



Publication Year	2024
Acceptance in OA	2026-01-14T13:22:51Z
Title	Euclid preparation: LI. Forecasting the recovery of galaxy physical properties and their relations with template-fitting and machine-learning methods
Authors	Euclid Collaboration, Enia, A., BOLZONELLA, Micol, POZZETTI, Lucia, Humphrey, A., Cunha, P. A. C., Hartley, W. G., Dubath, F., Paltani, S., Lopez Lopez, X., Quai, S., BARDELLI, Sandro, BISIGELLO, Laura, CAVUOTI, STEFANO, DE LUCIA, GABRIELLA, Ginolfi, M., GRAZIAN, Andrea, Siudek, M., TORTORA, CRESCENZO, Zamorani, G., Aghanim, N., Altieri, B., Amara, A., ANDREON, Stefano, AURICCHIO, NATALIA, Baccigalupi, C., Baldi, M., Bender, R., Bodendorf, C., BONINO, Donata, Branchini, Enzo, BRESCIA, Massimo, Brinchmann, J., Camera, S., Capobianco, Vito, CARBONE, Carmelita, Carretero, J., Casas, S., Castander, F. J., CASTELLANO, Marco, Castignani, G., Cimatti, A., Colodro-Conde, C., Congedo, G., Conselice, C. J., Conversi, L., Copin, Y., CORCIONE, Leonardo, Courbin, F., Courtois, H. M., Da Silva, A., Degaudenzi, H., DI GIORGIO, Anna Maria, Dinis, J., Dupac, X., Dusini, S., Fabricius, M., FARINA, Maria, Farrens, S., Ferriol, S., Fosalba, P., Fotopoulou, S., FRAILIS, Marco, FRANCESCHI, ENRICO, FUMANA, Marco, GALEOTTA, Samuele, Gillis, B., GIOCOLI, Carlo, Grupp, F., Haugan, S. V. H., Holmes, W., Hook, I., Hormuth, F., Hornstrup, A., Jahnke, K., Joachimi, B., Keihänen, E., Kermiche, S., Kiessling, A., Kubik, B., Kümmel, M., Kunz, M., Kurki-Suonio, H., LIGORI, Sebastiano, Lilje, P. B., Lindholm, V., Lloro, I., MAIORANO, Elisabetta, MANSUTTI, Oriana, Marggraf, O., Markovic, K., Martinelli, M., Martinet, N., Marulli, F., Massey, R., McCracken, H. J., Medinaceli, E., Mei, S., Melchior, M., Mellier, Y., MENEGHETTI, MASSIMO, MERLIN, Emiliano, Meylan, G., Moresco, M., Moscardini, L., MUNARI, Emiliano, Neissner, C., Niemi, S. -M., Nightingale, J. W., Padilla, C., Pasian, F., Pedersen, K., Pettorino, V., Polenta, G., Poncet, M., Popa, L. A., Raison, F., Rebolo, R., Renzi, A., Rhodes, J., RICCIO, GIUSEPPE, ROMELLI, Erik, Roncarelli, M., Rossetti, E., Saglia, R., Sakr, Z., Sapone, D., Schneider, P., Schrabback, T., SCODEGGIO, MARCO, Secroun, A., SEFUSATTI, Emiliano, Seidel, G., Serrano, S., Sirignano, C., Sirri, G., Stanco, L., Steinwagner, J., Surace, C., Tallada-Crespí, P., TAVAGNACCO, Daniele, Taylor, A. N., Teplitz, H. I., Tereno, I., Toledo-Moreo, R., Torradeflot, F., Tutusaus, I., VALENZIANO, Luca, Vassallo, T., Verdoes Kleijn, G., Veropalumbo, A., Wang, Y., Weller, J., ZUCCA, Elena, BIVIANO, ANDREA, Boucaud, A., BURIGANA, Carlo, Calabrese, M., Escartin Vigo, J. A., Gracia-Carpio, J., Mauri, N., Pezzotta, A., Pöntinen, M., Porciani, C., Scottez, V., Tenti, M., VIEL, Matteo, Wiesmann, M., Akrami, Y., Allevato, V., Anselmi, S., Ballardini, M., Bergamini, P., Bethermin, M., Blanchard, A., Blot, L., BORGANI, Stefano, Bruton, S., Cabanac, R., Calabro, A., Canas-Herrera, G., CAPPI, Alberto, Carvalho, C. S., Castro, T., Chambers, K. C., Contarini, S., Contini, T., Cooray, A. R., CUCCIATI, Olga, Davini, S., De Caro, B., Desprez, G., Díaz-Sánchez, A., Di Domizio, S., Dole, H., Escoffier,

Euclid preparation

II. Forecasting the recovery of galaxy physical properties and their relations with template-fitting and machine-learning methods

