**Universidad de La Laguna**

Facultad de Ciencias

# Deriving Star Formation Histories of Galaxies with Bayesian Deep Learning

*Master's Thesis*

Patricia Iglesias Navarro

Supervisors:
Ignacio Martín Navarro
Marc Huertas Company

July 2023

High-resolution galaxy spectra encode information about the stellar populations within galaxies. By investigating the properties of these stars, such as their ages, masses, and metallicities, we can gain insights into the underlying physical processes that drive the growth and transformation of galaxies over cosmic time, in particular, the triggers and quenching mechanisms of star formation. For this purpose, we explore an amortized implicit inference approach to estimate the posterior distribution of metallicity and non-parametric star formation histories (SFHs) of galaxies, i.e. star formation rate as a function of cosmic time, using their optical spectra. Fed with the spectroscopic predictions of the MILES stellar population models, we generate a sample of synthetic SFHs to train and test our model. We show that our approach is capable of reliably estimating the mass assembly of an integrated stellar population given its optical absorption spectrum with, crucially, well-calibrated uncertainties. Specifically, we achieve 94% accuracy for the time at which a given galaxy formed 50% of its stellar mass. We apply our pipeline to real observations of very massive elliptical galaxies and show that it recovers ranges of SFR($t$) consistent with the spectra, as well as the expected relation between age and velocity dispersion, demonstrating a good generalization to data. Not only being able to address a large number of galaxies, but also performing a thick sampling of the posteriors, it allows us to estimate both the deterministic trends and the inherent uncertainty of this highly degenerated inversion problem, so far inaccessible for more traditional methods. For this reason, we believe that our framework, a machine-learning-based implicit inference applied to full spectral fitting, is remarkably promising to deal with the size and complexity of upcoming massive spectroscopic surveys such as DESI, WEAVE or 4MOST.

*El espectro de una galaxia, es decir, el flujo de energía en función de la longitud de onda, nos proporciona información de sus poblaciones estelares. Al investigar las propiedades de estas estrellas, como sus edades, masas y metalicidades, podemos inferir los procesos físicos que impulsan el crecimiento y la transformación de las galaxias a lo largo del tiempo, en particular, los mecanismos desencadenantes y reguladores de la formación estelar. De hecho, la evolución de las galaxias es un equilibrio entre procesos que fomentan la formación estelar, y otros que tienden a impedirla expulsando o calentando el gas. Poder reconstruir con precisión las historias de formación estelar, es decir, la tasa de formación de estrellas en función del tiempo, es un paso imprescindible para comprender estos mecanismos. Sin embargo, inferirlas a partir de los espectros observados para una muestra estadísticamente significativa de galaxias es un problema complejo de inversión, sujeto a un gran número de degeneraciones no bien comprendidas, siendo la degeneración entre la edad y la metalicidad una de las más destacables.*

*Una herramienta comúnmente empleada para determinar las historias de formación estelar es el ajuste espectral. Este procedimiento consiste en comparar iterativamente los modelos con los espectros de galaxias observados, maximizando la semejanza entre los datos y los modelos. El éxito y la fiabilidad de este método dependen de la calidad de los espectros 'plantilla' y de la robustez del algoritmo de ajuste. En la última década, los métodos han evolucionado en el marco bayesiano hacia técnicas como los métodos Montecarlo basados en cadenas de Markov, permitiendo explorar eficientemente las degeneraciones asociadas al gran espacio de parámetros. Simultáneamente, los modelos basados en el aprendizaje automático se están haciendo cada vez más populares en la astronomía, proporcionando soluciones fiables y novedosas en múltiples ámbitos y disciplinas.*

*En esta línea, exploramos un enfoque de inferencia implícita amortizada para estimar la distribución 'posterior' de la metalicidad y de las historias de formación estelar de las galaxias utilizando sus espectros ópticos de absorción. Basándonos en los espectros de los modelos de poblaciones estelares MILES, generamos una muestra sintética de historias de formación estelar no paramétricas para entrenar y validar nuestro modelo. En particular, nos limitamos, dentro de los espectros de MILES, a una función de masa inicial (IMF por sus siglas en inglés) tal y como se observa en la Vía Láctea, a isocronas BaSTI y a unos modelos de enriquecimiento de elementos* $\alpha$ *escalados al entorno solar. Por otro lado, teniendo en cuenta las fuertes imposiciones sobre las medidas que genera emplear historias de formación estelar paramétricas (exponenciales, deltas, gaussianas, etc.), utilizamos el módulo GP-SFH para generar curvas no paramétricas basadas en modelos semi-analíticos, simulaciones hidrodinámicas y realizaciones estocásticas, minimizando así el efecto del 'prior' intrínseco en el conjunto de entrenamiento en las predicciones de nuestro modelo. Generamos espectros sintéticos de galaxias como combinación lineal de espectros de poblaciones estelares simples de MILES, donde los pesos vienen dados por las curvas de masa no paramétricas, y luego los normalizamos por su mediana. Así, obtenemos* 150.000 *galaxias sintéticas para las cuales conocemos su espectro, metalicidad e historia de formación estelar. Un* 90% *de estas galaxias se emplea para entrenar al modelo, mientras que el* 10% *restante,* 15.000 *galaxias, se emplea para evaluar su rendimiento.*

*Nuestro modelo consta de dos partes, un codificador del espectro y un estimador neuronal de densidad. La primera parte es indispensable para lidiar con la alta dimensión de los espectros (*$\sim$ 4.000 *puntos), empleando una red neuronal convolucional y un módulo de atención para comprimir el espectro en una representación latente de* 16 *componentes, que preserva la información específicamente asociada a las propiedades físicas que queremos predecir, en particular la metalicidad y el tiempo en formar el* 10%, 20%, ... 90% *de la masa estelar total de la galaxia (* 9 *percentiles de masa estelar). En el trabajo se demuestra que la distribución de estos vectores latentes está íntimamente relacionada con características del espectro empleadas en la literatura para realizar medidas similares, como las líneas de la serie de Balmer o el triplete de magnesio, y que además están optimizados para, de forma no lineal, capturar las historias de formación estelar y metalicidad.*

*A continuación, introducimos las representaciones latentes del espectro en la segunda parte del modelo, una red neuronal bayesiana capaz de estimar distribuciones 'posteriores' para los parámetros físicos a predecir. En concreto, se trata de un modelo de Flujo Autorregresivo Enmascarado (MAF, por sus siglás en inglés), que cuenta con dos principales ventajas. En primer lugar, no supone ninguna forma funcional para el 'likelihood' o verosimilitud bayesiana, ni para las distribuciones 'posteriores', práctica comúnmente utilizada en este tipo de inferencia y que en muchos casos favorece predicciones erróneas. Por otro lado, es un método amortizado, lo que significa que una vez el modelo ha sido entrenado, el proceso computacionalmente costoso, es posible generar predicciones adicionales del modelo, es decir, distribuciones 'posteriores' para los parámetros físicos de nuevas galaxias, de forma prácticamente instantánea. Esta característica, frente a los métodos Montecarlo basados en cadenas de Markov, permite no solo abordar un gran número de galaxias, sino ser capaces de evaluar con un número suficiente de realizaciones las distribuciones 'posteriores', lo que se traduce en una mejor determinación de las historias de formación estelar y metalicidades, pero sobretodo, de sus incertidumbres.*

*Mostramos que nuestro modelo es capaz de estimar de manera precisa las historias de formación estelar y metalicidad (supuesta fija a lo largo del tiempo) de una muestra sintética de galaxias, dado su espectro de absorción óptica, con incertidumbres bien calibradas. Concretamente, conseguimos una precisión del* 79.35%, 93.54% *y* 98.95% *para el tiempo en el que las galaxias forman el* 10%, 50%, *y* 90% *de su masa estelar total, y del* 75.83% *para su metalicidad* $[M/H]$. *Del mismo modo, realizamos un test de probabilidad de cobertura y demostramos que las incertidumbres que nuestro modelo proporciona están muy cercanas a la calibración perfecta, con una tendencia conservadora. Esto permite utilizar los errores proporcionados por el modelo como un límite superior de las incertidumbres, bajo la suposición de que los espectros a predecir son prácticamente indistinguibles de los espectros de entrenamiento, hipótesis fundamental que debe verificarse cautelosamente. Aplicando el modelo a los espectros sintéticos estudiamos cuales son las limitaciones intrínsecas del problema de inversión, recuperando las conocidas dificultades para inferir la edad de las primeras estrellas y la metalicidad en galaxias con formación estelar reciente, y cuantificándolas en las incertidumbres.*

*Finalmente, aplicamos nuestro modelo a observaciones reales de galaxias de tipo temprano (ETGs) con masas intermedias y altas. Obtenemos rangos de formación estelar en función del tiempo consistentes con los espectros, así como la relación esperada entre la edad y la velocidad de dispersión, demostrando una buena generalización de los datos. Comparamos nuestras medidas con las predicciones de un software de inversión consolidado, y validamos la capacidad de reproducir el espectro observado a partir de combinaciones lineales de espectros de MILES con edades y metalicidades iguales a las predichas por ambos métodos, con resultados muy favorables para nuestro modelo. De hecho, sin haber nunca optimizado las predicciones durante el entrenamiento para reconstruir los espectros, obtenemos un residuo medio del* 2% *en flujo con los espectros observados.*

*Así, nuestro modelo presenta una dualidad capaz de, no solo abordar una gran cantidad de galaxias, sino también realizar un muestreo amplio de las distribuciones 'posteriores'. Esto permite estimar tanto los valores deterministas como las incertidumbres intrínsecas al inferir las propiedades de las poblaciones estelares a partir de los espectros globales de las galaxias. Como continuación del proyecto, planeamos explorar los sezgos que la simulación de las galaxias de entrenamiento produce en las medidas, por ejemplo la suposición de una IMF fija, de una metalicidad constante o de modelos base de enriquecimiento de elementos* $\alpha$. *Una vez explorados estos efectos, el siguiente paso consistirá en medir las historias de formación estelar y la metalicidad de un conjunto más amplio de galaxias, con el objetivo final de aclarar los procesos que rigen la formación y evolución de las galaxias, haciendo uso de las observaciones espectroscópicas proporcionadas por proyectos como SDSS, DESI, WEAVE o 4MOST.*

# Contents

# 1 Motivation

At first order, galaxies are dark matter and baryonic overdensities which transform gas into stars. Understanding the physical processes that regulate star formation over cosmic time is one of the main challenges of galaxy studies, as their evolution is a balance between processes that trigger star formation and others that tend to prevent it by expelling or heating gas (Lilly et al., 2013; Martín-Navarro et al., 2020). Being able of reconstructing precisely the star formation histories is a fundamental step towards this direction. However, inferring them from the observed spectra for a statistically significant sample of galaxies is a complex inverse problem subject to a large number of degeneracies not well understood (Conroy et al., 2009).

A commonly employed tool is spectral fitting, the procedure of iteratively comparing models to the observed galaxy spectra, maximizing the resemblance between data and models (e.g. Cappellari, 2022). The success and reliability of this method depends on the quality of the template spectra, and of the robustness of the fitting algorithm. In the last decade, the methods have evolved into Bayesian statistics (Johnson et al., 2021), using primarily Markov Chain Monte Carlo (MCMC) for sampling the posterior distributions, which are assumed to be Gaussian, allowing to efficiently explore the degeneracies associated with the large parameter space. Simultaneously, machine-learning-based models are becoming more popular for astronomers (Huertas-Company et al., 2023). In contrast with spectral fitting, where the star formation history is built from some ensemble of simple stellar populations to recreate the spectra, machine learning directly learns the relationship between the observations and the entire star formation histories, carrying systematic uncertainties that are independent of those from spectral fitting, complementing and strengthening the results of these approaches, and learning from the population ensemble which star formation histories are common and which are unlikely, something analogous in Bayesian parameter estimation to the prior (Lovell et al., 2019).

In this work, we explore a novel approach based on probabilistic machine learning and likelihood-free inference to estimate the star formation histories of galaxies from their optical absorption spectra. Its main advantage over classical Bayesian inference methods is that it does not assume a functional form for likelihood, typically considered Gaussian, and that once the model is trained, it can be evaluated on different observations with minimal computational cost. The developed method will be compared with existing approaches of spectral fitting, and tested on observations of the Sloan Digital Sky Survey for early-type galaxies (La Barbera et al., 2013), helping to constrain the timescales involved in their evolution, past events and cosmological assembly histories.

# 2 Theoretical Background

## 2.1 Spectra and star formation histories

The spectra of galaxies, i.e. flux density as a function of wavelength, encode information about three of their main constituents: stars, gas and dust. While the absorption lines provide most of the information about the stellar component (ages, chemical composition and kinematics), the gas shows up in emission, in particular the very bright star-forming H II regions. The dust component, relatively cool, does not lead to any emission features in the optical spectrum ($\sim 10^3 - 10^4$ Å), but absorbs starlight (that re-emits in the far-infrared).

In the present work, as main baryonic drivers of galaxy formation and evolution, we will focus on the stellar population synthesis (SPS) from optical spectra, from which fundamental properties; such as the star formation history, the metallicity and abundances, or the initial mass function (IMF), can be inferred.

The starting point of a SPS model is a simple stellar population (SSP), which describes the evolution in time of the spectrum of a single stellar burst of fixed metallicity and chemical abundances, requiring three basic inputs: stellar evolution theory in the form of isochrones (e.g., Padova Girardi et al. (2000), BaSTI Pietrinferni et al. (2006), or MIST Choi et al. (2016)), empirical or theoretical stellar spectral libraries (e.g., CaT Cenarro et al. (2001), MILES Vazdekis et al. (2010), or XSL Gonneau et al. (2020)), and an IMF (Salpeter, 1955; Kroupa, 2001; Chabrier, 2003), each of which may in principle depend on metallicity and/or elemental abundance patterns. They are the building blocks for more complex stellar systems. Composite stellar populations (CSPs) differ from simple ones because they contain stars with a range of ages given by their SFHs, and with an interval of metallicities, as given by their time-dependent metallicity distribution function $P([M/H]$[1]$, t)$, and potentially other chemical abundances. An overview of the SPS technique is included in figure [1].

Thus, the SFH of a galaxy, i.e star formation rate (SFR) vs $t$, plays a key role in the SPS. A wrongly reconstructed SFH introduces significant biases in many parameters usually estimated through spectral fitting, such as stellar masses and ages, dust content, and redshift. In fact, several works have evaluated the impact of different sources on the errors when recovering physical quantities, concluding that the ones associated with the SFR($t$) are the most detrimental (Simha et al., 2014; Leja et al., 2017). Traditional parametric SFHs: Top-Hat (Tinsley, 1980), Exponentially Declining (Lee et al., 2010), or Log-Normal (Gladders et al., 2013), among others, have the advantages of being computationally fast and able to approximate short episodes of star formation (see figure [2]). However, they can not reproduce more complex behaviors like rejuvenation events, bursts or sudden quenching, failing to recover the true SFH when fit to galaxy observations. To solve this problem, efforts have been focused mainly on increasing the difficulty of the functional forms to give them more flexibility, instead of choosing well-motivated priors. Other well-known solutions are the so-called 'non-parametric' SFHs. Rather than not including parameters, they are defined as models which explicitly do not assume a functional form[2]. Some examples consist of fitting directly the stellar mass formed in fixed time-steps with constant values (Fernandes et al., 2008), using adaptive time binning (Tojeiro et al., 2007), obtaining the SFHs directly from theoretical libraries of galaxy formation (Pacifici et al., 2012), or assuming that the fractional specific SFR, sSFR, i.e. the SFR normalized by the stellar mass, for each time bin, follows a Dirichlet distribution (Leja et al., 2017). Despite being computationally more expensive, they constrain a wider range of behaviors in SFR($t$), offering explicit control over the density of curves, i.e. the prior, and promising the ability to describe the full breadth of complexity in galaxy SFHs (Iyer and Gawiser, 2017).

