The Astrophysical Journal, 942:91 (13pp), 2023 January 10 https://doi.org/10.3847/1538-4357/acacf5

A Machine-learning Approach to Predict Missing Flux Densities in Multiband Galaxy Surveys

Nima Chartab[1,2,3], Bahram Mobasher[3], Asantha R. Cooray[2], Shoubaneh Hemmati[4], Zahra Sattari[1,3], Henry C. Ferguson[5], David B. Sanders[6], John R. Weaver[7,8], Daniel K. Stern[9], Henry J. McCracken[10], Daniel C. Masters[4], Sune Toft[7,8], Peter L. Capak[4], Iary Davidzon[7,8], Mark E. Dickinson[11], Jason Rhodes[9], Andrea Moneti[10], Olivier Ilbert[12], Lukas Zalesky[6], Conor J. R. McPartland[6], István Szapudi[6], Anton M. Koekemoer[13], Harry I. Teplitz[4], and Mauro Giavalisco[14]

[1] The Observatories of the Carnegie Institution for Science, 813 Santa Barbara Street, Pasadena, CA 91101, USA; nchartab@uci.edu
[2] Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA
[3] Department of Physics and Astronomy, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA
[4] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
[5] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[6] Institute for Astronomy (IfA), University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA
[7] Cosmic Dawn Center (DAWN), Denmark
[8] Niels Bohr Institute, University of Copenhagen, Jagtvej 128, DK-2200 Copenhagen, Denmark
[9] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[10] Institut d'Astrophysique de Paris, UMR 7095, CNRS, and Sorbonne Université, 98 bis boulevard Arago, F-75014 Paris, France
[11] National Optical Astronomy Observatories, 950 N. Cherry Avenue, Tucson, AZ 85719, USA
[12] Aix Marseille Univ, CNRS, LAM, Laboratoire d'Astrophysique de Marseille, Marseille, France
[13] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[14] Department of Astronomy, University of Massachusetts, 710 N. Pleasant Street, Amherst, MA 01003, USA
Received 2022 January 27; revised 2022 August 16; accepted 2022 August 29; published 2023 January 17

Abstract

We present a new method based on information theory to find the optimal number of bands required to measure the physical properties of galaxies with desired accuracy. As a proof of concept, using the recently updated COSMOS catalog (COSMOS2020), we identify the most relevant wave bands for measuring the physical properties of galaxies in a Hawaii Two-0- (H20) and UVISTA-like survey for a sample of $i < 25$ AB mag galaxies. We find that with the available *i*-band fluxes, *r*, *u*, IRAC/*ch*2, and *z* bands provide most of the information regarding the redshift with importance decreasing from *r* band to *z* band. We also find that for the same sample, IRAC/*ch*2, *Y*, *r*, and *u* bands are the most relevant bands in stellar-mass measurements with decreasing order of importance. Investigating the intercorrelation between the bands, we train a model to predict UVISTA observations in near-IR from H20-like observations. We find that magnitudes in the *YJH* bands can be simulated/predicted with an accuracy of 1σ mag scatter $\lesssim$0.2 for galaxies brighter than 24 AB mag in near-IR bands. One should note that these conclusions depend on the selection criteria of the sample. For any new sample of galaxies with a different selection, these results should be remeasured. Our results suggest that in the presence of a limited number of bands, a machine-learning model trained over the population of observed galaxies with extensive spectral coverage outperforms template fitting. Such a machine-learning model maximally comprises the information acquired over available extensive surveys and breaks degeneracies in the parameter space of template fitting inevitable in the presence of a few bands.

Unified Astronomy Thesaurus concepts: Astronomy data analysis (1858); Astronomy data visualization (1968); Galaxy evolution (594)

1. Introduction

Future ground-based and spaceborne observatories, equipped with large aperture telescopes and sensitive large-format detectors, will provide broadband imaging data for more than a billion galaxies. These data are pivotal to a better understanding of the dark sectors of the universe (i.e., dark matter and dark energy) as well as the evolution of galaxies and large-scale structures over cosmic time. The challenge, however, is to obtain wide wave band coverage to constrain the spectral energy distributions (SEDs) of millions of galaxies and estimate their redshifts and physical parameters, such as stellar masses and star formation rates.

Template fitting is widely used to infer the photometric redshifts of galaxies and their physical properties (e.g., Arnouts et al. 1999; Bolzonella et al. 2000; Ilbert et al. 2006). However, theoretical synthetic templates may not be representative of the real parameter space of galaxies. For example, templates can include SEDs that do not have an observational analog. This will cause degeneracy in parameter measurement, especially when we reconstruct SEDs with few bands. Many of these degeneracies are mitigated by obtaining data with wide spectral coverage (e.g., with a larger number of wave bands). An example of such a data set is the Cosmic Evolution Survey (COSMOS; Scoville et al. 2007) that has been observed in more than 40 bands from X-ray to radio wavelengths. The wealth of information in this field provides very well-constrained SEDs for galaxies. However, not

all surveys have as many photometric bands as the COSMOS field. For instance, Euclid (Laureijs et al. 2011) will rely on near-infrared *Y*, *J*, and *H* bands (960–2000 nm), complemented by optical ground-based observations in *u*, *g*, *r*, *i*, and *z*, to measure photometric redshifts (Euclid Collaboration et al. 2020). It is therefore instructive to use the extensive data set in the COSMOS field to identify essential bands that carry most of the information regarding physical properties of galaxies.

The aim of this study is to transfer the information gained in the COSMOS field to fields such as the Euclid deep fields, where such extensive photometry does not exist. Using the concepts of information theory, we can find if there is any information shared between the bands and use these measurements to identify the most important bands (those that reveal most of the information about the physical properties of galaxies). Based on the machine-learning techniques we can then predict fluxes in the wave bands that are not observed in a survey but share information with other available (observed) bands. This allows us to carefully design future surveys and only observe in selected wave bands that include most of the information to significantly save in the observing time.

Machine learning has become popular in recent years to build models based on spectroscopic redshifts (e.g., Carrasco Kind & Brunner 2014; Masters et al. 2017) and train models based on synthetic templates (e.g., Hemmati et al. 2019) or mock catalogs generated from galaxy simulations (e.g., Davidzon et al. 2019; Simet et al. 2021). These methods are particularly useful as machine-learning algorithms can learn more complicated relations given a large and comprehensive training data set (Mucesh et al. 2021). Moreover, these models speed up parameter measurement, which is an important characteristic with the flood of data imminent from upcoming surveys (Hemmati et al. 2019).

In this paper, we develop a new technique based on information theory to quantify the importance of each wave band and identify essential bands to measure the physical properties of galaxies. We also develop a machine-learning model to predict fluxes in missing bands and thereby improve the wavelength resolution of existing photometric data. To demonstrate the application of these techniques, we apply our methods to a sample of galaxies drawn from the latest version of the COSMOS survey (COSMOS2020; Weaver et al. 2022), analogous to that planned by the Euclid deep fields. A new ground-based survey, Hawaii Two-0 (H20; C. McPartland et al. 2023, in preparation), has been designed to provide complementary photometric data for the Euclid mission. H20 will provide *u*-band observations from the MegaCam instrument on the Canada–France–Hawaii Telescope (CFHT) and *g*-, *r*-, *i*-, and *z*-band imaging from the Hyper Suprime-Cam (HSC) instrument on the Subaru Telescope over 20 deg^2 of the Euclid deep fields. Spitzer/IRAC observations from the Spitzer Legacy Survey (SLS) are also available in the same fields (Moneti et al. 2022). In this paper, we identify the importance of wave bands for an H20+UVISTA-like survey with similar wavelength coverage expected in Euclid deep fields, incorporating the near-IR *YJH* bands from UltraVISTA (McCracken et al. 2012) in addition to the H20 and SLS wave bands. We then predict fluxes in near-IR wave bands using the existing ground-based and mid-IR Spitzer/IRAC (H20-like) observations of the deep fields.

In Section 2, we briefly introduce the COSMOS2020 catalog and use that to build a sample of H20+UVISTA-like galaxies. Section 3 describes the concepts of information gain and quantifies the importance of each wave band based on them. In Section 4, we use dimensionality reduction techniques to visualize photometric data in two-dimensional space to explore the feasibility of predicting fluxes in near-IR fluxes based on *ugriz* and Spitzer/IRAC data. This is followed by Section 5 where we train a machine-learning algorithm, a random forest model, to predict fluxes in UVISTA/*YJH* wave bands using data in wave bands similar to the existing H20. In Section 6, we investigate the accuracy of the photometric redshifts and stellar masses given the limited number of bands available in H20-like and H20+UVISTA-like data. We discuss and summarize our results in Section 7.

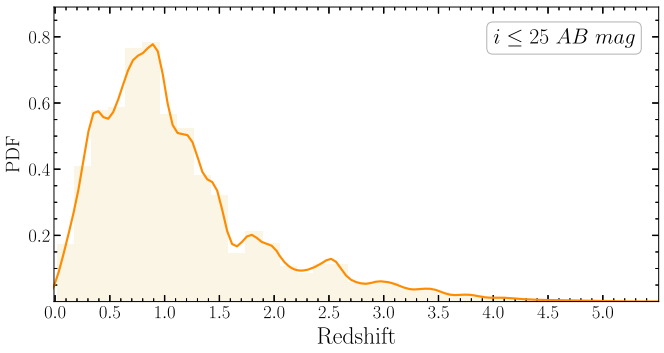


Figure 1. Redshift distribution for the subset of COSMOS2020 galaxies brighter than $i = 25$ AB magnitude (3σ). The entropy of the redshift calculated based on the distribution shown in this figure is less than the entropy of a uniformly distributed redshift. In other words, we become less surprised when we observe the redshift of a galaxy given this distribution (prior information).

