A&A, 706, A277 (2026)
https://doi.org/10.1051/0004-6361/202555894

Astronomy & Astrophysics

SHAPE

I. A SOM-SED hybrid approach for efficient galaxy parameter estimation leveraging JWST

Zihao Wang[1], Tao Wang[1,2,⋆], Ke Xu[1,2], Hanwen Sun[1,2], Ruining Tian[1], and Qi Hao[1,2]

[1] School of Astronomy and Space Science, Nanjing University, Nanjing, Jiangsu 210093, China
[2] Key Laboratory of Modern Astronomy and Astrophysics, Nanjing University, Ministry of Education, Nanjing 210093, China

Received 10 June 2025 / Accepted 22 November 2025

ABSTRACT

With the launch and application of next-generation ground- and space-based telescopes, astronomy has entered the era of big data, necessitating more efficient and robust data analysis methods. Most traditional parameter estimation methods do not have the capacity to reconcile differences between photometric systems. Ideally, we would like to optimally rely on high-quality observational data (e.g., from JWST) for calibrating and improving upcoming wide-field surveys, such as the Chinese Space Station Survey Telescope (CSST) and *Euclid*. To this end, we employed the self-organizing map (SOM) method and introduced a new approach that combines a SOM with a spectral energy distribution (SED). The resulting SOM-SED Hybrid Approach for efficient Parameter Estimation (SHAPE) is able to bridge different photometric systems and efficiently estimate key galaxy parameters, such as the stellar mass ($M_\star$) and star formation rate (SFR), leveraging data from a large and deep JWST/NIRCam and MIRI survey (PRIMER). As a test of the methodology, we focused on galaxies at z ∼ 1.5–2.5. To mitigate discrepancies between input colors and the training set, we replaced the default SOM weights with stacked SEDs from each cell, extending the applicability of our model to other photometric catalogs (e.g., COSMOS2020). By incorporating an SED library (SED Lib), we applied this JWST-calibrated model to the COSMOS2020 catalog. Despite the limited sample size and potential template-related uncertainties, SOM-derived parameters exhibit a good agreement with results from SED fitting using extended photometry. Under identical photometric constraints from CSST and *Euclid* bands, our method outperforms traditional SED fitting techniques in SFR estimation, exhibiting a reduced bias (–0.01 vs. 0.18) and a smaller $\sigma_{\rm NMAD}$ (0.25 vs. 0.35). With a computational efficiency capable of processing 10^6 sources per CPU per hour during the estimation phase, this JWST-calibrated estimator holds significant promise for next-generation wide-field surveys.

Key words. methods: data analysis – techniques: miscellaneous – astronomical databases: miscellaneous – galaxies: fundamental parameters – galaxies: stellar content

1. Introduction

The fundamental physical parameters of galaxies, such as the stellar mass ($M_\star$) and star formation rate (SFR), are crucial for understanding galaxy formation and evolution. These parameters serve as key diagnostics for tracing the diverse evolutionary pathways of galaxies across cosmic time (Madau & Dickinson 2014; Kennicutt & Evans 2012; Somerville & Davé 2015). Consequently, the accurate and efficient determination of these properties has remained a long-standing challenge in extragalactic astronomy.

In the absence of spectroscopic data, traditional methods infer galaxy properties through a spectral energy distribution (SED) fitting, which employs stellar population synthesis models to interpret multiwavelength photometry (Bruzual & Charlot 2003; Conroy 2013). Over the past two decades, several powerful SED fitting tools have been developed and widely adopted, including LePhare (Arnouts et al. 2002; Ilbert et al. 2006), PROSPECTOR (Johnson et al. 2021), CIGALE (Boquien et al. 2019), MAGPHYS (da Cunha et al. 2008), EAZY (Brammer et al. 2008), and BAGPIPES (Carnall et al. 2018). Among all, stellar mass estimates rely primarily on the integrated stellar luminosity, with low- to intermediate-mass stars contributing significantly in the rest-frame optical and near-infrared (NIR) bands. These stars, as they are less affected by extinction, enable relatively robust $M_\star$ estimates (Salvato et al. 2018) with an assumed initial mass function (IMF; e.g., Kroupa 2001). The SFR is primarily inferred from the radiation of young, massive stars, which emit strongly in the ultraviolet (UV) and whose emission is often absorbed by dust and re-emitted in the infrared (IR). To break the degeneracy between dust attenuation and stellar population age, SFR estimates typically combine UV and far-infrared (FIR) luminosities (e.g., Pannella et al. 2009; Buat et al. 2010; Hao et al. 2011; Riccio et al. 2021). Although effective, SED fittings are computationally expensive and sensitive to model assumptions regarding, for instance, stellar evolution, star formation histories and dust attenuation, particularly in the absence of FIR constraints (e.g., Michalowski et al. 2014; Wuyts et al. 2011).

In the coming decade, the launch of next-generation space-based telescopes and the ongoing accumulation of survey data from missions will statistically enhance our understanding of galaxy evolution. Among these, the China Space Station Survey Telescope (CSST; CSST Collaboration 2025) and *Euclid* (Euclid Collaboration: Moneti et al. 2022; Euclid Collaboration: Mellier

⋆ Corresponding author: taowang@nju.edu.cn

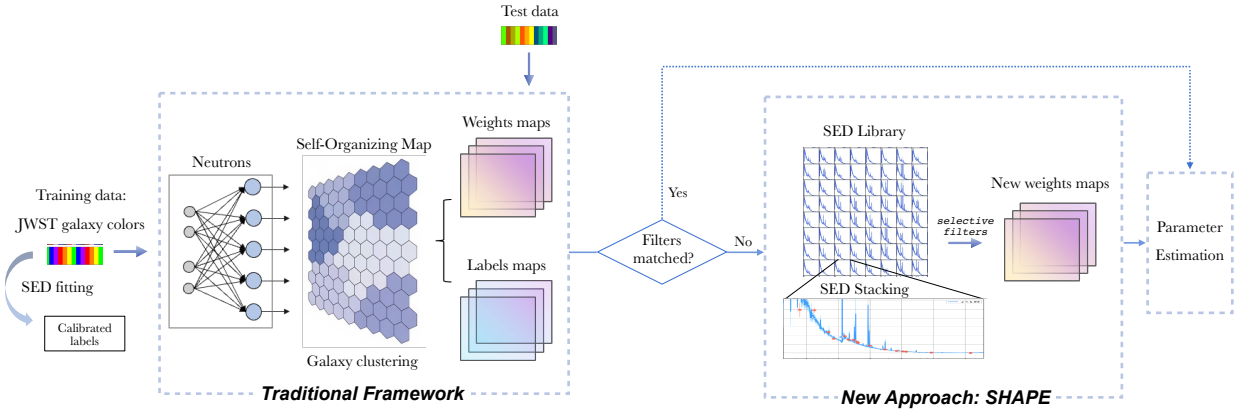

Fig. 1. Schematic diagram of SHAPE model. This method employs a SOM to cluster galaxies in the training set and assigns each SOM cell a representative average SED. When the photometric filters of the test set match those of the training set, galaxies can be directly projected onto the SOM for parameter estimation. Otherwise, the galaxy is matched to the SED Lib constructed from the SOM to determine its physical parameter.

et al. 2025) hold significant potential for these large-sample studies. The CSST aims to take the survey camera roughly 7 years of operation accumulated over 10 years of orbital time to image roughly 17 500 deg^2 (roughly 300 times wider view than the Hubble Space Telescope, HST) of the sky in the NUV, u, g, r, i, z, and y bands (Zhan 2021), enabling accurate photo-z estimations. It has the capacity to obtain more than one billion galaxy images and one hundred million galaxy spectra, while discovering millions of active galactic nuclei (AGNs) and other astronomical objects across a broad redshift range. When combined with *Euclid*'s three NIR bands of $Y_{\rm E}$, $J_{\rm E}$, and $H_{\rm E}$ (Laureijs et al. 2011), we can obtain robust estimates of $M_\star$. However, the influx of billions of sources will further exacerbate the computational burden of traditional template-fitting methods. Moreover, the traditional SED fitting method is sub-optimal for constraining SFR in the absence of MIR and FIR data (Pannella et al. 2009; Buat et al. 2010; Riccio et al. 2021), which both instruments lack, unless supported by external constraints or informed priors that help to break the dust–age degeneracy (e.g., Arango-Toro et al. 2025). Therefore, the challenge behind the large galaxy sample the CSST will provide lies in how to efficiently process billions of galaxies, how to robustly constrain the SFRs of large galaxy populations using only limited optical and NIR bands, and, more crucially, whether it is feasible to effectively leverage JWST data to calibrate other surveys.

In this context, machine learning (ML) techniques have emerged as powerful tools for analyzing complex, high-dimensional astronomical datasets (for a review, see e.g., Ball & Brunner 2010; Baron 2019; Longo et al. 2019; Salvato et al. 2018). Among these techniques, self-organizing maps (SOMs; Kohonen 1982) display distinct advantages in handling large astronomical datasets and visualizing high-dimensional data. A SOM is an unsupervised learning algorithm that allows us to project high-dimensional data onto a two-dimensional (2D) map, while preserving topological structure, making it particularly useful for analyzing large astronomical surveys. It can be employed as a computationally efficient alternative to traditional SED fitting, capable of processing vast amounts of data and exploring complex, nonlinear parameter spaces. Previous studies have successfully applied SOMs for tasks such as classifying stellar spectra (e.g. Mahdi 2011), classifying galaxy morphology (e.g. Galvin et al. 2019), estimating photometric redshifts and other parameters (e.g. Masters et al. 2015; Hemmati et al. 2019; Davidzon et al. 2019; Torre et al. 2024), as well as selecting specific targets (e.g. Masters et al. 2015; Hemmati et al. 2019). Davidzon et al. (2022) successfully applied this approach to physical parameter estimation using the COSMOS2020 dataset (Weaver et al. 2022). However, traditional SOM algorithms face challenges in handling missing values and require strict consistency between input color combinations and the training set, imposing stringent requirements on the input data.

In this study, we aim to prepare for upcoming CSST surveys and fully exploit the potential of JWST's comprehensive and high-quality data. We introduce a hybrid method that combines SOM with SED fitting techniques via SOM-SED Hybrid Approach for efficient Parameter Estimation (SHAPE), utilizing JWST's high-precision photometric data to establish a robust reference sample for future large-scale surveys. As a test of the methodology, we trained a SOM using deep data from the JWST PRIMER survey (Dunlop et al. 2021), initially focusing on galaxies at redshifts $z \sim 1.5 - 2.5$. SOM was employed to cluster galaxy samples and derive the average SED for each cell, thereby establishing an SED library (SED Lib hereafter) within the given redshift range. For any given galaxy, its colors were compared with the SEDs in the SED Lib to obtain the probability distribution over the SOM, followed by the corresponding parameter estimates. SHAPE extends the SOM-based approach from handling fixed discrete photometric points to a continuous functional framework and has been demonstrated to be efficient when applied to the COSMOS2020 catalog, which we employed as the test dataset. The basic workflow is described in Fig. 1 and more details are given in Sect. 3.