[1] i.e. metals over hydrogen, defined as $[M/H] = \log\left(\frac{M}{H}\right) - \log\left(\frac{M}{H}\right)_\odot$.

[2] Notice that as the number of parameters used becomes larger, the difference between parametric and 'non-parametric' will blur.

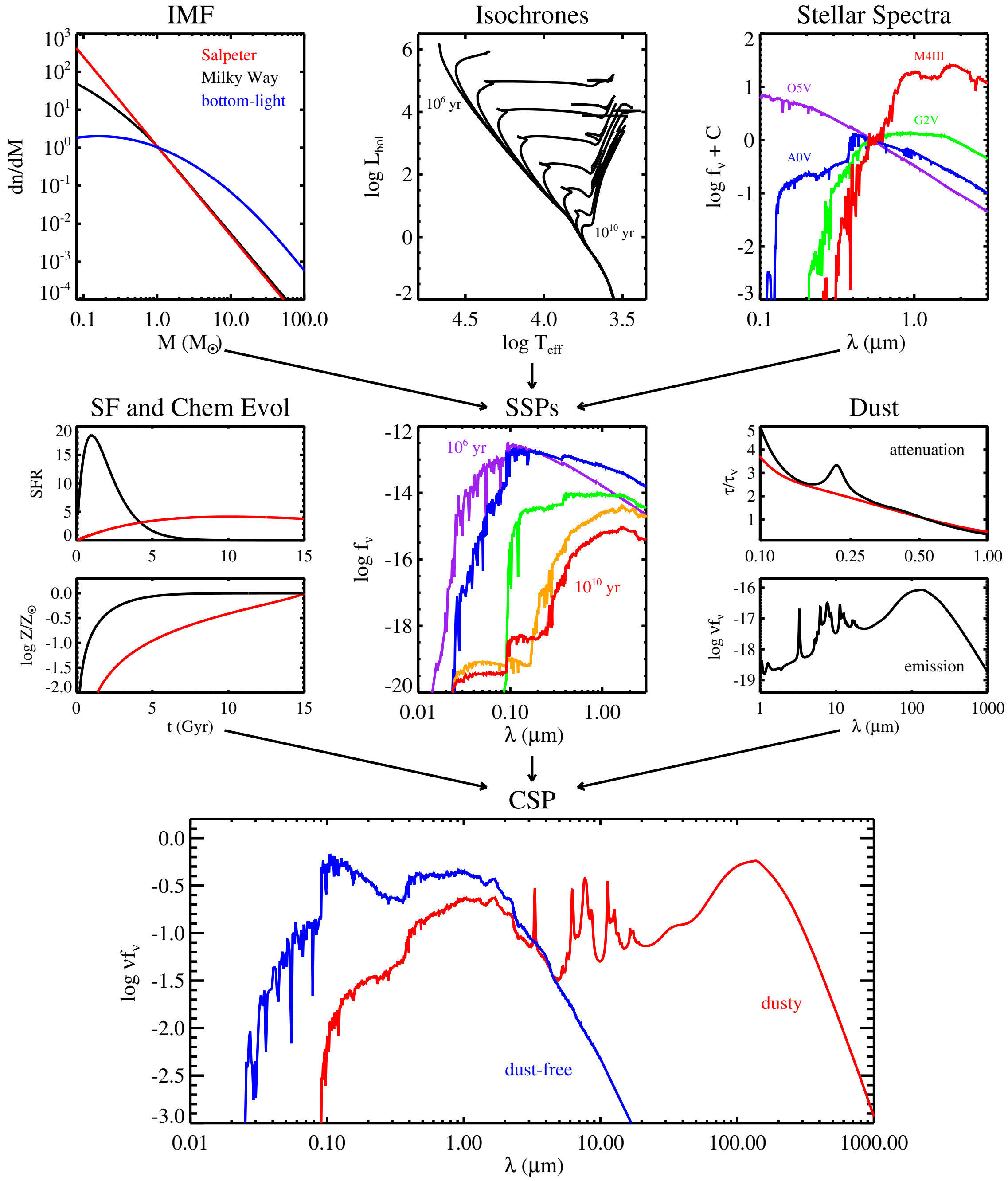

**Figure 1:** Overview of the SPS technique (Conroy, 2013). The upper panels highlight the ingredients necessary for constructing simple stellar populations (SSPs): an IMF, isochrones for a range of ages and metallicities, and stellar spectra spanning a wide range of parameters. The middle panels show the ingredients necessary for constructing composite stellar populations (CSPs): star formation histories and chemical evolution, SSPs, and a model for dust attenuation and emission. The bottom row includes the final CSPs both before and after a dust model is applied.

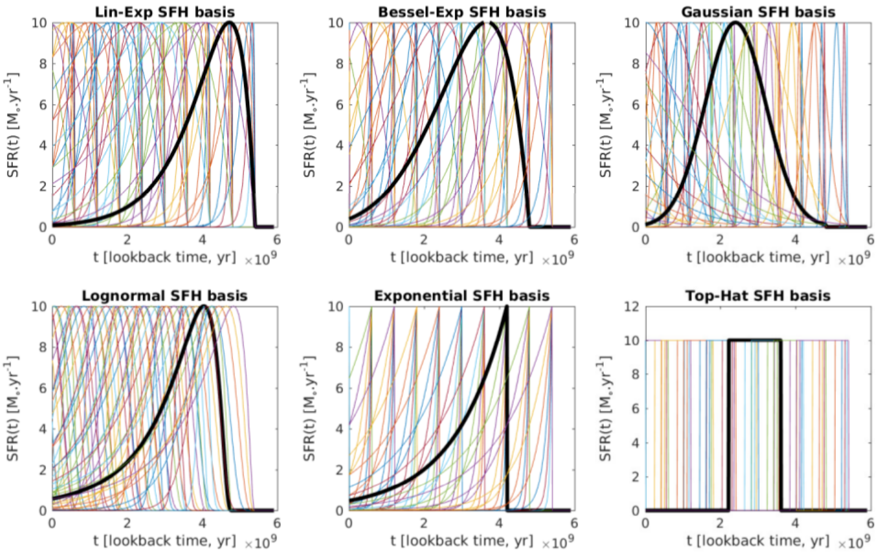

**Figure 2:** Representative examples of the SFHs at $z = 1$ for different functional families (Iyer and Gawiser, 2017). The full atlas for the Dense Basis method is constructed using all physical combinations of the linear-exponential, Bessel-exponential, Gaussian and log-normal families. A representative curve is shown in bold for each family.

In such a wide space of parameters, it is crucial to favor robust, temporally consistent, and above all, physically-motivated reconstructions. In Leja et al. (2019), they explore the effect of adopting different Bayesian priors when fitting non-parametric SFHs to photometry. They find strong impositions on the shape of the posteriors recovered, highlighting the necessity of a prior that best mirrors the distribution of galaxy SFHs in the real universe. This is especially relevant when the parameter to fit either has a relatively small effect in observable data or has a high degeneracy with other parameters (Carnall et al., 2019). In our case, the spectral signatures of multiple generations of stars can be mimicked by other physical effects, such as varying stellar metallicity, and older stellar populations with high mass-to-light ratios are easily hidden in observed spectra, an effect sometimes referred to as 'outshining' (Maraston and Strömbäck, 2011). So, we can use priors to guide our exploration of very large and degenerate parameter spaces.

To do that, it is required a statistical inference framework, which connects the model to the observations. Many techniques for SED-fitting in the literature have been based on maximum-likelihood optimization, a process sometimes called inversion (Walcher et al., 2010). These techniques are popular mainly because they are fast and simple. However, many of the likelihood spaces in SED-fitting are both highly non-Gaussian and ill-conditioned (Ocvirk et al., 2006), such that a slight change in the input (e.g., noise), can lead to a large change in the output (e.g., SFHs), resulting in unstable solutions that may cause severe difficulties in accurately assessing the uncertainties. While regularization can mitigate this amplification, it complicates interpretation. A solution to these problems can be found in Bayesian techniques, which are able to determine complicated, large, and correlated parameter uncertainties from galaxy observations, coupled with Markov chain Monte Carlo (MCMC) algorithms to efficiently explore the parameter space. This approach allows priors and fit parameters to be changed quickly for each sample, at the cost of longer per-object computational times (Johnson et al., 2021). However, there are other Bayesian methods, such as Amortized Neural Posterior Estimation (ANPE), which can, with similar performance, also decrease the time required for each prediction (Hahn and Melchior, 2022), as we will introduce later. With this in mind, we propose to apply ANPE to full spectral fitting using both MILES library (Vazdekis et al., 2010) and non-parametric SFHs (Iyer and Gawiser, 2017).

## 2.2 Simulation-based inference

Bayesian inference, applied to spectral fitting, tries to infer the posterior probability distributions $P(\theta \mid x)$ of galaxy properties, $\theta$, given observations, $x$. For a specific $\theta$ and $x$, we typically evaluate the posterior using Bayes' rule, $P(\theta \mid x) \propto P(\theta)\ P(x \mid \theta)$, where $P(\theta)$ denotes the prior distribution and $P(x \mid \theta)$ the likelihood, usually assumed to have a Gaussian functional form:

$$\log P(x \mid \theta) = -\frac{1}{2}(x - m(\theta))^T C^{-1}(x - m(\theta)), \tag{1}$$

where $m(\theta)$ is the theoretical model, in our case a galaxy spectral model, and $C$ is the covariance matrix of the observations.

Simulation-based inference (SBI), also known as likelihood-free inference, offers an alternative that requires no assumptions about the form of the likelihood. In a nutshell, it consists of creating synthetic data, which is often called the forward model, and then fitting them like real observations, the backward model, comparing against a known 'truth'. Therefore, SBI uses a generative model, i.e. a simulation $F$, to generate mock data $x'$ given parameters $\theta' : F(\theta') = x'$. It uses a large number of simulated pairs $(\theta', x')$ to directly estimate either the posterior $P(\theta \mid x)$, the likelihood $P(x \mid \theta)$, or the joint distribution of the parameters and data $P(\theta, x)$, using a neural density estimator, as shown in figure [3]. This technique has already been successfully applied to several Bayesian parameter inference problems in astronomy (Cameron and Pettitt, 2012; Mishra-Sharma, 2022; Hahn and Melchior, 2022), and many other fields in science (Cranmer et al., 2020).

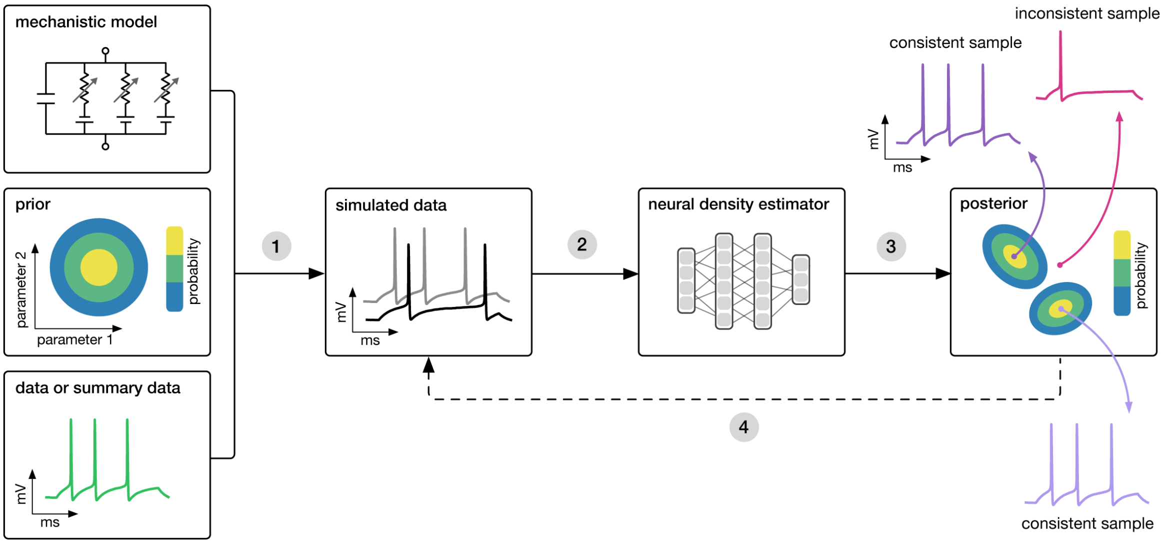

**Figure 3:** Workflow of the simulation-based inference approach (Tejero-Cantero et al., 2020): simulation of synthetic data (forward model) and prediction of the posterior distributions with a neural density estimator (backward model).

It is a convenient approach, first, because it tries to overcome the limitations of the parameter space coverage, that have nothing to do with the physics of stellar and galaxy evolution. Instead, they are mostly due to instrumental deficiencies that do not allow obtaining a broad and uniform dataset, usually required by inference algorithms. Moreover, SBI is especially helpful when testing single-effect biases (e.g., parametrizations of SFHs or noise), enabling to layer complexity. In contrast, it is likely to generate problems of domain shifts when generalizing results to real observations. Modeling considerations, like priors, wavelength coverage, or resolution, can have a significant impact on the inferred galaxy properties (Hahn and Melchior, 2022). Even uninformative uniform priors on spectral model parameters can impose undesirable priors on derived galaxy properties. To avoid strong biases, galaxy studies must carefully select priors and validate their results using multiple choices.

Connecting with our work, while the forward model used will be described in the section 3.1, to generate spectra given a SFH and a metallicity, at this point we continue discussing the basis of the two main ingredients of the backward model: an encoder of the spectra and an ANPE model. It allows us to obtain, given a galaxy spectrum, a reconstruction of its SFH and metallicity.

### 2.2.1 Convolutional encoders

High-quality observed spectra typically have, at least, a thousand data points. Because of this dimensionality, neither data-driven nor theoretical models have taken into account so far the full information content of galaxy spectra. Many physics-based approaches treat separately the continuum from the stellar absorption and nebular emission (Baldwin et al., 1981; Kewley et al., 2019). On the other hand, some are limited to nearby galaxies, where the cosmological redshift can be ignored (Brown et al., 2014), or to certain wavelength ranges, where a transformation to rest-frame can be easily applied (Portillo et al., 2020). The assumption that any given spectrum can be represented by a linear combination of a small number of basis vectors (Yip et al., 2004) or prototypical templates (Calzetti et al., 1994) further limits the complexity of spectral models.

However, the widespread adoption of template libraries suggests that galaxy spectra in fact occupy a low-dimensional manifold. In particular, Portillo et al. (2020) demonstrated that a high-fidelity reconstruction can be achieved by an autoencoder (AE) architecture (Hinton and Salakhutdinov, 2006), a non-linear dimensionality reduction technique, with a latent space of just six dimensions. Teimoorinia et al. (2022) and Melchior et al. (2022) improved the method by introducing convolutional elements into the AE to aid the extraction of correlated features from the spectra. In a few words, autoencoders are feed-forward neural networks that learn efficient encodings of data in an unsupervised manner. They consist of two parts: an encoder, which takes data as input, and compresses it to produce a low-dimensional latent representation, and a decoder, which takes the latent representations and decompresses them to reconstruct the original data. Due to their non-linear behavior, AEs can capture non-linear features, such as line widths, with fewer parameters than Principal Component Analysis (PCA) (Yip et al., 2004), one of the most commonly used techniques. Moreover, unlike line-ratio diagnosis, AEs use the continuum information in the spectra, resulting in an interpretable latent space that shows clear separations between different classes of galaxies, even though they were never given these classifications in training.

We take advantage of this approach, using the encoder part of the architecture implemented by SPENDER (Melchior et al., 2022) to obtain low-dimensional representations of the spectra, more suitable to introduce in a Bayesian inference model for a fully probabilistic treatment, and to perform meaningful summary statistics. Here we explain the design of the encoder, as well as the training pipeline, to reduce the number of components of the spectra, maximizing the information related to the stellar population ages (SFHs) and metal content.

