPXF within the PHANGS-MUSE DAP. Bottom row: Maps of the integrated flux, $H\alpha$ and $[O III]_{5007}$ emission, and the metallicity of the ionised gas (see [EM22](#)).

A clear correspondence is observed between regions with low predicted LW ages and enhanced $H\alpha$ emission, confirming that the CNN successfully identifies sites of recent star formation. These regions also tend to exhibit higher metallicities,

consistent with expectations from chemical evolution in star-forming environments.

Figure 14 presents a quantitative comparison between the CNN and the [EM22](#) estimates for both mass- and luminosity-weighted quantities. The median differences for the luminosity-

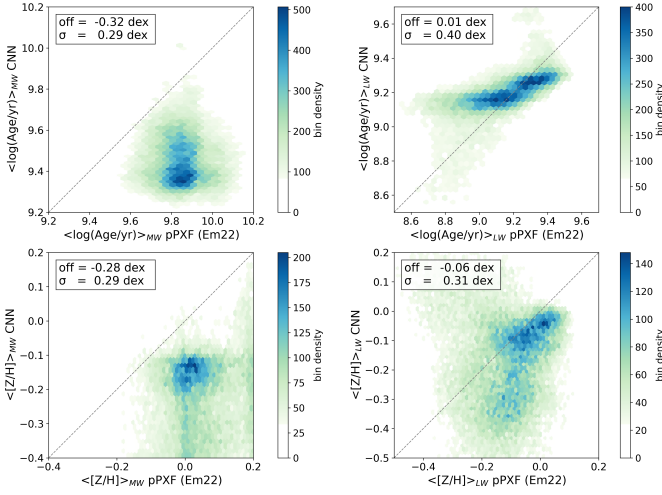


Fig. 14: Comparison of mass- and luminosity-weighted mean stellar ages and metallicities predicted in this work and by EM22 for the Voronoi-binned spectra of NGC 3627. Colours indicate the density of points. The mean offset (CNN–EM22) and the standard deviation of the residuals are reported in the upper-left corner. The one-to-one relation is shown as a dotted line for reference.

weighted quantities are relatively small, amounting to 2% and 12% for the mean LW age and metallicity, respectively, although with a substantial dispersion (0.4 and 0.31 dex). The mass-weighted quantities exhibit systematic differences of order $\sim 50\%$, with the CNN predicting lower values than EM22, corresponding to median offsets of -0.32 dex in age and -0.29 dex in metallicity.

These results indicate that both methodologies yield broadly consistent estimates for the more recent SFHs, while showing significant discrepancies in the properties of the older stellar populations. This behaviour is expected, as old populations can contribute only weakly to the observed spectra, particularly in star-forming galaxies, where even a small mass fraction of young stars can dominate the emitted light and outshine the flux of older components, making it more difficult to constrain their properties.

Several factors, apart from the analysis method (i.e. pPXF versus the CNN), may be contributing to the observed differences. First, the PHANGS-MUSE DAP analysis relies exclusively on MUSE spectra, whose wavelength coverage does not extend into the bluest regions. These regions are especially sensitive to intermediate and young stellar populations. This limitation makes it more difficult to break the age-metallicity degeneracy in older systems. Our approach incorporates additional photometric constraints that provide leverage in this regime. Second, the template library used in the PHANGS-MUSE DAP does not include models older than 14.2 Gyr or with metallicities below $[Z/H] < -0.7$. Additional discrepancies may also arise from differences in the treatment of dust attenuation and nebular emission.

A thorough investigation of these effects is beyond the scope of this work, and a more detailed analysis is deferred to a follow-up study (Galceran et al., in prep.).

6. Discussion

Recovering the SFH and the stellar metallicity distribution of galaxies remains a challenging inverse problem (see, e.g. Ocvirk

et al. 2006a,b; Cid Fernandes et al. 2013; Leja et al. 2019), where the conversion from observables to the physical solution is highly degenerate. Full spectral fitting techniques (Cid Fernandes et al. 2005; Ocvirk et al. 2006a,b; Cappellari 2012, 2023; Tojeiro et al. 2007; Wilkinson et al. 2017) have introduced regularisation techniques to minimise these problems, showing they can predict composite populations with different ages and metallicities. In parallel, stellar population inference has increasingly adopted Bayesian approaches (Martín-Navarro et al. 2018; Johnson et al. 2021), in which the parameter space is explored using techniques such as MCMC. These methods provide a principled way to characterise uncertainties and degeneracies, although they often rely on assumptions about the form of the posterior distribution and can be computationally demanding in high-dimensional parameter spaces.

A limitation of the full spectral fitting techniques is that regions of the spectra highly sensitive to stellar properties have equal weight in the fit compared to spectral regions containing far less information. In this context, ML approaches have recently emerged as a promising alternative, capable of learning informative spectral features and patterns that may be difficult for conventional full-spectral-fitting techniques to exploit efficiently (see e.g. Lovell et al. 2019; Liew-Cain et al. 2021; Iglesias-Navarro et al. 2024; Wang et al. 2024; Woo et al. 2024; Anwar et al. 2025; Domínguez Sánchez et al. 2026).

As noted by Walter et al. (2026), a key difference between ML-based and conventional methods lies in the definition of the optimisation target. Whereas traditional spectral-fitting codes minimise the residuals in flux space, CNNs allow minimising the error directly on the physical parameters. The minimisation of parameter-space errors allows the network to identify not only individual features but also higher-order correlations across the spectrum that may carry information relevant to the SFH. An important consequence of this difference in optimisation philosophy is its impact on the classical age-metallicity degeneracy. Several recent studies have shown that CNN-based approaches can substantially mitigate this degeneracy (e.g. Woo et al. 2024; Domínguez Sánchez et al. 2026). By leveraging subtle spectral signatures (such as small variations in the shapes of Balmer lines or weak metal-line blends that do not dominate the global χ^2), ML models can disentangle the effects of age and metallicity more effectively than traditional methods. Moreover, CNNs can, in principle, predict these parameters independently, with age and metallicity uncertainties showing weak correlations (Domínguez Sánchez et al. 2026), unlike classical fitting approaches where both parameters are strongly coupled through the continuum shape and line-depth trade-off.

The diversity of current methodologies includes transformer architectures optimised for capturing long-range spectral correlations (Anwar et al. 2025), simulation-based inference approaches that combine neural networks with explicit forward modelling to recover full posterior distributions (Iglesias-Navarro et al. 2024), CNNs tailored to extract local spectral features and infer SFHs or integrated stellar population properties (e.g. Lovell et al. 2019; Domínguez Sánchez et al. 2026), and hybrid frameworks that incorporate representation learning, such as autoencoders or latent-space embeddings, to compress spectral information before inference (e.g. Liew-Cain et al. 2021; Woo et al. 2024). The differences also extend to the training dataset, that includes real data with labels taken from the results of full spectral fitting analysis (e.g. Liew-Cain et al. 2021), synthetic data with SFH directly taken from cosmological simulations (e.g. Lovell et al. 2019; Wang et al. 2024), with single stellar populations or a combination of them (Anwar et al. 2025;

Domínguez Sánchez et al. 2026, e.g.), or with a similar technique such as the one used in this work (Iglesias-Navarro et al. 2024). Some networks are designed to predict single values of age and metallicities (Woo et al. 2024; Domínguez Sánchez et al. 2026), while others predict vectors quantifying the contribution and the mean metallicity of stars at different ages (Lovell et al. 2019; Anwar et al. 2025; Iglesias-Navarro et al. 2024).

All these works illustrate the breadth of tools now available. In this work, we present a new CNN capable of obtaining an accurate SFH and age-metallicity relation from combined spectroscopic and photometric inputs. It is of special interest to compare with those works that, similar to our model, are able to make predictions of the SFH and metallicities instead of predicting mean values, to investigate the pros and cons of the different configurations and training methods.

