Research in Astronomy and Astrophysics, 24:125019 (12pp), 2024 December https://doi.org/10.1088/1674-4527/ad95d7

Multi-layer Perceptron for Predicting Galaxy Parameters (MLP-GaP): Stellar Masses and Star Formation Rates

Xiaotong Guo (郭晓通)[1], Guanwen Fang[1], Haicheng Feng[2], and Rui Zhang[1]

[1] Institute of Astronomy and Astrophysics, Anqing Normal University, Anqing 246133, China; guoxiaotong@aqnu.edu.cn

[2] Yunnan Observatories, Chinese Academy of Sciences, Kunming 650011, China

Received 2024 October 01; revised 2024 October 27; accepted 2024 October 30; published 2024 December 11

Abstract

The large-scale imaging survey will produce massive photometric data in multi-bands for billions of galaxies. Defining strategies to quickly and efficiently extract useful physical information from this data is mandatory. Among the stellar population parameters for galaxies, their stellar masses and star formation rates (SFRs) are the most fundamental. We develop a novel tool, Multi-Layer Perceptron for Predicting Galaxy Parameters (MLP-GaP), that uses a machine learning (ML) algorithm to accurately and efficiently derive the stellar masses and SFRs from multi-band catalogs. We first adopt a mock data set generated by the Code Investigating GALaxy Emission (CIGALE) for training and testing data sets. Subsequently, we used a multi-layer perceptron model to build MLP-GaP and effectively trained it with the training data set. The results of the test performed on the mock data set show that MLP-GaP can accurately predict the reference values. Besides MLP-GaP has a significantly faster processing speed than CIGALE. To demonstrate the science-readiness of the MLP-GaP, we also apply it to a real data sample and compare the stellar masses and SFRs with CIGALE. Overall, the predicted values of MLP-GaP show a very good consistency with the estimated values derived from spectral energy distribution fitting. Therefore, the capability of MLP-GaP to rapidly and accurately predict stellar masses and SFRs makes it particularly well-suited for analyzing huge amounts of galaxies in the era of large sky surveys.

Key words: methods: data analysis – galaxies: fundamental parameters – galaxies: star formation

1. Introduction

Galaxies are fundamental building blocks of the cosmic large-scale structures and have played a crucial role in the evolution of baryons in the universe's history. We are entering the season of the Stage-IV all-sky surveys, which will observe billions of galaxies. Large sky surveys have become a fundamental tool for cosmology and galaxy formation studies over the past few decades. In particular, Stage-III surveys have provided multi-band data for tens to hundreds of millions of galaxies and other celestial objects, such as the Kilo-Degree Survey (KiDS; Wright et al. 2024), Hyper Suprime-Cam (HSC; Aihara et al. 2018), and the Dark Energy Survey (DES; Abbott et al. 2021). Over the past decade, these large-scale surveys have definitely pushed our understanding of the dark matter distribution in the universe via weak lensing (e.g., Laureijs et al. 2011; Hildebrandt et al. 2017; Abbott et al. 2018; Heymans et al. 2021; Joachimi et al. 2021; Mistele et al. 2023). However, they have also provided useful data to significantly grow our understanding of the formation and evolution of galaxies (e.g., Greco et al. 2018; Goulding et al. 2018; Roy et al. 2018; Adhikari et al. 2021; Xie et al. 2023). In this context, the ability to collect accurate stellar population properties of all examined galaxies, despite being crucial (e.g., Wright et al. 2019), has been a bottleneck in the science outcomes (e.g., Bilicki et al. 2021).

In the next decade, Stage-IV surveys will observe billions of galaxies, providing data that is both more in-depth and of higher quality and covering the wavelengths from ultraviolet (UV) to near-infrared (NIR). Observational programs for the Stage-IV surveys include Euclid (Laureijs et al. 2011), Vera Rubin Legacy Survey in Space and Time (VR/LSST; Ivezić et al. 2019), and the China Space Station Telescope (e.g., Zhan 2011; Zhan & Tyson 2018; Gong et al. 2019). The data for Stage-IV surveys will offer unprecedented insights into cosmology and galaxy evolution, including crucial revelations on dark matter, dark energy, and the formation and evolution of galaxies.

Galaxies are complex systems with numerous physical parameters, such as stellar mass ($M_\star$), size, morphology, star formation rate (SFR), age, metallicity and chemical composition. Accurately measuring these parameters is essential for understanding the formation and evolution of galaxies and their role in shaping the structure and evolution of the universe. However, obtaining unbiased stellar population parameters for galaxies remains a significant challenge due to the well-known age/metallicity degeneracies. Currently, there is no consensus on how to mitigate these degeneracies, and various attempts have been made to account for different phases of stellar evolution (e.g., Maraston 2005; Vazdekis et al. 2016) to solve this problem. On the other hand, combining different codes and stellar libraries, along with stellar population priors and star formation histories, can help reduce the impact of systematic errors (e.g., Xie et al. 2023). This approach, though, is time consuming and new strategies are needed to reduce the computational times for testing the largest variety of stellar population models and priors, especially over large data sets.

Among the different stellar population parameters for galaxies, $M_\star$ and SFR stand out as the most crucial ones. In particular, they have a rather tight relation, the so-called *main sequence* of galaxies (e.g., Brinchmann et al. 2004; Daddi et al. 2007; Elbaz et al. 2007; Noeske et al. 2007; Schreiber et al. 2015), that is a fundamental diagnostic for galaxy formation theories (e.g., Furlong et al. 2015; Donnari et al. 2019; Popesso et al. 2023). Nevertheless, estimating $M_\star$ and SFR is also a complex task (Kennicutt & Evans 2012) as they are directly or indirectly related to the observations of stars. $M_\star$ represents the total mass of stars within a galaxy. Low-mass, non-ionizing old stars are the most abundant within a galaxy and contribute significantly to its optical luminosities. Consequently, $M_\star$ is closely associated with its optical luminosity. Moreover, the estimation of $M_\star$ also depends on the stellar population model (e.g., Bruzual & Charlot 2003; Maraston 2005) and the Initial Mass Function (IMF; e.g., Salpeter 1955; Chabrier 2003), and the form of the star formation history (SFH), for which there is no consensus on the impact of systematics (e.g., Mitchell et al. 2013). The SFR represents the rate at which new stars are being formed within a galaxy. It is closely tied to the presence of young stars and the ionized gas that envelops massive stars. The emission from young and massive O/B-type stars is predominantly in the UV band. As a result, UV emission can serve as a valuable indicator of the SFR (e.g., Hao et al. 2011; Kennicutt & Evans 2012). Additionally, emission lines (such as Hα) from ionized gas can be observed in the optical and NIR bands, further tracing the SFR (e.g., Buat et al. 2002; Treyer et al. 2007). Dust also plays an important role, as this is heavily produced around new stars. This dust absorbs approximately half of the starlight and re-emits it in the far-IR, meaning that far-IR luminosities can also trace the SFR (e.g., Fixsen et al. 1998; Salim & Narayanan 2020).

Traditionally, the quantities $M_\star$ and SFR have been estimated primarily through techniques like optical spectroscopy, which involves fitting theoretical models to observed data. For instance, the Sloan Digital Sky Survey (SDSS) MPA–JHU Data Release 8 (DR8) catalog provides stellar masses and SFRs for 1 843 200 galaxies (Kauffmann et al. 2003; Brinchmann et al. 2004). However, many surveys lack spectroscopic observations and only provide photometry data, such as KiDS (e.g., Wright et al. 2024). Despite this, it is still feasible to derive the $M_\star$ and SFR by using their spectral energy distribution (SED) constructed from multi-band photometric data. For example, Gao et al. (2019) used SED fitting to obtain the SFRs and stellar masses of 145 635 galaxies in the Hawaii-Hubble Deep Field-North. When we look toward the future with the expected release of multi-band photometric data for billions of galaxies, as there is no way to solve the degeneracies among all the stellar population parameters, the only approach we have is to derive $M_\star$ and SFR using different set-up (e.g., Xie et al. 2023), make this dramatically time-consuming. Given the immense volume of data that will be available, developing a highly efficient method is crucial for extracting meaningful information from these data sets.