The input of the encoder will be galaxy spectra normalized by its median. As we expect correlations between sharp lines or breaks and the continuum, it is chosen the convolutional encoder architecture from Serrà et al. (2018). The network consists of three convolutional layers with progressively wider kernel sizes, trainable PReLU activations, and max-pooling. These are used to compress spectra as follows:

- Convolution layer: extracts the characteristic features of data by applying filters (kernels). These are generated by initializers that seek not to have elements (weights) that are too small or large (worse for learning). The set of matrices resulting from multiplying the data vectors by the kernels are the feature maps. Once a feature map is created, we can pass each of its values through a nonlinearity, such as PReLU (He et al., 2015). This is a very simple activation function: $f(y) = \max(0, y) + \alpha \min(0, y)$, where $\alpha$ is a learnable parameter, commonly used in Deep Learning. It helps to prevent vanishing or exploding gradients, which makes training unstable, very slow, or no convergent.

- Pooling Layer: extracts the most representative pixels from a delimited strip of an image. It is an operation of downsampling used to reduce the overall size of tensors. In particular, the max-pooling operation calculates the maximum value for patches of a feature map, and uses it to create a new downsampled map.

Then, an attention module is applied. It is a machine learning technique useful for searching patterns across the data domain, which involves weighting features from the input data that are beneficial for subsequent tasks (Vaswani et al., 2017). Following the implementation of Melchior et al. (2022), it is applied in the wavelength direction to the feature maps, splitting the channels into two parts, $h$ and $k$, and combining them as:

$$e = h \cdot \mathrm{softmax}(k) \equiv h \cdot a, \tag{2}$$

where the dot product and the softmax function[3] operate on the wavelength dimension. The vector $a$ contains the attention weights, indicating whether and where relevant signals have been found, so that their corresponding values are promoted to the attended features $e$. This architecture is capable of accounting for the apparent shift of spectral features in galaxies at different redshifts, and, because of the wide convolution kernels, naturally folds in continuum features to form a highly informative latent representation. It is included an extra multilayer perceptron (MLP), a fully connected class of neural network, which further compresses the attended features $e$ to latent variables $S \in \mathbb{R}^s$.

However, the main objective of the works we have cited concerning this architecture was to reconstruct the spectra from the latent representations. Instead, we want to optimize these low-dimensional vectors to estimate SFHs and metallicity. A convenient way of training the weights for that purpose is incorporating a small MLP which predicts, from the latent vectors, the SFH and metallicity, through a loss function. In particular, we use a log-cosh function[4], which works mostly like the mean squared error, but will not be so strongly affected by the occasional wildly incorrect predictions.

To obtain the optimal weights, and so the optimal latent representations, an attempt is made to minimize the loss function, evaluating the encoder as a whole. The values of the weights are updated seeking to reduce the value of that function using an Adam optimizer with a fixed learning rate, in a process commonly called 'backpropagation'.

At this point, it is important to distinguish between three different datasets: the training dataset, the sample of data used to fit the model, the validation dataset, used to provide an unbiased evaluation of a model fit on the training dataset while tuning the hyperparameters (helps during the development stage of the model), and the test dataset, used to provide an unbiased evaluation of a final model. So, once we have a trained encoder, we can evaluate its performance with test data, never seen before by the model. It should be noted that the predictions of the encoder are not relevant once we have trained it: they only allow us to verify that from the latent representations one can effectively estimate the galaxy properties, driving the training. Thus, the encoder is only used to extract optimal summary statistics for the Bayesian inference, in which an associated uncertainty is provided for each prediction, as it is explained in the following section.

[3] $\mathrm{softmax}(x_i) = \frac{\mathrm{e}^{x_i}}{\sum_j \mathrm{e}^{x_j}}$

[4] $L\left(x_i, x_i'\right) = \log\left(\cosh\left(x_i - x_i'\right)\right)$

### 2.2.2 Amortized Neural Posterior Estimation

The fundamental challenge of the simulation-based inference approach is that the likelihood $P(x \mid \theta)$, implicit on the forward model, is not tractable, as it corresponds to an integral over all the parameter space, i.e. all possible execution traces of the simulator. For real-life simulators with large parameter spaces, such as those related to the SFHs, it is clearly impossible to compute explicitly. Since the likelihood function is the central ingredient to both frequentist and Bayesian inference, this is a major challenge, typically addressed with Approximate Bayesian Computation (ABC) or Variational Inference (VI).

The most well-known method is ABC, commonly relying on Markov Chain Monte Carlo (MCMC) or Sequential Monte Carlo (SMC), and vastly used for astronomical purposes (Iyer and Gawiser, 2017; Johnson et al., 2021). Lacking space to fully explain them, we refer the reader to a review of these techniques (Sisson et al., 2018), limiting ourselves to the approach we use in this work: Variational Inference. VI is based on creating a model by estimating the distribution of simulated data with histograms or kernel density estimators. One of the main advantages over ABC is that it is 'amortized', which means, not focused on any particular observation, so, once we have trained it, new data can be evaluated without repeating the training: the computationally expensive step. This property makes VI especially suitable for problems with many independent observations. Naturally, we must take into account that if the diversity of observations is large, any of the inference methods will need to run a sufficiently wide simulation to perform well across these diverse observations.

In this field, density estimation techniques based on neural networks are becoming more and more popular, because of their flexibility as opposed to classical VI, which often assumes functional forms for the distributions to predict. One class of these neural density estimators are Normalizing Flows (Rezende and Mohamed, 2015), in which variables described by a simple base distribution $P(z)$, such as a multivariate Gaussian, are transformed through a parameterized invertible transformation $x = f(z)$, that has a tractable Jacobian. The target density $P_f(x)$ is then given by the change-of-variables formula as a product of the base density and the determinant of the transformation's Jacobian:

$$P_f(x) = P(z) \left|\det J_{f^{-1}(x)}\right| \tag{3}$$

Several such steps can be stacked, with the probability density 'flowing' through the successive variable transformations (see figure [4]). The parameters of the transformations during the training of the model are estimated by maximizing the log likelihood of the transformed data under the Gaussian distribution $P(z)$, which is indeed tractable, being this the reason why the transformations are required to be invertible:

$$\log P_f(x) = \log P(f^{-1}(x)) + \log \left|\det J_{f^{-1}(x)}\right| \tag{4}$$

Otherwise, it is not possible to compute the loss since $P(x)$ is unknown. Following this pipeline, Neural Flows have been generalized to model a conditional density such as the likelihood $P(x \mid \theta)$ or the posterior $P(\theta \mid x)$.

In this work, we will infer the posteriors through VI using Normalizing Flows, in the framework of Amortized Neural Posterior Estimation (ANPE). Learning the posterior directly provides the main target in Bayesian inference, but induces a prior dependence at every stage of the inference method, which we must deal with very carefully. According to the choice of Hahn and Melchior (2022), we use a Masked Autoregressive Flow (Papamakarios et al., 2017), a highly flexible design that exploits the chain rule to model a joint probability of a set of random variables as products of one-dimensional conditional probabilities, each of them obtained with neural networks.

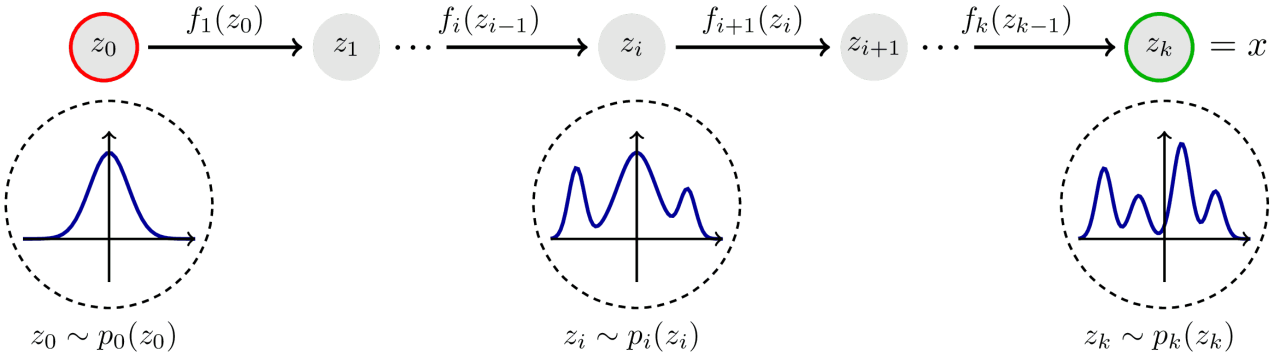

**Figure 4:** Overview of inference with Normalizing Flows (Riebesell, 2022). They are a general mechanism for estimating unknown probability distributions, only requiring the specification of a (usually simple) base distribution and a series of bijective transformations.

By using this workflow, we want to train our model with mock synthetic data, and then apply it to observations to quickly predict posteriors for SFHs and metallicity. It should be noted that we are making two important assumptions. First, the simulator $F$ is capable of generating mock data $x'$ that is practically indistinguishable from the observations, which is the same requirement of any probabilistic modeling approach, but unlike likelihood-based evaluations, such as conventional MCMC, data generated need to include all relevant features we can find in spectra, such as noise or outliers. Second, ANPE assumes that the neural density estimator is well trained, so $P_f(\theta \mid x')$ is a good approximation of $P(\theta \mid x')$, and therefore of $P(\theta \mid x)$. As opposed to standard VI, we are not assuming any functional form for the posterior. Therefore, since neural networks are universal approximators, we could eventually estimate the posterior with an arbitrarily small error.

# 3 Methodology

The purpose of this work is to derive the metallicity and star formation history (SFH) of galaxies from their absorption spectra. In this section, we will describe the steps to follow, and the tools to use such as a module for creating synthetic SFHs, an encoder of the spectra, and a neural density estimation pipeline that does not assume any functional form for the likelihood or the posteriors.

## 3.1 Forward model

First, we generate a synthetic dataset to train and test our model, for which the spectrum, the SFH, and metallicity are known. We work MILES SSP spectra (Vazdekis et al., 2010), whose parameters are included in table [1]. In figure [5], all the SSP MILES spectra for a fixed value of metallicity (corresponding to different ages) are plotted, as well as all the SSP MILES spectra fixing the age (with different $[M/H]$).

In order to reproduce physically-motivated non-parametric SFHs, we use Dense Basis (Iyer and Gawiser, 2017), in particular the module GP-SFH. It includes a combination of semi-analytic models (SAMs), hydrodynamical simulations and stochastic SFHs. In total, for the first training, 10.000 different SFHs are generated for each $[M/H]$ value (fixed along the galaxy's life). At this point, no specific prior on the stellar mass of the galaxies, or on any other feature, is included. In the next step, the stellar mass curves, i.e. SFR as a function of time, are integrated to get 9 stellar mass percentiles. We point out that throughout the entire work we use cosmic time, i.e. time since the Big Bang, except in the table [1] and in the figure [5], where we use lookback time according to MILES models. So, we focus on the cosmic time at which the 10%, 20%, ... 90% of the total stellar mass of the galaxy is formed. These quantities are more robust and smoother than their non-cumulative analogs, and can help to alleviate the effects imposed by the module's intrinsic priors, so the model can perform with more stability. In figure [6], 100 samples of SFHs are included, as well as their corresponding cumulative mass curves, from which we get the stellar mass percentiles.

Then, we interpolate MILES spectra in time to obtain a SSP spectrum each 0.01347 Gyr, with $t \in [0.00, 13.47]$ Gyr (1.000 bins fixed by GP-SFH). In addition, as the metallicity bins of the spectral library are not equally spaced, which may cause problems when introducing the spectra to the network, we interpolate them in $[M/H]$ too, obtaining 15 equally spaced values of this parameter in the range $[-2.3, 0.4]$. Finally, for each artificial SFH, given a value of $[M/H]$, all the MILES interpolated spectra (corresponding to different ages and with that metallicity) are combined as:

$$F_{\mathrm{gal}}\left(\lambda, M_{\mathrm{tot}}, [M/H], [\alpha/Fe]_{\mathrm{Base}}\right) = \sum_{t_i} \frac{M(t_i)}{M_{\mathrm{tot}}} \cdot F_{\mathrm{SSP}}\left(\lambda, t_i, [M/H], [\alpha/Fe]_{\mathrm{Base}}\right) \tag{5}$$

Then, each spectrum is normalized by its median. Finally, we obtain 10.000 SFHs for each value of $[M/H]$, so: 10.000 SFHs x 15 bins of metallicity = 150.000 samples.

## 3.2 Backward Model

### 3.2.1 Encoding the spectra

We want to encode 4.300-component vectors (the spectra) into a low-dimensional manifold to extract the most important features, so it is easier for the Bayesian model to learn (skipping useless information to accelerate training) and reconstruct SFHs, as well as metallicity. It should be noted that the size of the latent representations is a tunable parameter and it depends on what information in the spectrum is relevant to determine the mass percentiles and metallicity, as well as the dimension of the spectrum and the performance required. During this work, it is set to 16, even if several works justify that even 6 components for the vectors

| MILES SSP library parameters | |
|---|---|
| FWHM | 2.51 Å |
| IMF | Kroupa Universal[5] |
| Isochrones | BaSTI |
| $\lambda$ | $[3540.5, 7409.6]$ Å |
| $\Delta\lambda$ | 0.9 Å |
| Lookback Time | $0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.450.50, 0.60, 0.70, 0.80, 0.90, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00, 3.25, 3.50, 3.75, 4.00, 4.50, 5.00, 5.50, 6.00, 6.50, 7.00, 7.50, 8.00, 8.50, 9.00, 9.50, 10.00, 10.50, 11.00, 11.50, 12.00, 12.50, 13.00, 13.50, 14.00$ Gyr |
| $[M/H]$ | $-2.27, -1.79, -1.49, -1.26, -0.96, -0.66, -0.35, -0.25, 0.06, 0.15, 0.26, 0.40$ |
| $[\alpha/Fe]$ | Base model[6] |

**Table 1:** MILES SSP parameters used for generating the spectra of the synthetic dataset.

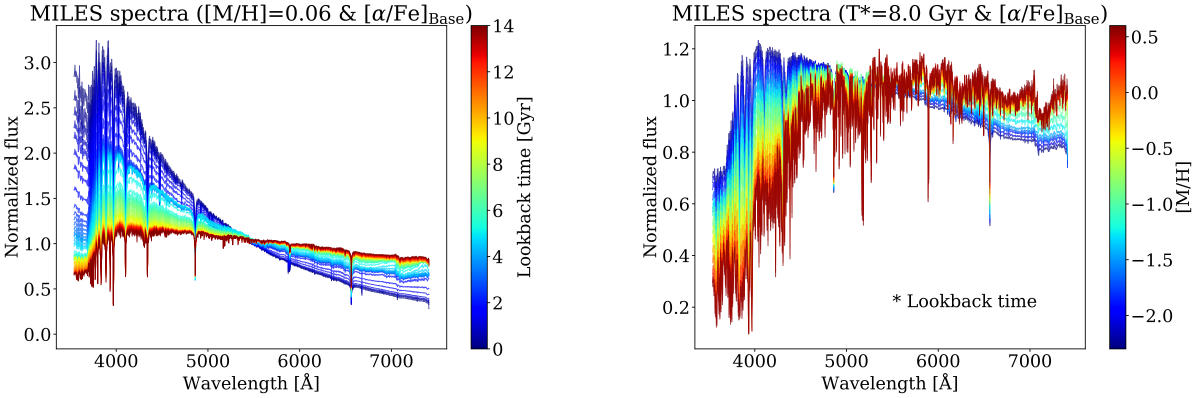

**Figure 5:** Normalized MILES SSP spectra according to different ages and metallicities. In the left panel, $[M/H] = 0.06$ is fixed, approximately solar metallicity, and ages vary between 0.03 and 14.00 Gyr (of lookback time). In the right panel, the age is fixed to 8 Gyr (of lookback time), and [M/H] values go from $-2.3$ to 0.4. Both correspond to MILES base models.

of the latent space can achieve high-fidelity reconstructions of the spectra (Portillo et al., 2020; Melchior et al., 2022). This choice has been made because most of the details of the spectra that contain information relevant to the early star formation of galaxies are extremely subtle.