A first natural comparison is with the pioneering work of Lovell et al. (2019), who trained CNNs on synthetic spectra derived from the EAGLE (Schaye et al. 2015) and Illustris (Genel et al. 2014) cosmological simulations. Their model learns the mapping between rest-frame optical spectra and piece-wise constant SFHs defined in eight logarithmically spaced lookback time bins using a single metallicity. They quantify the differences between the input and predicted SFH using SMAPE, achieving median SMAPE values of the order of 10–12% for dust-attenuated spectra at S/N=50. Unlike our approach, their CNN operates on spectra alone and assumes a single (effectively constant) metallicity for each galaxy. Architecturally, the Lovell et al. network is a relatively shallow 1D CNN with two convolutional layers, max pooling, and fully connected layers performing regression on the SFH bins. This setup is well-suited to capturing the dominant features imprinted by recent star formation. Our model, by comparison, employs a deeper architecture with an attention mechanism, a shared latent space, and a decoder jointly predicting mass fractions and mean metallicities across 16 age bins. It also ingests both spectra and NUV-optical photometry, which improves constraints on young populations.

A further relevant comparison is provided by the recent work of Woo et al. (2024), who also employs a deep-learning CNN (STARNET; Fabbro et al. 2018) to infer stellar population properties (mass-weighted ages, metallicities, and $M_\star L_r$ ratios) and colour excess from galaxy spectra. Their network is trained on synthetic spectra built from SFHs from the IllustrisTNG100-1 simulation and adapted to mimic the Sloan digital sky survey (SDSS; York et al. 2000). The authors include realistic noise that incorporates sky residuals and emission lines taken from real, randomly chosen SDSS spectra that minimise the domain shift.

Their model can predict the mass-weighted ages and metallicities with very small biases of 0.0032 and 0.017 dex, respectively, with scatters of 0.077 and 0.081 dex, respectively. In comparison, our predictions show a lower bias in the recovery of the mass-weighted age and metallicity of -0.0007 and 0.0003 dex, respectively, with dispersions of 0.117 and 0.042 dex. Overall, the values are comparable. However, there are important methodological differences between the two approaches. The model presented by Woo et al. is primarily designed to predict integrated summary quantities (e.g. mean ages and metallicities), rather than reconstructing the full star formation and chemical enrichment histories. As a result, their network architecture and training objective are optimised for capturing global spectral trends, rather than resolving the temporal structure of the SFH. In contrast, our model explicitly targets the recovery of the full SFH and age-metallicity relation across multiple age

bins, which requires a more expressive architecture and a training strategy tailored to high-dimensional outputs.

Our method incorporates emission lines self-consistently, linking their presence and strength to the population of young stars capable of ionising the gas. Woo et al. includes emission lines drawn from real spectra, which may result in more realistic line ratios and profiles. However, these emission features are added randomly, without any physical correlation with the underlying stellar population or ionising radiation field.

This work also presents a detailed comparison with several full spectral fitting methods. Consistent with our findings, the authors report that ML models recover stellar population parameters with reduced bias and weaker age-metallicity degeneracies. They further show that, while emission lines significantly affect the results of traditional full spectral fitting codes, they have minimal impact on the CNN predictions. We argue that this behaviour likely stems from the random inclusion of emission lines, which are not physically linked to the stellar populations that ionise the gas. As a result, their model cannot extract meaningful information from these features, as they do not carry physically relevant information.

Another comparison point is provided by the transformer-based architecture of GalProTe from Anwar et al. (2025). That model uses four parallel attention-based encoders with multi-scale kernels to process optical spectra and predicts the weights of a dense grid of stellar populations sampling 12 metallicities and 53 ages. The inclusion of self-attention and multi-scale convolution allows GalProTe to learn cross-dependencies between spectral regions and to capture correlations between age- and metallicity-sensitive features that are widely separated in wavelength. The model is trained with synthetic spectra tailored to characteristics of the PHANGS-MUSE survey, with SFH that are linear combinations of single stellar populations (from one to five), randomly selected from the 53×12 age-metallicity grid of templates with different levels of extinctions and noise. The authors include Gaussian noise, but they also test the effect of adding very high noise in regions of the spectra affected by strong sky residuals. The self-attention mechanisms help the network learn cross-dependencies across spectral regions and therefore help to break degeneracies between parameters, achieving small biases in mean stellar ages (~ 0.02 dex) with moderate scatter (~ 0.25 dex). However, the metallicity estimates are notably less accurate, likely reflecting the relatively simple SFH parametrisations used in its synthetic training set, with metallicities that can change rapidly between adjacent age bins.

A conceptually different approach is presented by Iglesias-Navarro et al. (2024), who adopts a simulation-based inference (SBI) framework. Their method compresses spectra using a convolutional autoencoder and then applies a normalising-flow density estimator to recover full posterior distributions for a set of cumulative stellar-mass percentiles and the metallicity. This approach provides formally Bayesian uncertainty quantification and captures the multi-modal posteriors that naturally arise in SFH inference. However, because the autoencoder is trained to optimise spectral reconstruction rather than direct recovery of physical parameters, its latent space may encode information that is not maximally informative for SFH inference. This differs from our CNN, where we trained end-to-end on the desired physical outputs (SFH and age-metallicity relation), ensuring that the latent representation is explicitly tuned to the astrophysical parameter space rather than to spectral fidelity.

The two approaches thus offer complementary strengths. SBI provides rigorous posterior estimation under controlled domain assumptions but is more sensitive to domain shifts between sim-

ulated training data and real spectra. Our CNN lacks a full probabilistic posterior and therefore cannot characterise multi-modal solutions in the same way; however, it achieves substantially higher accuracy in detailed SFH reconstruction, integrates photometric and spectroscopic information naturally, and mitigates age-metallicity degeneracy more effectively, which makes it particularly interesting for application to large spectroscopic surveys. Furthermore, our network parametrises the SFH with the SFR(t) vector rather than mass percentiles. This enables our network to better discriminate the variations of the SFH at young ages. For many studies that target mainly intermediate or older populations, a percentile-based parametrisation will suffice.

Despite these successes, our CNN also has limitations. The requirement to shift spectra to rest-frame ($z=0$) and broaden them to a fixed velocity dispersion ($\sigma = 150 \text{ km s}^{-1}$) introduces a preprocessing step that partly offsets the rapid inference speed of the CNN. More fundamentally, supervised ML models depend on the availability of labelled training data. Constructing such training sets requires explicit assumptions about extinction curves, the escape fraction of ionising photons, the IMF, and other astrophysical ingredients. Any mismatch between these assumptions and the true underlying physics may introduce biases in the predictions. This issue is not unique to ML (errors in stellar-population models propagate through all inference frameworks), but the impact is particularly direct for supervised systems. Previous studies (e.g. Kang et al. 2023; Ksoll et al. 2024) have emphasised the importance of addressing such model-dependence and have suggested that transfer-learning techniques (Farahani et al. 2021) may offer a way forward by adapting networks trained on synthetic data to better match real observations.

7. Summary

In this work, we have presented a CNN capable of recovering detailed SFHs and age-metallicity relations directly from combined spectroscopic and photometric inputs. The model was trained on synthetic observations generated from a wide range of SFH and chemical-enrichment histories. Dust attenuation was modelled following the two-component prescription of Charlot & Fall (2000), with independently sampled optical depths for birth clouds and the diffuse interstellar medium. Emission lines associated with young stellar populations were added using the MAPPINGS-III photoionisation models, and Poisson noise was incorporated at varying S/N.

When tested against synthetic data, the CNN attains higher accuracy and precision than pPXF, one of the most widely used full spectral-fitting codes. It produces lower biases and reduced scatter in the predicted mass- and luminosity-weighted ages and metallicities, while also delivering predictions at sub-millisecond speeds. Similar gains have been reported by recent machine-learning studies focused on mean stellar-population parameters (e.g. Liew-Cain et al. 2021; Wang et al. 2024; Woo et al. 2024; Domínguez Sánchez et al. 2026) and full SFHs (e.g. Lovell et al. 2019; Anwar et al. 2025; Iglesias-Navarro et al. 2024). Our model extends these efforts by predicting the full metallicity evolution and exploiting a combination of spectroscopic and photometric constraints.