In recent years, the rapid development of machine learning (ML) algorithms has brought revolutionary changes to various fields, including astronomy. ML has become an integral part of astronomical research, being widely used for a variety of classifications, including astronomical object categorization (e.g., Zeraatgari et al. 2024), galaxy morphology classification (e.g., Fang et al. 2023; Xu et al. 2023; Song et al. 2024), and much more. Moreover, ML is also used to predict various parameters and properties of astronomical objects. Li et al. (2022b) has developed an innovative ML tool (called GaZNet) capable of predicting galaxy redshifts by integrating both image data and multi-band photometric information. Wu & Boada (2019) have trained a deep residual convolutional neural network to predict the gas-phase metallicity of galaxies using three-band *gri* images from the SDSS. Bonjean et al. (2019) have used the Random Forest to estimate the stellar masses and SFRs of galaxies at redshifts in the range $0.01 < z < 0.3$. When compared to traditional methods, ML not only offers enhanced efficiency but also delivers accuracy. Therefore, to efficiently and accurately derive stellar population parameters for billions of galaxies, we plan to develop an ML algorithm called Multi-Layer Perceptron for Predicting Galaxy Parameters (MLP-GaP), which first is used to estimate the stellar masses and SFRs for galaxies with redshift $z < 3$ in this work.

The structure of this work is as follows. Section 2 presents the data used in this work and the process of generating a mock catalog. Section 3 describes the ML model used to build the MLP-GaP and outlines its training process. In Section 4, we comprehensively evaluate the performance of the MLP-GaP in the testing data set, offering a comparative analysis between MLP-GaP and traditional SED fitting techniques. In Section 5, we discuss the performance of the MLP-GaP on observational data of actual galaxies and provide its possible future application scenarios, as well as subsequent improvement methods for the algorithm. Finally, a brief summary is presented in Section 6.

2. Data

In this work, we want to develop an ML algorithm that uses multi-band photometric data (i.e., aperture fluxes or magnitudes) to predict the stellar population parameters of a given galaxy data set, similar to the standard SED fitting techniques (e.g., Boquien et al. 2019; Ilbert et al. 2006). To train and test such a "supervised" algorithm (see Section 3), it is essential to dispose of a galaxy catalog with "ground truth" stellar population parameters. Unfortunately, unlike other galaxy parameters that are straightforward to measure, the stellar population parameters of galaxies can only be indirectly derived from measured parameters. stellar population parameters of galaxies cannot obtain "ground truth" values. However, the derivation of these parameters relies on various assumptions, models, and inherent uncertainties, which means that the parameters obtained from different techniques are not uniform (see Xie et al. 2023) and do not represent their "ground truth" values.

In absence of real data set with "ground truth" values, a viable strategy is to produce a realistic mock data set, which include synthetic photometric data of the realistic galaxies and their known stellar population parameters. To construct such a catalog of galaxies or data set, we use the Code Investigating GALaxy Emission (CIGALE, V2022.1, Boquien et al. 2019) to generate a catalog of mock galaxies with known stellar population parameters (see Section 2.2). CIGALE is an open-source *Python* code designed to analyze the SEDs of galaxies across a wide range of wavelengths, from X-ray to radio. CIGALE models galaxy SEDs by employing composite stellar populations from simple stellar populations combined with highly flexible star formation histories (SFHs), and this approach can provide the flux densities for various bands, $M_\star$, SFR, attenuation, dust luminosity, and many other physical quantities. In particular, $M_\star$ and SFR which can be considered as reference values for developing MLP-GaP. For this work, we adopt a delayed SFH model (SFR $\propto t/\tau^2 \times \exp{-t/\tau}$), which is currently popular and aligns well with observational data. To derive the stellar population spectrum, we use the stellar population synthesis model from Bruzual & Charlot (2003, referred to as BC03), assuming that the IMF adopts Chabrier (2003). In addition, we also adopt attenuation law (Calzetti et al. 2000), and dust emission (Dale et al. 2014). As mentioned, we also use actual observational data as the basis for generating the catalog to ensure that the redshift and luminosity distributions of the mock galaxy sample are well-aligned with observations and to mimic the observation uncertainties on the photometry (i.e., data noise).

2.1. Observation Data

KiDS provides a unique data set, with 9-band photometry including four optical (*ugri*) and five NIR bands (ZYHJK$_s$), to study stellar populations of galaxies among Stage-III surveys, down to a limiting magnitude of $r \sim 24$. MLP-GaP is suitable for application in the KiDS data set. Therefore, we develop MLP-GaP based on KiDS data. We first have collected redshifts and 9-band magnitudes for 120,000 random galaxies, providing a solid foundation for generating a mock sample that closely aligns with actual observational data. We only used redshifts and r-band magnitudes from actual observation data to generate the mock sample. The redshifts used are "morphoto-z" obtained by GaZNet (Li et al. 2022b), ranging from 0 to 3. The "morphoto-z" is derived by combining imaging and multi-band photometric data. It offers superior accuracy, precision, and fewer outliers than traditional photometric redshifts. Figure 1 shows the distribution of redshifts and r-band magnitudes for 120,000 galaxies.

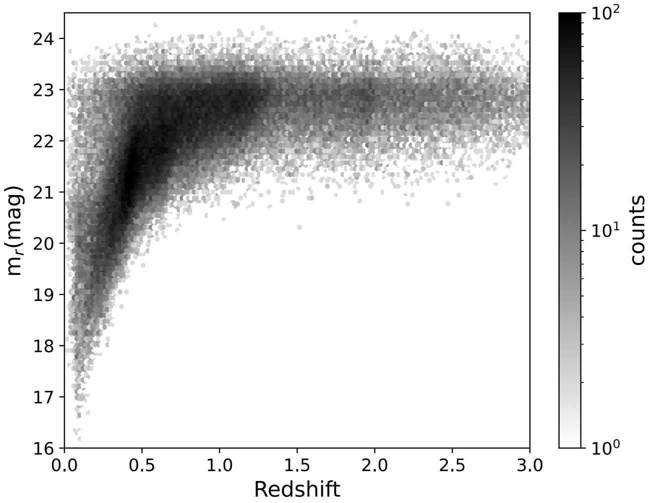

Figure 1. The distribution of r-band magnitude with redshift in real observation data.

Table 1
The Range of SED Parameters

Parameter	Unit	Range	Interval
tau	Myr	250–8000	50
age	Myr	250—Cosmology age at the redshift	50
metallicity	⋯	0.0001, 0.0004, 0.004, 0.008, 0.02, 0.05	⋯
A_V	mag	0–2	0.01
alpha	⋯	0.0625–4	0.0625

2.2. The Catalog for Mock Galaxies

To construct a catalog for mock galaxies, we need to simulate the generation of intrinsic parameters (e.g., age, extinction coefficient) and the observation information (photometry of each band). The intrinsic parameters are input parameters for SED fitting, and they are generated randomly within a predefined range (see Table 1) to emulate the natural variation in galactic properties. Among observation information, redshifts and r-band photometric data are from actual observation data. The aperture photometry from the other bands, their uncertainties, as well as corresponding stellar masses and SFRs are all generated through CIGALE.

Table 2
The Module Assumptions for Mock Dataset

Module	Parameters	Value
SFH(delayed)	Tau_main (Myr)	tau
	Age (Myr)	age
	f_burst	0
	Tau_burst (Myr)	50
	burst_Age (Myr)	100
BC03	IMF	1(Chabrier)
	Metallicity	metallicity
dustatt_modified_starburst	E_BV_lines (mag)	$\frac{A_V}{3.1 * 0.44}$
dale2014	Alpha	alpha

Note. Tau, age, metallicity, A_V, alpha are the parameters generated by the simulation of each galaxy, and their ranges are in Table 1.