Throughout this work, we assume flat ΛCDM cosmology with $H_0 = 70$ km s^{-1} Mpc^{-1}, $\Omega_{m_0} = 0.3$, and $\Omega_{\Lambda_0} = 0.7$. All magnitudes are expressed in the AB system, and the physical parameters are measured assuming a Chabrier (2003) initial mass function (IMF).

2. Data

Here we use the updated version of the COSMOS catalog, COSMOS2020, to build a sample of galaxies analogous to those that will be observed in the Euclid deep fields. Compared to the COSMOS2015 catalog (Laigle et al. 2016), COSMOS2020 provides much deeper near-IR and mid-IR (Spitzer) photometric data as well as two independent methods for photometric extraction—the conventional and profile-fitting (`The Farmer`; J. Weaver et al. 2023, in preparation) methods. We use `The Farmer` photometry that contains consistent photometric data in 39 bands from far-ultraviolet to mid-IR including broad, medium, and narrow filters. All the data are reduced to the same scale with appropriate point-spread functions. Photometric redshifts are calculated using LePhare (Arnouts et al. 1999; Ilbert et al. 2006) with a similar configuration described in Ilbert et al. (2013). Given the large number of bands with deep observations, photometric redshift solutions are accurate, reaching a normalized median absolute deviation ($\sigma_{\rm NMAD}$; Hoaglin et al. 1983) of 0.02 for galaxies as faint as $i \sim 25$ AB mag (Weaver et al. 2022). The redshifts of galaxies are then fixed on their estimated photometric redshifts, and the stellar masses were estimated. In this paper, we consider COSMOS2020 photometric redshifts and stellar masses as a "ground truth" since spectroscopic redshifts are only available for a limited number of galaxies, and using a mixture of photometric and spectroscopic redshifts can bias our sample toward specific populations of galaxies.

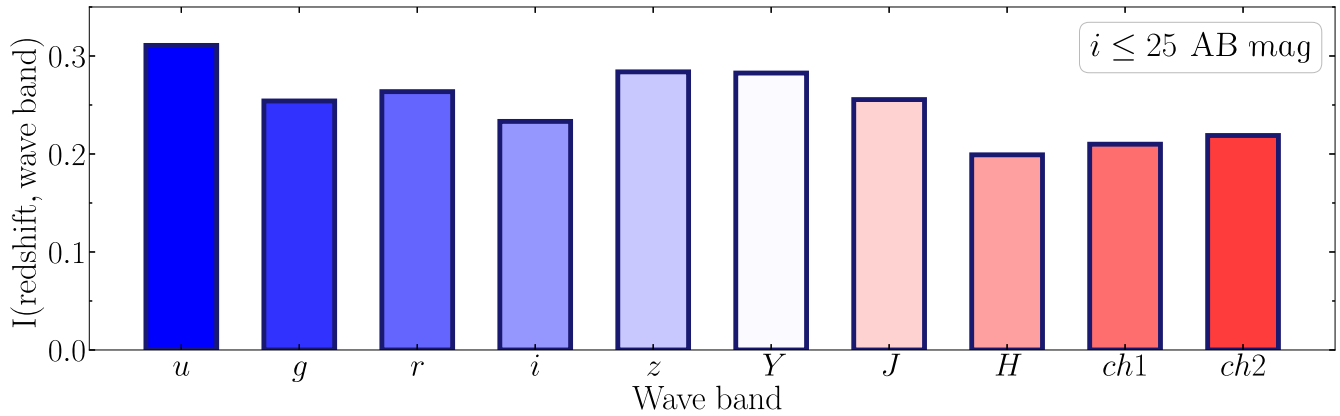
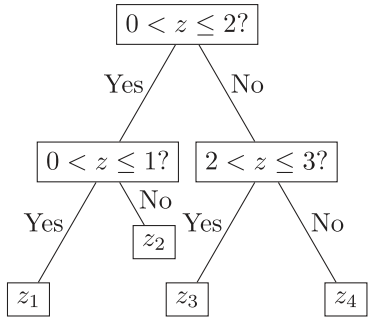


Figure 2. Mutual information of redshift and wave bands in bits per galaxy. With greater mutual information, the entropy of the redshift will decrease more if we include the band in photometric redshift measurements, thus increasing the importance of the band. Here, u is the most important followed by z band.

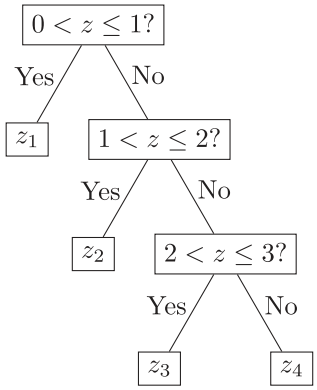
We use two sets of wave bands: (1) H20-like bands: $\boldsymbol{A} := \{u, g, r, i, z, ch1, \text{and } ch2\}$ and (2) H20+UVISTA-like bands: $\boldsymbol{B}:=\{u, g, r, i, z, Y, J, H, ch1, \text{and } ch2\}$. u-band observations are conducted by the MegaCam instrument at CFHT, and other optical bands (g, r, i, and z) are available from Subaru's HSC imaging. The Spitzer/IRAC channel 1, 2 ($ch1$, $ch2$) data are compiled from all the IRAC observations of the COSMOS field (Moneti et al. 2022). Near-IR photometry in Y, J, and H bands are obtained from the UltraVISTA survey (McCracken et al. 2012). We select a subset of the COSMOS2020 galaxies that are observed, but not necessarily detected, in all the aforementioned bands and have i-band AB magnitude $\leqslant$25 with 3σ detection. These selection criteria result in 165,807 galaxies out to $z \sim 5.5$. Photometric measurements in the COSMSOS2020 catalog are not corrected for Galactic extinction. We corrected them using the Schlafly & Finkbeiner (2011) dust map. Moreover, some sources have negative fluxes in the desired bands, which are due to the variation of background flux across the image. We set these fluxes to zero.

3. Information Gain

Let us suppose that we do not have any prior information about the redshift distribution of galaxies selected from the criteria mentioned in Section 2. We therefore assume a uniform distribution for the redshift. As an example, if we define four bins of redshifts ($\{z_1 = (0, 1]$; $z_2 = (1, 2]$; $z_3 = (2, 3]$; and $z_4 = (3, 4]\}$) and want to identify which bin a galaxy belongs to, we can encode it in two bits, as below:



Here, we need to ask two yes/no questions to identify the bin a galaxy belongs to. However based on the available observations of COSMOS2020, we know the redshift distribution of galaxies with $i \leqslant 25$ AB mag as background information. We therefore update the decision tree above, considering our prior information about the redshift distribution, to reduce the average number of questions we need to ask to identify the redshift bin of a galaxy. Based on the redshift distribution shown in Figure 1, the probability of a galaxy being in each redshift bin are $P(z_1) = 0.56$, $P(z_2) = 0.32$, $P(z_3) = 0.09$, and $P(z_4) = 0.03$. Thus, one possible decision tree to identify the redshift bin of a galaxy can be built as follows:



On average, $0.56 \times 1 + 0.32 \times 2 + (0.09 + 0.03) \times 3 = 1.56$ questions (bits) are required to identify the redshift bin of a galaxy. We find that the number of bits (questions) were reduced from 2 to 1.56 when we added information regarding the redshift distribution of galaxies. This decrease shows that we will become less surprised when we observe the redshift of a galaxy given that we know what the redshift distribution looks like.

Given the above example, the optimal number of bits required to store a variable called the Shannon Entropy (H) is defined as (Shannon 1948)

$$H(X) = -\sum_i P(x_i)\log_2 P(x_i), \tag{1}$$

where x_i is the possible outcome of a variable (X) that occurs with probability $P(x_i)$. In this formulation, $\log_2 P(x_i)$ represents the number of bits required to identify the outcome. Using Equation (1), the Shannon Entropy of redshift based on the probabilities in four bins is 1.45 bits. This means that we can still make our tree more optimal to encode the redshift values in 1.45 bits instead of 1.56. One possible way would be by building the tree to identify the redshift of two galaxies simultaneously, which makes the average number of questions