This paper is structured as follows. The employed JWST and COSMOS2020 photometric catalogs, along with their physical parameters, are summarized in Sect. 2. In Sect. 3.1, we thoroughly describe the mechanisms and setups of SOM, and replicate the parameter estimation method using SOM as presented in Davidzon et al. (2022), but with a SOM trained on JWST data. In Sect. 3.2, we introduce the SHAPE method, where we detail the construction of the SED Lib, synthesized based on SOM clustering results, and its application to parameter estimation. In Sect. 4, we present the parameter estimation results obtained by applying the traditional and hybrid methods to the JWST PRIMER and COSMOS2020 datasets and compare them with SED-derived estimates. In Sect. 5, we compare our work with previous work, and discuss the limitations and

outlook. In Sect. 6, we present our summary and conclusions. Throughout this work, we assume the Planck lambda cold dark matter (ΛCDM) cosmology with $H_0 = 70\ \mathrm{km\,s^{-1}/\,Mpc}$, $\Omega_m = 0.3$, and $\Omega_\Lambda = 0.7$. All magnitudes are given in the absolute bolometric (AB) system.

2. Data

To train a SOM that is both accurate and comprehensive in covering a wide variety of galaxies, we employed the PRIMER survey, which (at the time of writing) is the largest publicly available JWST survey. It provides extensive and deep NIR photometry as well as substantial MIRI coverage, making it an ideal choice for constructing the training dataset in our work. In addition, COSMOS2020 galaxies are used as a test sample to validate our methodology and assess its reliability.

2.1. JWST PRIMER photometric catalog

In this study, we utilized one of the largest and deepest JWST/NIRCam and JWST/MIRI surveys from public Treasury Programs, the Public Release IMaging for Extragalactic Research (PRIMER, GO 1837, Dunlop et al. 2021) program. The PRIMER survey targets the CANDELS-COSMOS and CANDELS-UDS fields with deep imaging in 10 bands: F090W, F115W, F150W, F200W, F277W, F356W, F444W, and F410M with NIRCam; along with F770W and F1800W with MIRI. All the JWST/NIRCam and previous HST images are publicly available from the Grizli Image Release v7.0 1[1], which were reduced and processed by the Grizli pipeline (Brammer et al. 2022). Due to a few issues in the default MIRI pipeline, the F770W and F1800W data were reprocessed using a custom-made pipeline to reconstruct the multiwavelength source catalog. The detailed technics are discussed in Wang et al. (2024a).

As the reference sample for SOM training and labeling, JWST galaxies are required to not only have broadband colors but also measurements of key physical parameters. We derived photometric redshifts using EAZY (Brammer et al. 2008), followed by the estimation of rest-frame colors, stellar mass and 100 Myr time-averaged SFR using BAGPIPES (Carnall et al. 2018). In summary, we adopted the delayed star formation history with the e-folding timescale between [0.01, 10] Gyr, the Calzetti extinction law (Calzetti et al. 2000) for young and old stellar populations separately, with A_V between [0, 5], nebular emission in BAGPIPES produced by CLOUDY (Ferland et al. 2017; Byler et al. 2017; Carnall et al. 2018), with log U between [–5, –2], a stellar metallicity between [0.01, 2.5] $Z_\odot$, and the time since star formation begins set between [0.03, 10] Gyr. Further details on catalog construction and parameter estimations are provided in Wang et al. (2024a).

2.2. COSMOS2020 photometric catalog

Since one of our goals is to bridge two different photometric systems using SOM, thereby leveraging JWST data to calibrate the physical parameters estimated from other photometric surveys, we employed the COSMOS2020 photometric catalog as the test sample and the basis of the mock CSST and *Euclid* catalog.

COSMOS2020 (Weaver et al. 2022) is an updated version of the previous COSMOS2015 catalog (Laigle et al. 2016). Source detections and multiwavelength photometry were performed for 1.7 million sources across the 2 deg^2 of the COSMOS field. Of these, ~966 000 were measured with all available broadband data, typically in twelve bands (u, g, r, i, z, y with MegaCam/CFHT, Y, J, H, K_s with VISTA/VIRCAM, and ch1, ch2 with *Spitzer*/IRAC). In addition, most sources have been provided with ancillary mediumband and narrowband photometry (e.g., NB118 with *VISTA*/VIRCAM, along with IB427, IB464, IA484, IB505, IA527, IB574, IA624, IA679, IB709, IA738, IA767, IB827, NB711, and NB816 with *Subaru*/Suprime-Cam). We incorporated a well-matched joint catalog (Wang et al. 2024b) as ancillary data, which extends the photometry to *Spitzer*/MIPS (24 μm), the Herschel PACS (100 and 160 μm) and SPIRE wavebands (250, 350, and 500 μm), further improving the robustness of parameter estimates.

Rather than utilizing the physical parameters derived from SED fitting with LePhare (Arnouts et al. 2002; Ilbert et al. 2006) in the original catalog, following the procedure in Sect. 2.1 and Wang et al. (2024a), we derived new parameters based on a subset of at least 12 broadbands, with a total of up to 33 bands. This approach ensures consistency by mitigating discrepancies introduced by different codes and templates in the comparison. Nevertheless, our estimates show excellent agreement with those provided in the COSMOS2020 public catalog.

2.3. The mock CSST and Euclid catalog

CSST (Zhan 2011; Cao et al. 2018; CSST Collaboration 2025) is a major science project initiated by the China Manned Space Program and is planned to be launched around 2027. The imaging bands consist of NUV, u, g, r, i, z, and y, covering a wide wavelength range from 2000 Å to 1.1 μm with a pixel scale of 0.″074. The slitless spectroscopy bands include GU (255–400 nm), GV (400–620 nm), and GI (620–1000 nm) with the same pixel scale. The planned surveys for both imaging and spectroscopy include (but are not limited to) a wide-field survey over about 17 500 deg^2 sky area and a deep-field survey of about 400 deg^2. The wide-field imaging survey will reach an average limiting magnitude better than 25.5 mag at 5σ for point sources.

Euclid (Euclid Collaboration: Mellier et al. 2025) is an ESA M-class astrophysics and cosmology mission. It is equipped with two instruments, namely, the optical VIS imager and the Near-Infrared Spectrograph and Photometer (NISP). The VIS imager comprises 36 4k × 4k CCDs with 0.″1/pixel, targeting at a broad 550–900 nm optical bandpass and reaching ~ 24.5 AB mag at 10σ for galaxies with a size larger than 1.25 times the full width at half-maximum (FWHM) of the PSF (Cropper et al. 2025). The NISP includes 16 arrays of 2k × 2k NIR-sensitive HgCdTe detectors with a pixel scale of 0.″3. It will obtain NIR imaging in Y_E, J_E, and H_E bands for photometric redshifts reaching 24 mag at 5σ for point sources in the Wide Survey (Jahnke et al. 2025). It will survey 15 000 deg^2 of the extragalactic sky avoiding the ecliptic plane due to increased zodiacal background in its Wide Survey, plus 40 deg^2 split over three deep fields with a depth increase by 2 mag. We refer to Liu et al. (2023) for further details on the scientific synergies between CSST and *Euclid*.

To simulate the application of our model on CSST data, we selected photometry from the COSMOS2020 catalog that corresponds to the nine bands available in the CSST and *Euclid* (i.e., u, g, r, i, z, y, Y, J, and H[2]). We then derived an additional set of physical parameters with the nine bands, adopting the same methodology. This set of parameters only serves as a baseline for

[1] https://dawn-cph.github.io/dja/imaging/v7

[2] In the following analysis, we use the Y, J, and H photometry from COSMOS2020 to mimic Euclid's Y_E, J_E, and H_E bands.

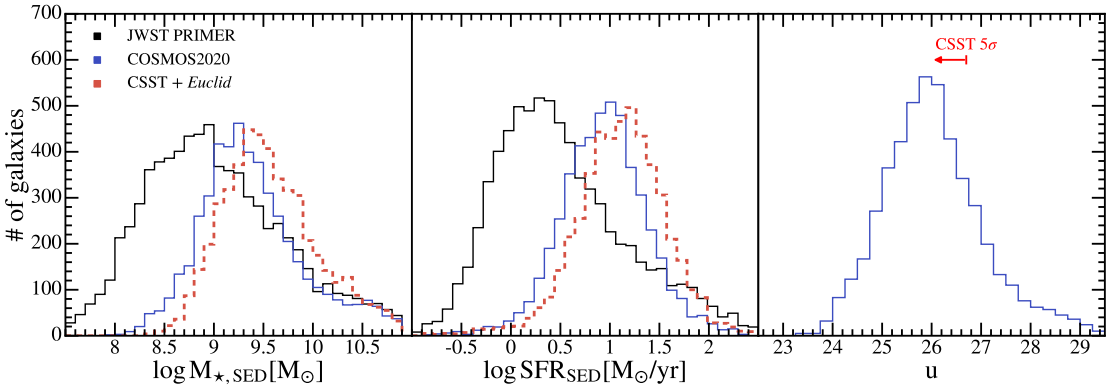

Fig. 2. Observational properties of the galaxy samples used throughout this work, selected according to the criteria in Sect. 2.4. *Left*: Stellar mass distribution of the JWST training set (black), the COSMOS2020 test sample (blue; 5000 sources randomly drawn from the full sample) and the mock CSST+*Euclid* catalog (dashed red; sharing the same sources as the blue sample but with parameters derived from reduced photometry). *Middle*: SFR distribution of the three samples. *Right*: The magnitude distribution of COSMOS2020 sample where we show the depth at 5σ for CSST.

comparison with the full-band photometry, allowing us to assess the extent to which SOM- and SED-based estimates can approximate the "true values" derived from comprehensive photometric datasets. Here, we focus solely on evaluating the performance of parameter estimation using photometry from the selected filters, without strictly requiring the source selection to match the target population of CSST observations. Nonetheless, we note that the majority of COSMOS2020 galaxies fall within the observable range of the CSST deep survey (see Fig. 2).