The good performance of the model can be attributed to the fact that the neural network is optimised to recover physically meaningful SFHs rather than to minimise spectral residuals alone. Traditional fitting methods based on explicit residual minimisation (such as χ^2 optimisation) can sometimes achieve deceptively small variances by adjusting broad continuum shapes while failing to reproduce age- and metallicity-sensitive spectral

features. The CNN is trained directly on the underlying SFHs and enrichment histories, favouring solutions that more accurately reflect the physical parameters even when they do not correspond to the smallest possible spectral residuals. Our analysis further shows that including photometric constraints, particularly in the NUV–optical range, plays a crucial role in breaking degeneracies and enhancing age sensitivity.

Our methodology is specifically designed to recover SFHs and age-metallicity relations from PHANGS-MUSE and PHANGS-HST data, and the network architecture reflects the characteristics of these observations. The resulting model is therefore not intended as a universal SFH inference tool that can be applied unchanged to arbitrary datasets. However, the approach itself is readily transferable to other spectral ranges or to different combinations of broad- and narrow-band photometry, provided that the architecture is adjusted and the network retrained accordingly. In this sense, the main contribution of this work is not only the trained model presented here, but also a reproducible framework for generating synthetic training data, defining an appropriate architecture, and validating the inference. Although developed in the context of PHANGS, this strategy is broadly applicable to other surveys or observations, and illustrates how convolutional architectures can be tailored to mixed spectroscopic and photometric datasets while preserving interpretability and training stability.

The application of the CNN to real PHANGS-MUSE data demonstrates its ability to recover physically meaningful stellar population properties from observed spectra. The spatial distributions of age and metallicity are consistent with the known structure and star formation activity of NGC 3627, and the expected abundance gradients. A follow-up study (Galceran et al., in prep.) will extend this application to 19 galaxies from the PHANGS-MUSE and PHANGS-HST surveys, enabling an exhaustive comparison of the CNN performance in real galaxies.

Importantly, the CNN offers significant advantages in terms of computational efficiency and robustness to noise, enabling the analysis of large spectroscopic datasets at a fraction of the computational cost of traditional fitting methods. Moreover, the smoother spatial behaviour of the inferred quantities suggests that the CNN is less affected by local degeneracies and noise-induced fluctuations.

These results contribute to a rapidly growing body of evidence demonstrating the potential of deep learning approaches as a complementary tool to classical full spectral fitting techniques for the analysis of spatially resolved stellar populations in nearby galaxies.

Acknowledgements. We thank the anonymous referee for a careful and thorough revision of the manuscript and for constructive comments that have significantly improved the final presentation of this paper. We acknowledge financial support by the Spanish Ministry of Science and Innovation through the research grants PID2019-107427-GB-31, funded by MICIU/AEI/10.13039/501100011033 and the European Union/FEDER, PID2022-138855NB-C31, funded by MICIU/AEI/10.13039/501100011033 and PRE2020-094548, funded by MICIU/AEI/10.13039/501100011033 and ESF Investing in your future. MB acknowledges support from the ANID BASAL project FB210003. This work was supported by the French government through the France 2030 investment plan managed by the National Research Agency (ANR), as part of the Initiative of Excellence of Université Côte d’Azur under reference number ANR-15-IDEX-01. FP acknowledges support from the Horizon Europe research and innovation programme under the Maria Skłodowska-Curie grant “TraNS-Late” No 101108180, and from the Agencia Estatal de Investigación del Ministerio de Ciencia e Innovación (MCIN/AEI/10.13039/501100011033) under grant (PID2021-128131NB-I00) and the European Regional Development Fund (ERDF) ‘A way of making Europe’. RSK acknowledges financial support from the ERC via Synergy Grant ‘ECOGAL’ (project ID 855130), from the German Excellence Strategy via the Heidelberg Cluster ‘STRUCTURES’ (EXC 2181 - 390900948), from the German Ministry for Economic Affairs and Climate Ac-

tion in project ‘MAINN’ (funding ID 50002206), and from DFG and ANR for project ‘STARCLUSTERS’ (funding ID KL 1358/22-1). This work was carried out as part of the PHANGS collaboration. Based on observations taken as part of the PHANGS-MUSE large program (Emsellem et al. 2021). Based on data products created from observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programme(s) 1100.B-0651, 095.C-0473, and 094.C-0623 (PHANGS-MUSE; PI Schinnerer), as well as 094.B-0321 (MAGNUM; PI Marconi), 099.B-0242, 0100.B-0116, 098.B-0551 (MAD; PI Carollo) and 097.B-0640 (TIMER; PI Gadotti). This research has made use of the services of the ESO Science Archive Facility. This research has made use of the services of the ESO Science Archive Facility. Science data products from the ESO archive may be distributed by third parties and disseminated via other services, according to the terms of the Creative Commons Attribution 4.0 International license. Credit to the ESO origin of the data must be acknowledged, and the file headers preserved.