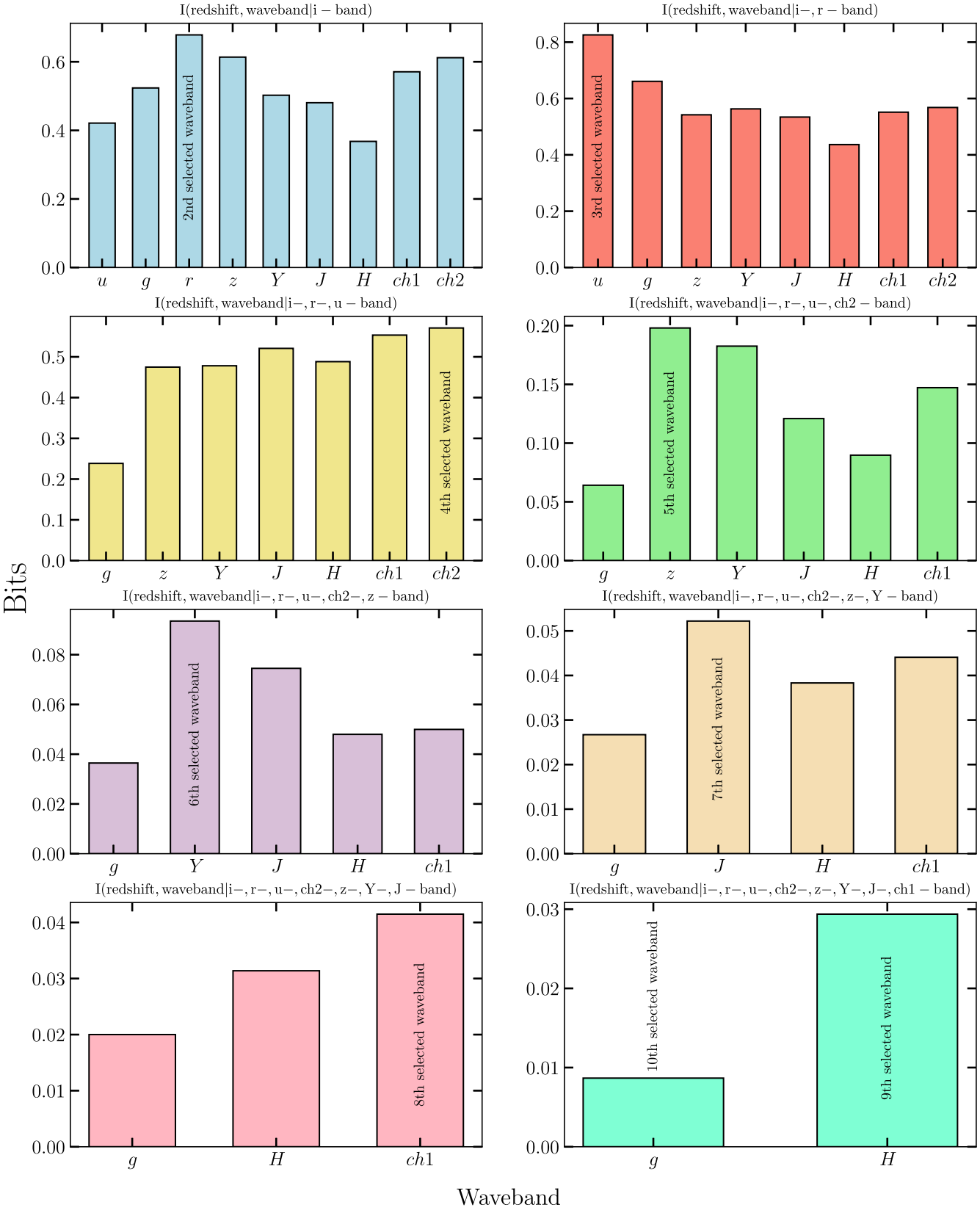


Figure 3. Conditional mutual information of redshift and wave bands in bits per galaxy. The most relevant bands can be selected based on their conditional mutual information. The sample is selected based on the magnitude of the *i* band, which implies that the first selected wave band is the *i* band. The top left panel shows the mutual information of redshift and wave bands given that *i*-band data are available. Therefore we select the *r* band as the second-most relevant band since it provides the most information. In the top right panel, we assume that *i*- and *r*-band data are available and find that the *u* band would be the third choice. We follow a similar procedure to find relevant bands in order of their importance. We note that these results depend on the selection criteria. For any new sample of galaxies with a different selection, these results should be remeasured.

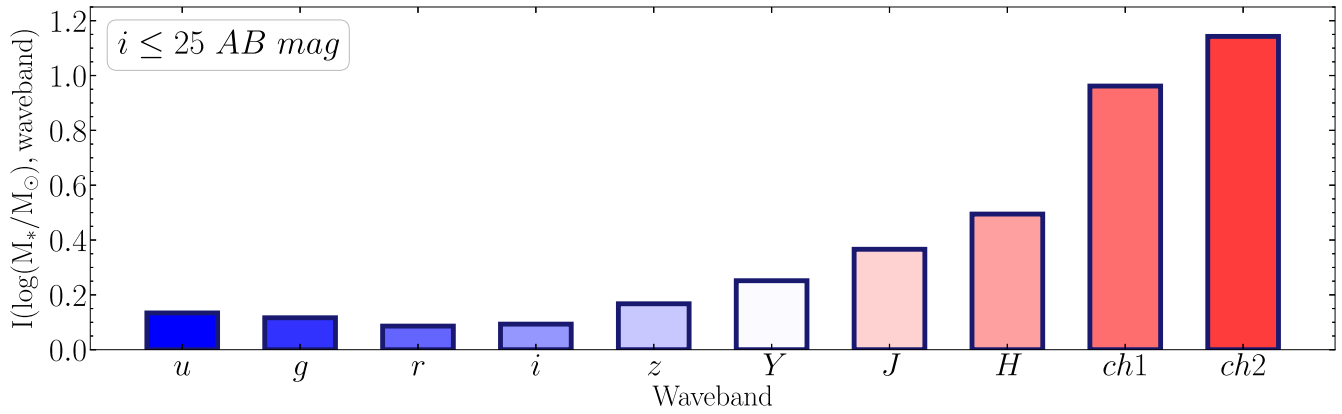


Figure 4. Similar to Figure 2 but for the stellar mass. Mutual information of stellar mass and wave bands in bits per galaxy is shown. With greater mutual information, the entropy of stellar mass will decrease more if we include the band in stellar-mass measurements, making the band more important.

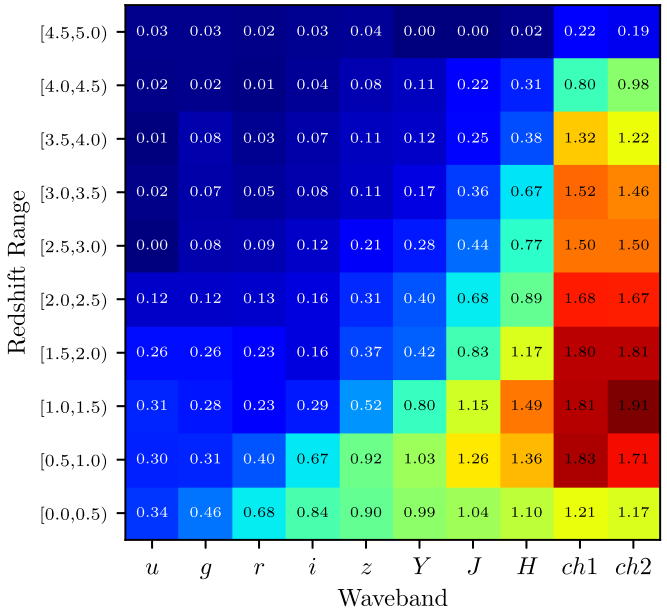


Figure 5. Mutual information of stellar mass and wave bands in bits per galaxy in the bins of redshift. The map is colored based on the value of mutual information, with red representing the most important band and blue representing the least important band. The role of low wavelength bands decreases as we approach higher redshift, as we would expect.

per galaxy even less than 1.56. However, we do not aim to find the optimal compression algorithm to encode the redshift information. We just use the Shannon Entropy to find the maximal compression rate.

In the presence of other information, such as observed fluxes in different bands, the entropy of the redshift decreases even more. The amount of uncertainty (entropy) remaining in X after we have seen Y is called conditional entropy and defined as

$$H(X|Y) = -\sum_{x\in X, y\in Y} P(x, y)\log_2 \frac{P(x, y)}{P(y)}, \quad (2)$$

where $P(x, y)$ is the joint probability distribution at (x, y). Moreover, mutual information between X and Y (i.e., the amount of uncertainty in X that is removed by knowing Y) is defined as

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X, Y), \end{aligned} \quad (3)$$

where $H(X, Y)$ is the joint entropy of a pair of variables (X, Y). In other words, $I(X, Y)$ is a measure of the amount of information (in bits) one can acquire about X by observing Y. This parameter can be used to identify the wave band that will be most useful for measuring galaxy properties (e.g., redshifts). For instance, the wave band with the highest I(*redshift*, *wave band*) carries the most information and decreases the entropy of the redshift the most.

The mutual information as in Equation (3) is defined for discrete variables. In the case of continuous variables (e.g., redshift, flux, and stellar mass), we need to properly discretize the data. Kraskov et al. (2004; hereafter KSG) introduced a k-nearest neighbor estimator to compute the mutual information of continuous variables. This method detects the underlying probability distribution of data by measuring distances to the kth nearest neighbors of points in the data set. There is nonzero mutual information when some points are clustered in the X–Y space, which allows us to predict $y \in Y$ given an $x \in X$ coordinate. We refer readers to the original KSG paper for details of the method. Figure 2 shows the mutual information between the redshift and each wave band based on the KSG algorithm with $k = 100$ nearest neighbors. It suggests that given the sample of $i < 25$ AB mag galaxies, the u band provides the largest information regarding the redshift compared to the rest of the H20+UVISTA-like bands. However, our sample is selected based on i-band magnitudes, so we assumed that i-band data are already available. Suppose that for our sample u-band fluxes are highly correlated with i-band data. In this case, the u band carries no information in the presence of i-band data. To take into account such an effect, we need to compute the conditional mutual information, defined as

$$I(X, Y|Z) = H(X|Z) - H(X|Y, Z), \quad (4)$$

where $I(X, Y|Z)$ is the mutual information of X and Y given that Z is observed. Following the KSG algorithm, we find the conditional mutual entropy to sort wave bands based on their importance. We compute I(*redshift*, *wave band*|*i band*) and choose the wave band with the highest conditional mutual information as the most important band. The conditional mutual information estimations reveal that the r band is the most important wave band given that i-band data are available. We continue computing conditional mutual information, I(*redshift*, *wave band*|*swave band*), where *swave band* is the previously selected wave band.