First, we create a data file for each galaxy, which includes observed photometric data and redshift. Among them, the redshift and flux density of the *r*-band are based on actual observation data from 120,000 galaxies. Assuming that the error of the *r*-band magnitude is 0.1 mag, it can convert that into the error of the flux density of the *r*-band. Using placeholder values, such as "−9999," represent data from other missing bands. Then, we initialize CIGALE in the directory where data files are located and generate configuration files. Next, we modify and update the configuration files. In this process, we write the file names, the specific modules used, and the corresponding parameters (i.e., tau, age, metallicity) into the configuration file. The modules and parameters for SED fitting are summarized in Table 2. Final, running CIGALE, to generate the photometric data for missing bands in the input data file, along with calculating the stellar masses and SFRs. To have realistic uncertainties on the photometry of the mock galaxies derived as above, we start from the errors reported in the catalog and perform a random sampling within a certain range. Suppose we need to generate a realistic error for the u-band photometry of a mock galaxy with a magnitude of 19.3. We first identify galaxies in the actual galaxy sample with u-band magnitudes close to this value, specifically within a ±0.01 mag range (i.e., from 19.29 to 19.31). If the magnitudes of all galaxies reported in the catalog do not fall within a specified range, we will expand our search to a broader range until we find galaxies that meet the criteria. Then, the error is randomly selected from among these errors. Figure 2 presents a flowchart that comprehensively outlines the process for generating the mock catalog for galaxies.

2.3. Partition of Datasets

The mock data set of 120,000 galaxies, as generated in the previous section, consists of redshifts, 9-band magnitudes, their associated errors, their stellar masses, and SFRs. Within this data set, the redshifts, 9-band magnitudes, and their associated errors are utilized as the input parameters, or "features," for our ML algorithm. Furthermore, the corresponding stellar masses and SFRs are the desired outcomes, or "targets," that our ML algorithm aims to predict.

This data set is used to train and test the MLP-GaP. To proceed with the training, validation and testing of the MLP-GaP, it is important to partition the data set appropriately. Hence, we split the data set of 120,000 mock galaxies into three separate samples:

1. *Training data set*. Consisting of 90,000 galaxies, this data set is used to train the MLP-GaP.
2. *Validation data set*. Comprising 10,000 galaxies, this data set is utilized to tune hyperparameters and prevent overfitting during the training process.
3. *Testing data set*. With 20,000 galaxies, this data set serves as unseen data to assess the MLP-GaP's performance and conduct error statistical analysis.

This partitioning strategy comprehensively evaluates the performance of the MLP-GaP while maintaining a balance between training, validation, and testing data sets.

3. MLP-GaP Architecture and Training

Due to our goal of constructing a mapping between photometric data of galaxies and their parameters, we use supervised ML algorithms. Given our algorithm is expected to predict continuous values, which aligns with a regression problem, using multi-layer perceptrons (MLPs) is the best choice. MLPs are a class of feedforward artificial neural networks characterized by their fully connected architecture and the use of nonlinear activation functions. MLPs include at least three layers: an input layer, one or more hidden layers, and an output layer.

3.1. Architecture

To accurately predict both $M_\star$ and SFR, MLP-GaP is built using an MLP model with 10 layers. Its architecture can be described as follows:

1. *Input Layer*. 19 nodes;
2. *Hidden Layers*. 512, 512, 512, 512, 256, 256, 128, 64, 32 nodes with Rectified Linear Unit activation function;
3. *Output Layer*. Two nodes for predicting $M_\star$ and SFR.

To enhance our model's performance and mitigate the risk of overfitting, we incorporate both training and validation data sets within our training domain. Our model uses the Huber loss (Friedman 1999) function, which provides a balanced approach for evaluating the performance of a regression model (e.g., Li et al. 2022a). The Huber loss is defined as

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2, & |a| \leqslant \delta \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise}. \end{cases} \qquad (1)$$

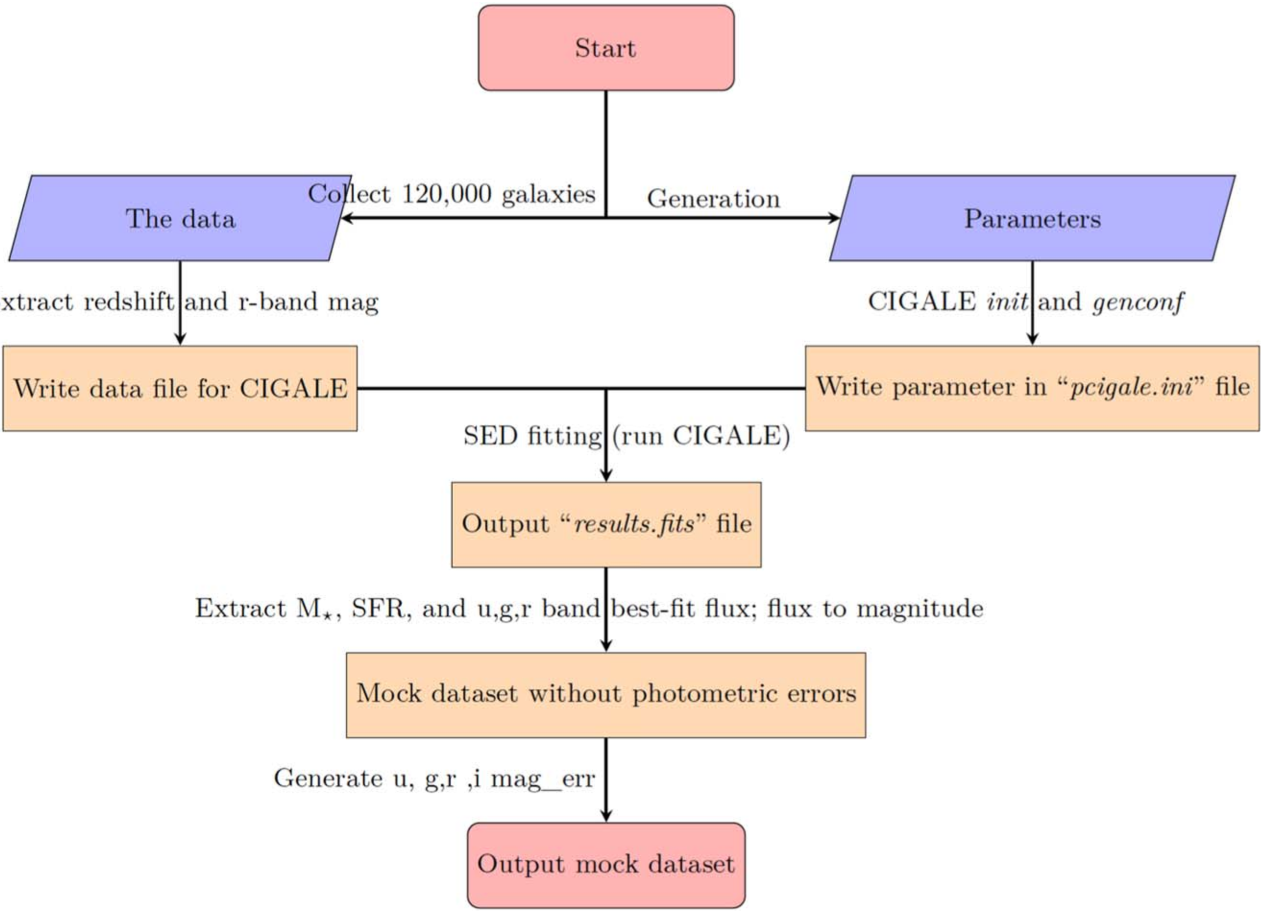
Start
Collect 120,000 galaxies
Generation
The data
Parameters
Extract redshift and r-band mag
CIGALE *init* and *genconf*
Write data file for CIGALE
Write parameter in "*pcigale.ini*" file
SED fitting (run CIGALE)
Output "*results.fits*" file
Extract M$_\star$, SFR, and u,g,r band best-fit flux; flux to magnitude
Mock dataset without photometric errors
Generate u, g,r ,i mag_err
Output mock dataset

Figure 2. The flowchart for generating mock data set for 120,000 galaxies.

Here $a = y_{\rm ture} - y_{\rm pred}$, $y_{\rm ture}$ is the reference values for the simulations, $y_{\rm pred}$ is the predicted values by the MLP model. δ is a threshold parameter that can be pre-setted and fixed to 0.001 in this work. Compared to traditional loss functions, the Huber loss function can provide better robustness, effectively mitigate the impact of outliers, maintain sensitivity when errors are small, and exhibit linear characteristics when errors are large, making the optimization process more stable and rapid. Additionally, we use the Adam optimizer (Kingma & Ba 2014) to facilitate the optimization process, ensuring efficient and effective model training.