The encoder used is inspired by SPENDER (Melchior et al., 2022). The network consists of a 3-layer CNN (moving to wider kernels and including max-pooling), plus an attention module (dot-product), and a 3-layer Multi-Layer Perceptron (MLP), to obtain the latent vectors. Furthermore, an additional 2-layer MLP is included to optimize the encoding for our final task: to obtain 9 stellar mass percentiles and a value of $[M/H]$, incorporating a convenient log-cosh loss function. This computes the difference between the predicted and actual values, driving the training's evolution. The architecture is shown in figure [7].

[5] We do not attempt to constrain the IMF. It is assumed universal and known a priori.

[6] Following the abundance pattern of the solar neighborhood, the MILES models are enhanced in $\alpha$-elements at low metallicity, whereas around solar metallicity these predictions are approximately scaled-solar.

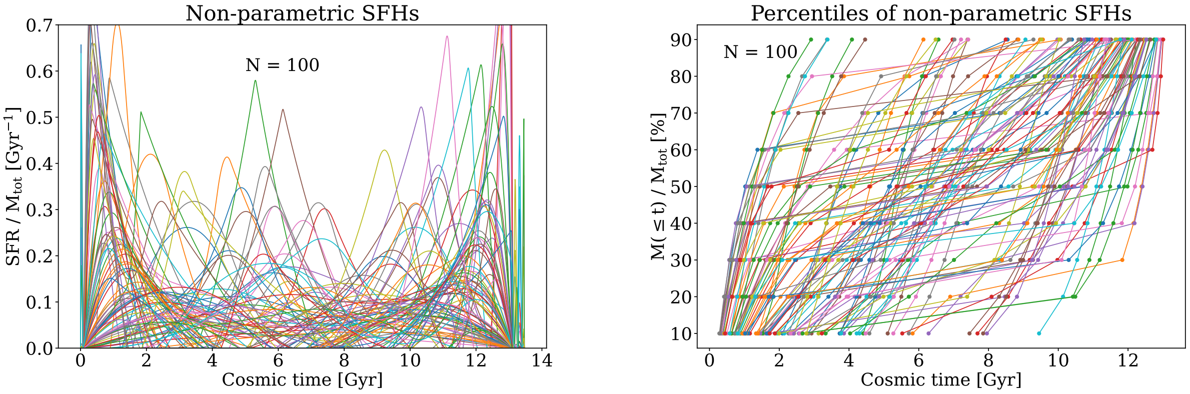

**Figure 6:** 100 non-parametric SFHs generated using GP-SFH (left panel) and their cosmic time values for 9 stellar mass percentiles (right panel). A normalization by the total stellar mass of each galaxy is performed to pursue clarity.

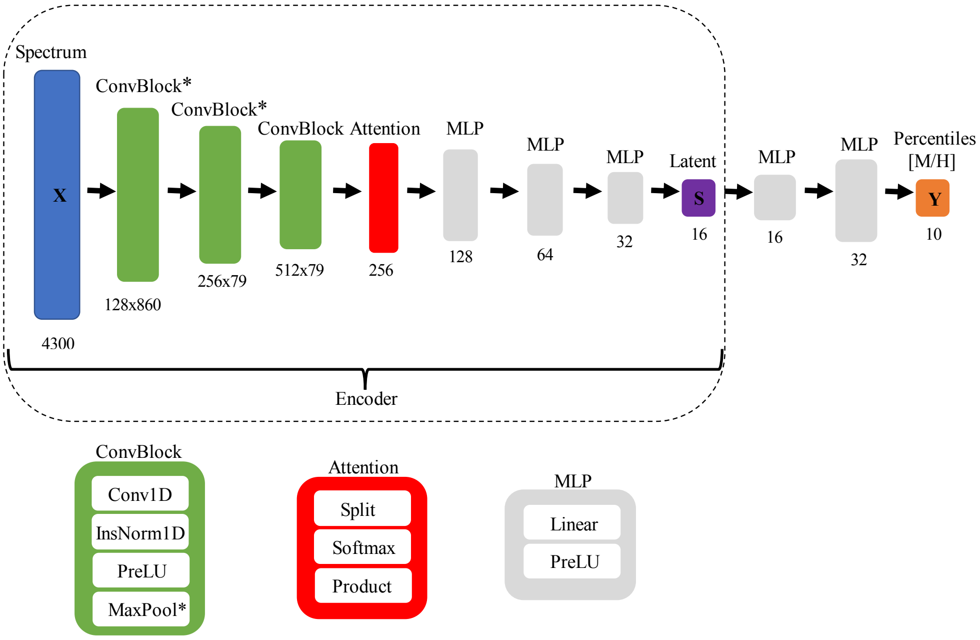

**Figure 7:** Encoder's architecture. It is composed of 3 convolutional blocks, a dot-product attention module and 3 MLP blocks. The inputs are the spectra and the outputs are 16-component latent vectors. 2 additional MLP blocks are included later to optimize the encoding, providing a prediction of 9 stellar mass percentiles and metallicity.

So, the encoder is trained with the training and validation sets (80% and 10% of the total number of samples), and the training is stopped when the validation loss reaches a plateau. Then, the model is applied to test data (remaining 10% of the total number of samples, never seen before by the network), getting not only the latent vectors we will use for the Bayesian inference, but also an initial prediction for the mass percentiles and metallicity, we can easily compare with the true ones. We remark again that these predictions will not be used once we have a trained encoder: they only drive the training process, while the proper measurements of the galaxy properties will be done by the neural density estimator.

We need to be sure that the latent representations contain relevant information, as they constitute the input of the Bayesian framework. However, due to its 16 components, they are difficult to interpret. We perform in section 4.1 several sanity checks, studying the distribution of their different components, correlations, and 2D projections. Once the training is completed, and the model evaluated with the test set, it is used to get the latent representations for the full dataset of 150.000 spectra.

### 3.2.2 Posterior estimation

We use Bayesian inference, in particular a kind of ANPE known as Normalizing Flows (see 2.2.2), to estimate the posterior probability distribution for each of the 9 stellar mass percentiles, and for the metallicity. The model is implemented with the module SBI (Tejero-Cantero et al., 2020).

We develop a Masked Autoregressive Flow (MAF) with 5 Masked Autoencoder for Distribution Estimation (MADE) blocks, each of them with 2 hidden layers of 128 hidden units. In total, the model has 50.560 free parameters $\phi$. Our goal is to determine $\phi$ for the MAF model, so that $P_{\phi}(\theta \mid x)$ accurately estimates the posterior probability distribution $P(\theta \mid x)$. In practice, we divide the training data into a training and validation set with a 90/10 split. We use the Adam optimizer with a learning rate of $5 \cdot 10^{-4}$. To prevent overfitting, we evaluate the likelihood of the data points under the base distribution, using the validation data at every training epoch, and stopping the training when that validation likelihood fails to increase after 20 epochs.

In section 4.2, 90% of the generated samples ( $x =$ latent vectors, $y = 9$ stellar mass percentiles, $[M/H]$) are set to train and validate the network. The remaining 10% of the samples is used to evaluate the performance of the model, by getting from their latent representations probability distributions for the values of each percentile and $[M/H]$, and comparing them with the true values. We also check the uncertainty estimation with a test of statistical coverage (Talts et al., 2018), and compute how much time it takes for the full model (encoder and neural density estimator) to predict the 10 distributions, given one spectrum. We further validate our model by measuring the SFHs and the metallicity for 18 ETG stacks (La Barbera et al., 2013) and comparing them with the results obtained through a consolidated spectral fitting method (Cappellari, 2022).

# 4 Results

## 4.1 Low-dimensional representations of the spectra

We encode the spectra to obtain low-dimensional representations, optimized to introduce them in the Bayesian inference model. The architecture seeks to preserve the relevant information to determine the SFHs and metallicity. The encoder is trained for 250 epochs, when the loss function stops decreasing for the validation set (see figure [8]). Its performance is evaluated with the test sample and eventually, the latent representations for the full dataset are obtained. We proceed to check the correlations between their components and with the spectral regions, which allows us to measure which features of the spectra provide the network with the most information on the stellar mass percentiles and the metal content.

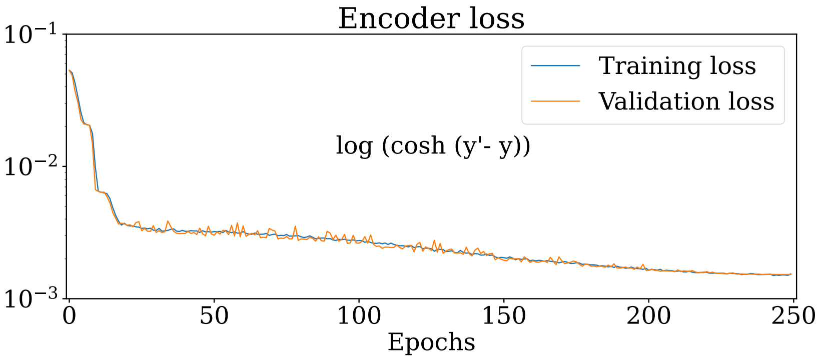

**Figure 8:** Training and validation loss of the encoder, computed as $\log\left(\cosh\left(y' - y\right)\right)$, where $y'$ is a vector with the predicted values for the mass percentiles and $[M/H]$, and $y$ a vector with the true values. The training is stopped at 250 epochs, when the loss functions reach a plateau. The evolution towards lower values is indicative that the latent representations given by the model preserve the information of the spectra related to the SFH and metallicity.

In figure [9], we observe that between the 16 components of the latent vectors, some correlations exist, as opposed to more traditional methods like PCA, where the components are orthogonal by definition. These slight correlations are precisely what allows encoders to capture non-linearities. The exhaustive analysis of how many components are optimal to encode the spectra is beyond the scope of this work, limiting ourselves to 16 dimensions, although we do point out again to the works of Portillo et al. (2020) and Melchior et al. (2022), specifying that for our purpose it is enough, but strictly necessary, to keep the information related to the SFH of the galaxy, and its average metallicity. As we go back in time, trying to infer information about the first stars to form, it becomes an arduous task, requiring subtle information that is difficult to determine a priori, so, cautiously, we use a larger number than the typically used in the bibliography, which is also larger than the number of quantities we want to predict.

In figure [10], the correlation between four different components of the latent vectors and the spectra is plotted, by including the absolute value of the Pearson's coefficient computed with the component value and the normalized flux in each wavelength. We do observe that different components focus on different features, such as $H_\beta$ (4863 Å), the NaI doublet (5890, 5896 Å), and $H_\alpha$ (6564 Å). It is also noticeable an outlier around $5000 - 5500$ Å, both very high correlation or not at all. This intermediate region typically has a unit of flux, as a consequence of normalization, and contains the magnesium triplet, relevant for the study of metallicity, so this effect can be justified. Thus, the most correlated spectral regions, taking into account the 16 components, are at first glance those that provide the network with the most relevant information about the physical

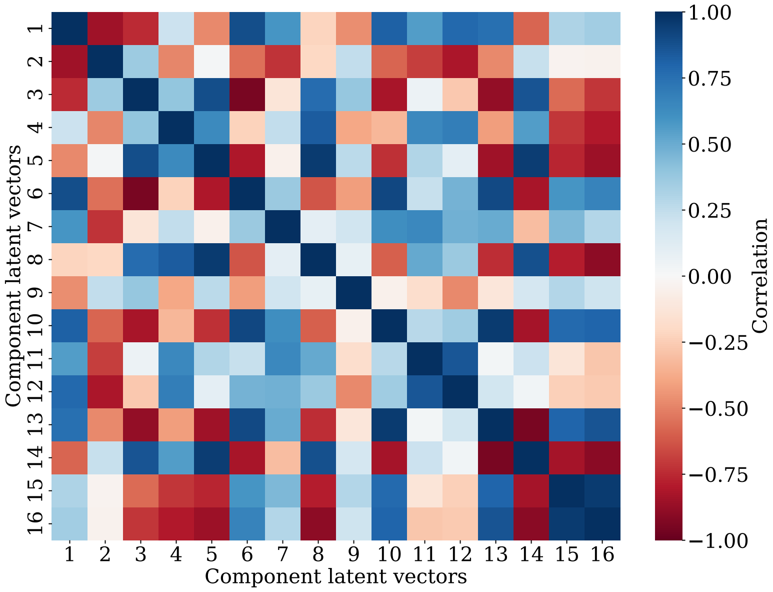

**Figure 9:** Correlation between the 16 different components of the latent vectors, computed as the Pearson's coefficient. We observe how different components of the vectors are correlated with each other (off-diagonal elements) both positively and negatively. This feature allows the encoder to capture non-linearities, such as the width of the absorption lines, in contrast to methods like PCA, where the components are orthogonal by definition.

quantities we want to predict, recovering the fundamental role of the mentioned absorption lines, widely used in measurements with line-strength indices (Vazdekis et al., 2010).

In figure [11], it is performed a Uniform Manifold Approximation and Projection (UMAP) of the latent vectors, only for visualization purposes. It consists of a non-linear dimensionality reduction technique that produces two-dimensional maps topologically equivalent to the representation of the high-dimensional data, the latent vectors of 16 components (we refer the reader to McInnes et al. (2018) for more details). The shape of the UMAP, where each dot corresponds to a synthetic galaxy from the test sample, can be intuitively explained according to the physical quantities we want to recover, demonstrating a good encoding of the spectral information. The map is shown using five different colormaps: the value of the cosmic time for the 10% and 90% mass percentiles, the metallicity, and the line indices $H_{\beta_0}$, an indicator of the age, and Mgb5177, an indicator of the metallicity (Vazdekis et al., 2010). First, we see in the percentile 90% and $H_{\beta_0}$ plots how quenched galaxies are clustered in the lower right zone, and young ones (with more recent SF) in the left and upper right zones. These two different clusters for young galaxies can be justified if we check the percentile 10% plot, as in the left only galaxies with a late start in forming stars are included (large values of cosmic time for the percentile 10%), so with late and brief star formation, while in the upper right we have galaxies with an early and extended star formation (low 10% percentile and high 90% percentile values referring to cosmic time). In addition, it is observed that although in the lower right region there is no difference in ages, there is in metallicity, dividing old galaxies into two groups: the upper group is metal-rich and the lower one is metal-poor (see both metallicity and Mgb5177 plots). Finally, we highlight that in the youngest galaxies there is no clear distinction in metallicity. The complexity of measuring metallicity in young populations is already known (Conroy, 2013) and it is based on stellar physics, as the high temperatures of the atmospheres of massive stars cause the metallic lines to be very faint, and the parameter that fundamentally governs the spectra is age. So, the network, without any