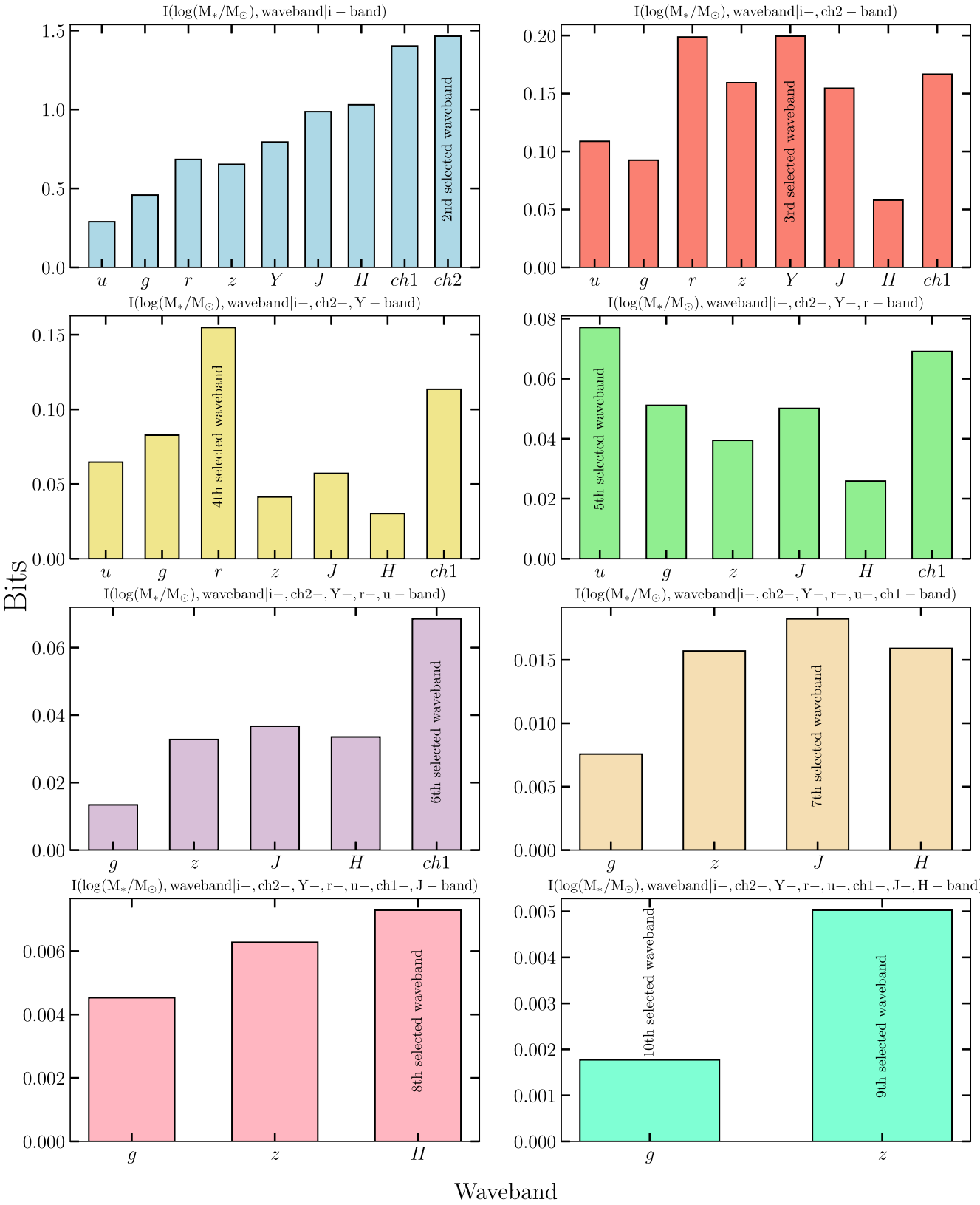


Figure 6. Similar to Figure 3 but for the stellar mass. Each panel shows the conditional mutual information of stellar mass and wave bands given that all the previously selected bands are available. We find that for the *i*-band selected sample, the *ch*2, *Y*, *r*, and *u* bands are the four most relevant bands with decreasing order of importance. The top left panel shows that IRAC data are essential for stellar-mass measurements.

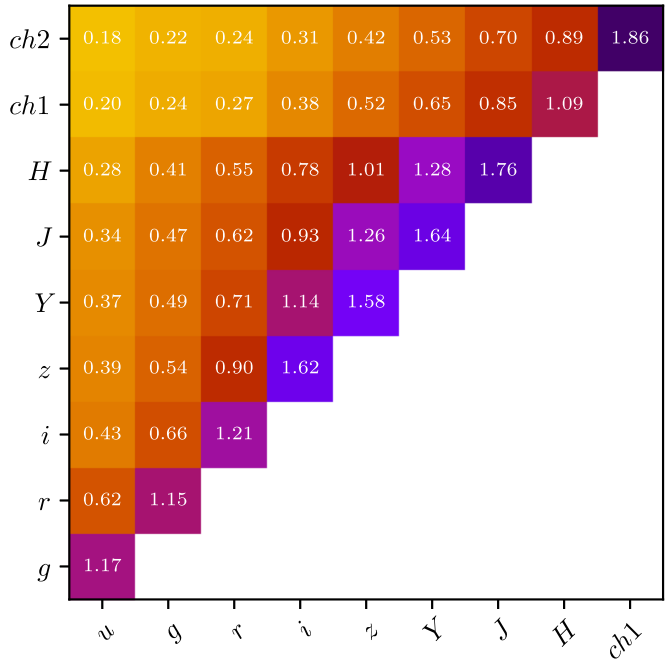


Figure 7. Visual representation of the mutual information between different wave bands for a sample of $i < 25$ AB mag galaxies. The map is colored based on the value of mutual information, with purple representing the most correlated bands and yellow representing the least correlated bands (mostly independent). For instance, the mutual information of *ch*1 and *ch*2 quantifies the bits of information about the IRAC/*ch*1 flux obtained by observing IRAC/*ch*2 flux. It is similar to the correlation coefficient, but it is able to capture nonlinear relationships.

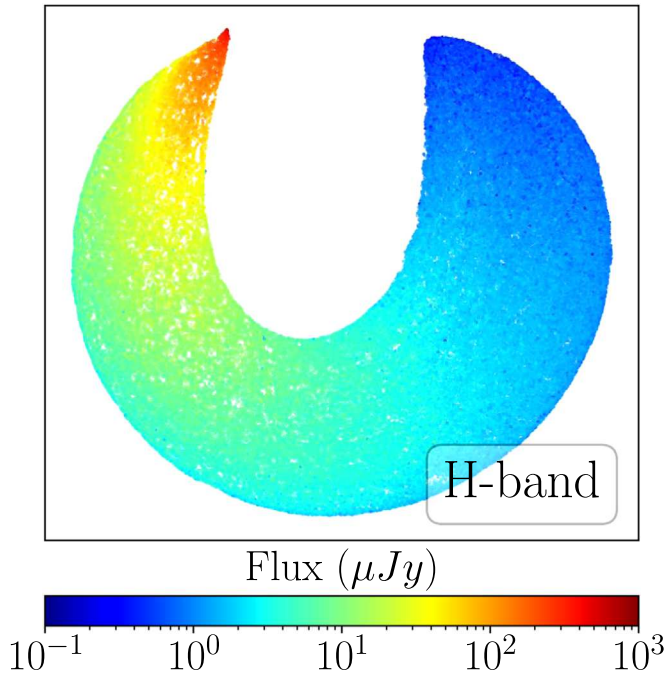


Figure 8. 2D visualization of the sample with H20-like bands using the UMAP technique. The mapped data are color coded by the *H*-band fluxes. The smooth gradient of *H*-band fluxes in the 2D representation reassures us that galaxies with similar fluxes in H20-like bands have similar *H*-band fluxes as well.

Figure 3 shows the nonzero conditional mutual information as we select relevant wave bands. We find that for $i < 25$ AB mag galaxies, *r*, *u*, *ch*2, and *z* bands are the bands that provide most of the information about the redshift with decreasing importance from *r* band to *z* band. We repeat these analyses for stellar-mass measurements. In Figure 4, we measure the mutual information between stellar mass and each wave band for the whole sample, and in Figure 5 we measure the same quantity, $I(\log(M_*/M_\odot), \text{wave band}|i \text{ band})$, in the bins of redshifts. As we expect, the role of short wavelength bands decreases as we approach higher redshifts. We further compute the important wave bands given the availability of *i*-band data in Figure 6. We find that *ch*2, *Y*, *r*, and *u* bands are the most relevant bands in the stellar-mass measurements with decreasing order of importance. One can constrain the redshift and repeat the analysis to find the optimal bands for stellar-mass measurements in the desired redshift range given the availability of *i*-band data.

One should note that these conclusions depend on the selection criteria of the sample. This method provides a powerful tool in designing future surveys and quantifying the importance of each wave band. An efficient observation can be conducted by prioritizing important wave bands identified by the information gain-based method.