References

- Allen, M. G., Groves, B. A., Dopita, M. A., Sutherland, R. S., & Kewley, L. J. 2008, *ApJS*, 178, 20–55
- Angthopo, J., Granett, B. R., La Barbera, F., et al. 2024, *A&A*, 690, A198
- Anwar, O., Groves, B., Cortese, L., & Watts, A. B. 2025, *PASA*, 42, e102
- Asa’d, R. S., Vazdekis, A., Cerviño, M., et al. 2017, *MNRAS*, 471, 3599
- Asari, N. V., Cid Fernandes, R., Stasińska, G., et al. 2007, *MNRAS*, 381, 263
- Bertelli, G., Bressan, A., Chiosi, C., Fagotto, F., & Nasi, E. 1994, *A&AS*, 106, 275
- Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, *A&A*, 622, A103
- Breiman, L. 2001, *Machine Learning*, 45, 5
- Cabayol, L., Sevilla-Noarbe, I., Fernández, E., et al. 2019, *MNRAS*, 483, 529
- Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, *ApJ*, 533, 682
- Cappellari, M. 2012, pPXF: Penalized Pixel-Fitting stellar kinematics extraction, *Astrophysics Source Code Library*, record ascl:1210.002
- Cappellari, M. 2023, *MNRAS*, 526, 3273
- Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, *ApJ*, 345, 245
- Carnall, A. C., McLure, R. J., Dunlop, J. S., & Davé, R. 2018, *MNRAS*, 480, 4379
- Cenarro, A. J., Cardiel, N., Gorgas, J., et al. 2001, *MNRAS*, 326, 959
- Chabrier, G. 2001, *ApJ*, 554, 1274
- Chandar, R., Barnes, A. T., Thilker, D. A., et al. 2025, *AJ*, 169, 150
- Charlot, S. & Fall, S. M. 2000, *ApJ*, 539, 718
- Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2020, *MNRAS*, 493, 4209
- Cid Fernandes, R., Gu, Q., Melnick, J., et al. 2004, *MNRAS*, 355, 273
- Cid Fernandes, R., Mateus, A., Sodré, L., Stasińska, G., & Gomes, J. M. 2005, *MNRAS*, 358, 363
- Cid Fernandes, R., Pérez, E., García Benito, R., et al. 2013, *A&A*, 557, A86
- da Cunha, E., Charlot, S., & Elbaz, D. 2008, *MNRAS*, 388, 1595
- Dale, D. A., Helou, G., Magdis, G. E., et al. 2014, *ApJ*, 784, 83
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, *MNRAS*, 450, 1441
- Domínguez Sánchez, H., Coelho, P., Bruzual, G., et al. 2026, *A&A*, 705, A219
- Domínguez Sánchez, H., Margalef, B., Bernardi, M., & Huertas-Company, M. 2022, *MNRAS*, 509, 4024
- Domínguez Sánchez, H., Martin, G., Damjanov, I., et al. 2023, *MNRAS*, 521, 3861
- Dumoulin, V. & Visin, F. 2016, *A Guide to Convolution Arithmetic for Deep Learning*, arXiv preprint arXiv:1603.07285
- Emsellem, E., Schinnerer, E., Santoro, F., et al. 2022, *A&A*, 659, A191
- Fabbro, S., Venn, K. A., O’Brien, T., et al. 2018, *MNRAS*, 475, 2978
- Farahani, A., Pourshojae, B., Rasheed, K., & Arabnia, H. R. 2021, arXiv e-prints, arXiv:2104.02144
- Fisher, A., Rudin, C., & Dominici, F. 2019, *Journal of Machine Learning Research*, 20, 1
- Ge, J., Yan, R., Cappellari, M., et al. 2018, *MNRAS*, 478, 2633
- Genel, S., Vogelsberger, M., Springel, V., et al. 2014, *MNRAS*, 445, 175
- Girardi, L., Bressan, A., Bertelli, G., & Chiosi, C. 2000, *A&AS*, 141, 371
- Glort, X. & Bengio, Y. 2010, in *Proceedings of Machine Learning Research*, Vol. 9, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, ed. Y. W. Teh & M. Titterton (Chia Laguna Resort, Sardinia, Italy: PMLR), 249–256
- Graham, G., Dale, D., & Chase, S. 2025, *ApJ*, submitted
- Gregg, M. D., Silva, D., Rayner, J., et al. 2006, in *The 2005 HST Calibration Workshop: Hubble After the Transition to Two-Gyro Mode*, ed. A. M. Koekemoer, P. Goudfroi, & L. L. Dressel, 209
- Hannon, S., Whitmore, B. C., Lee, J. C., et al. 2023, *MNRAS*, 526, 2991
- Hausen, R. & Robertson, B. E. 2020, *ApJS*, 248, 20
- He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints, arXiv:1502.01852
- Hopkins, P. F., Somerville, R. S., Hernquist, L., et al. 2006, *ApJ*, 652, 864
- Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, *ApJS*, 221, 8
- Iglesias-Navarro, P., Huertas-Company, M., Martín-Navarro, I., Knapen, J. H., & Pernet, E. 2024, *A&A*, 689, A58
- Ioffe, S. & Szegedy, C. 2015, arXiv e-prints, arXiv:1502.03167
- Iyer, K. G., Gawiser, E., Faber, S. M., et al. 2019, *ApJ*, 879, 116
- Johnson, B. D., Leja, J., Conroy, C., & Speagle, J. S. 2021, *ApJS*, 254, 22
- Kang, D. E., Ksoll, V. F., Itrich, D., et al. 2023, *A&A*, 674, A175
- Kaviraj, S., Rey, S.-C., Rich, R. M., Yoon, S.-J., & Yi, S. K. 2007, *MNRAS*, 381, L74
- Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. 2020, *Artificial Intelligence Review*, 53, 5455
- Kingma, D. P. & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980
- Kotz, S., Balakrishnan, N., & Johnson, N. L. 2000, *Continuous multivariate distributions. vol. 1, models and applications*, Vol. 1 (A Wiley-Interscience Publication John Wiley and Sons)
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2017, *Commun. ACM*, 60, 84–90
- Ksoll, V. F., Reissl, S., Klessen, R. S., et al. 2024, *A&A*, 683, A246
- LeCun, Y., Bengio, Y., & Hinton, G. 2015, *Nat*, 521, 436
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, *Proceedings of the IEEE*, 86, 2278
- Lee, J. C., Sandstrom, K. M., Leroy, A. K., et al. 2023, *ApJ*, 944, L17
- Lee, J. C., Whitmore, B. C., Thilker, D. A., et al. 2022, *ApJS*, 258, 10
- Leja, J., Carnall, A. C., Johnson, B. D., Conroy, C., & Speagle, J. S. 2019, *ApJ*, 876, 3
- Leroy, A. K., Hughes, A., Liu, D., et al. 2021, *ApJS*, 255, 19
- Levesque, E. M., Kewley, L. J., & Larson, K. L. 2010, *AJ*, 139, 712
- Li, Z., Yang, W., Peng, S., & Liu, F. 2022, *IEEE Transactions on Neural Networks and Learning Systems*, 33, 6999
- Liew-Cain, C. L., Kawata, D., Sánchez-Blázquez, P., Ferreras, I., & Symeonidis, M. 2021, *MNRAS*, 502, 1355
- Lovell, C. C., Acquaviva, V., Thomas, P. A., et al. 2019, *MNRAS*, 490, 5503
- Lu, L., Dao, M., Kumar, P., et al. 2020, *Proceedings of the National Academy of Science*, 117, 7052
- Magris, G., Mateu, J., Mateu, C., et al. 2015, *PASP*, 127, 16
- Maksymowicz-Maciata, M., Spiniello, C., Martín-Navarro, I., et al. 2024, *MNRAS*, 531, 2864
- Marino, R. A., Rosales-Ortega, F. F., Sánchez, S. F., et al. 2013, *A&A*, 559, A114
- Martín-Navarro, I., Vazdekis, A., Falcón-Barroso, J., et al. 2018, *MNRAS*, 475, 3700
- Maschmann, D., Lee, J. C., Thilker, D. A., et al. 2024, *ApJS*, 273, 14
- McInnes, L., Healy, J., & Melville, J. 2018, arXiv e-prints, arXiv:1802.03426
- Melchior, P., Liang, Y., Hahn, C., & Goulding, A. 2023, *AJ*, 166, 74
- Noll, S., Burgarella, D., Giovannoli, E., et al. 2009, *A&A*, 507, 1793
- Ocvirk, P., Pichon, C., Lançon, A., & Thiébaud, E. 2006a, *MNRAS*, 365, 74
- Ocvirk, P., Pichon, C., Lançon, A., & Thiébaud, E. 2006b, *MNRAS*, 365, 46
- Panther, B., Heavens, A. F., & Jimenez, R. 2003, *MNRAS*, 343, 1145
- Pessa, I., Schinnerer, E., Sánchez-Blázquez, P., et al. 2023, *A&A*, 673, A147
- Pietrinferni, A., Cassisi, S., Salari, M., & Castelli, F. 2004, *ApJ*, 612, 168
- Pols, O. R., Tout, C. A., Eggleton, P. P., & Han, Z. 1995, *MNRAS*, 274, 964
- Robotham, A. S. G., Bellstedt, S., Lagos, C. d. P., et al. 2020, *MNRAS*, 495, 905
- Ruiz-Lara, T., Pérez, I., Gallart, C., et al. 2015, *A&A*, 583, A60
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, in *Nat*, Vol. 323 (Macmillan Education), 533–536
- Saleh, R. A. & Ehsanes Saleh, A. K. M. 2022, arXiv e-prints, arXiv:2208.04564
- Salvador-Rusiñol, N., Vazdekis, A., La Barbera, F., et al. 2020, *Nature Astronomy*, 4, 252
- Sánchez, S. F., Pérez, E., Sánchez-Blázquez, P., et al. 2016, *Rev. Mexicana Astron. Astrofis.*, 52, 21
- Sánchez-Blázquez, P., Ocvirk, P., Gibson, B. K., Pérez, I., & Peletier, R. F. 2011, *MNRAS*, 415, 709
- Sánchez-Blázquez, P., Peletier, R. F., Jiménez-Vicente, J., et al. 2006, *MNRAS*, 371, 703
- Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, *MNRAS*, 446, 521
- Tojeiro, R., Heavens, A. F., Jimenez, R., & Panther, B. 2007, *MNRAS*, 381, 1252
- Tolstoy, E., Hill, V., & Tosi, M. 2009, *ARA&A*, 47, 371
- Turner, J. C., Dale, D. A., Lee, J. C., et al. 2021, *MNRAS*, 502, 1366
- Valdes, F., Gupta, R., Rose, J. A., Singh, H. P., & Bell, D. J. 2004, *ApJS*, 152, 251
- Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, arXiv e-prints, arXiv:1706.03762
- Vazdekis, A., Koleva, M., Ricciardelli, E., Röck, B., & Falcón-Barroso, J. 2016, *MNRAS*, 463, 3409
- Vazdekis, A., Sánchez-Blázquez, P., Falcón-Barroso, J., et al. 2010, *MNRAS*, 404, 1639
- Vega-Ferrero, J., Domínguez Sánchez, H., Bernardi, M., et al. 2021, *MNRAS*, 506, 1927
- Walmsley, M., Slijepcevic, I., Bowles, M. R., & Scaife, A. 2022, in *Machine Learning for Astrophysics*, 29
- Walter, D., Ksoll, V. F., Klessen, R. S., et al. 2026, *A&A*, 706, A201
- Wang, L.-L., Yang, G.-J., Zhang, J.-L., et al. 2024, *MNRAS*, 527, 10557
- Werle, A., Cid Fernandes, R., Vale Asari, N., et al. 2019, *MNRAS*, 483, 2382
- Wilkinson, D. M., Maraston, C., Goddard, D., Thomas, D., & Parikh, T. 2017, *MNRAS*, 472, 4297
- Woo, J., Walters, D., Archinuk, F., et al. 2024, *MNRAS*, 530, 4260
- Worthey, G., Faber, S. M., & Gonzalez, J. J. 1992, *ApJ*, 398, 69
- Worthey, G., Faber, S. M., Gonzalez, J. J., & Burstein, D. 1994, *ApJS*, 94, 687
- Wu, Y. & He, K. 2018, arXiv e-prints, arXiv:1803.08494
- York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, *AJ*, 120, 1579
- Zhao, X., Ma, L., Liu, J., Sun, H., & Wang, Y. 2024, *Artificial Intelligence Review*, 57, 481