3.2. Training

To enhance the training process and ensure that the model can converge effectively to the global optimum while avoiding entrapment in local optima, we have adopted a segmented training method. This method is more flexible than the decay rate strategy, permitting us to make necessary adjustments to the learning rate at various training stages based on the model's performance. Furthermore, considering that adjustments may need to be made to the model during the training process, the segmented training method provides us with increased control, thereby significantly improving the efficiency and effectiveness of our model training. Next, we will provide a detailed exposition of our training process.

Initially, the model was trained for 20 epochs with a learning rate of 10^{-3}. Starting with a high learning rate can lead to a faster reduction of the loss and accelerate the convergence of the model by making larger updates to the weights. Then, the learning rate is set to 10^{-4}, the pre-trained model is re-loaded, and the model is trained for 50 epochs. Next, the learning rate is reduced again to 10^{-5}, and the model is trained for another 50 epochs. If the model has not fully converged, we repeat the previous training until the model is well-trained. Ultimately, our model converged with a loss function value of 7.63×10^{-6}.

4. Evaluation of MLP-GaP on Testing Dataset

After building and training the MLP-GaP, to further assess its performance, it will be applied to estimate the stellar masses and SFRs of the galaxies in the testing data set. First, the predictive values of the MLP-GaP are compared with their reference values. Subsequently, the predictive values of the MLP-GaP are compared with the estimated values of CIGALE.

4.1. The Evaluator Metrics

To assess the performance of the MLP-GaP in terms of accuracy and precision, we use three different statistical estimators:

1. The coefficient of determination R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y^i_{\mathrm{pred}} - y^i_{\mathrm{true}})^2}{\sum_{i=1}^{N}(y^i_{\mathrm{pred}} - \bar{y}_{\mathrm{true}})^2}. \tag{2}$$

2. The Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y^i_{\mathrm{pred}} - y^i_{\mathrm{true}}|. \tag{3}$$

3. The Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y^i_{\mathrm{pred}} - y^i_{\mathrm{true}})^2. \tag{4}$$

Here N is total number of the data points, y^i_{pred} is the predictive or estimated values, y^i_{true} is the reference values, and $\bar{y}_{\mathrm{true}}$ is the mean value of the reference values.

R^2 is used to measure the goodness of fit of a regression model. It ranges from 0 to 1 and is a proportion of the variance in the dependent variable that is predictable from the independent variables. In practical applications, R^2 values close to 1 are desirable, indicating that the model fits the data well. However, a high R^2 value does not necessarily mean that the model is accurate or precise, as it only measures the fitting goodness of the model and not the accuracy of its predictions. MAE and MSE measure the average absolute and squared differences between the predicted and actual values. Both MAE and MSE are non-negative values, and lower values for both metrics indicate better performance of the ML model. Therefore, the comprehensive evaluation using R^2, MAE, and MSE can serve as a better measure of the MLP-GaP's performance. In our analysis, we assess the performance of the MLP-GaP by comparing its predictions to reference values. The closer the R^2 value is to 1, and the lower the MAE and MSE values are, the better the performance of the MLP-GaP is considered to be.

4.2. Comparing the MLP-GaP Predictions with Reference Values

We initially assess the performance of the MLP-GaP using the mock testing data set. In the top-left panel of Figure 3, we illustrate the predictions of stellar masses plotted against their reference values. The comparison reveals a strong correlation, with data points closely clustered around the 1:1 line, indicating high accuracy in the predictions. The R^2, MAE and MSE are 0.994, 0.041 and 0.0036, respectively, suggesting that the MLP-GaP could accurately predict stellar masses of mock galaxies. To check whether the predictions are affected by the redshifts, in the top-right panel of Figure 3, we present the variation of the predicted stellar masses, in the form of $\log(M_{\star,\mathrm{MLP-GaP}}/M_\star)$, against the redshifts. We find that the variation is relatively minor, with an overall standard deviation of $\sigma_{M_\star} = 0.060$ dex, which translates to a factor of $10^{\sigma_{M_\star}} = 1.15$ relative to the established values. We also notice that the scatter tends to increase at high redshift, but only with a maximum standard deviation of 0.1 dex, suggesting that the MLP-GaP experiences a decrease in accuracy when predicting stellar masses as redshift increases. Overall, the MLP-GaP performs well in the $M_\star$ prediction in the redshift range of $0 < z < 3$.

Let us move on to SFR prediction. In our mock sample, some passive galaxies show statistically negligible and very low SFRs, which we can exclude from our analysis. We use a standard threshold based on the specific star formation rate ($\mathrm{sSFR} = \mathrm{SFR}/M_\star$) to filter out these passive galaxies. The threshold we have adopted is $\log \mathrm{sSFR} > -12$ (e.g., Katsianis et al. 2021), which is regarded as the minimum value above which the star formation activity in galaxies cannot be ignored. The following all analyses regarding SFR have filtered out these passive galaxies. The predicted values of SFRs versus their reference values are plotted in the bottom-left panel of Figure 3. Similarly to the $M_\star$, most of the points are distributed around the 1:1 line, although with a larger scatter. The evaluation indices for the SFRs indicate a very good accuracy, with a $R^2 = 0.984$, $\mathrm{MAE} = 0.065$, and $\mathrm{MSE} = 0.0134$. Although these indices are not as good as those for $M_\star$, they still suggest a high predictive accuracy for SFRs. In the bottom-left corner of this panel, where $\log \mathrm{SFR} < 0$, a certain discrepancy in the predicted values is observed, indicating that MLP-GaP's accuracy has declined in the region. In the bottom-right panel of Figure 3, we plot the variation, $\log(\mathrm{SFR}_{\mathrm{MLP-GaP}}/\mathrm{SFR})$ of the prediction, against the redshifts. From a general perspective, we can see a relatively minor variation, with a standard deviation of $\sigma_{\mathrm{SFR}} = 0.116$ dex, which is equivalent to a factor of $10^{\sigma_{\mathrm{SFR}}} = 1.306$, relative to established values.

In summary, our research has established that the MLP-GaP is a robust and accurate tool for predicting the stellar masses and SFRs of galaxies in the testing data set. It can exhibit consistent performance across a wide range of redshifts.