2.4. Selection functions

The JWST catalog is limited to sources with $S/N_{F444W} > 7$ and requires nonzero detections in all nine bands. The former criterion is adopted empirically to exclude spurious detections, while the latter is imposed due to the current model's inability to handle missing data (see Sect. 5.2). We also exclude objects with $S/N_{F770W} < 2$, which is crucial in determining stellar mass and constraining SFR (Wang et al. 2024a). Additionally, we constrained the training sample to the redshift range $1.5 < z < 2.5$, as there is a lack of rest-frame UV-optical bands for lower-redshift galaxies, which are essential for reliable SFR estimation. Notably, we did not incorporate HST bands to train our SOM, as the photometric depth differences and measurement uncertainties between the HST and JWST could potentially affect clustering results. However, we did include them during SED fitting to derive accurate physical properties. In principle, it would be feasible to extend the sample coverage to lower redshifts using available optical and UV bands and to higher redshifts with larger fields and more samples (see Sect. 3.1.2). After applying the selection criteria, the JWST training sample consisted of 7507 galaxies.

For the COSMOS2020 catalog and the mock catalog, we applied a 3σ cut, corresponding to a $K_s < 24.2$ and require non-zero detection in all twelve broadband filters. To maintain structural consistency between the two datasets, we limited the sample to the redshift range $1.5 < z < 2.5\,(\pm 0.1)$. A random subset of 5000 sources was extracted from the selected sample as the test set and the basis of the mock catalog. For the mock catalog, despite the differences in observational depths between CSST and COSMOS2020 due to instrument variations, the majority of our sample lies within the 5σ detection limit of the u-band for CSST's deep field survey. Therefore, in this study, the mock catalog and the COSMOS2020 catalog share the same set of sources, differing only in the selection of filters used during SED fitting. Unless explicitly stated, all parameter estimates from COSMOS2020 are derived from the full-band photometry (the "true values"), which serves as the reference for comparison with our SOM-based estimates.

In Fig. 2, we show the parameter distribution of galaxy samples used throughout this work. In the left and middle panels, we show that the composition of the observational samples from JWST and COSMOS2020 differs in parameter space. The latter tends to favor high-$M_\star$, high-SFR galaxies, with median values exceeding those of the JWST sample by 0.3 dex in $M_\star$ and 0.6 dex in SFR. Overall, however, its distribution range is well represented by the JWST sample. In the absence of K_s, ch1, and ch2 bands, the SED fitting results for $M_\star$ and SFR from the mock CSST+*Euclid* catalog show some degree of overestimation, which we discuss in Sect. 4.2.

3. Methods

3.1. Parameter estimation in a traditional SOM-based method

3.1.1. Self-organizing maps

Self-organizing maps (SOMs; Kohonen 1982) are typically employed for classifying and analyzing multidimensional galaxy photometric data. SOM-based parameter estimation involves two main steps: unsupervised classification and supervised labeling. Below, we outline the algorithm used during the unsupervised phase. In the unsupervised training phase, the SOM is governed by a well-defined mathematical framework consisting of initialization, competition, cooperation, adaptation, and iteration. First, a 2D grid of nodes is constructed, each associated with a prototype vector of galaxy colors. These prototype vectors, $\boldsymbol{w}_j = (w_{1,j}, w_{2,j}, \ldots, w_{N_{\rm dim},j})$, hereafter referred to as "weights," are initialized using principal component analysis (PCA) and iteratively refined through a competitive learning process. During each iteration, an input vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_{N_{\rm dim}})$ is presented, and each neuron computes its discriminant function to measure similarity via

$$\Delta = \sqrt{\sum_{N_{\rm dim}} (c_i - w_i)^2}\,, \tag{1}$$

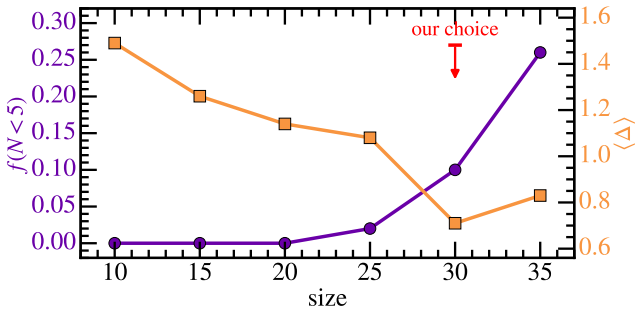

Fig. 3. Undersampling fraction (the fraction of SOM cells containing fewer than five galaxies; purple) and quantization error (the average Δ in Eq. (1); orange) as a function of SOM size, given a limited data volume of 7507 galaxies. While the undersampling fraction increases monotonically with size, the quantization error first decreases and then increases, reaching an optimal value at a SOM size of 30×30.

where the neuron with the smallest Δ is identified as the best matching unit (BMU).

Following the competition step, a cooperation phase is introduced. In this phase, not only the best matching unit (BMU) but also its neighboring neurons in the 2D lattice are updated to reduce their distance from the input vector to better fit the training data. The degree of adjustment decreases with both spatial distance from the BMU and iteration time, following a Gaussian function where the neighborhood radius T shrinks exponentially,

$$T_{j,I(c)}(t) = \exp\left(-\frac{S^2_{j,I(c)}}{2\sigma(t)^2}\right). \quad (2)$$

Here, $S_{j,I(c)}$ is the distance between the neuron, j, and the BMU $I(c)$ of input $\boldsymbol{c}$. The learning rate $\sigma(t) = \sigma_0 \exp(-t/\tau_\sigma)$ decreases over time, where t denote the current number of iterations, τ_σ and σ_0 are free parameters controlling the rate of decay. This ensures that distant nodes are updated less significantly. These updates progressively align the SOM with the underlying structure of the input data, resulting in a topologically ordered grid where similar galaxy colors are expected to be clustered together.

Finally, the process iterates up to the maximum number of iterations, during which the learning rate gradually approaches zero, ultimately yielding a topologically ordered and stable map. Galaxies associated with the same BMU are expected to exhibit similar SEDs, aside from differences in brightness (i.e., accounted for by a normalization factor) and the scatter introduced by photometric uncertainties. In this work, we utilized the Python library miniSOM (Vettigli 2018) for SOM construction and training.

3.1.2. SOM training

In this work, the input parameter space consists of eight dimensions, with each representing a observed-frame adjacent color of galaxies. These colors, derived from the photometry of the JWST PRIMER catalog (see Sect. 2.1), are: F090W–F115W, F115W–F150W, F150W–F200W, F200W–F277W, F277W–F356W, F356W–F410W, F410W–F444W, and F444W–F770W. The colors were normalized to have zero mean and unit variance, ensuring that each color contributes equally during training, independent of its intrinsic range or scatter.

The performance of a SOM depends on the user-defined map size and geometry, which should be chosen to balance adequate data sampling with sufficient resolution (i.e., minimizing quantization error). For the SOM geometry, Davidzon et al. (2019) demonstrated that a square topology yields the most effective parameter estimates and we followed their recommendation accordingly. To determine the optimal map size, we started with a 10×10 grid and gradually increased the size in increments of five. For each configuration, we assessed the quantization error (i.e., the average Δ of the SOM) and the fraction of the undersampling cells (i.e., the cells that associate to fewer than five galaxies). In general, the average number of galaxies per cell decreases with increasing SOM size due to the limited data volume; whereas the quantization error decreases initially, but rises again beyond a map size of 30 (see Fig. 3). Therefore, we adopted a relatively moderate SOM size of 30×30, smaller than those used in Davidzon et al. (2022), Torre et al. (2024), to maintain a reasonable trade-off between parameter resolution and robust sampling.

Our final 30×30 SOM is shown in Fig. 4. The galaxy distribution across the map is relatively uniform, with 90% of cells containing statistically sufficient samples (more than five galaxies), and only two cells being empty (see Fig. 4, left panel). To evaluate the accuracy of the SOM in representing the JWST data manifold, we computed the Euclidean distance between each galaxy and its corresponding cell weight using Eq. (1). Across the cell the average Δ per cell is < 1 (Fig. 4, middle panel).

For marginal cells with relatively high Δ values, we find that they predominantly contain extreme or low-S/N galaxies. This phenomenon is consistent with the boundary effect observed in previous studies (Davidzon et al. 2019), where irregular samples tend to populate the outermost regions of the map. An additional metric to evaluate the clustering performance, incorporating photometric uncertainties, is the reduced χ^2 distance, defined as

$$\chi^2_{\rm SOM} = \sum_{N_{\rm dim}} \frac{(c_i - w_i)^2}{\sigma_i^2}, \quad (3)$$

where σ_i is the normalized photometric error of the ith color. Despite the presence of large uncertainties concentrated near the edges, our analysis indicates significantly lower χ^2 values compared to Davidzon et al. (2022). This demonstrates that, despite the smaller sample size, reduced SOM dimensions, and fewer colors, the precise photometry and broad wavelength coverage of JWST substantially enhance the clustering accuracy of SOM. As shown in Sect. 3.1.3, regions with high quantization error, primarily located in the lower-right corner of the SOM, correspond to massive, older, and quiescent galaxies. The clustering of these galaxies is significantly influenced by the diversity of dust emission in the mid-infrared, which, given the current JWST sample size, cannot yet be resolved in fine detail.

We further verified that the distributions of the eight colors in the input dataset are consistent across SOM cells, ensuring that the SOM accurately represents the input data. Figure A.1 shows the distribution of the eight weight dimensions (representing the normalized colors) across the map. The deviation between the median colors and the SOM vectors within each cell (see Fig. A.2) is defined as the difference between the two, normalized by the scatter of the respective color. Across all cells, the deviations remain significantly smaller than 1σ, with no notable differences among colors, confirming that each color contributes equally to the classification.

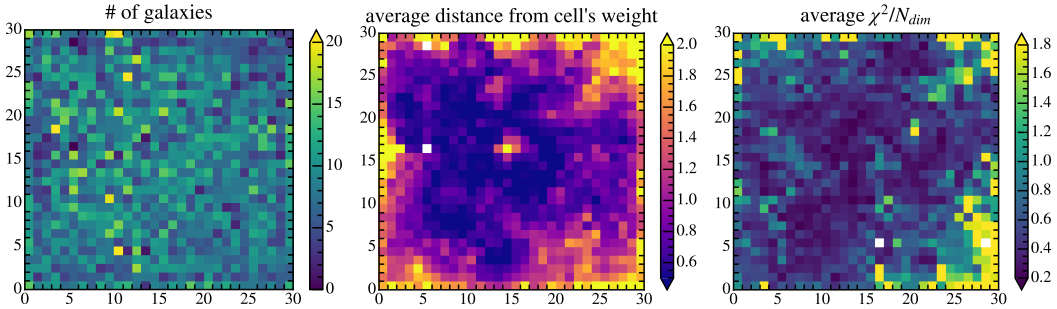

Fig. 4. SOM of JWST PRIMER galaxies at $z \sim 1.5$–2.5 selected as described in Sect 2.4. *Left*: Number of galaxies per cell. *Middle*: Similarity between galaxies in a given cell and the corresponding SOM weight, quantified using Eq. (1). *Right*: Similarity when incorporating photometric errors. Compared to Davidzon et al. (2022), the scatter is significantly reduced when applying the JWST catalog.