Appendix A: Stability of the loss function to the training size

Figure A.1 shows the final value of the loss function reached by our CNN when trained with random samples of different sizes. The final loss decreases monotonically with increasing dataset size up to $\approx 125\,000$, reaching a plateau beyond this point, except for the small fluctuations expected in the training process. This behaviour indicates that the adopted training-set size is sufficient to constrain the model and mitigates concerns about overfitting driven by an undersampled training distribution.

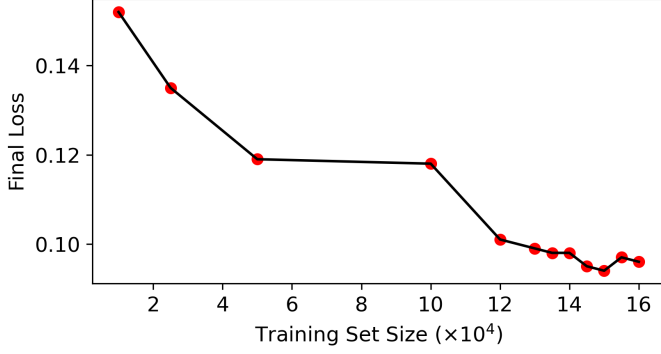


Fig. A.1: Final converged validation loss as a function of the number of training samples used.

Appendix B: Age-metallicity correlations

Figure B.1 shows the mean light-weighted (top row) and mass-weighted (bottom row) metallicities as a function of age for the synthetic test set (i.e. unseen during training). The left panels show the ground-truth values, and the right panels show the CNN predictions. Each panel displays a logarithmic hexbin density map. The predicted relations closely follow the input ones, indicating that the network preserves the underlying trends present in the synthetic data and that these relations are not washed out by dust, noise, or the range of chemical enrichment parameters.

Appendix C: UMAP representations

To assess the structure of the test dataset and the behaviour of the CNN across different regions of parameter space, we projected the 40-dimensional latent space onto a 2-D manifold using uniform manifold approximation and projection (UMAP; McInnes et al. 2018). Figures C.1 and C.2 present the resulting UMAP projection with the colour representing the $\langle \log(\text{Age}/\text{yr}) \rangle$ and $\langle \log([Z/H]) \rangle$, respectively. The bottom row represents the difference in the input and predicted values (in dex). The resulting embedding exhibits a smooth and continuous geometry, with no disconnected clusters, indicating that the latent space forms a coherent family rather than a mixture of isolated modes. Physical quantities such as mass- and luminosity-weighted age and metallicity display clean, monotonic gradients across the manifold, demonstrating that the latent space organises according to physically meaningful variations. Importantly, the residual maps show no coherent regions of systematic over- or under-prediction, indicating that the CNN performs consistently across the full latent space manifold. In Fig. C.3, we present the dust extinction and the strength of the H_α and H_β Balmer lines. The extinction is homogeneously distributed along the space and does not drive the

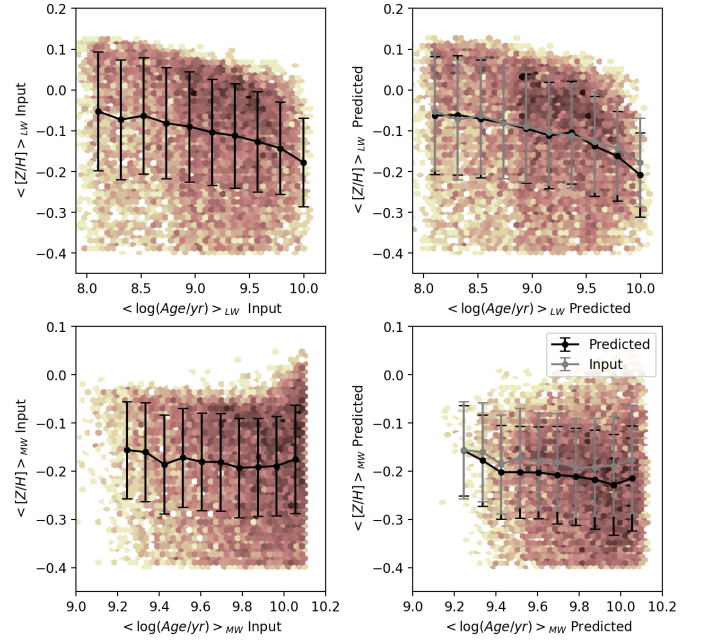


Fig. B.1: Age-metallicity relation for the input (left column) and predicted (right column) data. The panels show the relation between the mean $\langle \log(\text{Age}/\text{yr}) \rangle$ and mean $\langle [Z/H] \rangle$ weighted with luminosity (top row) and mass (bottom row). The colours represent the density of points in hexbins.

topology, confirming that the extinction plays a secondary role in defining the intrinsic structure of the dataset. The only slightly isolated clustered regions of the UMAP projection are located in the small section with high Balmer line measurements. Together, these diagnostics demonstrate that the test data populate a well-behaved physical manifold and that the network learns this structure robustly, without topology-dependent biases.

Appendix D: Performance of the CNN to data with realistic noise

To assess the robustness of the model under more realistic observational conditions, we perform an additional test in which synthetic spectra are perturbed using noise derived from real data. Specifically, we constructed noise from the RMS of the residuals obtained by fitting PHANGS-MUSE spectra with pPXF and alter each pixel of the spectra randomly with Gaussian distributions of width given by these residuals. This approach captures not only photon noise but also systematic features present in real observations, such as residuals from strong sky emission lines and other instrumental or reduction-related imperfections. Incorporating this type of noise provides a more faithful representation of the data domain and allows us to evaluate the impact of potential domain shifts between the synthetic training set and real observations. Figure D.1 compares the mean mass- and luminosity-weighted ages and metallicities recovered by the model under these conditions with the ground truth values. The bias and the dispersion in the predicted values do not show any statistically meaningful difference with those obtained with the data with Poisson noise (see Fig. 5), demonstrating that our model is robust to those imperfections of the spectra not correlated with the physical outputs of the CNN.