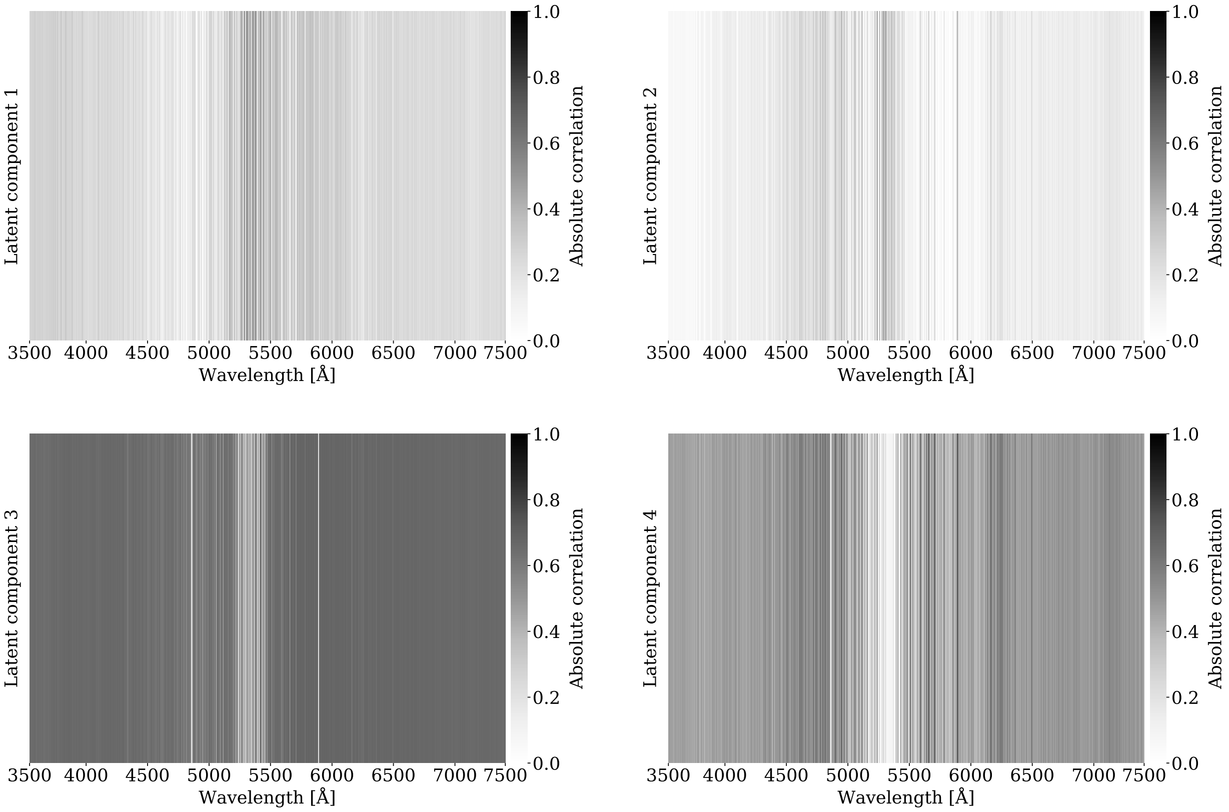

**Figure 10:** Correlation between the first 4 components of the 16-component latent vectors codifying the spectra, and the actual flux along the wavelengths, computed with the absolute value of the Pearson's coefficient. The spectra are encoded to optimize the information regarding their SFHs and their metallicity, and we observe how the components of their low-dimensional representations focus on different spectral features, such as the Mg I triplet (5167 Å, 5172 Å, 5183 Å), the Balmer lines (4863 Å, 6564 Å) or the Na I doublet (5890 Å, 5896 Å), as well as the continuum.

constraints in the input, naturally recovers the underlying physics of spectral fitting.

## 4.2 Recovering SFHs of synthetic galaxies

The model is trained with the synthetic data and then applied to the test sample. Once the training is finished ($\sim$ 4 hours: 1 hour for the encoder and 3 hours for the Normalizing Flows), each posterior estimation is performed with 1.000 samples, taking $\sim$ 0.4 s to get the predictions for the 10 quantities of each galaxy. These time estimations have been done on a NVIDIA Tesla P100 PCIe GPU with 12GB.

We show 4 predictions for the stellar mass growth in figure [12], observing how the medians of the posteriors are located very close to the true values, inside the confidence interval given by two standard deviations. Thus, the model is able to efficiently recover from the latent vectors the SFHs of these four galaxies, with meaningful error intervals. In figure [13], we include the medians of the posterior distributions predicted for the 15.000 test galaxies against the true values, for the percentiles 10%, 50%, 90%, and for the metallicity. Both seem to agree, close to the one-to-one relation, as well as the distributions shown in the edges of the panels, reaching high accuracy values: 79.35%, 93.54%, 98.95%, and 75.83%, respectively. A

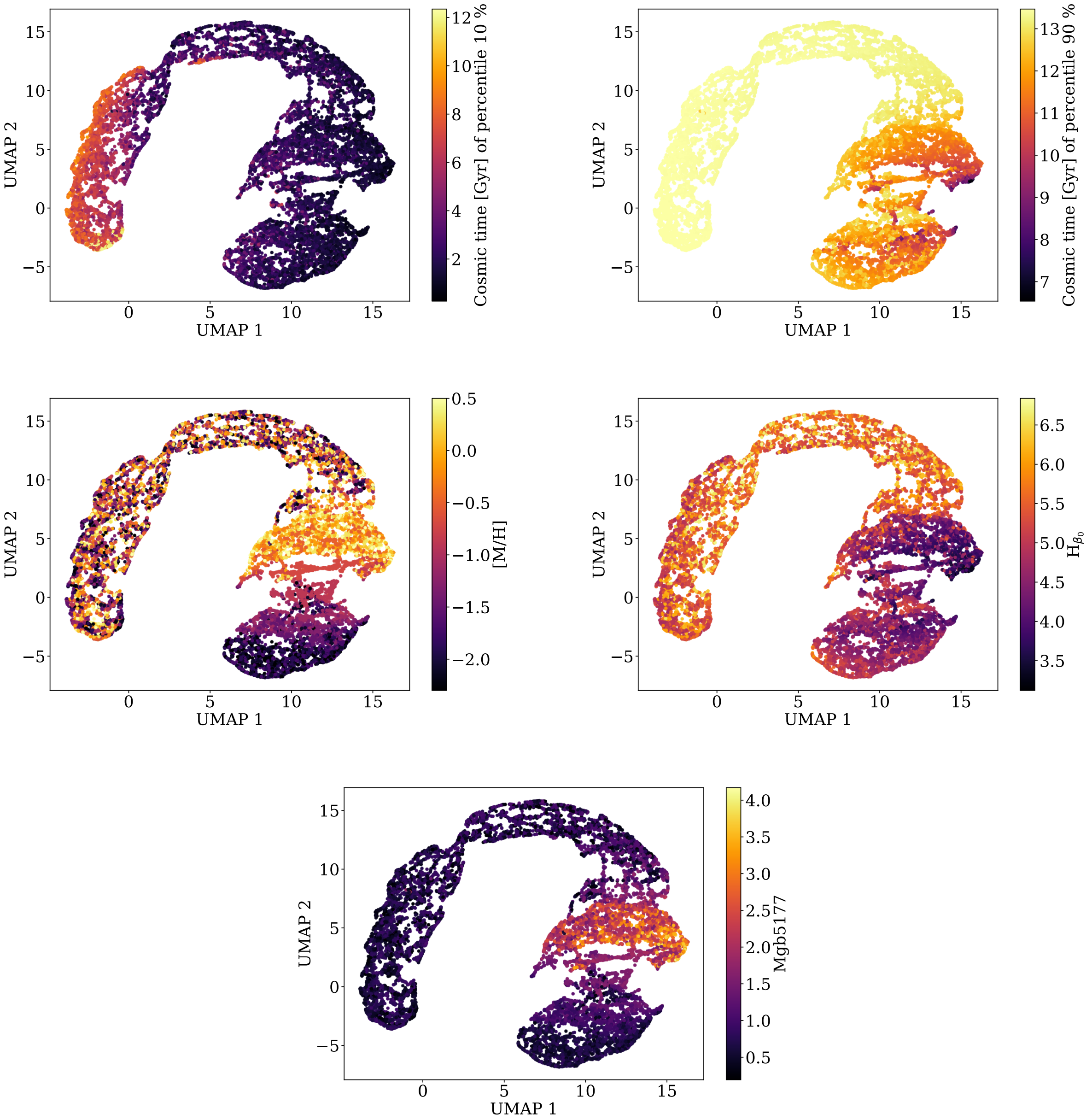

**Figure 11:** UMAP embedding for the latent representations. Each dot corresponds to a synthetic galaxy from the test sample. The closer two galaxies are on the map, the more similar are their latent representations. Different colormaps are set according to the cosmic time at which the percentiles 10% and 90% of the total stellar mass are reached, $[M/H]$, and the two line indexes $H_{\beta_0}$ and Mgb177. Clear trends can be observed, demonstrating that the information encoded in the latent vectors is optimal for recovering the SFHs and metallicity.

larger scatter is observed for earlier percentiles, which is expected as the luminosity of young stars ‘outshines’ the spectra, hiding the information of the oldest. These two figures show that the model is indeed capable of

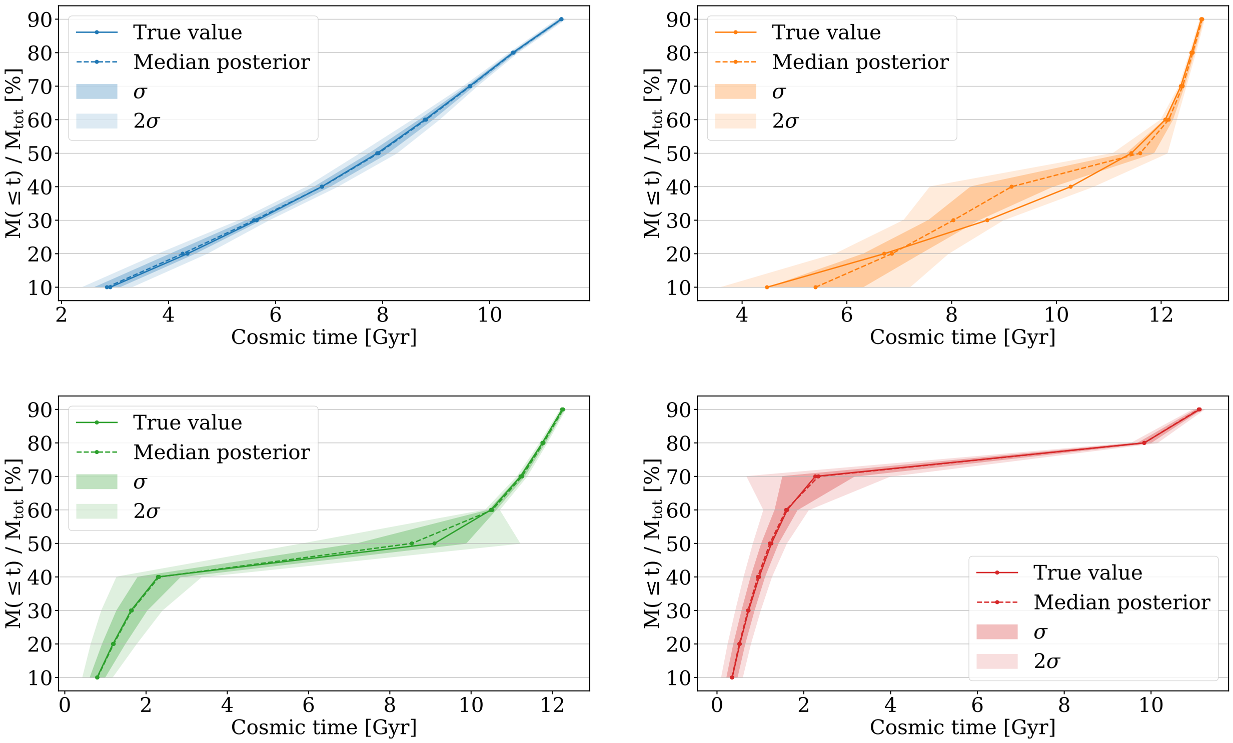

**Figure 12:** Percentile predictions for four different synthetic galaxies. The cumulative mass curves are shown as the time at which the 9 stellar mass percentiles are reached across the cosmic time, in Gyr. The solid lines correspond to the true values and the dashed ones to the predictions (medians of the posterior distributions). The $\sigma$ and $2\sigma$ intervals of confidence are shaded dark and light, respectively. We observe that the model performs a reliable reconstruction for all four galaxies.

recovering the stellar mass growth and metallicity with uncertainties associated in part with the complexity of the inversion problem, due to the very nature of the spectra and to the observational imprints left by galaxies on them.

In figure [14], we include a corner plot for the posterior predictions of a test galaxy from the synthetic sample. We observe that the posterior distributions are centered in the true values, shown with solid lines, demonstrating again an optimal performance. Correlations between the distributions, captured by the model without any external condition, are mainly in consecutive percentiles. They indicate that the model is capable of detecting that these are not independent quantities, since omitting processes that cause sudden changes in star formation (e.g., wet mergers), mass growth often occurs smoothly and continuously.

Hereunder, the uncertainties are validated in figure [15], where a coverage probability test is performed for all the predictions. On it, we verify we have in consecutive confidence intervals the proper number of samples, assuming a Gaussian behavior for the posterior distributions, for the first time in this work. This assumption of unimodal posteriors has been made after checking that the distributions predicted by the neural density estimator do not show very strong multimodalities in the corner plot. We do observe the model is close to the perfect calibration line, in the conservative half for every feature we are predicting, which allows us to estimate

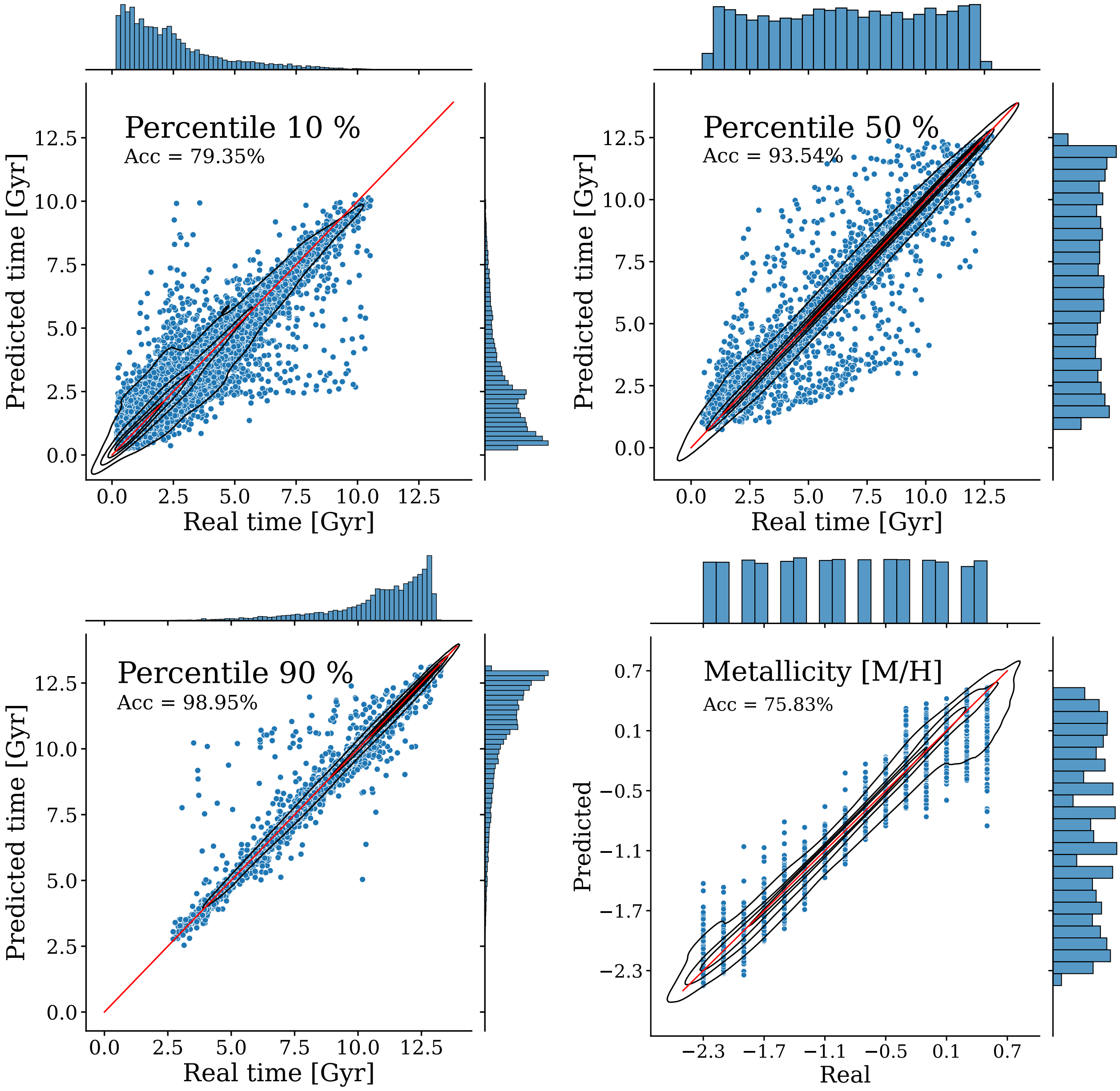

**Figure 13:** Medians of the posterior distributions estimated for the percentiles 10%, 50%, 90%, and $[M/H]$, compared to the true values. The accuracy achieved for each prediction is 79.35%, 93.54%, 98.95%, and 75.83%, respectively. Each blue dot is a different sample from the test set. The red line shows the one-to-one relation, the histograms at the right of each panel show the marginal distributions of the predictions, and the histograms of the real data are shown at the top. Kernel Density Estimation (KDE) contours are drawn in black at iso-proportions of the density of samples.

an upper limit for the errors of the model. These errors are able to capture the intrinsic uncertainties of the degenerated inversion problem we are dealing with, so far inaccessible for traditional methods.