Moreover, different wave band fluxes can be intercorrelated for a specific sample of galaxies. For instance, the top left panel in Figure 6 shows that IRAC/*ch*1 and *ch*2 provide a comparable amount of information for stellar-mass measurements, which suggests that these bands are intercorrelated for our sample with $i < 25$ AB mag. Figure 7 visualizes the mutual information between different bands. A greater value of mutual information indicates that the wave bands are more correlated. Intercorrelation between wave bands allows us to predict/simulate fluxes of galaxies in missing bands. In the following, we investigate the possibility of predicting/simulating near-IR UVISTA/*YJH* fluxes based on H20-like data for a sample of galaxies with $i < 25$ AB mag.

4. Data Visualization

Fluxes of galaxies in *N* wave bands are used to measure the photometric redshifts and physical parameters of galaxies. For example, the H20-like data with $N = 7$ bands occupy a seven-dimensional space, where the position of each galaxy is determined by its fluxes in seven bands. Therefore, galaxies with similar positions in *N*-dimensional space are expected to have similar redshifts and physical parameters if *N* is large enough to fully sample the observed SED of galaxies. Similarly, it is expected that they will have similar fluxes in the $(N+1)$th wave band. However, showing galaxy fluxes in a high-dimensional space (e.g., seven-dimensional space) is impossible, and thus, we use dimensionality reduction techniques to present them in 2D space such that the information of higher dimension is maximally preserved. In this work, we use the Uniform Manifold Approximation and Projection (UMAP; McInnes et al. 2018) technique to visualize our sample in a two-dimensional space. UMAP is a nonlinear dimensionality reduction technique that estimates the topology of high-dimensional data and uses this information to construct a low-dimensional representation of data that preserves structure information on local scales. It also outperforms other dimensional reduction algorithms such as t-Distributed Stochastic Neighbor Embedding (van der Maaten & Hinton 2008) used in the literature (Steinhardt et al. 2020) since it preserves structures on global scales as well. In a simple sense, UMAP constructs a high-dimensional weighted graph by extending a radius around each data point and connecting points when their

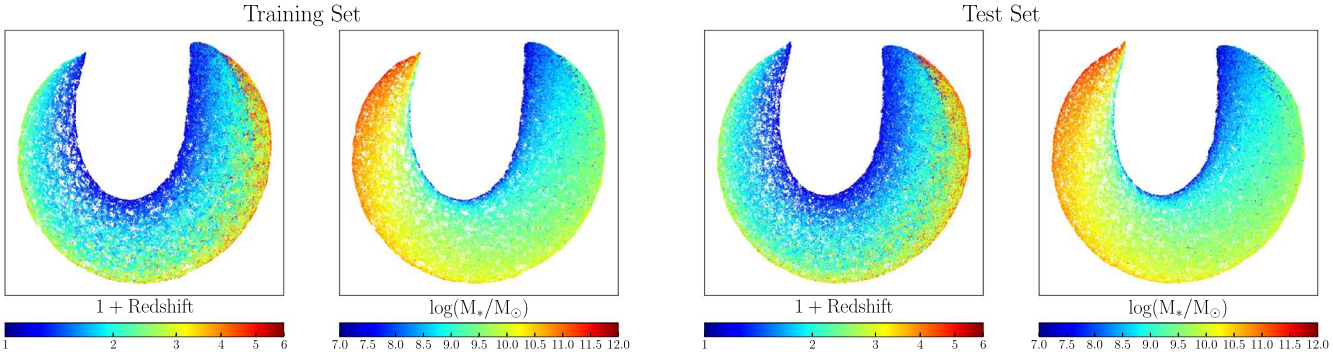


Figure 9. Similar to Figure 8 but for training (two left panels) and test (two right panels) samples. Maps are color coded with photometric redshifts and stellar masses. We find that the training and test samples share the same properties, so the randomly selected training sample is representative of the galaxies in the COSMOS field.

radii overlap. This radius varies locally based on the distance to the nth nearest neighbor of each point. The number of the nearest neighbor (n) is the hyperparameter in UMAP that should be fixed to construct high-dimensional graphs. Small (large) values for n will preserve more local (global) structures. Once the high-dimensional weighted graph is constructed, UMAP optimizes the layout of a low-dimensional map to be as similar as possible to the high-dimensional graph.

We use the UMAP Python library[15] to map seven-dimensional flux space of H20-like data to two dimensions considering 50 of the nearest neighbors to provide a balance between preserving local and global structures. We do not map magnitudes or colors since nondetected values cannot be handled properly when using them. Multiwave band fluxes contain all the information regarding colors, but using colors misses information regarding fluxes or magnitudes. Therefore, mapping fluxes of galaxies from that space to two dimensions is a better way than using colors. Since fluxes in different bands have fairly similar distributions, no normalization is needed before applying UMAP. In the case of significantly distinct distributions, normalization is needed to avoid the dominance of a wave band with a larger dynamic range. Figure 8 shows a 2D visualization of the sample with H20-like bands using the UMAP algorithm. As an example, the mapped data are color coded by the H-band fluxes (not present in H20 photometry) in μJy. The smooth transition of the H-band fluxes in the 2D representation in Figure 8 reassures us that galaxies with similar fluxes in H20-like bands also have similar H-band fluxes. We note that the H20-like data set does not include H-band data.

Visualized data in Figure 8 show qualitatively that the H-band fluxes are predictable to some extent using H20-like data. To perform a quantitative assessment on how accurately one can predict fluxes in the UVISTA YJH bands given H20-like observations, we train a random forest (Breiman 2001) model with half of our sample and evaluate the model's performance with the other half. A random forest consists of an ensemble of regression trees. The algorithm picks a subsample of the data set, builds a regression tree based on the subsample, and repeats this procedure numerous times. The final value is the average of all the values predicted by all the trees in the forest. Having numerous decision trees based on subsampled data makes this algorithm unbiased and unaffected by overfitting. Another advantage of this method is that the inputs do not need to be scaled before feeding into the model. In the following section, we train a random forest model and evaluate its accuracy.

5. Flux Predictions

We split the sample (described in Section 2) randomly into a training and a test sample. To evaluate if the training sample is representative, we construct a 2D projection of H20-like fluxes similar to Figure 8 for both training and test samples. Figure 9 shows the 2D visualizations color coded by the properties of galaxies (photometric redshift and stellar mass). We find that the training and test samples share the same properties, so the training sample is representative of the galaxies in the COSMOS field. With 82,903 galaxies as a training sample, we build a random forest model with 100 regression trees to predict the UVISTA YJH bands from the H20-like band fluxes. We use Python implementation of the algorithm (Scikit-learn; Pedregosa et al. 2011)[16] with its default parameters to build the model. The true (observed) fluxes in the YJH bands are available in the COSMOS2020 catalog. Using the trained random forest model, we then predict the expected fluxes for galaxies not included in the training set, with the results compared in Figure 10. For each band, we compare the predicted magnitudes ($\mathrm{Mag_{Predicted}}$) with the true observed magnitudes ($\mathrm{Mag_{True}}$). We find that the random forest model predicts unbiased YJH fluxes with high accuracy. The bottom panel in each figure shows the scatter of the $\mathrm{Mag_{Predicted}} - \mathrm{Mag_{True}}$ as a function of true magnitudes. With median magnitude discrepancy (Δ) of $\sim$0.01, we find that the offset is comparable with discrepancies that arise from different methods of photometric data reduction. Weaver et al. (2022) found that the median tension between the magnitudes derived from aperture photometry and profile-fitting extraction is $\Delta \sim 0.002$ in YJ bands and $\Delta \sim 0.02$ in the H-band for sources brighter than the 3σ depth of each band. Thus, such small offsets in the random forest regressor are within the intrinsic uncertainties of the data reduction techniques. The green solid and dashed lines in the subpanels of Figure 10 show the median of Δ and 1σ (68%) scatter, respectively. The scatter in the prediction is $<$0.17 mag for galaxies brighter than 24 AB mag. This shows that YJH near-IR observations of UVISTA can be simulated with acceptable accuracy from the available observations of H20 for a sample of galaxies with $i < 25$ AB mag. Our results remain consistent when we rebuild a new random forest with different randomly selected training samples. While our focus in this paper is on the UVISTA/YJH