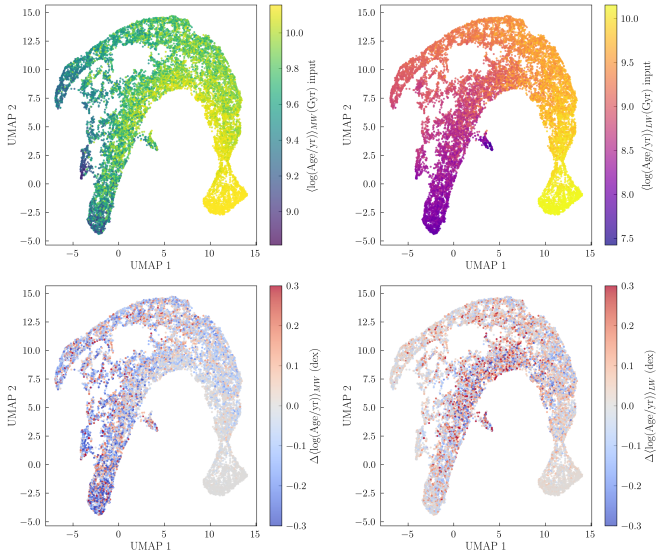


Fig. C.1: UMAP projection of the 40-dimensional latent space in the test set and corresponding CNN predictions. Each panel shows the same 2-D embedding, coloured by either the CNN’s prediction (top row) or the relative error (bottom row) for the $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{MW}}$ (left columns) and $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{LW}}$ (right columns).

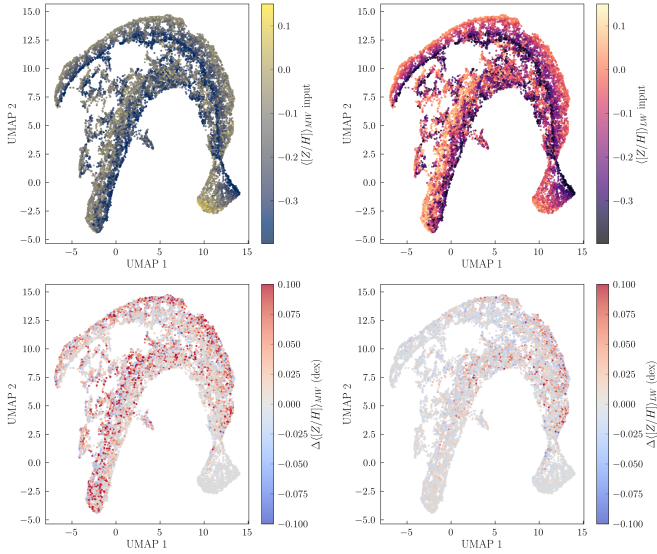


Fig. C.2: UMAP projection of the 40-dimensional latent space in the test set and corresponding CNN predictions. Each panel shows the same 2-D embedding, coloured by either the CNN’s prediction (top row) or the relative error (bottom row) for the $\langle [Z/H] \rangle_{\text{MW}}$ (left columns) and $\langle [Z/H] \rangle_{\text{LW}}$ (right columns).

Appendix E: Spectral segments and transmission responses used in the ablation tests

For clarity, Figs. E.1 and E.2 illustrate the location of the spectral segments and the photometric transmission bands used in the ablation analysis. These figures are intended to facilitate the interpretation of the results presented in Sect. 4.3 by showing the correspondence between the perturbed wavelength regions, the main spectral features, and the HST photometric filters.

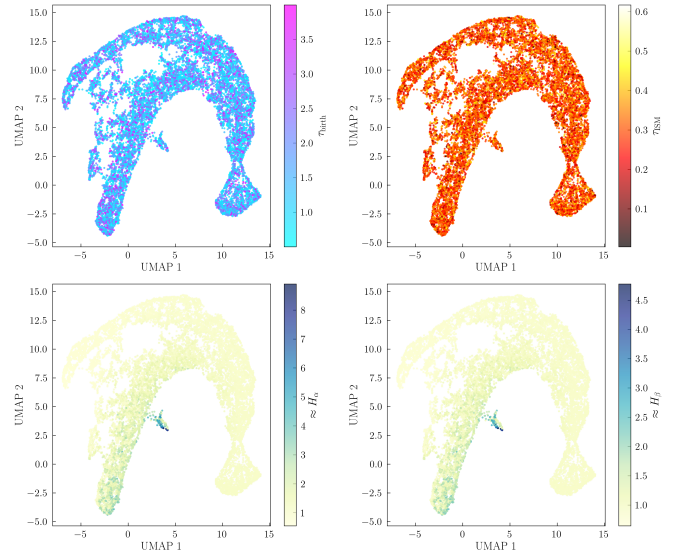


Fig. C.3: UMAP projection of the 40-dimensional latent space in the test set and corresponding CNN predictions. Each panel shows the same 2-D embedding, coloured (from top to bottom, and left to right) by the extinction values τ_{birth} and τ_{ISM} , and by the relative strength of H_α and H_β Balmer lines.

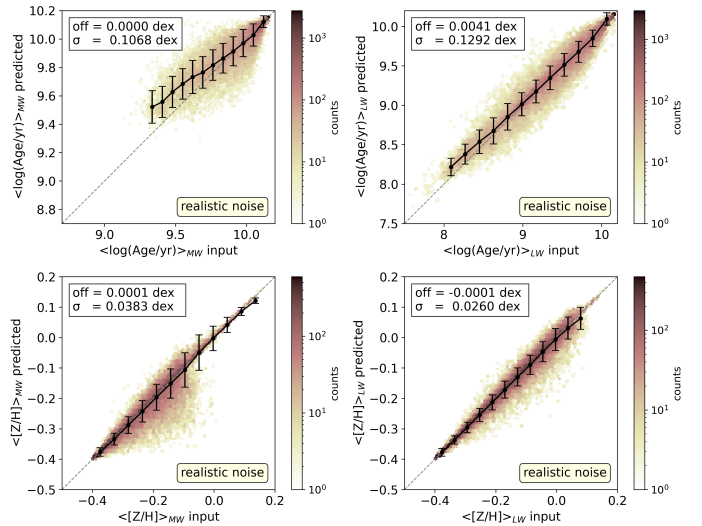


Fig. D.1: Comparison between the ground truth and the predicted mean mass- and luminosity-weighted ages (top row) and metallicities (bottom row) for test data altered with “realistic” noise (see text). The insets display the mean logarithmic differences between the input and predicted values ($\langle \log(\text{Age}/\text{yr})_{\text{input}} - \log(\text{Age}/\text{yr})_{\text{predicted}} \rangle$ and $\langle [Z/H]_{\text{input}} - [Z/H]_{\text{output}} \rangle$), together with their standard deviations. Each hexbin is coloured according to its point density, as indicated by the colour bar. Black symbols denote the median predicted values within bins of the corresponding input quantities.