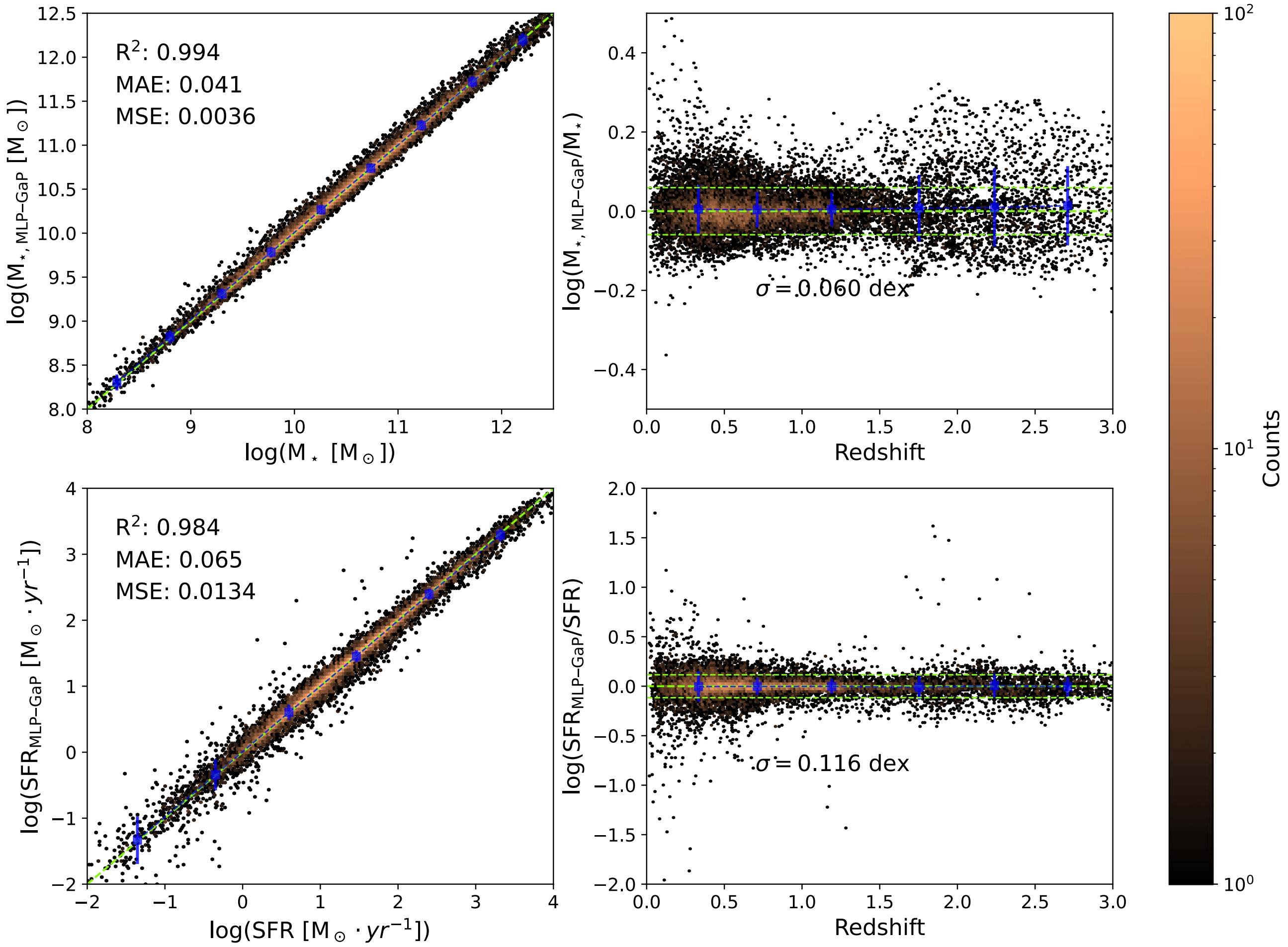

Figure 3. The comparison between the MLP-GaP's predictions and reference values in the testing data set (20,000 galaxies). Top-Left panel: The stellar masses of the MLP-GaP are compared with the reference values. Top-Right panel: Errors of the MLP-GaP results obtained for the testing data set as a function of redshift for $M_{\star}$. Bottom-Left panel: The SFRs of the MLP-GaP are compared with the reference values. Bottom-Right panel: Errors of the MLP-GaP results obtained for the testing data set as a function of redshift for SFR.

4.3. Comparing the MLP-GaP Predictions with CIGALE Estimations

To thoroughly evaluate the performance of MLP-GaP, we use CIGALE to estimate the stellar masses and SFRs of the mock testing data set. This analysis involves fitting multi-band photometric data, and the relevant modules and parameters used in the configuration file of CIGALE are detailed in Table 3. The results obtained are then compared with those predicted from MLP-GaP to assess the accuracy and precision of the predictions.

In Figure 4, we present the CIGALE fitting results versus their reference values for both the $M_{\star}$ (top-left panel) and SFR (bottom-left panel), and their variations as a function of redshifts are also appended in the right panels. In terms of estimating stellar masses, CIGALE is comparable performance to MLP-GaP. This conclusion is supported by the evaluator indices of R^2, MAE and MSE, with the deviation being similarly close to that of MLP-GaP. In the top-right panel of Figure 4, we find that the stellar masses derived using CIGALE exhibit a slight advantage at high redshifts ($z > 1.5$). This is evidenced by the relatively smaller error bars within this redshift range, and a greater concentration of data points around the line where $\log(M_{\star,\mathrm{CIGALE}}/M_{\star}) = 0$. Moving on to SFR estimation, the evaluator indices suggest that MLP-GaP outperforms CIGALE in estimating SFRs. The standard deviation of the differences between the SFRs derived by CIGALE and the reference values is $\sigma = 0.320$ dex, a value that is notably higher compared to the standard deviation of the differences between the SFRs predicted by MLP-GaP and the reference values. The primary discrepancy arises at low redshifts ($z < 1.5$), where CIGALE not only exhibits greater variability in its deviations but also deviates from the line representing $\log(\mathrm{SFR}_{\mathrm{CIGALE}}/\mathrm{SFR}) = 0$. These findings suggest that MLP-GaP can provide more accurate predictions

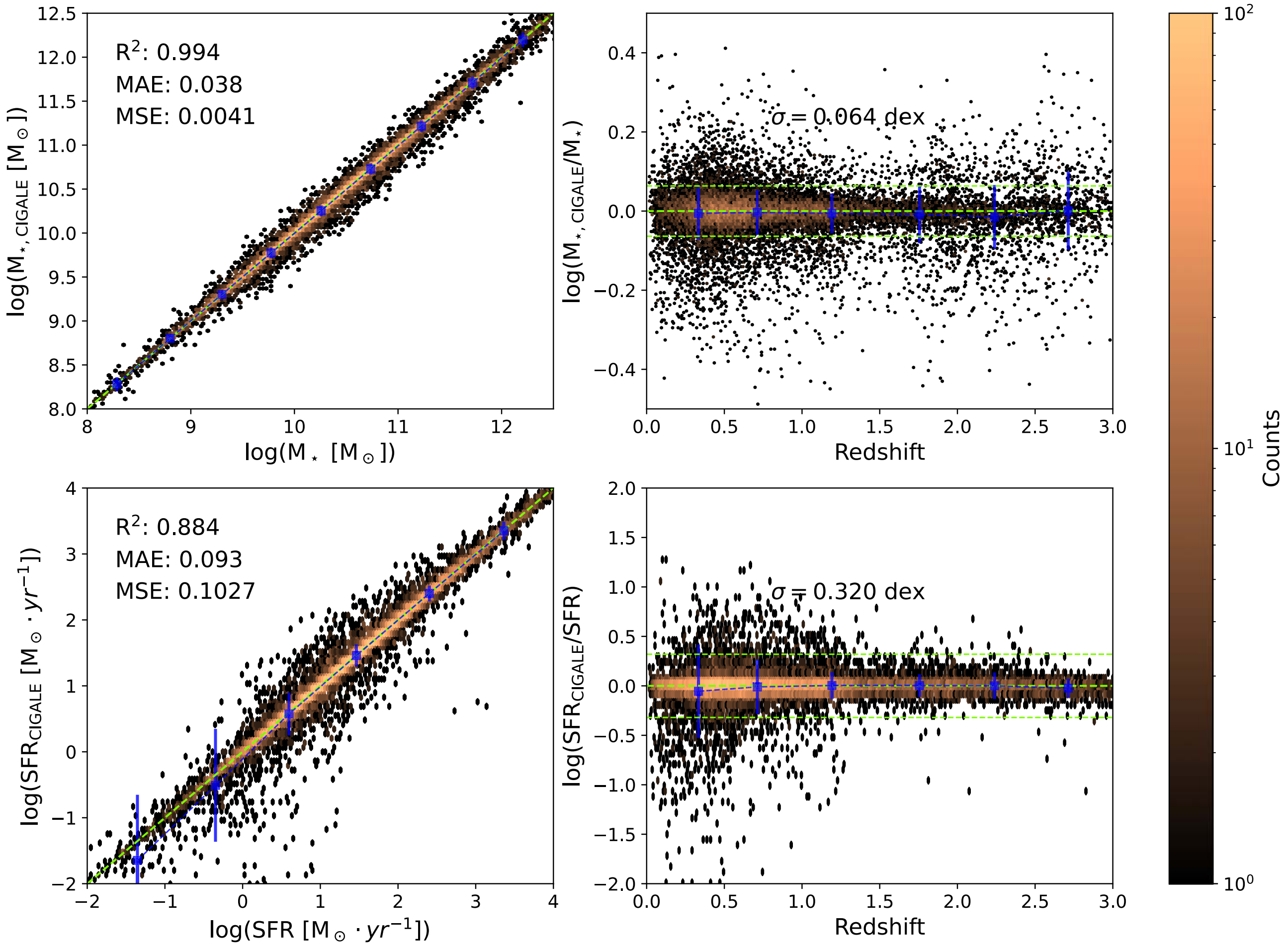

Figure 4. The comparison between the CIGALE predictions and reference values on the testing data set. Top-Left panel: The stellar masses of the CIGALE are compared with the reference values. Top-Right panel: Errors of the CIGALE results obtained for the testing data set as a function of redshift for $M_\star$. Bottom-Left panel: The SFRs of the CIGALE are compared with the reference values. Bottom-Right panel: Errors of the CIGALE results obtained for the testing data set as a function of redshift for SFR.