[15] https://github.com/lmcinnes/umap

[16] https://scikit-learn.org/stable

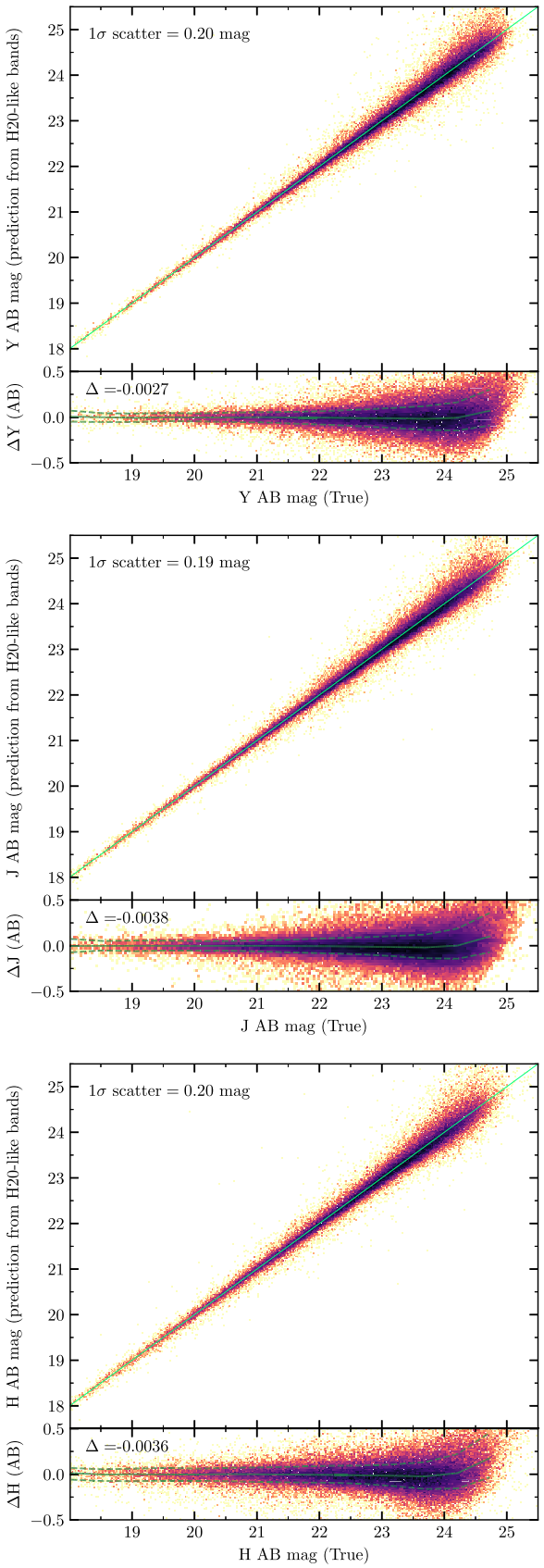


Figure 10. The performance of the random forest model on the 82,904 test galaxies not used for the training of the model. The model is trained based on H20-like bands (u, g, r, i, z, $ch1$, and $ch2$) and predicts UVISTA YJH bands. Bottom panels show the scatter of $\mathrm{Mag_{Predicted}} - \mathrm{Mag_{True}}$ as a function of true magnitudes, and Δ is the median offset in these scatter plots.

and H20 bands, the method we present is general and directly applicable to other surveys.

6. Photometric Redshift and Stellar Mass

In the previous section, we showed that given the observations of the H20 survey, near-IR observations of UVISTA can be constrained to some extent. In other words, observations of the COSMOS field provide valuable information regarding the distribution of galaxies in the flux space even if we do not observe galaxies as extensively as it is done in the COSMOS field in terms of the spectral coverage. When we use the template-fitting code with synthetic templates, we usually do not take into account this constraint. There are two approaches to incorporate this information in the photometric redshifts or physical parameters measurements. First, add a prior to fluxes in the bands that are not observed in the survey. For instance, when we perform SED fitting using H20-like bands, we can add priors to the YJH bands based on a random forest model, which is trained over the population of galaxies from the COSMOS observations. Second, train a model based on SED-fitting results calculated with a large number of bands. In this case, when we feed our model with H20-like data, it will decide the best value of a parameter based on both the existence of similar observations in the COSMOS field (information from galaxy populations) and the SED-fitting solution for that galaxy.

In this section, we employ the latter approach to train a model to predict the photometric redshifts and the stellar masses of galaxies based on H20-like and H20+UVISTA-like bands. We train a random forest model based on a training sample of observed galaxies. The inputs of the model are H20-like fluxes, and the output is either photometric redshift or stellar mass computed from SED fitting over 29 bands available in the COSMOS2020 catalog. We also train another similar model where the inputs are H20+UVISTA-like bands. Figure 11 shows the performance of trained models on the test sample with 82,904 galaxies. We find that both models recover photometric redshifts and stellar masses with comparable accuracy, although a model trained on H20+UVISTA-like data has slightly higher accuracy. Normalized median absolute deviation (σ_{NMAD}) of $\Delta z/(1+z)$ is $\sim$0.03 for both models with $\sim$4% outlier fraction. Outlier galaxies are defined as galaxies with $\Delta z/(1+z) > 0.15$. The median absolute deviation of $\log(M_*/M_\odot)$ is $\sim$0.1 dex for both models. We explain this similar performance using the results of Sections 3 and 5. The random forest model with H20-like bands comprises most of the information regarding UVISTA bands as we trained the model with the population of observed COSMOS galaxies. Therefore, it should recover photometric redshifts and stellar masses as accurately as the model that includes near-IR (YJH) observations.

We repeat a similar analysis starting with only i-band data and adding other important bands in the same order as we identified in Section 3. Figure 12 shows the the normalized median absolute deviation of $\Delta z/(1+z)$ and $\log(M_*/M_\odot)$ as a function of bands used to measure the parameter. We find that i, r, u, $ch2$, and z bands are the minimal number of bands to reach an acceptable accuracy of $\sigma^{\mathrm{NMAD}}_{\Delta z/(1+z)} = 0.03$ to measure photometric redshifts of $i < 25$ AB mag. For the same sample, i, $ch2$, Y, r, and u bands are the optimal bands for stellar-mass measurements reaching an accuracy of $\sigma^{\mathrm{NMAD}}_{\log(M_*/M_\odot)} = 0.15$ dex.

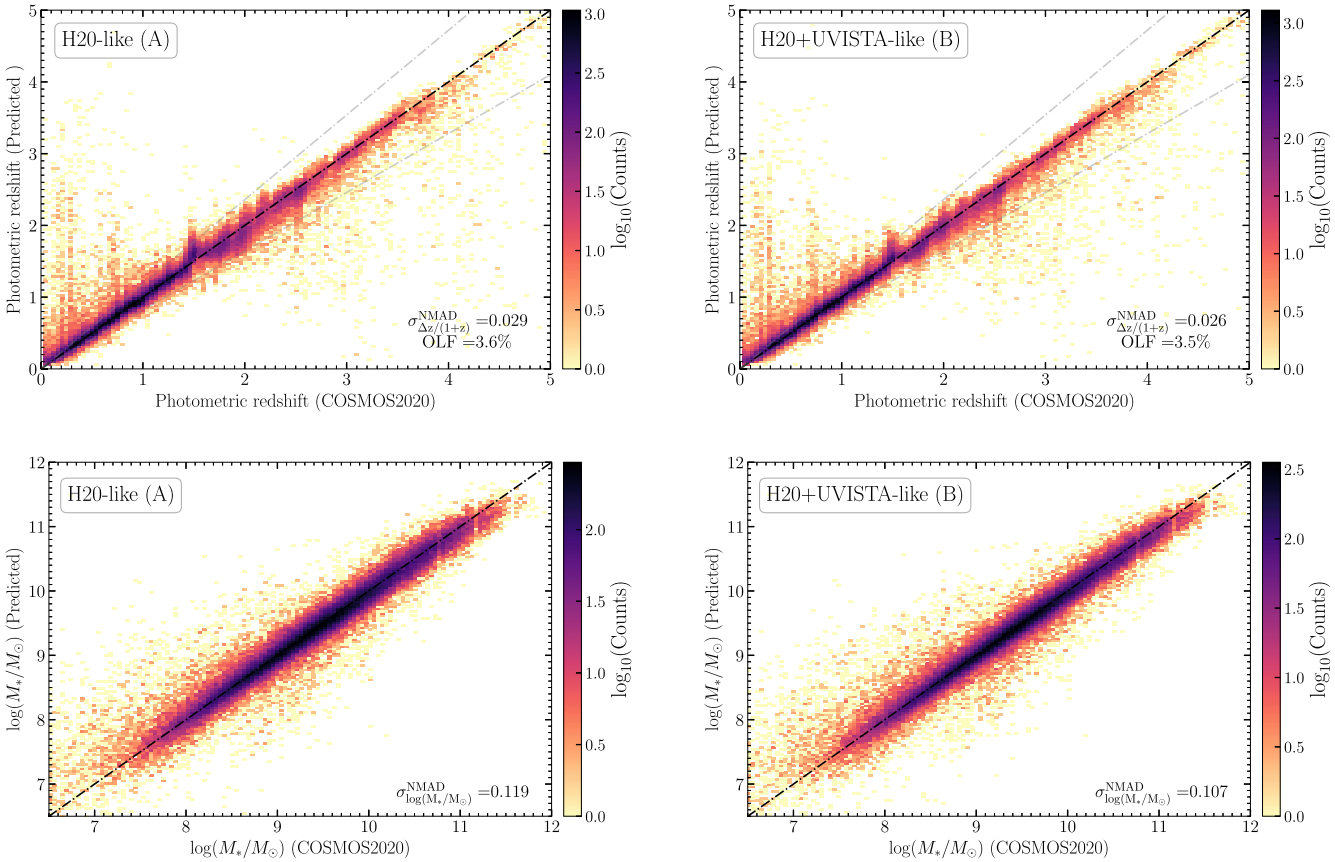


Figure 11. Performance of the random forest model to predict photometric redshifts and stellar masses when the model is trained by H20-like bands (left panels) and H20+UVISTA-like bands (right panels). Both trained models recover photometric redshifts and stellar masses with high accuracy. The similar performance of the model with and without *YJH* bands originates from the fact that the H20-like bands capture most of the information available in *YJH* bands as shown in Figure 10. The black dashed−dotted lines show one-to-one relation, and the gray dashed−dotted lines correspond to the predicted redshifts at $\pm 0.15(1+z)$ (outlier definition boundaries).