Appendix F: Ablation test effect on the mass-weighted ages

Figure F.1 shows the mean differences between the input and predicted ages when the photometric fluxes or sections of the spectra are substituted by other values, randomly taken from a different SFH of the training set. Figures contain the same infor-

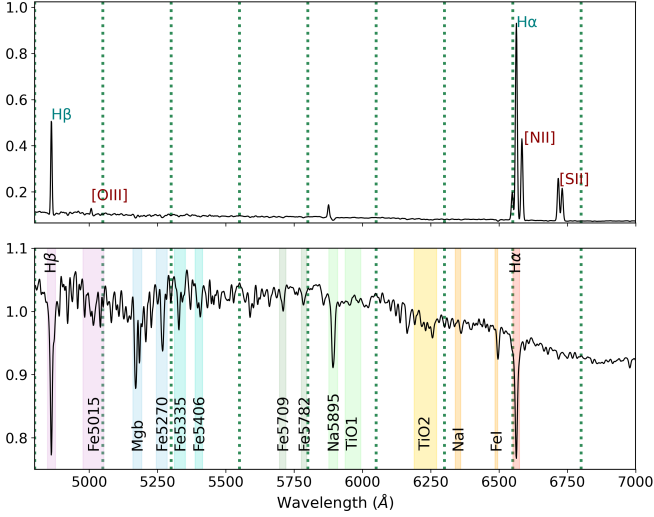


Fig. E.1: Main emission (top panel) and absorption (bottom panel) features in the spectral range 4780–7000 Å. The central passbands of the Lick/IDS indices (Worthey et al. 1994) are shown as shaded regions. The dotted vertical lines mark the nine spectral segments used in the ablation tests.

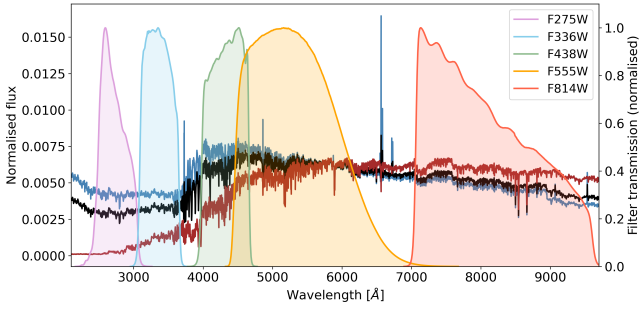


Fig. E.2: Transmission curves of the HST/WFC3 filters used in this work (F275W, F336W, F438W, F555W, and F814W), overlaid on three example spectra from the test sample. The spectra have similar mean metallicity ($\langle [Z/H] \rangle_{LW} = 0.17$) but different light-weighted ages: $\langle \log(\text{Age}/\text{yr}) \rangle_{LW} = 8.9$ (blue), 9.6 (black), and 10.1 (red). This comparison illustrates how the relative flux sampled by the different photometric bands varies with stellar population age.

mation as Figs. 9 and 10 but for the MW version of the mean ages. Figs. F.3 and F.4 show the results of the ablation tests for the mass-weighted metallicity predictions.

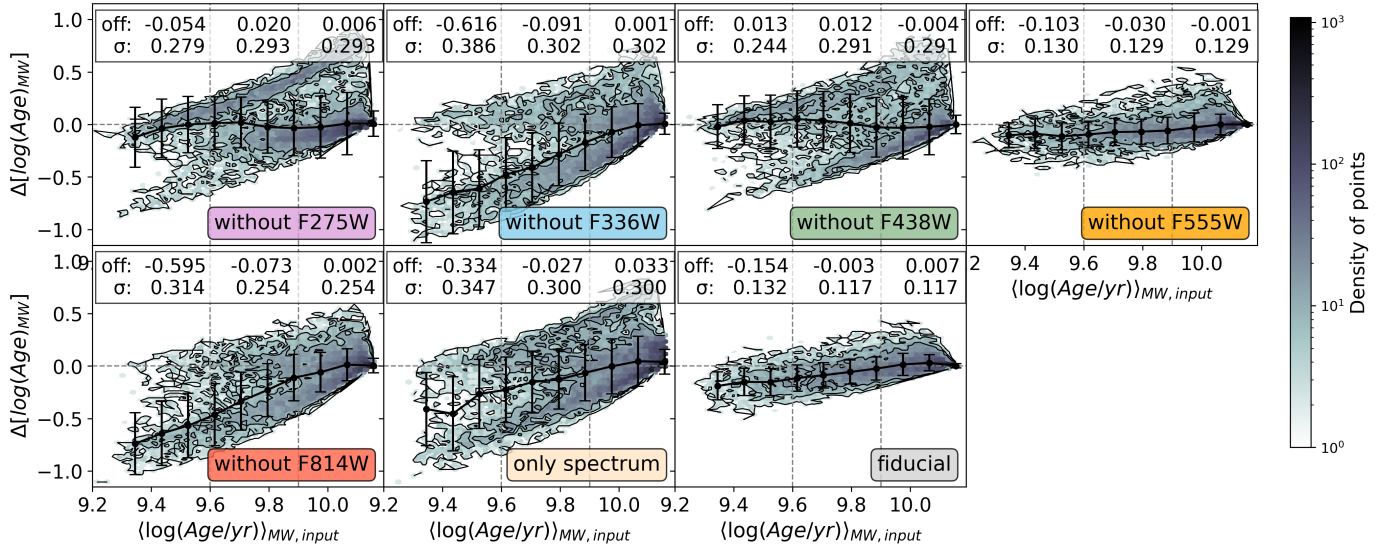


Fig. F.1: Impact of removing photometric fluxes from the input data on the recovery of the MW age. Each panel shows the difference between the predicted and input MW age as a function of the input value, when individual photometric fluxes are replaced by values randomly drawn from the empirical distribution of the whole sample. Additionally, the full photometric input can be replaced with random values, leaving only the spectrum unchanged. For reference, the fiducial case in which all photometric fluxes are included is also shown. The running median and standard deviation are plotted as black circles with error bars. The insets report the mean biases and standard deviations in three age intervals, whose boundaries are marked by vertical lines: (1) $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{MW}} \leq 9.6$; (2) $9.6 < \langle \log(\text{Age}/\text{yr}) \rangle_{\text{MW}} \leq 9.9$; and (3) $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{MW}} > 9.9$.