3.1.3. SOM pixel labeling

Galaxies with similar colors assigned to the same or neighboring cells are expected to exhibit similar physical properties (e.g., $M_\star$ and SFR), which can be used to characterize the properties of each cell (hereafter referred to as the "label"). Unlike Davidzon et al. (2019), where the median value was adopted for labeling, we chose to label the SOM pixels using the mode of each parameter distribution (i.e., the value corresponding to the peak of the distribution within each cell). Since $M_\star$ and SFR vary with the amplitude of a galaxy's spectrum, which is not explicitly modeled in our color-trained SOM, we normalized these values by rescaling them to a reference magnitude of F444W = 26 via

$$\mathrm{nX} = \mathrm{X} \times 10^{-0.4\times(26-\mathrm{F444W})}, \tag{4}$$

where X is the galaxy parameter to be normalized and F444W represents the F444W-magnitude for each individual galaxy. The F444W band was chosen for the normalization throughout the analysis, as for galaxies in the redshift range $1.5 < z < 2.5$, it contains fewer emission lines and is predominantly continuum-dominated. Additionally, the PRIMER survey provides the deepest imaging in this band, yielding the highest S/N ratio for sources.

The left panels of Fig. 5 show the distribution of normalized stellar mass ($\mathrm{n}M_\star$) and SFR (nSFR) across the SOM. We also checked the distribution width δ for each parameter as shown in the right panels of Fig. 5. The width is quantified as the difference between the 84th and 16th percentiles the galaxy properties within each pixel. The $\delta/2$ represents the systematic uncertainties $\sigma_{\rm sys}$ associated with the redshift and stellar population parameters derived using SOM. Specifically, we set a lower limit of 10^{-1} $\mathrm{M}_\odot$/yr for nSFR labels. This threshold is motivated by the fact that quiescent galaxies can exhibit arbitrarily low SFRs, causing their labels to deviate significantly from the main sequence and often appear as outliers in logarithmic space. Including such extreme values in the label maps may bias the weighted averaging process and destabilize subsequent parameter estimation. Nonetheless, we find that the SOM effectively distinguishes between star-forming and passive populations, while the choice of this lower limit primarily affects only a small fraction ($\lesssim$5%) of outlier galaxies in the SFR estimates, with a negligible impact on the bulk of the population.

It is evident that the scatter of physical parameters within the cells is generally less than 0.5 dex, with most central cells exhibiting a scatter below 0.2 dex, indicating good clustering performance across the SOM. Comparison with Fig. 4 reveals that cells with larger scatter are primarily located at the periphery, where observational uncertainties dominate. The most significant scatter is concentrated around low-SFR and high-$M_\star$ cells, where the inherent challenges in estimating SFR make robust measurements impractical and beyond the scope of this study.

3.1.4. Parameter estimation in the traditional method and its limitation

Before applying the method to the COSMOS2020 catalog, we first conducted a sanity check by randomly dividing the JWST sample into two subsets with a 7:3 ratio, designated as the training and test sets, respectively. Following the procedure described above, we trained a 25×25 SOM given the reduced size of the training data. We follow Davidzon et al. (2022) to estimate $M_\star$ and SFR for the test sample by utilizing the labels of the five cells with the closest weights to each galaxy (Davidzon et al. 2019, 2022). The parameter estimates were derived from the weighted mean of the target labels and denormalized from the reference magnitude F444W=26, following the equation,

$$\mathrm{X_{SOM}} = \frac{\sum_{j=1}^{5} \frac{1}{\Delta_j} \mathrm{nX_{label}}}{\sum_{j=1}^{5} \frac{1}{\Delta_j}} \times 10^{+0.4\times(26-\mathrm{F444W})}, \tag{5}$$

where Δ_j is defined in Eq. (1) and represents the j th neighbor among the five nearest cells (including the BMU).

The left panel of Fig. 6 compares the stellar mass and SFR estimates obtained using SED fitting and the SOM-based method. We also present five statistical metrics for each parameter: normalized median absolute deviation (NMAD), outlier fraction (η), root mean squared error (RMSE), and bias. The outlier fraction, η, is the percentage of objects satisfying $|\log \mathrm{X_{SOM}} - \log \mathrm{X_{SED}}| > 2\sigma$, where σ is the standard deviation of the residuals; namely, log(predicted) – log(true). It is evident that the parameter estimates derived using the SOM method closely match those obtained through traditional SED fitting, with dispersions of $\sigma_{\rm NMAD} \sim 0.12$ and 0.18 for the stellar mass and SFR, respectively. Moreover, the computational cost of the SOM approach is reduced to the order of a few CPU minutes, demonstrating its efficiency (see Table 1).

Despite its computational efficiency and the agreement of its estimates with those from SED fitting, the traditional SOM-based method has notable limitations. It is important to recall that in SOM, galaxy colors are used to generate corresponding weights, with each component representing a specific color. During the parameter estimation, the input colors are compared with the weights assigned to each cell. Consequently, the applicability

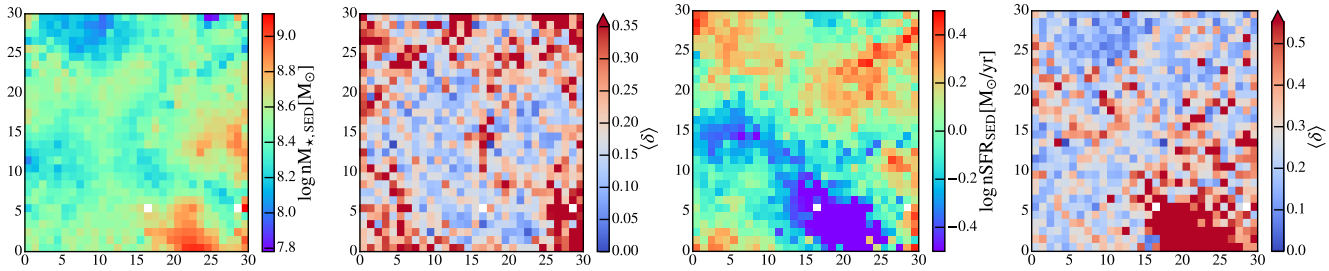

Fig. 5. Label maps of normalized stellar mass and SFR. For each pair, the left panel displays the SOM grid color-coded by the mode of the corresponding parameter (stellar mass or SFR), while the right panel shows the distribution width, defined as the 84th–16th percentile range $\langle\delta\rangle = \langle 84\% - 16\% \rangle$, within each cell.

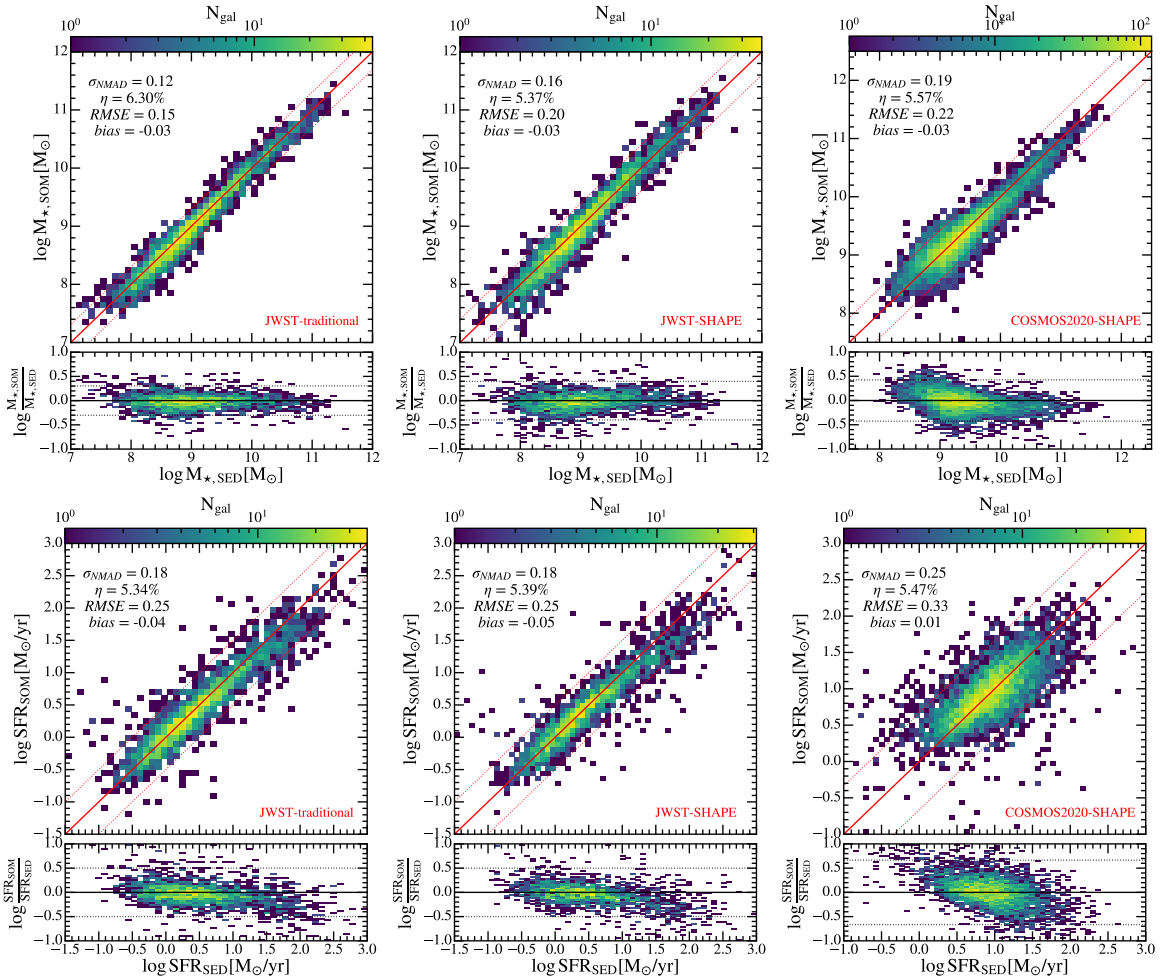

Fig. 6. Comparison between stellar mass, $M_\star$, and SFR obtained through standard template fitting ($X_{\rm SED}$) and different SOM-based methods presented in this work ($X_{\rm SOM}$, with the specific model highlighted in red in the lower right corner). From left to right, the panels show the application of the traditional SOM method to the 7:3 JWST sample, the application of the 25×25 hybrid method to the 7:3 JWST sample, and the application of the 30×30 hybrid method to the COSMOS2020 sample. The solid red line represents the bisection line, while the red dashed lines indicate the boundaries of the outlier.

of this method is restricted to galaxy samples that meet the following criteria: (1) they used JWST filters and (2) all nine bands used in the training set must have detections. As a result, high-precision training results have limited generalizability to other datasets. At the final stages of our work, Torre et al. (2024) and Rejeb et al. (2023) independently proposed methods to mitigate the issue of missing values during training and parameter estimation. However, these methods cannot fully address the challenge posed by mismatched filters and thus have not been incorporated into this study. We plan to integrate their approaches in future work to address the missing data issue during the training phase.