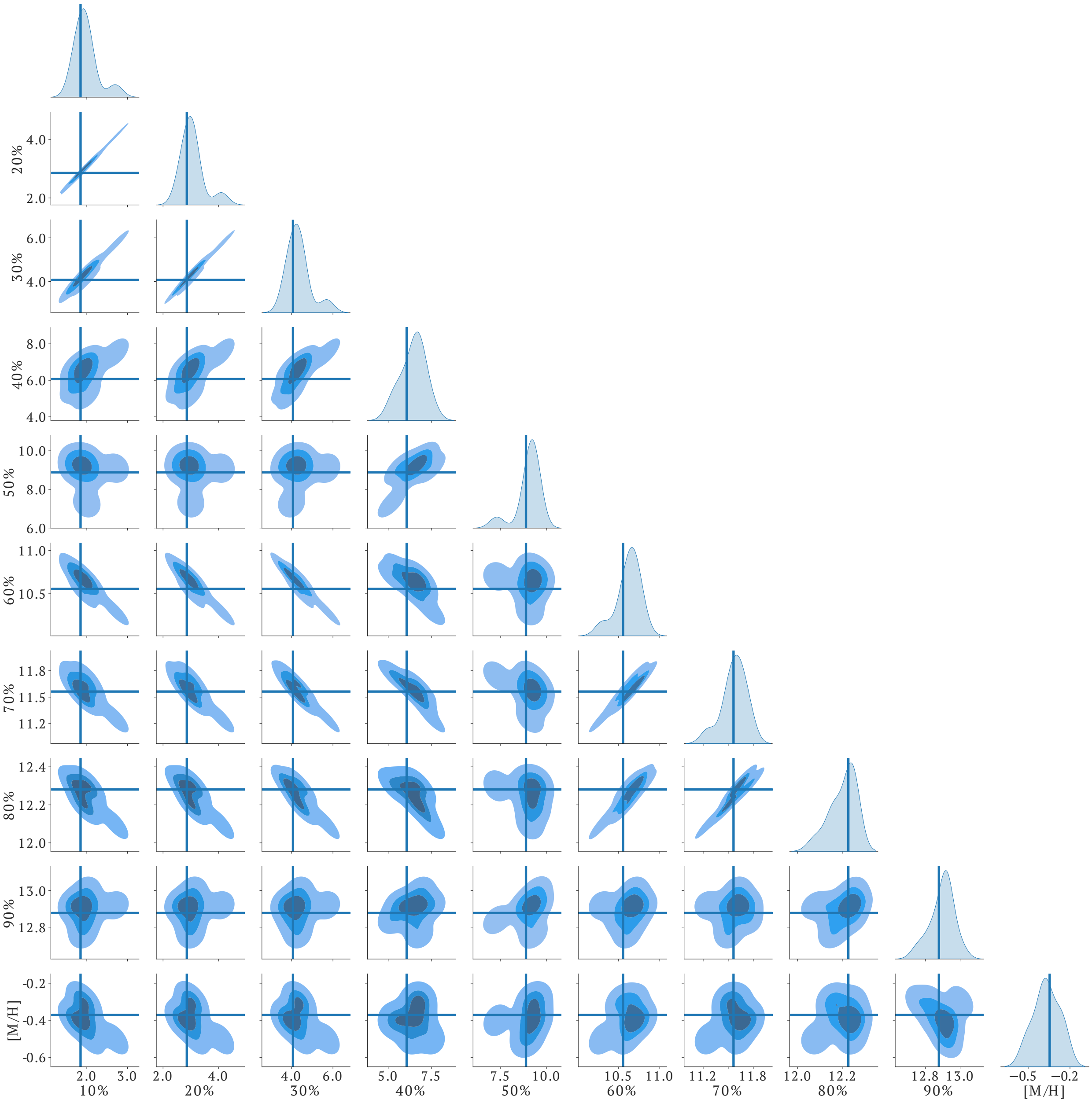

**Figure 14:** Corner plot for the cosmic time in Gyr of the 9 stellar mass percentiles and the metallicity value, predicted for a synthetic test galaxy, that includes the posterior distribution of every single feature in the right margins, together with its behavior in relation with the others in the remaining panels. The contours correspond to consecutive confidence intervals, and the solid lines to the true values. We observe the distributions are mainly centered on the real values, demonstrating good performance, and also that there are some correlations between consecutive percentiles, proving the model can capture that the growth of stellar mass often occurs smoothly and continuously, despite unknowing the nature and direct relationship of the percentiles.

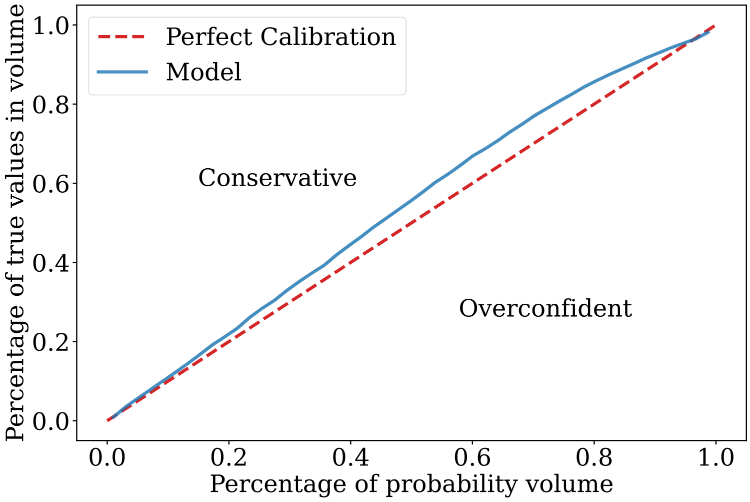

**Figure 15:** Coverage probability test including the totality of the predictions (9 stellar mass percentiles and metallicity) for the test set. If the uncertainties are properly calibrated, the nominal coverage probability (percentage of probability volume), on the $x$-axis, would be equal to the coverage probability (percentage of actual values in such volume), on the $y$-axis, assuming a Gaussian distribution. If they are not, the actual coverage probability could either be less than or greater than the nominal coverage probability. When the actual coverage probability is greater than the nominal coverage probability, the model is called 'conservative', if it is less than the nominal coverage probability, is called 'overconfident'. The solid blue line corresponds to our final model, while the dashed red line shows the one-to-one relation, the perfect calibration. Both are in good agreement, and the slight differences towards the conservative half allow the use of the model uncertainties as an upper limit.

## 4.3 Testing with observations

We repeat the training with synthetic spectra and non-parametric SFHs to generate a final model able to measure the stellar mass percentiles and metallicity for 18 SDSS stacks of early-type galaxies at $z \sim 0$ from their spectra (with velocity dispersion in the range $100 - 320$ km/s). For a more detailed description we refer the reader to La Barbera et al. (2013). We highlight that the main advantage of working with these stacks is their very high S/N, and the absence of emission lines.

First, all the observations are convolved to emulate the maximum velocity dispersion of the dataset, with a kernel of width $\sigma_i^2 = \sigma_{320}^2 + \sigma_{\rm SDSS}^2 - \left(\sigma_{{\rm v}_i}^2 + \sigma_{\rm SDSS}^2\right) = \sigma_{320}^2 - \sigma_{{\rm v}_i}^2$. Then, the wavelength range is clipped to $[4023, 6000]$ Å. On the other hand, the MILES spectra used for training are processed in order to simulate the conditions of the observations, convolving with a kernel of width $\sigma^2 = \sigma_{320}^2 + \sigma_{\rm SDSS}^2 - \sigma_{\rm MILES}^2$, then interpolating to get $\Delta\lambda = 1$ Å as in the stacks, and eventually clipping to $[4023, 5500]$ Å. Once all the artificial spectra have the same resolution and wavelength range as the observations, the training is repeated, and then the model performance is tested again, analyzing the impact of these changes on the performance.

The accuracy on the synthetic test set, as a result of the processing, decreases to 65.94%, 91.04% and 98.43% for the estimation of the percentiles 10%, 50% and 90%, respectively, and to 66.97% for the metallicity. These losses in performance, mainly for the first percentiles and metallicity, are a direct consequence of reducing resolution in the spectral lines, and clipping the Balmer jump in the bluer region of the spectra.

Then, we apply the trained model to the stacks' spectra, sampling the posteriors with 10.000 evaluations. In figure [16], we show the medians of the measured distributions for the mass percentiles, as a function of cosmic time and redshift[7], with a color map based on the velocity dispersions. The trend evidences that the most massive galaxies (highest velocity dispersions) build up their stellar masses more abruptly, up to 90% of their total stellar mass 1 Gyr after the Big Bang, while the growth of stellar mass is softened as we move to less massive ones.

In figure [17], we include the measurements of the star formation histories for the stacks with velocity dispersion of 105, 205, and 300 km/s. It is shown the medians of the distribution predicted for the stellar mass percentiles, and error bars correspond to the $2\sigma$ confidence interval. We observe a higher uncertainty, compared to the results on the synthetic test sample. This is expected due to the limitations of the simulation, such as the bias introduced by the inherent prior in training SFHs, the absence of chemical evolution, the observational gaps of the MILES library, or the effects of the stack combination. We remark that the uncertainties given by the model are as accurate as the observations are close to the synthetic data used for training. As in any Bayesian inference method, the posterior cannot be calibrated for observations that are outside of the prior introduced during the forward model. In the same figure, we plot the predictions for the stellar mass percentiles obtained with Penalized Pixel-fitting software (pPXF) (Cappellari, 2022), performing full-spectrum fitting through a linear combination of MILES single stellar population models and obtaining luminosity-weighted ages and metallicities. In general, our measurements and pPXF ones are closer for medium and high velocity dispersions, revealing a noticeable discrepancy for the less massive galaxy stacks. We also point out that the predictions of our model are smoother: it is a technical issue as we interpolate the MILES spectra uniformly every 0.013 Gyr, whereas pPXF sampling is 0.5 Gyr for the first 10 Gyr cosmic time. The metallicities measured are also shown in the figure, in agreement with pPXF results inside the high estimated uncertainties, and with a closer averaged value for intermediate and high velocity dispersions.

For a more detailed analysis, we include in table [2] the medians of the posterior predictions, together with their uncertainties given by one standard deviation, for all the 18 observations. The quantities shown are the

[7] Assuming a Planck13 cosmology (Ade et al., 2014).

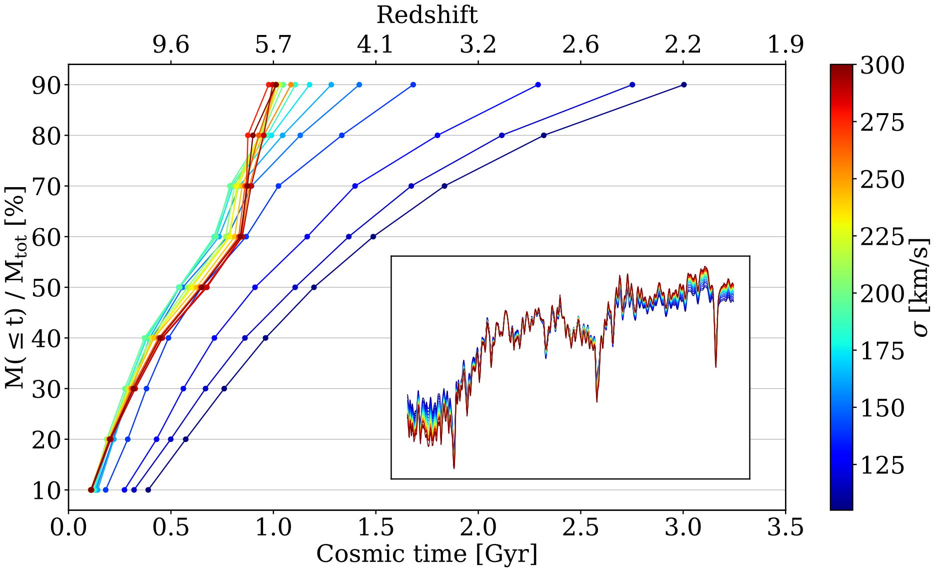

**Figure 16:** Medians of the posterior distributions predicted for the stellar mass percentiles of 18 stacks, as a function of the cosmic time in Gyr and redshift, colored according to their velocity dispersions in km/s. As shown in the color map, redder colors are assigned to higher velocity dispersions (more massive galaxies), while bluer ones point to lower velocity dispersions (less massive ones). In the inner frame, we plot the observed spectra, normalized and in the wavelength range [4023, 6000] Å, with the same color map.

time at which 10%, 50% and 90% of the total stellar mass are formed, as well as the metallicity. It should be noted that we do not find a clear trend between the metallicity and the velocity dispersion, however, the margin of uncertainty is too large to make any further assumptions, as already introduced by the poor performance determining $[M/H]$ on the synthetic test when processed with the observational features.

Finally, we carry out a spectral reconstruction with our model in figure [18], again for the stacks with velocity dispersion 105, 205, and 300 km/s. This allows us to verify that our measurements are indeed compatible with the observed spectra, something that a priori is not trivial, since in our approach, once the synthetic sample of galaxies is created, we lose the information of the forward model, and the predictions are not optimized to reproduce the observed spectra like in classical spectral fitting. Even so, it is the most direct way we have of determining the feasibility of the measurements, and we do it by performing a linear combination of 9 SSP MILES spectra of ages obtained for the percentiles of stellar mass, and metallicity fixed and equal to the predicted. By definition, all the 9 spectra are weighted with 1/9. The results are very positive, showing a 13% of maximum discrepancy in normalized flux with the observed spectra, for the bluer region of the stack of 105 km/s, and mean residuals of 2 % averaging for all the stacks and wavelengths. We also observe how the uncertainties are propagated into the spectra, within the gray-shadowed stripe corresponding to the interval of two standard deviations. Moreover, in the same figure, we include another reconstruction of the spectra as a linear combination of MILES templates of different 34 ages and 12 metallicities, where the weights are given by pPXF. Again we observe relatively close spectra, with higher residuals for intermediate velocity dispersions. To sum up, our model achieves lower residuals for 13 of the 18 stacks, performing better for intermediate and high velocity dispersions. We believe this is a consequence of the binning in ages and metallicity, among other technical aspects. In particular, the fitting in pPXF has been done in luminosity-weighted spectra, and it is performed without the continuum through a multiplicative polynomial of degree 15 in the range [4000, 6600] Å.