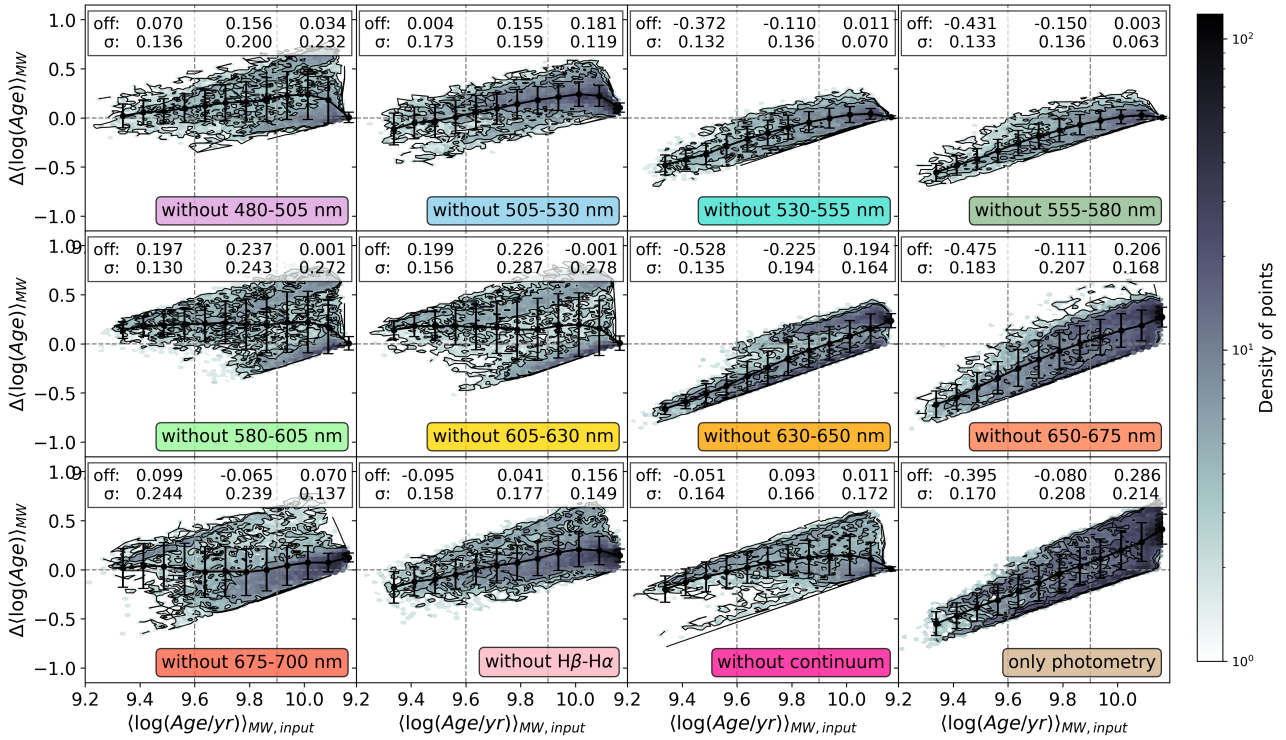


Fig. F.2: Impact of different spectral regions on the MW-age predictions. Each panel shows the difference between the ground-truth and predicted $\langle \log(\text{Age}/\text{yr}) \rangle_{\text{MW}}$ for different perturbations of the input spectrum. The first nine panels (from left to right and top to bottom) show the results obtained when individual wavelength segments are replaced by random values drawn from the empirical distribution of the same segment across the full dataset. The panel labelled ‘without continuum’ shows the results when the continuum shape is removed from the spectra. The final panel shows the case in which the full spectrum is replaced by random values. The running median and standard deviation are plotted as black circles with error bars. The insets report the mean biases and standard deviations in three age intervals (see Fig. F.1).

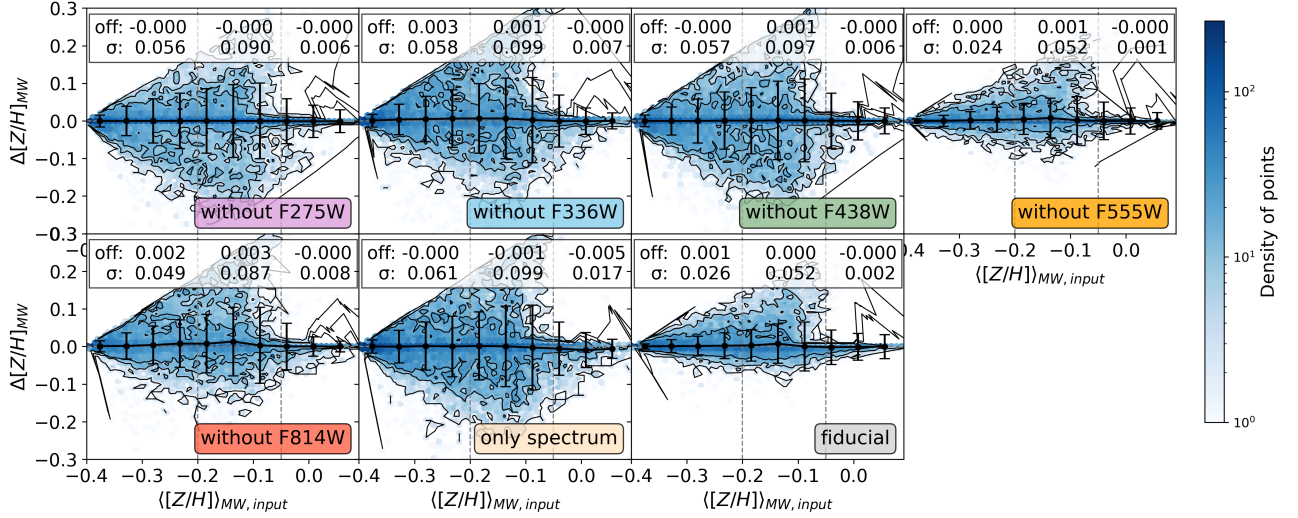


Fig. F.3: Same as Fig. F.1, but for the mass-weighted metallicity. The insets report the mean bias and standard deviation in three metallicity intervals, whose boundaries are marked by vertical lines: (1) $\langle [Z/H] \rangle_{\text{LW}} \leq -0.2$; (2) $-0.2 < \langle [Z/H] \rangle_{\text{LW}} \leq 0.05$; and (3) $\langle [Z/H] \rangle_{\text{LW}} > 0.05$.

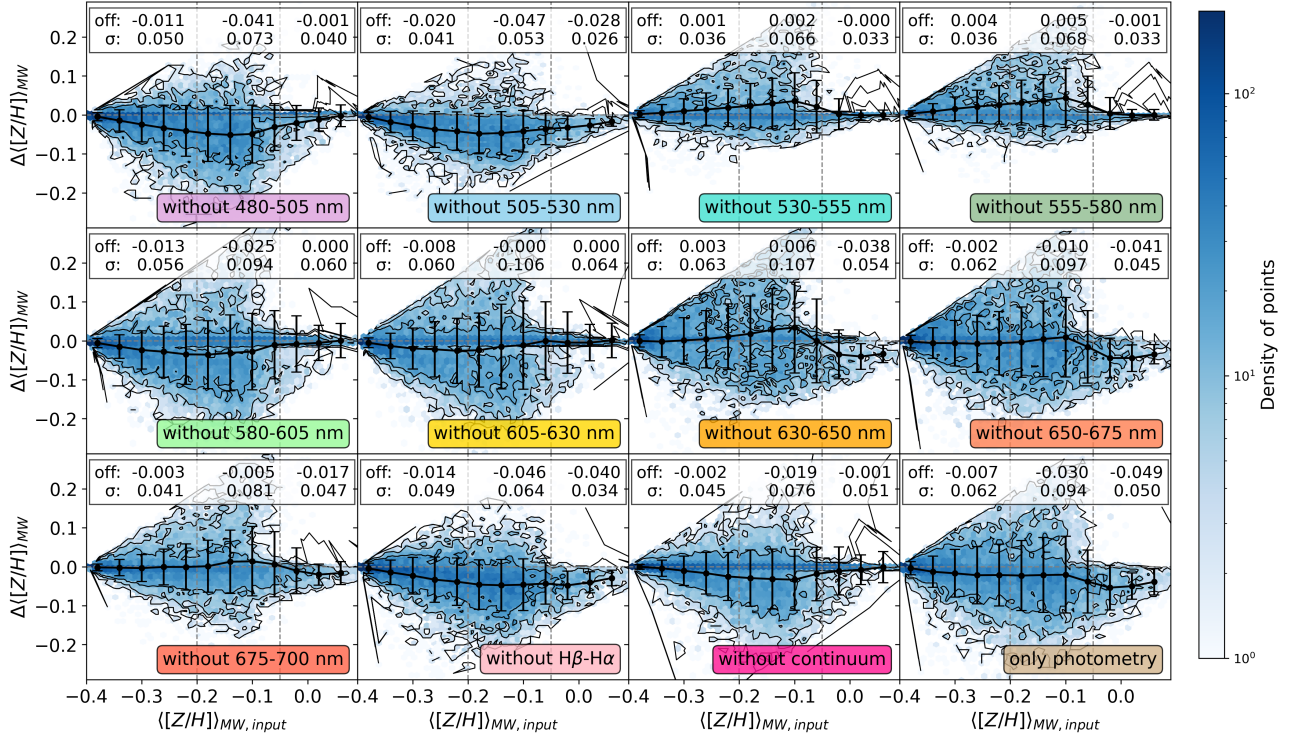


Fig. F.4: Impact of different spectral regions on the predictions of the mean MW metallicity. The different panels are equivalent to those in Fig. F.2, but for $\langle [Z/H] \rangle_{\text{MW}}$. Masking most wavelength intervals produces only minor changes in accuracy, while removing broad continuum information or relying on photometry alone leads to larger scatter, demonstrating the complementary role of spectral and photometric data in constraining metallicity. The insets report the mean biases and standard deviations in three metallicity intervals (see Fig. F.3).