The core limitation of the traditional method lies in its simplification of continuous SEDs into discrete color information. However, for galaxies with similar colors, their SEDs are expected to be correspondingly similar. To overcome this limitation, we extended the discrete model to a continuous framework by fitting SEDs for each cell based on the results of galaxy clustering. This extension allows us to generalize beyond the discrete

colors used in the training phase, offering a more flexible and robust approach to parameter estimation.

3.2. Parameter estimation with SHAPE

3.2.1. Introduction of the hybrid approach of SOM and SED fitting

Building on the above concept, we introduced SHAPE, which utilizes the SOM to cluster galaxies from the JWST catalog and generate representative average SEDs for each SOM cell. In Fig. 1, we show a schematic diagram of our model. The model can be roughly divided into three parts, the galaxy automatic classification, construction of the SED Lib and derivation of a new SOM from SED Lib and selective filters. The first part follows the framework of the traditional method, where, if the filters of the test set match those of the training set, galaxies can be directly mapped onto the SOM for parameter estimation (as discussed in Sect. 3.1).

The second part builds upon the galaxy clustering in the first step by incorporating SED templates. From this step onward, the method is no longer purely data-driven, as theoretical models are incorporated. Here, we briefly summarize the several advantages of this hybrid method compared to direct SED fitting applied to individual galaxies: (1) by clustering the observational data, it eliminates unphysical or unrealistic template combinations; (2) stacking the fluxes enhances both the S/N ratio and wavelength coverage of the photometric points, yielding more accurate SEDs within the SED Lib; and (3) the method retains the computational efficiency of the SOM, following the same principles during parameter estimation as in traditional SOM methods.

Once the SED Lib is constructed, each cell corresponds to a normalized SED. In the third step, the filters used in the test set are applied to the SEDs to derive a corresponding set of colors. These colors replace the original SOM weights, forming a new SOM (marked as SOM* hereafter) with the components perfectly matching the colors of the test sample. This approach allows for flexible selection of photometric bands to be used during the parameter estimation. Below, we provide a detailed explanation of each step in the process.

3.2.2. Construction of the SED Lib and derivation of the new SOM

The construction of the SED Lib follows a series of steps: first, we normalized the flux and error in all filters by the flux in the F444W band for each galaxy. Next, we calculated the error-weighted flux ($\bar{f}_i$) and error ($\bar{\sigma}_i$) in each filter as the stacked photometry via

$$\bar{f}_i = \frac{\sum_j (w_{i,j} f_{i,j})}{\sum_j w_{i,j}}, \quad \bar{\sigma}_i = \sqrt{\sum_j 1/w_{i,j}}, \tag{6}$$

where $w_{i,j} = 1/\sigma^2_{i,j}$, $f_{i,j}$, and $\sigma_{i,j}$ denote the normalized flux and its associated uncertainty in the i th filter of j th galaxy[3]. We emphasize that this weighted approach, though enhancing the S/N ratio, may potentially bias the stacked SEDs toward brighter and, thus, more massive galaxies. In Fig. B.1, we quantify the differences between the mode stellar mass and SFR of individual galaxies in each SOM cell and the corresponding values derived from the stacked SED templates, denoted as $\Delta_{\rm mode-stack}$. We find only a slight offset typically below 0.1 dex, indicating that this effect has a negligible impact on our results.

Following the setup described in Sect. 2.1, we performed an SED fitting with BAGPIPES, ultimately synthesizing a total of 30×30 SEDs. Then we perform photometry on the SEDs in the SED Lib using selective filters, constructing an equivalent new SOM that aligns with the test data's filter set. In this test, we select 12 bands (e.g., u, g, r, i, z, y, Y, J, H, K_s, ch1, and ch2) from the COSMOS2020 catalog, consistent with those used in Davidzon et al. (2022), Torre et al. (2024), for demonstration and direct comparison. The new component map and color input are normalized by the standard deviation of the colors corresponding to the 900 SEDs in the SED Lib (including two empty cells; see Fig. A.3).

As previously mentioned, when labeling physical quantities such as the $M_\star$ and SFR (both of which are influenced by the amplitude of the spectrum), these quantities are normalized to F444W = 26. Then during parameter estimation, they are denormalized using the observed F444W magnitude. Since practical applications may lack F444W detections, the SED Lib is employed to correct the label maps via

$$\begin{aligned} X_{[x,y]} &= \mathrm{n}X_{[x,y]} \times 10^{+0.4\times(26-\mathrm{F444W}_{[x,y]})}, \\ \mathrm{n}X^*_{[x,y]} &= X_{[x,y]} \times 10^{-0.4\times(m_0-m_{[x,y]})}, \end{aligned} \tag{7}$$

where, in the pixel $[x, y]$, $\mathrm{nX}_{[x,y]}$, and $\mathrm{nX}^*_{[x,y]}$ represent the normalized parameters in the original and corrected label maps, respectively. Then, $\mathrm{F444W}_{[x,y]}$ and $m_{[x,y]}$ correspond to the photometric magnitude of the SED assigned to the given pixel, while m_0 is the reference magnitude. We tested different filter choices for normalization and found that using the band closest to F444W yields the most accurate estimates, likely because it minimizes errors introduced by the fitted SED during the label map correction process. Accordingly, we adopted the ch2 band as the normalization reference, with a fixed magnitude of $m_0 = 26$.

3.2.3. Parameter estimation from the SED Lib

Once we have an adaptive SOM that galaxies can be directly mapped onto, then we move on to the mapping method. Following Torre et al. (2024), we considered the impact of photometric uncertainties. Although in this work, we find that the final statistic metrics of the parameter estimates are broadly consistent with those obtained using the five-point matching method in Davidzon et al. (2019) and Sect. 3.1.4, this probabilistic framework allows new samples to be matched to the SOM in the form of a probability distribution, obtaining a representative average SED. For any given galaxy, the likelihood, $\mathcal{L}$, of being mapped to a SOM pixel at coordinates [x,y] is defined via

$$\mathcal{L}_{[x,y]} = e^{-\frac{\chi^2_{[x,y]}}{2}},$$

with

$$\chi^2_{[x,y]} = \sum_{N^*_{dim}} \left(\frac{(c_i - w^*_i)^2}{\sqrt{\sigma^2_{c_i} + \sigma^2_{\rm sys}}}\right)_{[x,y]}, \tag{8}$$

where w^*_i is the i th component of the updated SOM weights after SED replacement, σ_{c_i} represents the photometric uncertainties of c_i, and $\sigma_{\rm sys}$ represent the parameter

[3] An alternative approach involves stacking the best-fit SEDs of individual galaxies. However, in certain sub-classified SOM cells, this method can yield unphysical composite SEDs, thereby compromising the reliability and generalizability of the resulting SED library. For this reason, we did not adopt this strategy.

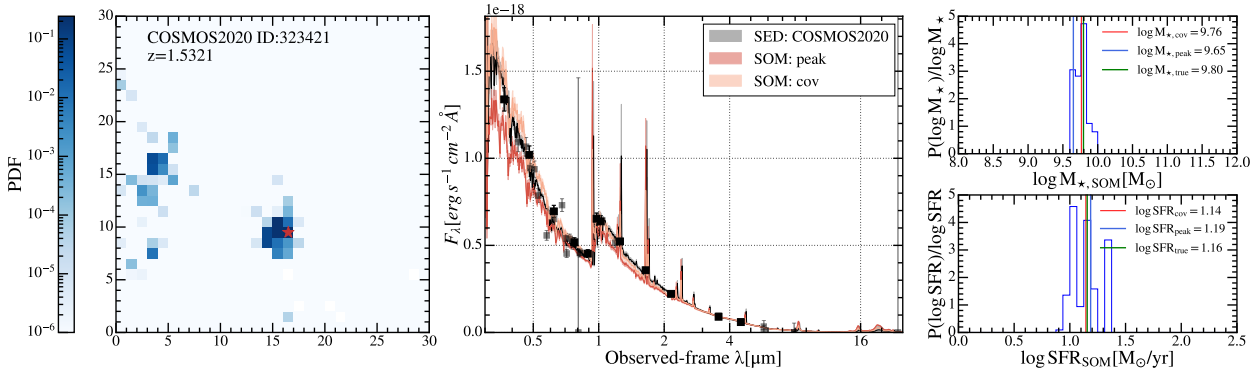

Fig. 7. Example of parameter estimation using SHAPE. *Left*: Probability distribution of galaxy ID323421. The target is mapped onto this surface according to the likelihood, with the red star marking the best-matching grid. *Middle*: Comparison between the SEDs obtained via SED Lib matching (dark and light red) and the synthetic templates (black). The shaded regions represent the $16^{\rm th}$ – $84^{\rm th}$ uncertainty range. Black dots with error bars denote the photometric points used as input (u, g, r, i, z, y, Y, J, H, $K_{\rm s}$, ch1, and ch2), while transparent dots represent additional bands used in SED fitting but not included in SOM. *Right*: Probability distribution of parameter estimates. The vertical red line represents the estimates obtained by convolving the probability distribution with the label maps, the blue line represents the estimate from the best-matching cell, and the green line corresponds to the SED fitting result.

Table 1. Comparisons of parameter estimates for different catalogs using various methodologies.

Catalog	Band	Method	Bias	$\sigma_{\rm NMAD}$	η^* [%]	$t_{\rm proc}$ [h/CPU/1000]
JWST/PRIMER	F090W, …F770W (9)	SOM25	–0.03 | –0.04	0.12 | 0.18	0.7 | 7.3	∼0.001
JWST/PRIMER	F090W, …F770W (9)	SHAPE25	–0.03 | –0.05	0.16 | 0.18	2.8 | 7.3	
COSMOS2020/full-phot	u, g, …ch2 (12)	SHAPE30	–0.03 | 0.01	0.19 | 0.25	3.1 | 11.6	
COSMOS2020/miss-phot	u, g, … H (9)	SHAPE30	–0.02 | –0.01	0.20 | 0.25	3.5 | 11.2	
COSMOS2020/miss-phot	u, g, … H (9)	BAGPIPES	0.19 | 0.18	0.22 | 0.35	19.2 | 14.1	∼10

Notes. Three key statistical metrics for the derived parameters are reported in the order of $M_\star$ and SFR. For a fair comparison, the outlier fraction η^* is defined as the fraction of objects with $|\log {\rm X_{SOM}} - \log {\rm X_{SED}}| > 0.5$ dex. The computational efficiency ($t_{\rm proc}$) of each model inevitably depends on factors such as CPU architecture and the complexity of the SED fitting templates, thus, we provide only an order-of-magnitude estimate as a reference.

systematic uncertainty, which in this context, is defined as $(\delta_{\rm M}/\log {\rm nM_{label}})_{[x,y]}$. The likelihood surface is normalized by dividing $\mathcal{L}_{[x,y]}$ by its sum. From these likelihood surfaces, combined with the pixel labels for a given parameter, we can derive probability distribution functions for various physical parameters.