Table 3
The Module Assumptions for SED Fitting

Module	Parameters	Value
SFH(delayed)	Tau_main (Myr)	250, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 7000, 8000
	Age (Myr)	250, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 7000, 8000, 9000, 10000, 11000, 12000, 13000
	f_burst	0
BC03	IMF	1(Chabrier)
	Metallicity	0.0001, 0.0004, 0.004, 0.008, 0.02, 0.05
dustatt_modified_starburst	E_BV_lines (mag)	0.0, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5
dale2014	Alpha	1.0, 2.0, 3.0, 4.0

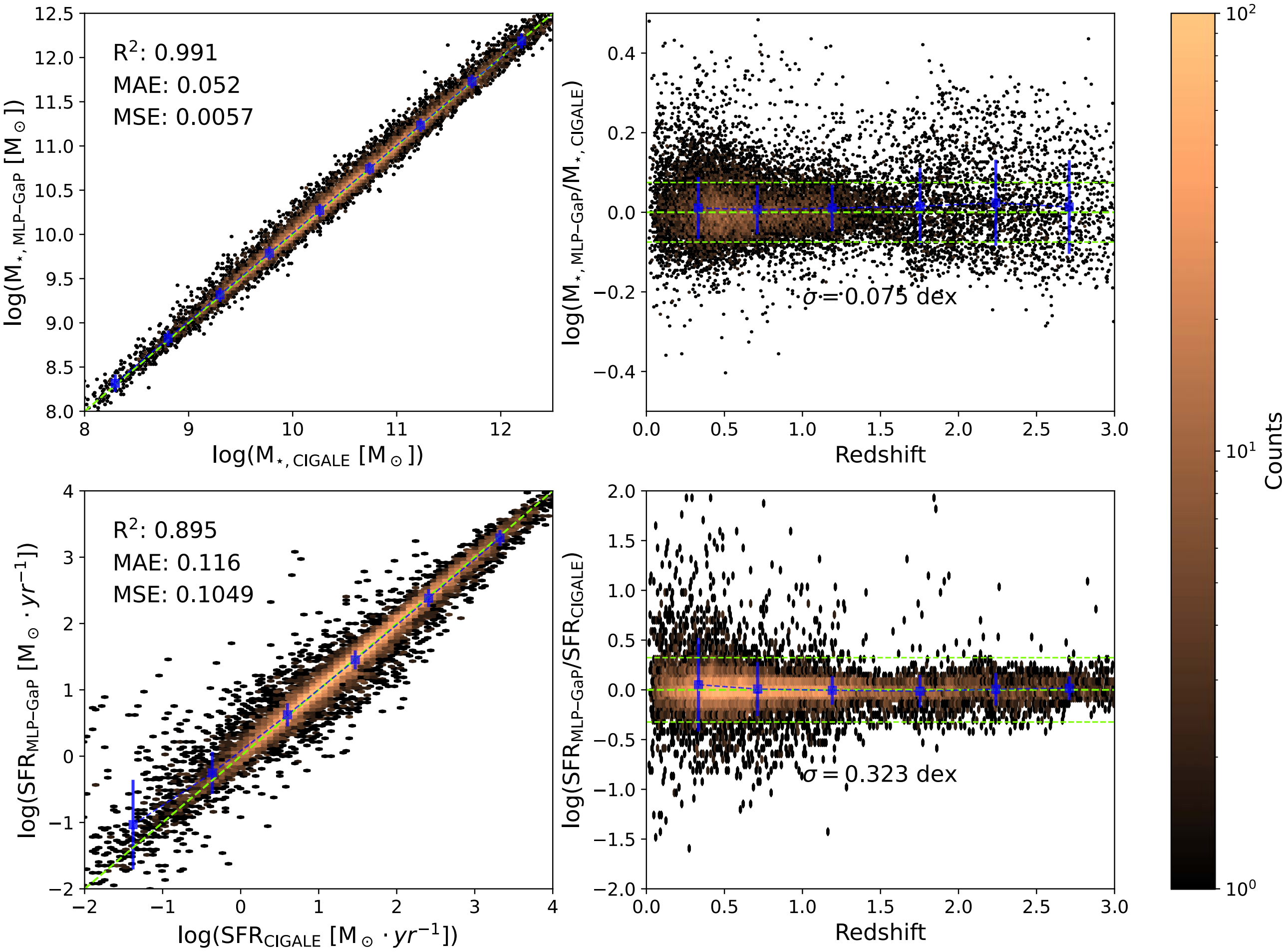

Figure 5. The comparison between the MLP-GaP predictions and CIGALE predictions on the testing data set. Top-Left panel: The stellar masses of the MLP-GaP are compared with those of CIGALE. Top-Right panel: Errors between MLP-GaP and CIGALE results obtained for the testing data set as a function of redshift for $M_\star$. Bottom-Left panel: The SFRs of the MLP-GaP are compared with those of CIGALE. Bottom-Right panel: Errors between MLP-GaP and CIGALE results obtained for the testing data set as a function of redshift for SFR.

of SFRs for mock galaxies in the testing data set compared to CIGALE.

Subsequently, a direct comparison will be made between the predicted values from MLP-GaP and those derived using CIGALE. Such a comparison is pivotal because, in actual astronomical observations, the "ground truth" values of the stellar population parameters are unattainable. All values are dependent on various assumptions, models, and inherent uncertainties, which means that error is an inevitable aspect of the process. Therefore, the performance of a novel tool like MLP-GaP must be evaluated by comparing its results with those obtained from standard tools or methodologies. In the left panels of Figure 5, the vertical axis represents the results from MLP-GaP, while the horizontal axis displays the corresponding results from CIGALE. The majority of sources are concentrated near the diagonal line, indicating a good consistency between the predictions made by the two tools. Additionally, the results from their evaluators also demonstrate this point. The right panels of Figure 5 illustrate the deviation of their results as a function of redshifts. The top-right panel indicates that the $M_\star$ estimates derived from both MLP-GaP and CIGALE are in close alignment, exhibiting a standard deviation of $\sigma = 0.075$ dex. The bottom-right panel also suggests that the SFRs estimated from MLP-GaP and CIGALE agree, with a standard deviation of $\sigma = 0.323$ dex. These mean that MLP-GaP is capable of delivering predictions that align with those of standard tools, demonstrating its reliability and potential as a viable alternative for estimating stellar masses and star formation rates in galaxies.

Finally, an evaluation is conducted on the computational efficiency of MLP-GaP compared to the traditional standard tools. The two primary indicators for assessing computational efficiency are the time consumption and the utilization of computational resources. To ensure a fair and unbiased comparison, both CIGALE and MLP-GaP are run on identical hardware platforms. Specifically, the tests are run on a computing system equipped with an Intel Core i7-11700F processor, featuring 12 cores operating at a base frequency of 2.5 GHz. For MLP-GaP, the estimation of stellar masses and SFRs for 20,000 galaxies is conducted using a single core, with the entire process being completed in 11.018 seconds. In contrast, despite utilizing 10 cores, the SED fitting method still requires a longer computational time, approximately 200 minutes. When not considering the number of cores used, the time expended by the SED fitting method is already 985 times greater than that of MLP-GaP. Should MLP-GaP also use 10 cores, it would undoubtedly achieve an even more rapid execution speed. Therefore, MLP-GaP demonstrates outstanding performance in terms of running speed, showcasing its potential as a highly efficient tool for astronomical data analysis.

In conclusion, MLP-GaP demonstrates a superior ability to predict the stellar masses and SFRs of galaxies with higher precision in comparison to CIGALE. Furthermore, its computational performance is exceptionally impressive, providing a remarkable running speed that substantially surpasses that of traditional tools.