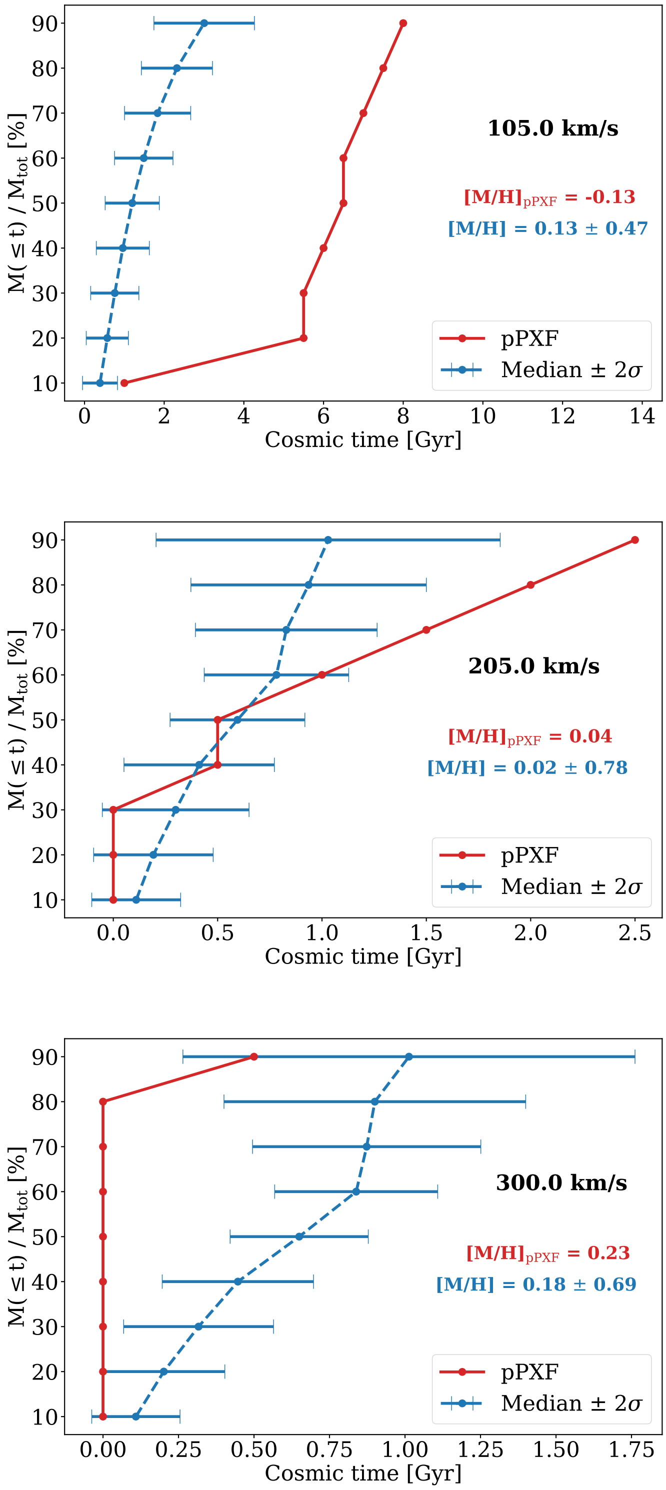

**Figure 17:** Predictions of stellar mass percentiles for observations as a function of the cosmic time in Gyr. We include the medians of the posteriors measured, in blue, for the stacks with velocity dispersion of 105 km/s, 205 km/s, and 300 km/s, from top to bottom, with an error bar corresponding to the 2$\sigma$ confidence interval. The percentile predictions obtained with pPXF are shown in red. It is also included the values for the average metallicity given by pPXF, and the predicted by our method again with the 2$\sigma$ interval of confidence, in red and blue respectively.

| $\sigma$ [km/s] | P10% [Gyr] | P50% [Gyr] | P90% [Gyr] | [M/H] |
|---|---|---|---|---|
| 100-110 | 0.42±0.22 | 1.21±0.34 | 2.89±0.63 | 0.11±0.24 |
| 110-120 | 0.36±0.22 | 1.12±0.32 | 2.63±0.66 | 0.13±0.23 |
| 120-130 | 0.32±0.21 | 0.95±0.33 | 2.21±0.75 | -0.05±0.32 |
| 130-140 | 0.23±0.17 | 0.72±0.29 | 1.66±0.70 | -0.12±0.40 |
| 140-150 | 0.16±0.13 | 0.63±0.24 | 1.43±0.57 | 0.06±0.38 |
| 150-160 | 0.17±0.13 | 0.61±0.25 | 1.31±0.61 | -0.05±0.39 |
| 160-170 | 0.15±0.12 | 0.59±0.21 | 1.18±0.54 | 0.01±0.40 |
| 170-180 | 0.15±0.11 | 0.60±0.21 | 1.11±0.53 | -0.05±0.40 |
| 180-190 | 0.13±0.10 | 0.59±0.20 | 1.08±0.50 | -0.07±0.41 |
| 190-200 | 0.13±0.09 | 0.62±0.18 | 1.04±0.46 | 0.01±0.37 |
| 200-210 | 0.13±0.11 | 0.63±0.16 | 1.02±0.41 | -0.03±0.39 |
| 210-220 | 0.13±0.10 | 0.64±0.17 | 1.02±0.44 | -0.04±0.37 |
| 220-230 | 0.12±0.09 | 0.65±0.15 | 1.02±0.42 | -0.07±0.37 |
| 230-240 | 0.12±0.12 | 0.67±0.14 | 1.06±0.39 | 0.07±0.37 |
| 240-250 | 0.12±0.07 | 0.66±0.12 | 1.02±0.39 | 0.03±0.38 |
| 250-260 | 0.12±0.07 | 0.69±0.11 | 0.96±0.36 | 0.05±0.39 |
| 260-280 | 0.12±0.06 | 0.69±0.12 | 0.99±0.41 | -0.15±0.38 |
| 280-320 | 0.12±0.07 | 0.66±0.11 | 1.02±0.37 | 0.11±0.34 |

**Table 2:** Predictions of the model for 18 stacks with mean velocity dispersions in the range $105-300$ km/s. In particular, we include the time in Gyr at which 10%, 50%, and 90% of the total stellar mass are formed, as well as the metallicity $[M/H]$. The values shown correspond to the medians of the predicted posterior distributions, and the uncertainties to the standard deviations.

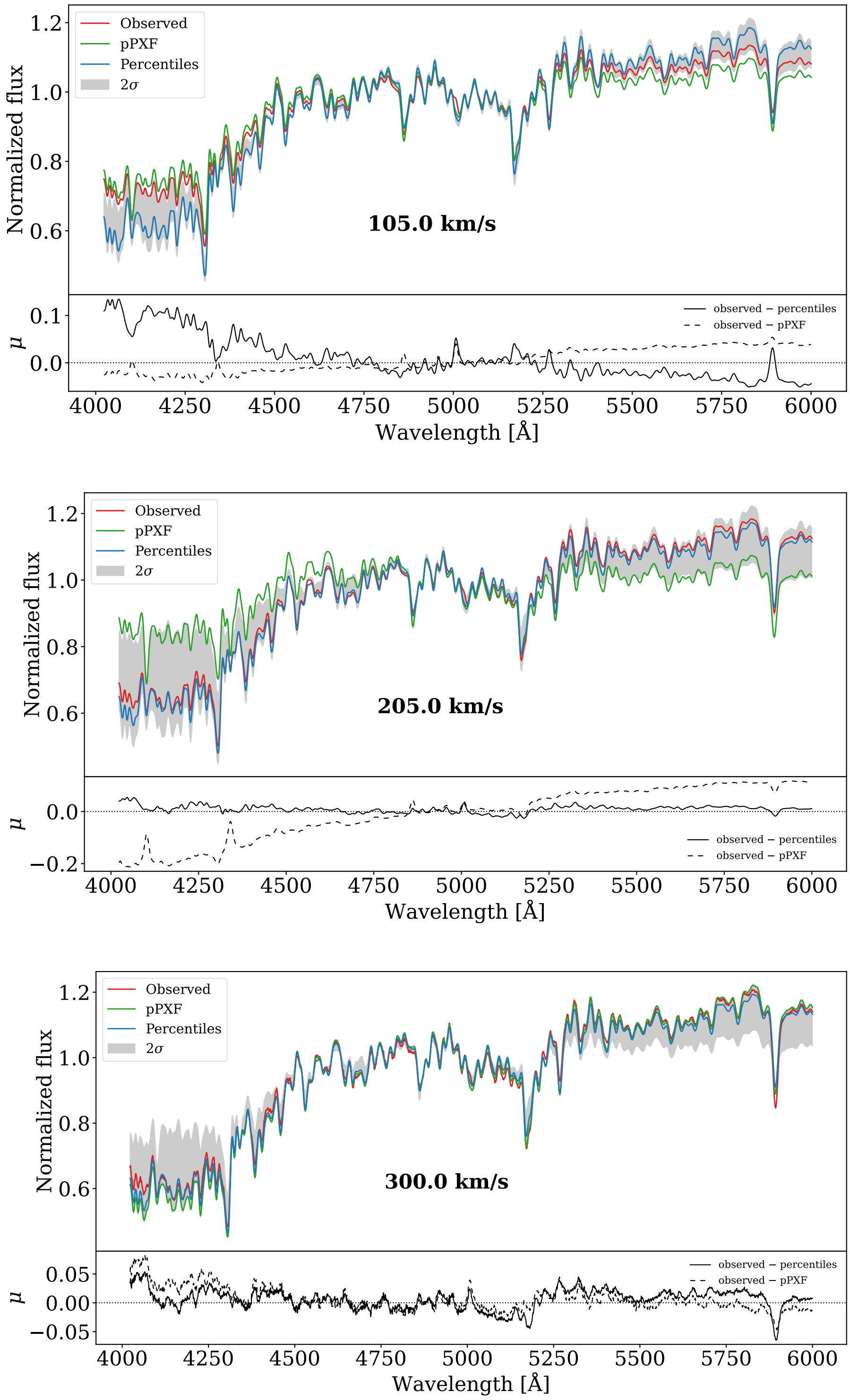

**Figure 18:** Reconstructions of the observed spectra for the stacks with velocity dispersion 105 km/s, 205 km/s and 300 km/s, from top to bottom. In red we show the normalized observed spectra, in the wavelength range $[4023, 6000]$ Å and downgraded to the lowest resolution (320 km/s). In blue, we show the spectra obtained as a linear combination of MILES SSP spectra according to our model, using the medians of the posteriors predicted. We combine MILES interpolated spectra with the metallicity predicted, and the ages coming from the 9 times predicted for the stellar mass percentiles. We repeat the procedure for metallicity and ages coming from the median minus two standard deviations, and from the median plus two standard deviations, to get the gray-shadowed stripe, where we can observe how the uncertainties in the quantities measured are manifested in the spectra. In green, it is plotted the spectra obtained as a linear combination of 408 MILES SSP spectra, with 34 different ages and 12 metallicities, where the weights are given by pPXF. In the lower panels, we show the residuals between the observed spectra and the one reconstructed from the percentiles and metallicity predicted by our model, with a solid black line, and between the observed spectra and the one recovered by pPXF, with a dashed black line. We also include a dotted line pointing to the zero level.

# 5 Discussion

We have developed a new approach to estimating star formation histories of galaxies from their optical absorption spectra, using simulation-based inference to obtain posteriors in a fully probabilistic treatment. From interpretable latent representations of the spectra, we predict the stellar mass growth and metallicity, reaching high accuracy for the synthetic sample, as well as properly calibrated uncertainties, evaluated through a probability coverage test. Furthermore, the predictions made require less than half a second for each galaxy, sampling the posteriors with 1.000 evaluations. This not only makes it possible to address a large dataset, but also to increase the number of evaluations for the posteriors, raising the precision of the measurements and their uncertainties.

Likewise, we obtain measurements for stacks of ETGs, consistent with their spectra in the uncertainty range. The model recovers the very known relationship between the age and the velocity dispersion, showing that the most massive galaxies ($\sigma \sim 300$ km/s) build up their stellar masses more abruptly, up to 90% of their total stellar mass 1 Gyr after the Big Bang, while the growth of stellar mass is softened as we move to less massive ones. How these massive galaxies, up to $M_* \sim 10^{12}\,\mathrm{M}_\odot$ for the highest velocity dispersion stack, can form up most of their masses so early is a question that is still open, given that a priori it requires a very high star formation efficiency. We remark that these are rare galaxies in the known Universe, with rapid bursts of star formation and quenching soon. Thus, the stars we see may have formed in different progenitor galaxies, which were located in the highest density peaks of the Cosmic Web, and later assemble in major mergers (Conselice, 2008). To attend to these processes, hierarchical assembly models are necessary, commonly studied through large-scale cosmological simulations (Angeloudi et al., 2023).

Although the model is substantially faster than other Bayesian inference methods, such as MCMC, and provides well-calibrated uncertainties, unlike classical inversion methods applied to spectral fitting, it is subject to different systematics and modeling assumptions, which we will discuss in detail here, as well as possible extensions in future work.

One of the obvious limitations is using the forward model to train the network, because as our understanding of galaxy evolution is incomplete, the considerations taken when generating the synthetic data will never be fully constrained by a typical galaxy spectrum. Therefore, the prior, set by the distribution of parameters in the training set, will always have at least a moderate role in determining the answer. In particular, the simplifications taken in chemical evolution by combining MILES SSP spectra with a fixed metallicity, instead of taking into account the metallicity histories (ZHs), together with working base $\alpha$-enhancement models, can affect the reliability of our model predictions, being aware of the intrinsic degeneration between age and metal content in the spectra. Even if considering chemical evolution, there is currently no consensus in the stellar evolution (Conroy et al., 2009) or IMF of galaxies (Martín-Navarro et al., 2015), and our model training has been restricted to BaSTI isochrones and a Kroupa Universal IMF, limiting the range of the training data, and doomed to fail in observations of galaxies that differ from these considerations.

With regards to the non-parametric SFHs used, it is known that they have a strong impact on the posteriors predicted, as shown by Leja et al. (2019), and models that mimic the breadth distribution of SFR($t$) in the real Universe are required. One key question, which determines the behavior of SFHs, is on what timescales the SFR changes in true galaxies: either this timescale is primarily set by the halo dynamical time (i.e., the gas accretion rate), or it is set by the timescales of star formation feedback (Torrey et al., 2018). This dichotomy is taken into account by two different kinds of priors in the non-parametric models: concentrated or dispersive, respectively. In this work, in the pipeline provided by the GP-SFH module, we have selected a dispersive prior: a Dirichlet distribution with $\alpha = 1.0$ for the fractional specific star formation rate in each time bin (see Leja et al. (2019) for more details), positioning ourselves in favor of short-term variations in the SFHs (with smaller characteristic times). However, this selection must be explored with caution in future works, remarking the existence of observational signatures that a priori can discriminate between the two scenarios, such as the ratio $H_\alpha$/UV (Smit et al., 2016).

It is clear that constraining the theoretical models of stellar physics, mass functions, halo dynamical times and feedback processes is beyond the scope of our work. However, our model can help with its predictions, from the reconstructions of the spectra, to study the systematic effects that they produce. Thus, the dependence on the forward model, if used consciously, becomes a profitable feature of our approach, and not a bug, that we plan to explore as a continuation of the project.