In Fig. 7, we present an example of parameter estimation using SHAPE. In the left panel of Fig. 7, we show the probability distribution for a given galaxy. The middle panel presents a comparison of SEDs obtained using two different matching methods with 12 colors as input, against the SED derived from a full-band SED fitting. The solid dark red line represents the SED from the SED library that best matches the input colors, while the light solid red line corresponds to the probability-weighted SED. The solid black line represents the SED fitting result. The shaded region encompasses the 1σ uncertainty range. Black dots with error bars represent the photometric points used as input, while transparent dots indicate additional photometric points used in the SED fitting but not included in SOM. In the right panel, we present the probability distribution of parameter estimates. The vertical red line represents the estimates obtained by convolving the probability distribution with the label maps, the blue line represents the estimate from the best-matching cell, and the green line corresponds to the SED fitting result. The comparison demonstrates that the matched SEDs and estimated parameters closely align with the SED fitting results, further validating the effectiveness of our approach.

4. Results

4.1. Stellar mass and SFR estimates

In this section, we present the results of parameter estimation using SHAPE. Direct comparisons are shown in Fig. 6 and Table 1.

Before directly applying SHAPE to the COSMOS2020 test data, we first check whether it's still effective compared to the traditional SOM method by evaluating the estimation quality in the 7:3 sample split of the JWST catalog in Sect. 3.1.4. Strictly following the procedures outlined in Sect. 3.2, we construct a 25×25 SED Lib and select the JWST bands (F090W, F115W, F150W, F200W, F277W, F356W, F444W, F410M, and F770W) to reconstruct the component maps. For the batch processing, we performed parameter estimation by convolving the probability distribution with the label map. To mitigate potential biases, we excluded cells with probabilities below 1%. In the middle panels of Fig. 6, we present a comparison of the stellar mass and SFR estimates obtained using the 25×25 SOM* with those derived from SED fitting. For each estimate, four statistical metrics are evaluated. The boundaries of outliers are indicated by dashed red lines. Both estimates show good agreements with the SED fitting results and only a very slight increase in $\sigma_{\rm NMAD}$ and bias is observed, indicating that the error introduced by the SED Lib is minimal.

In the right panels of Fig. 6, we present the estimation results of applying 30×30 SOM* in Sect. 3.2.2 to the COSMOS2020

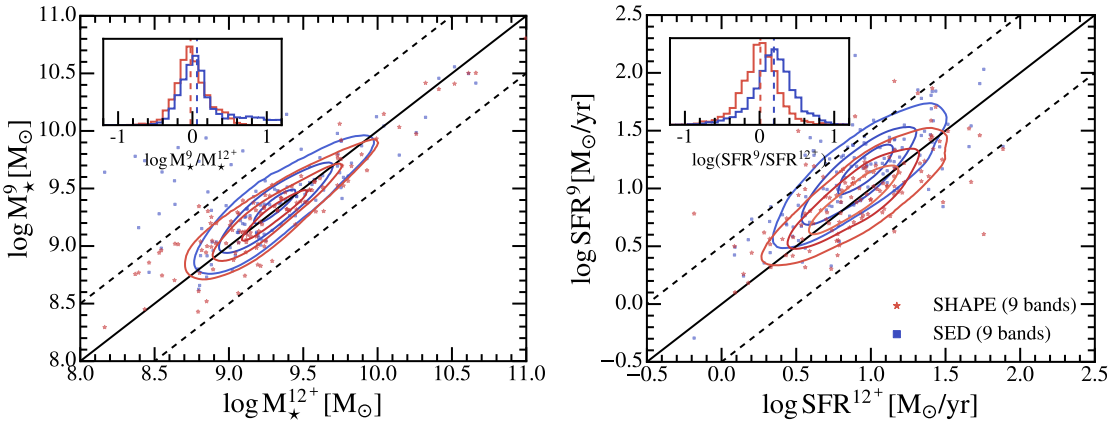

Fig. 8. Comparison between SOM- and SED-derived estimates using the nine-band photometry (red and blue, respectively), with 12+-band photometry values as a reference. The solid line represents the zero offset, while the two dotted lines indicate ±0.5 dex boundaries. A random selection of 100 galaxies is used for visualization. The same color scheme is applied to the histograms in the inset, showing the respective distributions.

dataset through SHAPE. In the left panel, the $M_\star$ estimates exhibit a strong agreement with the SED fitting results, despite the SOM method utilizing fewer photometric bands. The standard deviation of the differences is within 0.2 dex, with only 5.6% of outliers. No significant bias is observed, which can be attributed to the completeness of the training sample based on the JWST PRIMER catalog. This ensures that even low-mass galaxies in the COSMOS2020 catalog can find well-matched SEDs within the SED Lib.

In the right panel, the SFR estimates show slightly larger deviations compared to the stellar mass estimates but still demonstrate a strong 1:1 correlation with the SED fitting results, with 0.25 of $\sigma_{\rm NMAD}$ and 5.5% of outliers. A decline in accuracy is observed at the high-SFR end. This behavior arises due to the fact that, within SOM, the classification accuracy for specific galaxy types and the resolution of parameter estimation (i.e., the ability to distinguish between different parameter combinations) are influenced by the representation of such galaxies in the training set. Galaxies with lower representation correspond to less SOM cells, leading to decreased precision in parameter estimates. Specifically, the JWST catalog contains a larger proportion of low-$M_\star$ and low-SFR galaxies, enhancing the sensitivity of the model to such populations. However, for galaxies experiencing strong dust attenuation (primarily located in the lower right region of the SOM), their SEDs are affected by prominent emission lines in certain bands, introducing intrinsic complexity. A sufficiently large number of such galaxies is required to accurately account for dust emission effects. Another challenge arises from the incomplete coverage of galaxies in the FIR and sub-mm, as not all galaxies have detections in these wavelengths. Consequently, SFR estimates derived from SED fitting may exhibit inaccuracies in such cases.

4.2. Comparison to synthetic templates

To prepare for future large-scale surveys such as CSST, we assess the performance of SHAPE and SED fitting for parameter estimation using a fixed set of nine photometric bands that correspond to the observational bands of CSST and *Euclid*. These estimates are then compared to those obtained from SED fitting using full photometric coverage. In this analysis, we select the nine bands (e.g., u, g, r, i, z, y, Y, J, H) and repeat the procedures outlined in Sect. 3.2.3 to obtain SOM-based estimates. Here we adopt the H band for label map correction via Eq. (7), as the ch2 band is not available in CSST and *Euclid* photometric systems. For a fair comparison, we also perform SED fitting with the same nine bands with no ancillary data to derive a second set of SED-based estimates (see Sect. 2.3). In this analysis, FIR and sub-mm detections are not explicitly required, as such selection criteria would introduce a bias toward high-SFR and heavily dust-attenuated galaxies, which are underrepresented in the JWST PRIMER field. However, by incorporating additional bands such as Ks, ch1, ch2, and medium/narrow bands, SFR estimates derived from 12+ bands can be considered more robust compared to those obtained using nine-band photometry alone.

In Fig. 8, we compare the SED- and SHAPE-derived estimates using 12 (or more) band SED fitting results as a reference. The red contours represent the distribution of SHAPE-based estimates, while the blue contours denote those obtained from SED fitting. To avoid excessive overlap, we randomly select 100 galaxies for visualization. For each estimate, we calculate the discrepancies between the estimated values and the true values, present them in the inserted panel. In the left panel, both SED- and SHAPE-derived $M_\star$ show no significant offset from the reference values, although SED fitting displays a subset of galaxies with overestimated $M_\star$, a trend not observed in the SOM-based estimates. In the right panel, SFR$_{\rm SOM}$ aligns well with the reference values at the low-mass end with a bias of -0.01, whereas SED fitting shows a systematic bias of 0.18 dex due to the exclusion of the $K_{\rm s}$, ch1, and ch2 bands. For both parameters, the scatter of SOM-based estimates is slightly lower than that of the SED-based estimates.

Specifically, in Fig. 9, we present an example comparing the SED obtained through SOM matching with the SED derived from SED fitting. Since our SOM is imprinted with MIRI band information during clustering and labeling, the SOM-derived SED (red lines) demonstrates better consistency with the reference SED (black lines) at longer wavelengths compared to the SED derived from template fitting (blue lines). Quantitatively, when using the exact same filter set, 60% of galaxies have SOM-derived SFR values closer to the reference than those obtained from SED fitting. These results confirm the effectiveness of the SOM-based method for future large-scale survey projects.

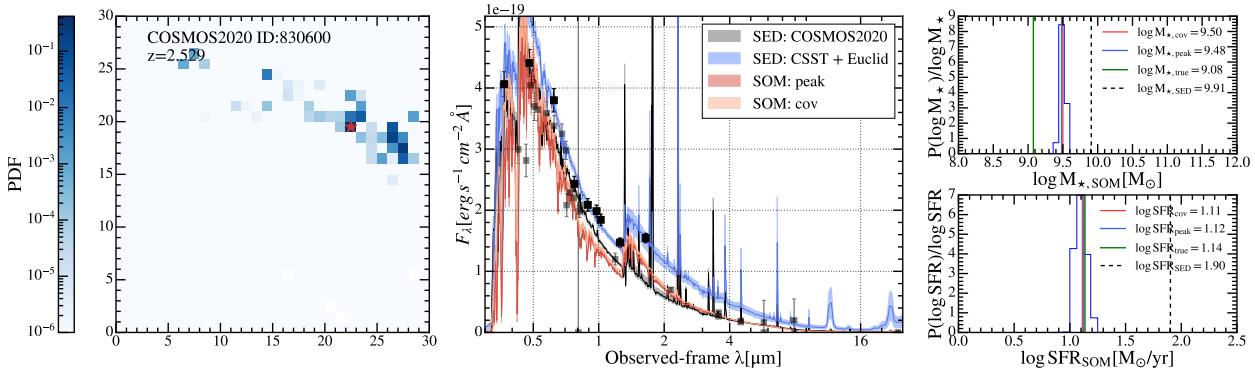

Fig. 9. Similar to Fig. 7, but using only the u, g, r, i, z, y, Y, J, H bands in SOM. The blue line in the middle panel represents the synthetic templates obtained using the same nine-band photometry. The dashed lines in the right panel correspond to the nine-band SED fitting estimates.