5. Discussion

5.1. Performance on Actual Dataset

Although the mock data set closely approximates actual observational data, inherent differences may still exist, particularly in the distribution of parameters and the interrelationships among them. Therefore, relying solely on the mock data set to test MLP-GaP may not fully confirm its reliability. To demonstrate the science-readiness of the MLP-GaP, we will advance to assess its performance using actual observational data. The catalog provided by Xie et al. (2023) is highly suitable for serving as a test data set for MLP-GaP. This catalog includes the observational data for 288,809 galaxies and the stellar population parameters for their galaxies as outputted by LePhare and CIGALE. Here, we will apply MLP-GaP and CIGALE to the catalog to estimate the stellar masses and SFRs of the galaxies, and then compare these estimations. The redshifts used in this comparison are "morphoto-Z" obtained by GaZNet (Li et al. 2022b).

Figure 6 presents a detailed comparison of the performance of MLP-GaP and CIGALE on actual observational data of galaxies. The top-left panel demonstrates a good consistency in the stellar masses estimation between MLP-GaP and CIGALE, as evidenced by $R^2 = 0.952$, MAE $= 0.107$, MSE $=0.0195$. The top-right panel indicates that the deviations in their stellar masses exhibit some variation with redshift, but these fluctuations are minor than their standard deviation $\sigma = 0.132$ dex. Compared to the results from the testing data set, the consistency between MLP-GaP and CIGALE in estimating the stellar masses of actual galaxies has slightly deteriorated. The bottom-left panel illustrates the agreement between the SFR estimates by MLP-GaP and CIGALE, albeit with a certain degree of dispersion. The bottom-right panel demonstrates that variations in SFR estimation exhibit some redshift dependency, yet these fluctuations are relatively minor compared to the standard deviation. In estimating the SFRs of actual galaxies, there is also some degradation in consistency. This may arise from inherent limitations within MLP-GaP itself. Given that the training data is mocked, there are certain discrepancies in the distribution of parameters compared to actual galaxies. For instance, the fraction of massive galaxies in both the training and testing data sets may be significantly higher than that found in actual galaxies. This could lead MLP-GaP to learn patterns that do not align with those of actual galaxies, particularly in the case of massive mock galaxies. Regardless, MLP-GaP still demonstrates good consistency with CIGALE in estimating stellar masses and SFRs of actual galaxies. Therefore, MLP-GaP can serve as an alternative to traditional SED fitting tools for predicting stellar masses and SFRs. In particular, MLP-GaP is significantly faster than conventional methods in terms of computation speed, so it is more suitable for estimating the stellar masses and SFRs of billions of galaxies in large-scale surveys.

5.2. Application and Improvement in Future

Shortly, astronomy will enter a new era of development, where an unprecedented wealth of observational data can be obtained from various large-scale and deep-area surveys. This new age will be marked by the availability of data from major projects. These surveys cover vast sky areas, providing multiband (UV, optical, and NIR) photometric data and images for billions of galaxies, thus presenting unique opportunities for scientific inquiry. Our MLP-GaP, leveraging the power of ML, offers significant advantages over traditional SED fitting techniques. It is uniquely positioned to swiftly and accurately estimate the stellar masses and SFRs of the galaxies observed in these large-scale surveys. As the volume of observational data continues to expand, the implementation of efficient and accurate ML algorithms is poised to become increasingly invaluable. The capability of MLP-GaP to rapidly and accurately predict the physical parameters for billions of galaxies expands our comprehension of the cosmos, unveiling new perspectives on the evolution of galaxies throughout the universe, thereby significantly advancing our knowledge of the astrophysical processes that shape the cosmos.

While MLP-GaP has demonstrated its potential in predicting stellar masses and SFRs for huge volumes of galaxies, it requires further and profound enhancements to fully realize its capabilities. Our roadmap for future improvements includes the following aspects:

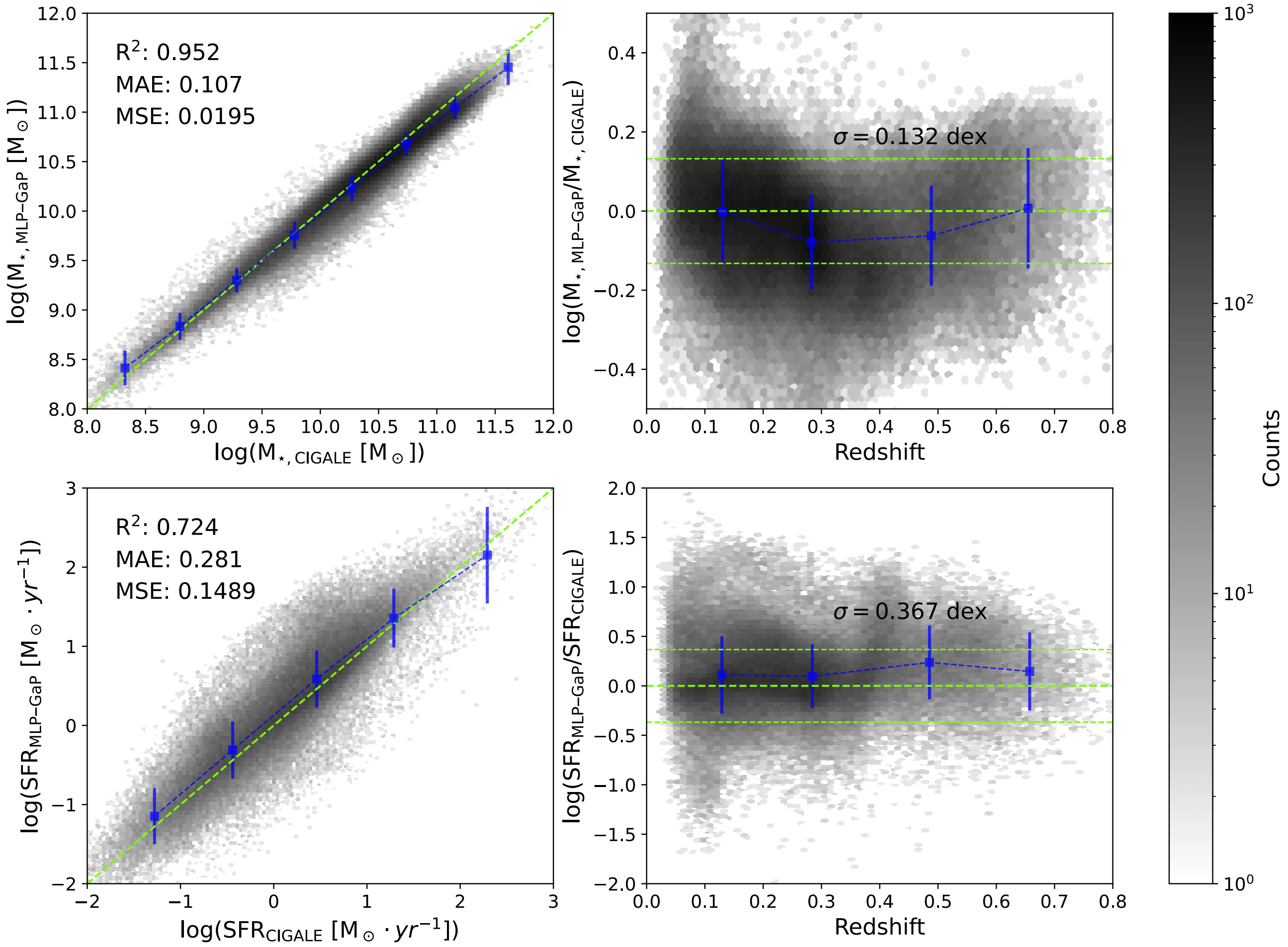

Figure 6. The comparison between the MLP-GaP predictions and CIGALE predictions on the actual data set. Top-Left panel: The stellar masses of the MLP-GaP are compared with those of CIGALE. Top-Right panel: Errors between MLP-GaP and CIGALE results obtained for the testing data set as a function of redshift for $M_\star$. Bottom-Left panel: The SFRs of the MLP-GaP are compared with those of CIGALE. Bottom-Right panel: Errors between MLP-GaP and CIGALE results obtained for the testing data set as a function of redshift for SFR.