It is also possible to use the model to measure SFHs and metallicity for Late Type Galaxies (LTGs), for which we expect more troubles due to a larger number of emission lines and 'over-shinning' effects when recovering the metallicity and ages of their first stars to form, decreasing in performance. However, the observations analyzed in this work are also quite a challenge for the model, not only rare in the training data because of their high masses, but also intrinsically difficult to infer from the spectra, since the oldest stars have the most subtle observational signatures and therefore, the SFHs predicted are more prior-dominated, so we believe it is worth it to try it. However, so far we have not incorporated noise models, dust attenuation or emission lines. Including these aspects in the future is essential to be able to manage a wide range of observations, moving from stacks to spectra of single galaxies, using the Bayesian framework to include observational uncertainties as conditioning variables. Moreover, it is possible to complement the inference based on stellar populations with information extracted from the gaseous component, by measuring the strength of emission lines produced by the galaxy (Robert C. Kennicutt, 1998), and the metal content of ionized gas, which can be traced by emission line ratios (Sánchez et al., 2012), and offers a snapshot of the chemical composition of the material from which stars are currently forming, possibly constraining our measurements.

Finally, we point out some modifications and improvements for the backward model. First of all, a tuning of the model hyperparameters could be performed, among which we highlight the number of components of the latent representations in which we encode the spectrum, as well as the number of blocks used in the Masked Autoregressive Flow. On the other hand, further tests can be performed on the uncertainties predicted in the synthetic data, continuing with the methodology started by the coverage probability test as indicated in Talts et al. (2018).

Ultimately, several directions can be followed. We believe that our approach can contribute, given its speed and error handling, to study the biases that produce features of the forward model that have traditionally been assumed in similar analyses. Following the first steps taken in this project for the stacks of ETGs from the SDSS, the model could be used to perform a deep sampling of current or upcoming spectroscopic surveys such as DESI (Hahn et al., 2023), WEAVE (Dalton et al., 2012), 4MOST (Jong et al., 2014), or MOONS (Cirasuolo et al., 2020), which will observe billions of galaxies and more than $10^8$ TB of data (Smith and Geach, 2023). We emphasize the necessity of state-of-the-art models capable of dealing with large volumes of data, whose size is doubled every 16 months (Zhang and Zhao, 2015), and whose complexity increases as instrumental efforts are made to observe a universe in more detail and farther and farther away from our own galaxy.

# 6 Conclusion

By analyzing the spectrum of a galaxy we can infer physical properties such as its stellar mass, star formation rates and chemical abundances, key ingredients of our understanding of galaxy evolution. State-of-the-art spectral fitting methods use MCMC sampling to perform Bayesian statistical inference, deriving posterior probability distributions of galaxy properties given observations. However, obtaining a posterior sampled with $\sim$ 10.000 evaluations with these methods requires $\gtrsim 1 - 10$ GPU hours per galaxy, being a major bottleneck for addressing large galaxy surveys.

We demonstrate in this work that an amortized simulation-based inference, with a previous encoding of the spectra, provides an alternative approach for spectral fitting through a neural density estimator. By using MILES templates and non-parametric SFHs, we construct a flexible model able to recover SFHs and metallicities with their uncertainties for observed stacks of spectra from the SDSS, taking $\sim$ 4 s for galaxy to perform 10.000 evaluations of the posteriors for the 10 quantities we predict. To finish we present the key results from our project.

- We construct a model composed of an encoder for the spectra based on a dot-product attention model, plus a Masked Autoregressive Flow, to estimate the posterior distributions for the cosmic time at which 9 stellar mass percentiles are reached, and a fixed metallicity. We train it during $\sim$ 4 hours with 135.000 synthetic samples obtained from SSP MILES templates and non-parametric SFHs from the GP-SFH module.

- We test the model with 15.000 synthetic samples, reaching 79.35%, 93.54%, 98.95%, and 75.83% of accuracy when estimating the percentiles 10%, 50%, 90%, and the metallicity, respectively. We find the model learns correlations between consecutive percentiles, and working with synthetic data is able to recover uncertainties associated with the degeneration of the inversion problem, having more difficulties in assessing the first percentiles and metallicity in galaxies with recent star formation, but reflecting it adequately in the errors as indicated by the coverage probability test performed.

- We estimate with our model the SFHs and metallicities of 18 stacks of ETGs from the SDSS, in the range of $105 - 300$ km/s velocity dispersions. We obtain very early bursts of star formation in the most massive galaxies, and a smoother growth of stellar mass when moving to intermediate masses, recovering a well-known relation between age and velocity dispersion. Moreover, we accurately reconstruct the spectra from MILES templates, using the 9 ages and the mean metallicity measured, in good agreement with the real spectra with an averaged error of 2%.

# Bibliography

Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., et al. (2014). Planck 2013 results. XVI. Cosmological parameters, *Astronomy and Astrophysics*.

Angeloudi, E., Falcón-Barroso, J., Huertas-Company, M., et al. (2023). ERGO-ML: Towards a robust machine learning model for inferring the fraction of accreted stars in galaxies from integral-field spectroscopic maps, *Monthly Notices of the Royal Astronomical Society*.

Baldwin, J. A., Phillips, M. M., and Terlevich, R. (1981). Classification parameters for the emission-line spectra of extragalactic objects, *Publications of the Astronomical Society of the Pacific*.

Brown, M. J. I., Moustakas, J., Smith, J.-D. T., et al. (2014). An atlas of galaxy spectral energy distributions from the ultraviolet to the mid-infrared, *The Astrophysical Journal Supplement Series*.

Calzetti, D., Kinney, A. L., and Storchi-Bergmann, T. (1994). Dust extinction of the stellar continua in starburst galaxies: The ultraviolet and optical extinction law, *The Astrophysical Journal*.

Cameron, E. and Pettitt, A. N. (2012). Approximate Bayesian Computation for astronomical model analysis: A case study in galaxy demographics and morphological transformation at high redshift, *Monthly Notices of the Royal Astronomical Society*.

Cappellari, M. (2022). Full spectrum fitting with photometry in pPXF: Non-parametric star formation history, metallicity and the quenching boundary from 3200 LEGA-C galaxies at redshift $z \approx 0.8$, *Monthly Notices of the Royal Astronomical Society submitted*.

Carnall, A. C., Leja, J., Johnson, B. D., et al. (2019). How to measure galaxy star formation histories. I. Parametric models, *The Astrophysical Journal*.

Cenarro, A. J., Cardiel, N., Gorgas, J., et al. (2001). Empirical calibration of the near-infrared Ca II triplet - I. The stellar library and index definition, *Monthly Notices of the Royal Astronomical Society*.

Chabrier, G. (2003). Galactic stellar and substellar initial mass function, *Publications of the Astronomical Society of the Pacific*.

Choi, J., Dotter, A., Conroy, C., et al. (2016). Mesa Isochrones and Stellar Tracks (MIST). I. Solar-scaled models, *The Astrophysical Journal*.

Cirasuolo, M., Fairley, A., Rees, P., et al. (2020). MOONS: The new multi-object spectrograph for the VLT, *The Messenger*.

Conroy, C. (2013). Modeling the panchromatic spectral energy distributions of galaxies, *Annual Review of Astronomy and Astrophysics*.

Conroy, C., Gunn, J. E., and White, M. (2009). The propagation of uncertainties in stellar population synthesis modeling. I. The relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties, *The Astrophysical Journal*.

Conselice, C. J. (2008). The assembly history of massive galaxies: What do we know?, *Astronomical Society of the Pacific Conference Series*.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of Simulation-Based Inference, *Proceedings of the National Academy of Sciences*.

Dalton, G., Trager, S. C., Abrams, D. C., et al. (2012). WEAVE: The next generation wide-field spectroscopy facility for the William Herschel Telescope, *Ground-based and Airborne Instrumentation for Astronomy IV*.

Fernandes, R. C., Schoenell, W., Gomes, J. M., et al. (2008). The star formation histories of galaxies: A tour through the STARLIGHT-SDSS database, *Revista Mexicana de Astronomia y Astrofisica*.

Girardi, L., Bressan, A., Bertelli, G., and Chiosi, C. (2000). Evolutionary tracks and isochrones for low and intermediate mass stars, *Astronomy and Astrophysics Supplement Series*.

Gladders, M. D., Oemler, A., Dressler, A., et al. (2013). The IMACS cluster building survey. IV. The log-normal star formation history of galaxies, *The Astrophysical Journal*.

Gonneau, A., Lyubenova, M., Lancon, A., et al. (2020). The X-shooter Spectral Library (XSL): Data release 2, *Astronomy and Astrophysics*.

Hahn, C., Kwon, K. J., Tojeiro, R., et al. (2023). The DESI PRObabilistic Value-added Bright Galaxy Survey (PROVABGS) mock challenge, *The Astrophysical Journal*.

Hahn, C. and Melchior, P. (2022). Accelerated Bayesian SED modeling using Amortized Neural Posterior Estimation, *The Astrophysical Journal*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, *arXiv e-prints*.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks, *Science*.

Huertas-Company, M., Sarmiento, R., and Knapen, J. (2023). A brief review of contrastive learning applied to astrophysics, *arXiv e-prints*.

Iyer, K. and Gawiser, E. (2017). Reconstruction of galaxy star formation histories through SED fitting: The Dense Basis approach, *The Astrophysical Journal*.

Johnson, B. D., Leja, J., Conroy, C., and Speagle, J. S. (2021). Stellar population inference with Prospector, *The Astrophysical Journal Supplement Series*.

Jong, R. S., Barden, S., Bellido-Tirado, O., et al. (2014). 4MOST: 4-metre Multi-Object Spectroscopic Telescope, *Ground-based and Airborne Instrumentation for Astronomy V*.

Kewley, L. J., Nicholls, D. C., and Sutherland, R. S. (2019). Understanding galaxy evolution through emission lines, *Annual Review of Astronomy and Astrophysics*.

Kroupa, P. (2001). On the variation of the initial mass function, *Monthly Notices of the Royal Astronomical Society*.

La Barbera, F., Ferreras, I., Vazdekis, A., et al. (2013). SPIDER VIII – constraints on the stellar initial mass function of early-type galaxies from a variety of spectral features, *Monthly Notices of the Royal Astronomical Society*.

Lee, S.-K., Ferguson, H. C., Somerville, R. S., et al. (2010). The estimation of star formation rates and stellar population ages of high-redshift galaxies from broadband photometry, *The Astrophysical Journal*.

Leja, J., Carnall, A. C., Johnson, B. D., et al. (2019). How to measure galaxy star formation histories. II. Nonparametric models, *The Astrophysical Journal*.

Leja, J., Johnson, B. D., Conroy, C., et al. (2017). Deriving physical properties from broadband photometry with Prospector: Description of the model and a demonstration of its accuracy using 129 galaxies in the local universe, *The Astrophysical Journal*.

Lilly, S. J., Carollo, C. M., Pipino, A., et al. (2013). Gas regulation of galaxies: The evolution of the cosmic specific star formation rate, the metallicity-mass-star-formation rate relation, and the stellar content of halos, *The Astronomical Journal*.

Lovell, C. C., Acquaviva, V., Thomas, P. A., et al. (2019). Learning the relationship between galaxies spectra and their star formation histories using convolutional neural networks and cosmological simulations, *Monthly Notices of the Royal Astronomical Society*.

Maraston, C. and Strömbäck, G. (2011). Stellar population models at high spectral resolution, *Monthly Notices of the Royal Astronomical Society*.

Martín-Navarro, I., Burchett, J. N., and Mezcua, M. (2020). Black hole feedback and the evolution of massive early-type galaxies, *Monthly Notices of the Royal Astronomical Society*.

Martín-Navarro, I., La Barbera, F., Vazdekis, A., et al. (2015). Radial variations in the stellar initial mass function of early-type galaxies, *Monthly Notices of the Royal Astronomical Society*.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction, *arXiv e-prints*.

Melchior, P., Liang, Y., Hahn, C., and Goulding, A. (2022). Autoencoding galaxy spectra I: Architecture, *arXiv e-prints*.

Mishra-Sharma, S. (2022). Inferring dark matter substructure with astrometric lensing beyond the power spectrum, *Machine Learning: Science and Technology*.

Ocvirk, P., Pichon, C., Lançon, A., and Thiébaut, E. (2006). STECKMAP: STEllar Content and Kinematics from high resolution galactic spectra via Maximum A Posteriori, *Monthly Notices of the Royal Astronomical Society*.

Pacifici, C., Kassin, S. A., Weiner, B., et al. (2012). The rise and fall of the star formation histories of blue galaxies at redshifts $0.2 < z < 1.4$, *The Astrophysical Journal*.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked Autoregressive Flow for density estimation, *arXiv e-prints*.

Pietrinferni, A., Cassisi, S., Salaris, M., and Castelli, F. (2006). A large stellar evolution database for population synthesis studies. II. Stellar models and isochrones for an $\alpha$-enhanced metal distribution, *The Astrophysical Journal*.

Portillo, S. K. N., Parejko, J. K., Vergara, J. R., and Connolly, A. J. (2020). Dimensionality reduction of SDSS spectra with variational autoencoders, *The Astronomical Journal*.

Rezende, D. and Mohamed, S. (2015). Variational inference with Normalizing Flows, *Proceedings of Machine Learning Research*.

Riebesell, J. (2022). Random TikZ Collection.

Robert C. Kennicutt, J. (1998). The global Schmidt law in star-forming galaxies, *The Astrophysical Journal*.

Salpeter, E. E. (1955). The luminosity function and stellar evolution, *The Astrophysical Journal*.

Serrà, J., Pascual, S., and Karatzoglou, A. (2018). Towards a universal neural network encoder for time series, *arXiv e-prints*.

Simha, V., Weinberg, D. H., Conroy, C., et al. (2014). Parametrising star formation histories, *arXiv e-prints*.

Sisson, S. A., Fan, Y., and Beaumont, M. A. (2018). Overview of Approximate Bayesian Computation, *arXiv e-prints*.

Smit, R., Bouwens, R. J., Labbé , I., et al. (2016). Inferred $H_\alpha$ flux as a star formation rate indicator at $z \sim 4 - 5$: Implications for dust properties, burstiness, and the $z = 4 - 8$ star formation rate functions, *The Astrophysical Journal*.

Smith, M. J. and Geach, J. E. (2023). Astronomia ex machina: a history, primer and outlook on neural networks in astronomy, *Royal Society Open Science*.

Sánchez, S. F., Rosales-Ortega, F. F., Marino, R. A., et al. (2012). Integral field spectroscopy of a sample of nearby galaxies, *Astronomy and Astrophysics*.

Talts, S., Betancourt, M., Simpson, D., et al. (2018). Validating Bayesian inference algorithms with Simulation-Based Calibration, *arXiv e-prints*.

Teimoorinia, H., Archinuk, F., Woo, J., et al. (2022). Mapping the diversity of galaxy spectra with deep unsupervised machine learning, *The Astronomical Journal*.

Tejero-Cantero, A., Boelts, J., Deistler, M., et al. (2020). SBI - a toolkit for Simulation-Based Inference, *The Journal of Open Source Software*.

Tinsley, B. M. (1980). Evolution of stars and gas in galaxies, *Fundamental Cosmic Physics*.

Tojeiro, R., Heavens, A. F., Jimenez, R., and Panter, B. (2007). Recovering galaxy star formation and metallicity histories from spectra using VESPA, *Monthly Notices of the Royal Astronomical Society*.

Torrey, P., Vogelsberger, M., Hernquist, L., et al. (2018). Similar star formation rate and metallicity variability time-scales drive the fundamental metallicity relation, *Monthly Notices of the Royal Astronomical Society: Letters*.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need, *arXiv e-prints*.

Vazdekis, A., Sánchez-Blázquez, P., Falcón-Barroso, J., et al. (2010). Evolutionary stellar population synthesis with MILES - I. The base models and a new line index system, *Monthly Notices of the Royal Astronomical Society*.

Walcher, J., Groves, B., Budavári, T., and Dale, D. (2010). Fitting the integrated spectral energy distributions of galaxies, *Astrophysics and Space Science*.

Yip, C. W., Connolly, A. J., Szalay, A. S., et al. (2004). Distributions of galaxy spectral types in the Sloan Digital Sky Survey, *The Astronomical Journal*.

Zhang, Y. and Zhao, Y. (2015). Astronomy in the Big Data era, *Data Science Journal*.