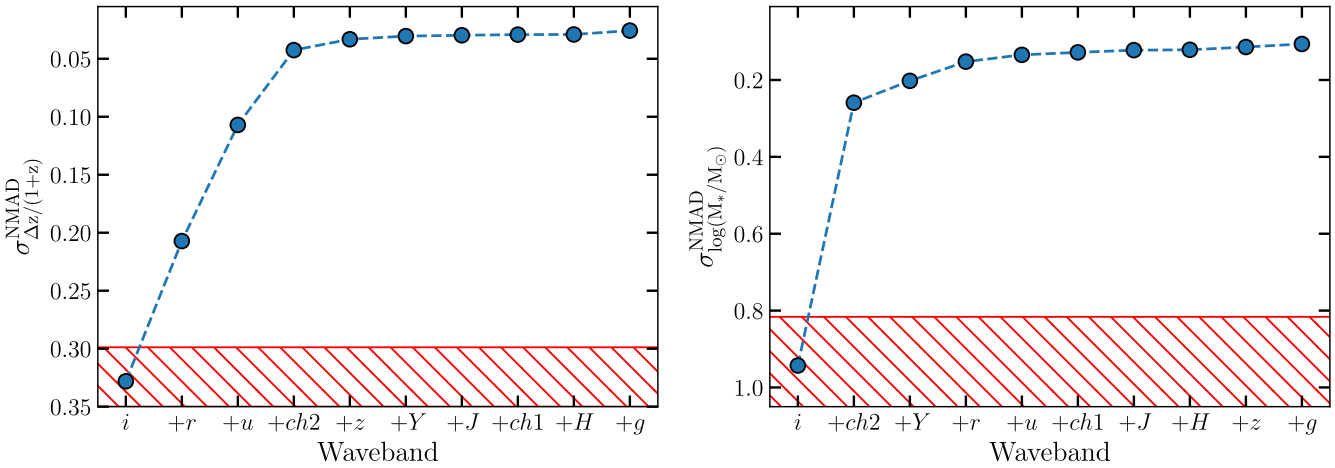


Figure 12. The normalized median absolute deviation of $\Delta z/(1+z)$ (left panel) and $\log(M_*/M_\odot)$ (right panel) as a function of bands used to measure the parameter. As the sample is selected based on the *i*-band magnitude of galaxies, we start with training a random forest model based on only *i*-band data, and then we add other bands following the same order of importance we find in Figures 3 and 6. The red hatched region represents a region where the normalized median absolute deviation exceeds the scatter of the data relative to their mean value.

6.1. Synthetic Templates

In the following, we use UMAP to visualize the photometry of synthetic SED models commonly used in template-fitting procedures. We build a set of theoretical templates using the 2016 version of a library of Bruzual & Charlot (2003), considering Chabrier (2003) IMF. Star formation histories are modeled with an exponentially declining function (SFR $\propto e^{-t/\tau}$), where τ is the star formation timescale. Dust attenuation is applied using the Calzetti et al. (2000) law, and solar stellar metallicity is assumed for all templates. We build ∼750,000 theoretical templates assuming $\tau \in (0.1,\ 10)$ Gyr, $t \in (0.1,\ 13.7)$ Gyr, $A_V \in (0,\ 2)$ mag, and $z \in (0,\ 5.5)$. t and A_V are the stellar age and the extinction in the visual band,

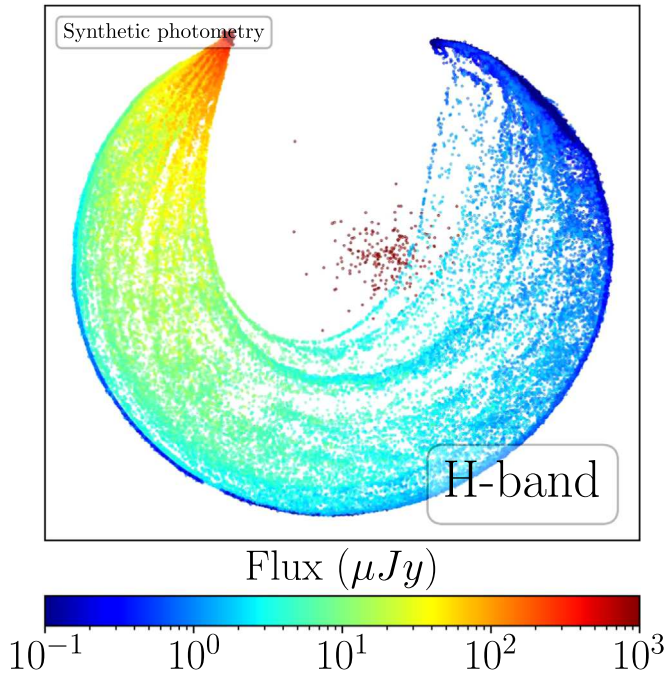


Figure 13. Similar to Figure 8 but for synthetic photometric data. The high-dimensional synthetic H20-like data are transformed to the space learned in Figure 8. The map is color coded by the synthetic *H*-band fluxes. Existing dissimilarities between this figure and Figure 8 show that synthetic models lack the observed information.

respectively. We then calculate the synthetic photometry in both the H20-like and H20+UVISTA-like bands by applying the corresponding filter response function.

As we learned the topology of fluxes in the H20-like bands for real observed galaxies in the COSMOS2020 catalog (Figure 8), we can transform the H20-like band fluxes of synthetic photometry into the learned space. Figure 13 shows the 2D visualization of the theoretical templates with H20-like bands in that learned space. As an example, data points in the reduced dimension are color coded by their synthetic *H*-band fluxes in μJy. Comparing theoretical templates with the observed data shown in Figure 8 reveals that model galaxies encounter degeneracies. In this specific example, we show that templates with similar H20-like fluxes have more diverse *H*-band fluxes than real observations, which can produce degenerate results when template fitting is performed based on H20-like bands. Adding the information of the COSMOS2020 observations as a prior imposes a strong correlation between the observed and missing bands and makes the theoretical templates less degenerate as shown in Figure 8. For example, the dark blue arc on the left side of Figure 13 mismatches with the observational counterpart. In other words, synthetic templates predict the *H*-band flux of $\sim$0.1 μJy for galaxies in that vicinity (i.e., the dark blue arc), but real observations show that they have in fact an *H*-band flux of $\sim$10 μJy. This shows that the extra information that exists in the previous observations can add valuable information to template-fitting analyses.

If one adds a predicted band in the template-fitting procedure, the errors should be assigned based on the 1σ scatter of the predicted flux (dashed lines in Figure 10). It is particularly important to properly take into account the systematic scatter of the predicted bands in template fitting and ensure that the predicted bands are not overweighted in best-template selection. In the following section we perform a simple template fitting to evaluate the values added by predicted fluxes. However, it is worth highlighting that the better approach would be using a machine-learning model that is trained based on template-fitting results of a galaxy population with well-constrained SEDs such as COSMOS2020 (Figure 11).

6.2. Template Fitting

We perform template fitting for three cases using 1) H20-like bands, 2) H20-like+predicted YHJ bands, and 3) H20 +UVISTA-like bands. For this purpose, we split the test sample used in Section 5 into half to have a validation set as well as a new test sample. The validation sample is used to measure the 1σ scatter of the predicted flux (similar to the dashed lines in Figure 10). We assign errors to the predicted fluxes of the new test sample based on 1σ scatter of the validation sample at a given magnitude. We use a template-fitting code `LePhare` with the same configuration as Ilbert et al. (2015). This configuration differs from the templates used for COSMOS2020 redshift measurements. In the COSMOS2020 catalog, the photometric redshifts are measured based on templates employed by Ilbert et al. (2013), followed by stellar masses measured in the same manner as Ilbert et al. (2015) at fixed photometric redshifts, but here we fit both photometric redshifts and stellar masses simultaneously. Figure 14 presents the results of the template fitting for these three cases. We find that the lack of observed near-IR fluxes in template fitting increases the $\sigma_{\rm NMAD}$ and outlier fraction by 50% and 80%, respectively. We also find that adding predicted fluxes improves the $\sigma_{\rm NMAD}$ and outlier fraction by 10% and 25%, respectively. Predicted fluxes also improve the scatter of the stellar-mass measurements by 7%.

Improvement in template-fitting results by adding predicted fluxes suggests that observationally driven priors on near-IR fluxes can help reduce both scatter and outlier fraction of SED-derived properties. Moreover, we find that adding observed near-IR data significantly ($\sim$50%) improves the template-fitting results, but this is not the case for the random forest model shown in Figure 11 ($\sim$10% improvement). This suggests that machine-learning models are able to fully incorporate the information gathered from extensive surveys and avoid the degeneracies in template-fitting parameters that are inevitable when a few bands are present.