1. *Enhancing Training Data Diversity*. Our current MLP-GaP is trained on mock data sets generated by CIGALE. Moving forward, we want to achieve that the parameter distributions of mock galaxies closely mirror those of actual galaxies. Moreover, we intend to enrich the training data sets with more intricate galaxy formation histories, a wider array of dust attenuation models, and different stellar population models. Additionally, we will integrate various astrophysical processes and observational noise to better simulate real-like data, thereby enhancing the generalizability and accuracy of the MLP-GaP.
2. *Optimizing Model Architecture*. Although the MLP model utilized in this study has shown its power, there is ample scope for optimization. Our future endeavours will not only seek to augment the model's computational speed but also refine its predictive accuracy. We will investigate alternative network architectures, including Self-Attention and Transformer Models, which have demonstrated remarkable performance in various domains.
3. *Expanding Parameter Prediction*. We aim to extend MLP-GaP to predict more parameters of galaxies. This will encompass characteristics such as age, metallicity, IMF, and so on, providing a more comprehensive understanding of the properties of galaxies.
4. *Uncertainty Quantification*. In the realm of scientific inquiry, the accurate estimation of parameter uncertainties is critical for robust error analysis. Recognizing this, we will adopt Bayesian methods to determine the uncertainties of different parameters. By integrating Bayesian networks within our ML algorithm, we can better quantify the uncertainties associated with our predictions. It will provide not only point estimates of the parameters but also their probabilistic distributions. Such

an approach is essential for credible scientific discourse and for making informed decisions in the light of inherent uncertainties in observational data.

6. Summary

In the era of large scale surveys, there is a massive amount of observational data available. However, the challenge remains of how to rapidly and accurately derive various stellar population parameters for billions of galaxies. Among the numerous parameters that characterize galaxies, the $M_\star$ and SFR are considered the most critical. In response to the aforementioned challenge, we have developed an ML algorithm called MLP-GaP, which can rapidly and accurately predict the stellar masses and SFRs for a massive amount of galaxies.

First, we used CIGALE to generate a mock data set, which is a catalog consisting of 120,000 mock galaxies. This catalog provided redshift, $M_\star$, SFR, and photometric data of nine bands for each mock galaxy. The mock data set was meticulously partitioned into three distinct subsets, with a training data set comprising 90,000 galaxies, a validation data set of 10,000 galaxies, and a testing data set of 20,000 galaxies. Subsequently, we used an MLP model with 10 layers to build MLP-GaP. Through rigorous training and validation processes, MLP-GaP was optimized to yield predictions that were consistent with the reference values on the testing data set. Furthermore, MLP-GaP demonstrated a significant faster in processing speed compared to CIGALE. To demonstrate the science-readiness of the MLP-GaP, we applied to actual galaxy samples. The predicted values from MLP-GaP exhibited a commendable level of consistency with the estimated values derived using SED fitting. This consistency suggested MLP-GaP could serve as an alternative to traditional SED fitting tools for predicting stellar masses and SFRs. Given the outstanding processing speed of MLP-GaP, it can be considered an essential tool for estimating the parameters of billions of galaxies in the era of large scale surveys.

Acknowledgments

We sincerely thank the anonymous reviewer for useful suggestions. We acknowledge the support of the National Nature Science Foundation of China (Nos. 12303017, and 12203096). This work is also supported by Anhui Provincial Natural Science Foundation project No. 2308085QA33. This work is supported by the science research grants from the China Manned Space Project.

ORCID iDs

Xiaotong Guo (郭晓通) https://orcid.org/0000-0002-2338-7709

References

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., et al. 2018, PhRvD, 98, 043526

Abbott, T. M. C., Adamów, M., Aguena, M., et al. 2021, ApJS, 255, 20

Adhikari, S., Shin, T., Jain, B., et al. 2021, ApJ, 923, 37

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4

Bilicki, M., Dvornik, A., Hoekstra, H., et al. 2021, A&A, 653, A82

Bonjean, V., Aghanim, N., Salomé, P., et al. 2019, A&A, 622, A137

Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, A&A, 622, A103

Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000

Buat, V., Boselli, A., Gavazzi, G., & Bonfanti, C. 2002, A&A, 383, 801

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682

Chabrier, G. 2003, PASP, 115, 763

Daddi, E., Dickinson, M., Morrison, G., et al. 2007, ApJ, 670, 156

Dai, Y., Xu, J., Song, J., et al. 2023, ApJS, 268, 34

Dale, D. A., Helou, G., Magdis, G. E., et al. 2014, ApJ, 784, 83

Donnari, M., Pillepich, A., Nelson, D., et al. 2019, MNRAS, 485, 4817

Elbaz, D., Daddi, E., Le Borgne, D., et al. 2007, A&A, 468, 33

Fang, G., Ba, S., Gu, Y., et al. 2023, AJ, 165, 35

Fixsen, D. J., Dwek, E., Mather, J. C., Bennett, C. L., & Shafer, R. A. 1998, ApJ, 508, 123

Friedman, J. H. 1999, AnSta, 29, 1189

Furlong, M., Bower, R. G., Theuns, T., et al. 2015, MNRAS, 450, 4486

Gao, F.-Y., Li, J.-Y., & Xue, Y.-Q. 2019, RAA, 19, 039

Gong, Y., Liu, X., Cao, Y., et al. 2019, ApJ, 883, 203

Goulding, A. D., Greene, J. E., Bezanson, R., et al. 2018, PASJ, 70, S37

Greco, J. P., Greene, J. E., Strauss, M. A., et al. 2018, ApJ, 857, 104

Hao, C.-N., Kennicutt, R. C., Johnson, B. D., et al. 2011, ApJ, 741, 124

Heymans, C., Tröster, T., Asgari, M., et al. 2021, A&A, 646, A140

Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454

Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841

Ivezić, Ž, Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111

Joachimi, B., Lin, C. A., Asgari, M., et al. 2021, A&A, 646, A129

Katsianis, A., Xu, H., Yang, X., et al. 2021, MNRAS, 500, 2036

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 33

Kennicutt, R. C., & Evans, N. J. 2012, ARA&A, 50, 531

Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193

Li, R., Napolitano, N. R., Roy, N., et al. 2022a, ApJ, 929, 152

Li, R., Napolitano, N. R., Feng, H., et al. 2022b, A&A, 666, A85

Maraston, C. 2005, MNRAS, 362, 799

Mistele, T., McGaugh, S., & Hossenfelder, S. 2023, JCAP, 2023, 004

Mitchell, P. D., Lacey, C. G., Baugh, C. M., & Cole, S. 2013, MNRAS, 435, 87

Noeske, K. G., Weiner, B. J., Faber, S. M., et al. 2007, ApJL, 660, L43

Popesso, P., Concas, A., Cresci, G., et al. 2023, MNRAS, 519, 1526

Roy, N., Napolitano, N. R., La Barbera, F., et al. 2018, MNRAS, 480, 1057

Salim, S., & Narayanan, D. 2020, ARA&A, 58, 529

Salpeter, E. E. 1955, ApJ, 121, 161

Schreiber, C., Pannella, M., Elbaz, D., et al. 2015, A&A, 575, A74

Song, J., Fang, G., Ba, S., et al. 2024, ApJS, 272, 42

Treyer, M., Schiminovich, D., Johnson, B., et al. 2007, ApJS, 173, 256

Vazdekis, A., Koleva, M., Ricciardelli, E., Röck, B., & Falcón-Barroso, J. 2016, MNRAS, 463, 3409

Wright, A. H., Hildebrandt, H., Kuijken, K., et al. 2019, A&A, 632, A34

Wright, A. H., Kuijken, K., Hildebrandt, H., et al. 2024, A&A, 686, A170

Wu, J. F., & Boada, S. 2019, MNRAS, 484, 4683

Xie, L., Napolitano, N. R., Guo, X., et al. 2023, SCPMA, 66, 129513

Zeraatgari, F. Z., Hafezianzadeh, F., Zhang, Y., et al. 2024, MNRAS, 527, 4677

Zhan, H. 2011, SSPMA, 41, 1441

Zhan, H., & Tyson, J. A. 2018, RPPh, 81, 066901