5. Discussions

5.1. Comparisons to previous work

This study builds upon the foundational work of Davidzon et al. (2019, 2022), with several key extensions. A major distinction from previous studies lies in our focus on utilizing JWST as a benchmark and bridging data-driven techniques with traditional SED fitting, integrating observational data with template-based modeling.

Despite employing a smaller SOM size that corresponds to a lower parameter resolution and incorporating SED fitting, which introduces certain model-fitting uncertainties, our model performs comparably well as the approach presented in Davidzon et al. (2022). The improvement is attributed to JWST's superior imaging quality and the inclusion of MIRI data. Specifically, for the $M_\star$ estimates, Davidzon et al. (2022) reported a symmetric scatter of 0.25 dex and that approximately 3% of galaxies exhibited significant underestimation (by more than a factor of 3); whereas in our work, the scatter is 0.20 dex and no significant bias or extreme outliers are observed. We cannot directly compare the quality of SFR estimates, as the labeling methods differ significantly.

For the parameter estimation, we adopted the probabilistic distribution approach proposed by Torre et al. (2024), effectively addressing the issue of missing data at this stage. However, during the model training phase, we do not follow their method of randomly sampling fluxes. This study primarily focuses on validating the methodology and, in future work, we plan to incorporate their approach to better address missing data during training, thereby expanding both the training sample and the SOM size.

In our study, the introduction of the SED Lib appears to act as an additional source of uncertainty. This is largely due to the limited sample size, which results in many SOM grid cells representing mixed or under-resolved subclasses of galaxies (i.e., different types of galaxies are not fully separated into distinct cells) and some types are too sparsely populated to form independent, well-defined nodes. When constructing the SED Lib, stacking SEDs within these under-resolved cells makes it difficult to capture the intrinsic features of each subclass. However, as the dataset expands and galaxy classification becomes more refined, SED Lib may transition from being a source of error to serving as a physical constraint, mitigating the impact of photometric uncertainties and potentially improving the overall accuracy of parameter estimation.

5.2. Limitations of this work

The primary limitation of this study arises from the restricted size of the training sample. Both Davidzon et al. (2022) and Torre et al. (2024) have demonstrated that an 80×80 SOM provides an optimal compromise. Given the higher redshift accuracy and broader stellar mass range of the JWST sample, a larger SOM size would ideally be required. However, due to our stringent selection criteria, the number of sources in our training set is significantly smaller than in previous studies (7507 galaxies compared to 174 522 and 228 524, respectively). We adopt a relatively small 30×30 SOM in the initial clustering step. This choice sacrifices classification accuracy and reduces the resolution of parameter estimation. The impact is most pronounced at the high-SFR end, where the number of massive dusty star-forming or starburst galaxies is only sufficient to populate approximately 50 grids. Consequently, dust emission is not well resolved in the current model. However, we anticipate that the model will be continuously refined as more data become available.

Another limitation associated with the relatively small data sample is that, although JWST reaches much deeper imaging depths and is theoretically capable of detecting smaller and fainter galaxies, the COSMOS2020 survey covers a substantially larger field. As a result, the latter includes a significantly higher number of very massive or extreme galaxies compared to the JWST training set, implying that the training sample may not be fully representative. Nevertheless, this issue is secondary, as SOMs are inherently limited in accurately estimating the properties of such extreme objects; these galaxies typically fall into peripheral cells with suboptimal clustering performance, commonly referred to as the boundary effect (Davidzon et al. 2019).

In addition, challenges may arise form the treatment of missing data and upper limits during the training process. This lack of a dedicated solution restricts the size of the training sample and potentially biases the selection, which affects the representativeness of the trained SOM. Specifically, requiring a non-zero detection in F090W may inadvertently exclude heavily dust-obscured galaxies, while imposing a high S/N requirement for F770W may result in the exclusion of low-mass galaxies. However, as we constrain our sample to the range $1.5 < z < 2.5$, the selection effects should be relatively limited. Moreover, addressing missing data ultimately involves the prediction of absent fluxes and clustering itself can be utilized as a potential method for flux prediction. In our forthcoming SHAPE II framework (Wang in prep.), we are developing an iterative approach wherein the parameter estimation and flux prediction are refined

cyclically through clustering, progressively enhancing the accuracy and completeness of the model.

5.3. Redshift estimation through SOM

Throughout this work, we evaluated the capability of SOM and SHAPE to estimate stellar mass and SFR, while noting that both methods underperform in recovering photometric redshift. Although some simulation-based studies (Davidzon et al. 2019; Torre et al. 2024) have argued that SOMs can yield reliable redshift estimates, empirical applications trained on observational data tend to face similar challenges. For example, Davidzon et al. (2022) did not show the redshift estimates but fixed the redshifts of their sources prior to estimating other physical properties. Abedini et al. (2025) also reported that recovering photometric redshifts using SOMs remains challenging.

Our strategy is also to treat photometric redshift as a prior and use it to bin the data, rather than as a target parameter for direct inference. This approach is motivated by two key considerations. First, SOMs are primarily sensitive to the continuum shape of galaxy SEDs and identify redshifts through broad spectral features such as the Lyman and Balmer breaks; they are generally less effective than SED fitting at capturing narrow emission-line information[4]. Second, reliable redshift estimation via SOM requires sufficient sampling density in the 3D z–$M_\star$–SFR space. That is, a SOM node must be populated by at least two galaxies with nearly identical SEDs and redshifts to define a meaningful cluster, a requirement not met by the current JWST training sample. An alternative approach would be to incorporate spectroscopic redshifts to refine photometric redshift estimates (e.g. Masters et al. 2015; Hemmati et al. 2019; Zhang et al. 2025).

6. Conclusions

In this study, we explore the application of SOM to bridge different photometric systems, thereby leveraging the exceptional depth, quality, and wavelength coverage of the JWST PRIMER survey. We calibrated the estimation of physical parameters in upcoming wide-field surveys. We summarize our main findings below:

1. The SOM trained on JWST data provides a high level of accuracy in the parameter estimation. The SOM achieves parameter estimates that are nearly identical to those obtained from SED fitting, with $\sigma_{\rm NMAD} < 0.2$ for both $M_\star$ and SFR, while offering a significantly higher computational efficiency.
2. We introduce a hybrid method called SHAPE which replaces the default SOM weights with SED Lib, derived from galaxies in each cell. Utilizing the SED Lib, the SOM model can be extended to samples from different filter sets, such as COSMOS2020 and CSST. This enables parameter estimation even when the input color combinations do not exactly match the training data, as the SED Lib provides a continuous framework to replace discrete photometric points. This advancement significantly enhances the generalizability of the SOM method across multiple surveys, marking a major step forward in multisurvey parameter estimation.
3. We applied this method to JWST and COSMOS2020 test sample. For the JWST catalog, $\sigma_{\rm NMAD}$ increased slightly to 0.16 for stellar mass estimates. Therefore, the error introduced by SED Lib is shown to be minimal. For the COSMOS catalog, it demonstrates a strong agreement between $M_{\star,\,\rm SOM}$ and $M_{\star,\,\rm SED}$. While SFR estimates exhibit slightly larger discrepancies at the high end, the overall deviations remain within 0.5 dex, validating the reliability of our approach.
4. The SOM suggests its promising application in upcoming large-scale surveys, with its outstanding robustness and efficiency. With a limited set of bands (e.g., u, g, r, i, z, y, Y, J, H), our method produces unbiased SFR estimates than traditional SED fitting (with a bias of –0.01 and 0.18, respectively).

Overall, although this work is based on a relatively small JWST sample, it already demonstrates the potential of SHAPE in efficiently processing and analyzing large, high-dimensional astronomical datasets. In our next-generation of SHAPE II, future efforts will focus on addressing challenges from missing data and extending the model to other redshift bins, particularly through the incorporation of more comprehensive datasets as they become available from JWST (e.g., COSMOS-Web; Casey et al. 2023; Shuntov et al. 2025). With a hopefully more complete and robust SHAPE II, we aim to explore its utility in guiding future observations, identifying peculiar objects, and predicting fluxes across multiple bands.

Acknowledgements. This work was supported by National Natural Science Foundation of China (Grant No.12525302, 12173017 and 12141301), Natural Science Foundation of Jiangsu Higher Education Institutions of China(Grant No. BK20250001), National Key R&D Program of China (Grant no. 2023YFA1605600), Scientific Research Innovation Capability Support Project for Young Faculty (Project No. ZYGXQNJSKYCXNLZCXM-P3), the Fundamental Research Funds for the Central Universities with Grant no.KG202502, and the China Manned Space Program with grant no. CMS-CSST-2025-A04.

[4] Masters et al. (2015) and Zhang et al. (2025) successfully employed SOMs to calibrate photometric redshifts, but adopted a distinct methodology, and is therefore beyond the scope of this study.