7. Discussion and Summary

In this paper, we present an information gain-based method to quantify the importance of wave bands and find the optimal set of bands needed to be observed to constrain the photometric redshifts and physical properties of galaxies. To demonstrate the application of this method we build a subsample of galaxies from the COSMOS2020 catalog with similar wave band coverage (*ugrizYJH* and IRAC/*ch*1, *ch*2) that will be available in Euclid deep fields. For a sample of galaxies with $i < 25$ AB mag, we find that given the availability of *i*-band fluxes, *r*, *u*, IRAC/*ch*2, and *z* bands provide most of the information for measuring the photometric redshifts with importance decreasing from the *r* band to the *z* band. We also find that for the same sample, IRAC/*ch*2, *Y*, *r*, and *u* bands are the most relevant bands in stellar-mass measurements with decreasing order of importance. We note that these results should be remeasured

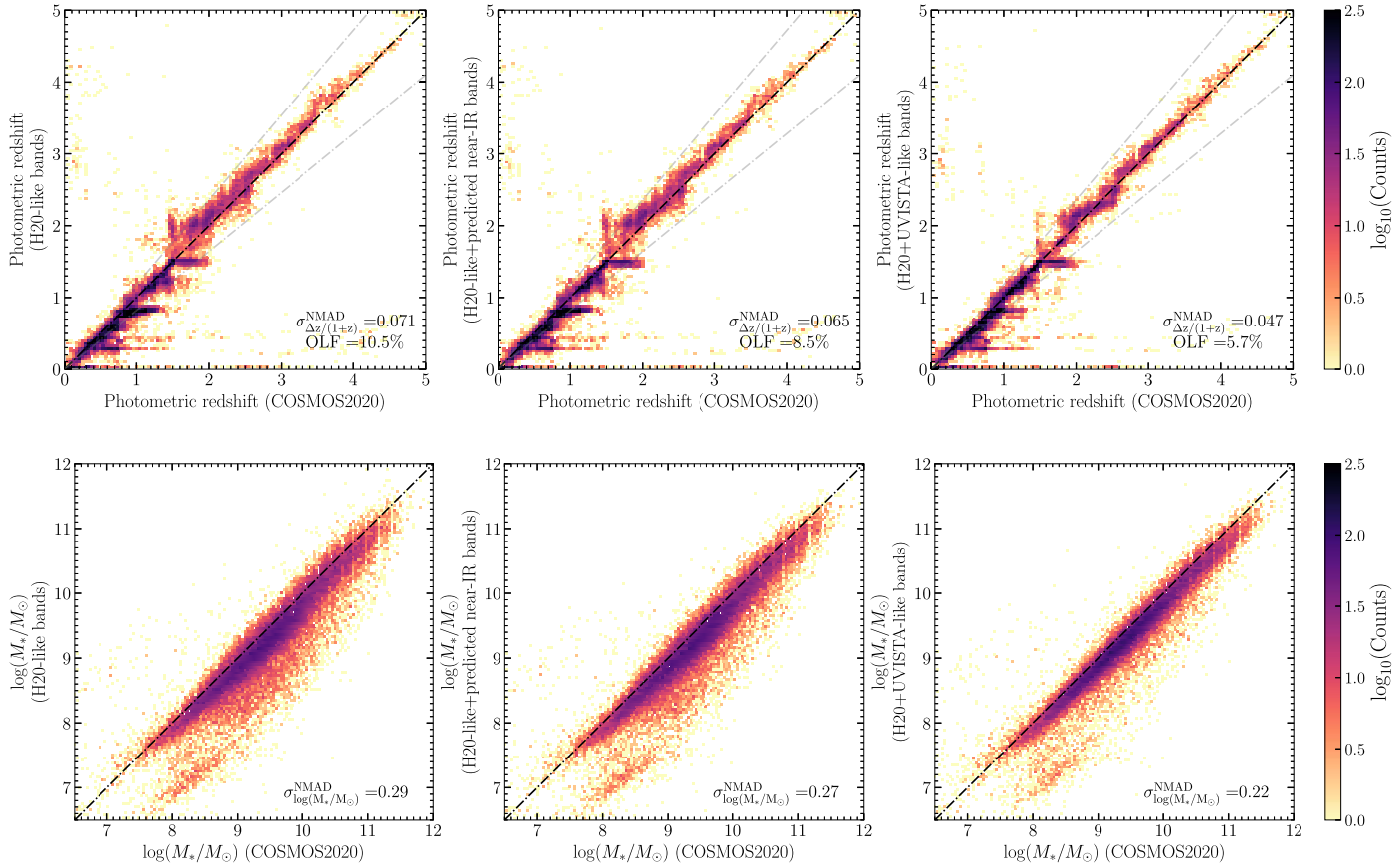


Figure 14. Template-fitting results are compared against photometric redshifts and stellar masses of the COSMOS2020 catalog (derived from 29 bands) for three cases: 1) using observed *ugrizch1ch2* bands (left panels), 2) using observed *ugrizch1ch2*+predicted *JHK* bands (middle panels), and 3) using observed *ugrizYJHch1ch2* bands (right panels).

for any new sample with different selection criteria. Moreover, we present the relative importance of wave bands for stellar-mass measurements in the bins of redshifts since their importance depends on the redshift. We also investigate the intercorrelation between the flux in different wave bands and use a machine-learning technique to predict/simulate missing fluxes from a survey. To prove the concept, we apply the method trained on the COSMOS2020 data to predict UVISTA near-IR observations based on the H20-like survey data, which include *ugriz* and Spitzer/IRAC observations. We find that near-IR bands (*YJH*) can be predicted/simulated from ground-based (*ugriz*) and mid-IR Spitzer (IRAC/*ch*1, *ch*2) observations with an accuracy of 1σ mag scatter $\lesssim$0.2 for galaxies brighter than 24 AB mag in near-IR bands. We demonstrate that theoretical templates lack such valuable information already observed through numerous bands in the COSMOS field. We conclude that degeneracies in template fitting can be alleviated if one trains a model based on template-fitting solutions for observed galaxies with extensive observations instead of using conventional SED fitting. We show that a model trained on H20-like bands has comparable accuracy to a model that is trained over H20+UVISTA-like bands, given that the model is trained over the observed galaxy population with a vast number of wave bands.

Masters et al. (2015) mapped the high-dimensional color space of COSMOS galaxies in UVISTA bands using the self-organizing map (SOM) technique (Kohonen 1982) and proposed a spectroscopy survey to fully cover regions in reduced color space with no spectroscopic redshifts. This survey, C3R2, was awarded 44.5 nights on the Keck Telescope to map the color−redshift relation necessary for weak lensing cosmology (Masters et al. 2017, 2019). Later on, Hemmati et al. (2019) used SOM to map the color space of theoretical models and used the reduced map as a fast template-fitting technique. In the present work, we use a new technique, UMAP, to create a two-dimensional representation of a high-dimensional flux distribution. This technique can also be utilized to map the color space of galaxies and study their physical properties (similar to Figure 9), providing an opportunity for further analyses that can be performed in the future.

Acquiring data for galaxy surveys over wide areas and a range of wavelengths with a large number of wave bands is costly. A new method based on machine-learning algorithms is presented in this paper to supplement the present and future surveys in their missing bands with information from previous extensive surveys (e.g., COSMOS). It can be used to optimize observations of future surveys, as well as to predict the photometry of observatories that have ceased operation (Dobbels et al. 2020).

We thank the anonymous referee for providing insightful comments and suggestions that improved the quality of this work. N.C. and A.C. acknowledge support from NASA ADAP 80NSSC20K0437. I.D. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 896225.

ORCID iDs

Nima Chartab https://orcid.org/0000-0003-3691-937X
Shoubaneh Hemmati https://orcid.org/0000-0003-2226-5395
Zahra Sattari https://orcid.org/0000-0002-0364-1159
Henry C. Ferguson https://orcid.org/0000-0001-7113-2738
David B. Sanders https://orcid.org/0000-0002-1233-9998
John R. Weaver https://orcid.org/0000-0003-1614-196X
Daniel K. Stern https://orcid.org/0000-0003-2686-9241
Henry J. McCracken https://orcid.org/0000-0002-9489-7765
Daniel C. Masters https://orcid.org/0000-0001-5382-6138
Sune Toft https://orcid.org/0000-0003-3631-7176
Peter L. Capak https://orcid.org/0000-0003-3578-6843
Iary Davidzon https://orcid.org/0000-0002-2951-7519
Mark E. Dickinson https://orcid.org/0000-0001-5414-5131
Jason Rhodes https://orcid.org/0000-0002-4485-8549
Olivier Ilbert https://orcid.org/0000-0002-7303-4397
Lukas Zalesky https://orcid.org/0000-0001-5680-2326
Anton M. Koekemoer https://orcid.org/0000-0002-6610-2048
Harry I. Teplitz https://orcid.org/0000-0002-7064-5424
Mauro Giavalisco https://orcid.org/0000-0002-7831-8751

References

Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476
Breiman, L. 2001, Mach. Learn., 45, 5
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Carrasco Kind, M., & Brunner, R. J. 2014, MNRAS, 438, 3409
Chabrier, G. 2003, PASP, 115, 763
Davidzon, I., Laigle, C., Capak, P. L., et al. 2019, MNRAS, 489, 4817
Dobbels, W., Baes, M., Viaene, S., et al. 2020, A&A, 634, A57
Euclid Collaboration, Desprez, G., Paltani, S., et al. 2020, A&A, 644, A31
Hemmati, S., Capak, P., Pourrahmani, M., et al. 2019, ApJL, 881, L14
Hoaglin, D. C., Mosteller, F., & Tukey, J. W. 1983, Understanding Robust and Exploratory Data Analysis (New York: Wiley)
Ilbert, O., Arnouts, S., Le Floc'h, E., et al. 2015, A&A, 579, A2
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55
Kohonen, T. 1982, Biol. Cybern., 43, 59
Kraskov, A., Stögbauer, H., & Grassberger, P. 2004, PhRvE, 69, 066138
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53
Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2017, ApJ, 841, 111
Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2019, ApJ, 877, 81
McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, A&A, 544, A156
McInnes, L., Healy, J., & Melville, J. 2018, arXiv:1802.03426
Moneti, A., McCracken, H. J., Shuntov, M., et al. 2022, A&A, 658, A126
Mucesh, S., Hartley, W. G., Palmese, A., et al. 2021, MNRAS, 502, 2770
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Shannon, C. E. 1948, BSTJ, 27, 379
Simet, M., Chartab, N., Lu, Y., & Mobasher, B. 2021, ApJ, 908, 47
Steinhardt, C. L., Weaver, J. R., Maxfield, J., et al. 2020, ApJ, 891, 136
van der Maaten, L., & Hinton, G. 2008, J. Mach. Learn. Res., 9, 2579
Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, ApJS, 258, 11