References

Abedini, F., Gozaliasl, G., Zonoozi, A. H., et al. 2025, arXiv e-prints [arXiv: 2506.04138]
Arango-Toro, R. C., Ilbert, O., Ciesla, L., et al. 2025, A&A, 696, A159
Arnouts, S., Moscardini, L., Vanzella, E., et al. 2002, MNRAS, 329, 355
Ball, N. M., & Brunner, R. J. 2010, Int. J. Mod Phys D, 19, 1049
Baron, D. 2019, arXiv e-prints [arXiv: 1904.07248]
Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, A&A, 622, A103
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brammer, G., Strait, V., Matharu, J., & Momcheva, I. 2022, https://doi.org/10.5281/zenodo.6672538
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Buat, V., Giovannoli, E., Burgarella, D., et al. 2010, MNRAS, 409, L1
Byler, N., Dalcanton, J. J., Conroy, C., & Johnson, B. D. 2017, ApJ, 840, 44
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Cao, Y., Gong, Y., Meng, X.-M., et al. 2018, MNRAS, 480, 2178
Carnall, A. C., McLure, R. J., Dunlop, J. S., & Davé, R. 2018, MNRAS, 480, 4379
Casey, C. M., Kartaltepe, J. S., Drakos, N. E., et al. 2023, arXiv e-prints [arXiv:2211.07865]
Conroy, C. 2013, ARA&A, 51, 393
Cropper, M. S., Al-Bahlawan, A., Amiaux, J., et al. 2025, A&A, 697, A2
CSST Collaboration (Gong, Y. et al.) 2025, arXiv e-prints [arXiv: 2507.04618]
da Cunha, E., Charlot, S., & Elbaz, D. 2008, MNRAS, 388, 1595
Davidzon, I., Laigle, C., Capak, P. L., et al. 2019, MNRAS, 489, 4817
Davidzon, I., Jegatheesan, K., Ilbert, O., et al. 2022, A&A, 665, A34
Dunlop, J. S., Abraham, R. G., Ashby, M. L. N., et al. 2021, JWST Proposal. Cycle 1, ID. #1837,
Euclid Collaboration (Moneti, A., et al.) 2022, A&A, 658, A126
Euclid Collaboration (Mellier, Y., et al.) 2025, A&A, 697, A1
Ferland, G. J., Chatzikos, M., Guzmán, F., et al. 2017, Rev. Mex. Astron. Astrofis., 53, 385
Galvin, T. J., Huynh, M., Norris, R. P., et al. 2019, PASP, 131, 108009
Hao, C.-N., Kennicutt, R. C., Johnson, B. D., et al. 2011, ApJ, 741, 124
Hemmati, S., Capak, P., Masters, D., et al. 2019, ApJ, 877, 117
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841

Jahnke, K., Gillard, W., Schirmer, M., et al. 2025, A&A, 697, A3
Johnson, B. D., Leja, J., Conroy, C., & Speagle, J. S. 2021, ApJS, 254, 22
Kennicutt, R. C., & Evans, N. J. 2012, ARA&A, 50, 531
Kohonen, T. 1982, Biol. Cybern., 43, 59
Kroupa, P. 2001, MNRAS, 322, 231
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints [arXiv: 1110.3193]
Liu, D. Z., Meng, X. M., Er, X. Z., et al. 2023, A&A, 669, A128
Longo, G., Merényi, E., & Tino, P. 2019, PASP, 131, 100101
Madau, P., & Dickinson, M. 2014, ARA&A, 52, 415
Mahdi, B. 2011, arXiv e-prints [arXiv: 1108.0514]
Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53
Michalowski, M. J., Hayward, C. C., Dunlop, J. S., et al. 2014, A&A, 571, A75
Pannella, M., Carilli, C. L., Daddi, E., et al. 2009, ApJ, 698, L116
Rejeb, S., Duveau, C., & Rebafka, T. 2023, arXiv e-prints [arXiv:2202.07963]
Riccio, G., Małek, K., Nanni, A., et al. 2021, A&A, 653, A107
Salvato, M., Ilbert, O., & Hoyle, B. 2018, arXiv e-prints [arXiv:1805.12574]
Shuntov, M., Akins, H.B., Paquereau, L., et al. 2025, arXiv e-prints [arXiv:2506.03243]
Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51
Torre, V. L., Sajina, A., Goulding, A. D., et al. 2024, arXiv e-prints [arXiv:2403.18888]
Vettigli, G. 2018, MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map, https://github.com/JustGlowing/minisom/
Wang, T., Sun, H., Zhou, L., et al. 2024a, arXiv e-prints [arXiv: 2403.02399]
Wang, J., La Marca, A., Gao, F., et al. 2024b, A&A, 688, A20
Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, ApJS, 258, 11
Wuyts, S., Förster Schreiber, N.M., van der Wel, A., et al. 2011, ApJ, 742, 96
Zhan, H. 2011, Sci. Sin. Phys. Mech. Astron., 41, 1441
Zhan, H. 2021, Chin. Sci. Bull., 66, 1290
Zhang, Y.H., Zuntz, J., Moskowitz, I., et al. 2025, arXiv e-prints [arXiv:2508.20903]

Appendix A: Component maps of the SOMs and the deviations

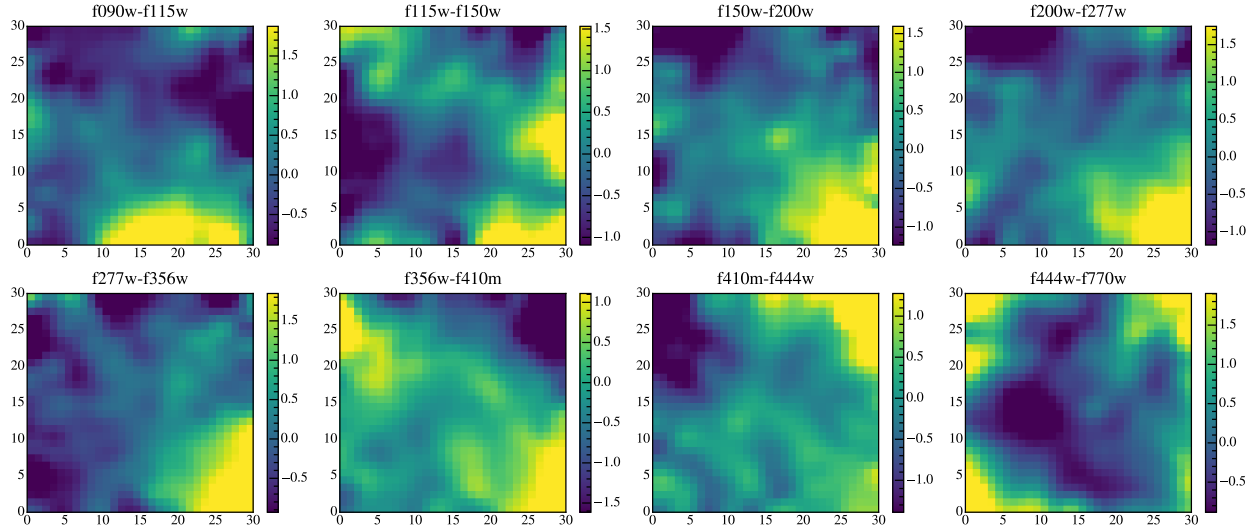

Fig. A.1. Component maps showing each of the 8 input colors across our trained SOM. The color bar indicates the normalized color values.

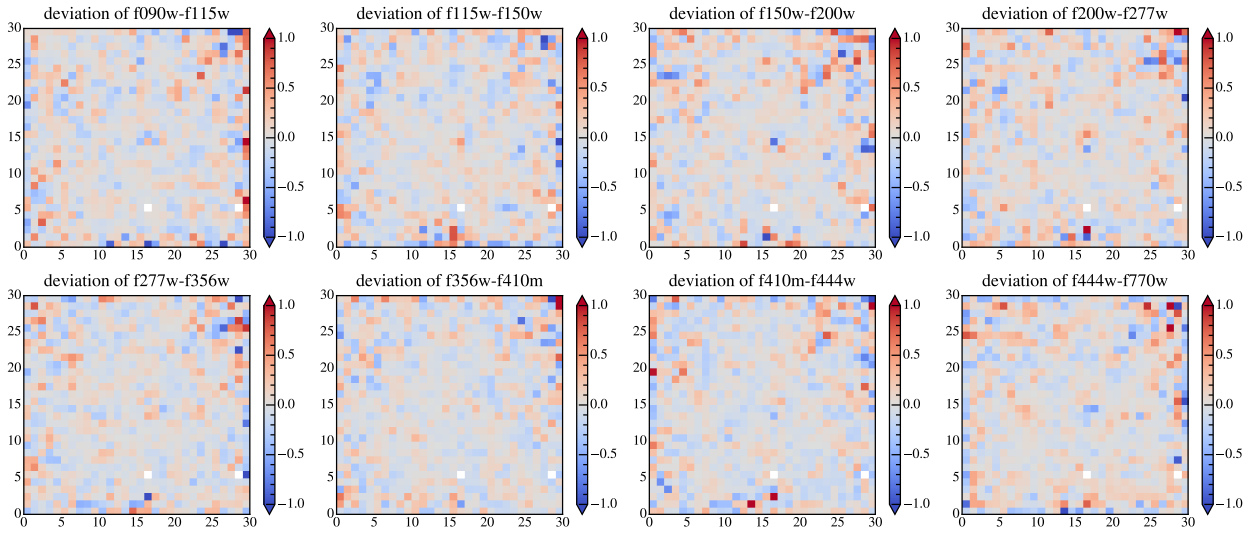

Fig. A.2. Deviation of SOM-generated weights and median normalized colors of galaxies of each.

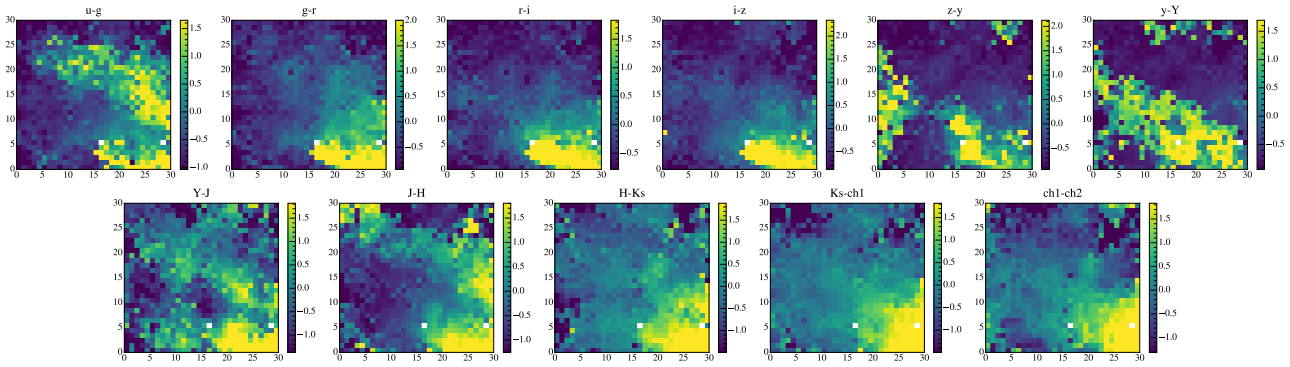

Fig. A.3. Component maps of synthetic SOM*. The color bar indicates the normalized color values.

Appendix B: Discrepancy between individual-galaxy parameters and SED-stacked estimates

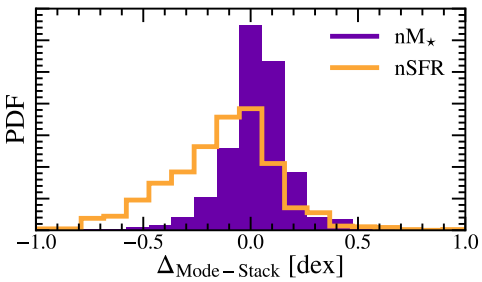

Fig. B.1. Distribution of $\Delta_{\rm mode-stack}$, defined as the difference between the mode $M_\star$ and SFR of individual galaxies in each SOM cell and the corresponding values derived from the stacked SED template.