Euclid Collaboration: A. Enia^{1,2,*}, M. Bolzonella², L. Pozzetti², A. Humphrey^{3,4}, P. A. C. Cunha^{5,3}, W. G. Hartley⁶, F. Dubath⁶, S. Paltani⁶, X. Lopez Lopez^{1,2}, S. Quai^{1,2}, S. Bardelli², L. Bisigello^{7,8}, S. Cavauoti^{9,10}, G. De Lucia¹¹, M. Ginolfi^{12,13}, A. Grazian¹⁴, M. Siudek^{15,16}, C. Tortora⁹, G. Zamorani², N. Aghanim¹⁷, B. Altieri¹⁸, A. Amara¹⁹, S. Andreon²⁰, N. Auricchio², C. Baccigalupi^{21,11,22,23}, M. Baldi^{24,2,25}, R. Bender^{26,27}, C. Bodendorf²⁶, D. Bonino²⁸, E. Branchini^{29,30,20}, M. Brescia^{31,9,10}, J. Brinchmann³, S. Camera^{32,33,28}, V. Capobianco²⁸, C. Carbone³⁴, J. Carretero^{35,36}, S. Casas³⁷, F. J. Castander^{16,38}, M. Castellano³⁹, G. Castignani², A. Cimatti⁴⁰, C. Colodro-Conde⁴¹, G. Congedo⁴², C. J. Conselice⁴³, L. Conversi^{44,18}, Y. Copin⁴⁵, L. Corcione²⁸, F. Courbin⁴⁶, H. M. Courtois⁴⁷, A. Da Silva^{48,49}, H. Degaudenzi⁶, A. M. Di Giorgio⁵⁰, J. Dinis^{48,49}, X. Dupac¹⁸, S. Dusini⁵¹, M. Fabricius^{26,27}, M. Farina⁵⁰, S. Farrens⁵², S. Ferriol⁴⁵, P. Fosalba^{38,53}, S. Fotopoulou⁵⁴, M. Frailis¹¹, E. Franceschi², M. Fumana³⁴, S. Galeotta¹¹, B. Gillis⁴², C. Giocoli^{2,55}, F. Grupp^{26,27}, S. V. H. Haugan⁵⁶, W. Holmes⁵⁷, I. Hook⁵⁸, F. Hormuth⁵⁹, A. Hornstrup^{60,61}, K. Jahnke⁶², B. Joachimi⁶³, E. Keihänen⁶⁴, S. Kermiche⁶⁵, A. Kiessling⁵⁷, B. Kubik⁴⁵, M. Kümmel²⁷, M. Kunz⁶⁶, H. Kurki-Suonio^{67,68}, S. Ligorì²⁸, P. B. Lilje⁵⁶, V. Lindholm^{67,68}, I. Lloro⁶⁹, E. Maiorano², O. Mansutti¹¹, O. Marggraf⁷⁰, K. Markovic⁵⁷, M. Martinelli^{39,71}, N. Martinet⁷², F. Marulli^{1,2,25}, R. Massey⁷³, H. J. McCracken⁷⁴, E. Medinaceli², S. Mei⁷⁵, M. Melchior⁷⁶, Y. Mellier^{77,74}, M. Meneghetti^{2,25}, E. Merlin³⁹, G. Meylan⁴⁶, M. Moresco^{1,2}, L. Moscardini^{1,2,25}, E. Munari^{11,21}, C. Neissner^{78,36}, S.-M. Niemi⁷⁹, J. W. Nightingale^{80,73}, C. Padilla⁷⁸, F. Pasian¹¹, K. Pedersen⁸¹, V. Pettorino⁷⁹, G. Polenta⁸², M. Poncet⁸³, L. A. Popa⁸⁴, F. Raison²⁶, R. Rebolo^{41,85}, A. Renzi^{8,51}, J. Rhodes⁵⁷, G. Riccio⁹, E. Romelli¹¹, M. Roncarelli², E. Rossetti²⁴, R. Saglia^{27,26}, Z. Sakr^{86,87,88}, D. Sapone⁸⁹, P. Schneider⁷⁰, T. Schrabback⁹⁰, M. Scodeggio³⁴, A. Secroun⁶⁵, E. Sefusatti^{11,21,22}, G. Seidel⁶², S. Serrano^{38,91,16}, C. Sirignano^{8,51}, G. Sirri²⁵, L. Stanco⁵¹, J. Steinwagner²⁶, C. Surace⁷², P. Tallada-Crespi^{35,36}, D. Tavagnacco¹¹, A. N. Taylor⁴², H. I. Teplitz⁹², I. Tereno^{48,93}, R. Toledo-Moreo⁹⁴, F. Torradeflot^{36,35}, I. Tutusaus⁸⁷, L. Valenziano^{2,95}, T. Vassallo^{27,11}, G. Verdoes Kleijn⁹⁶, A. Veropalumbo^{20,30,97}, Y. Wang⁹², J. Weller^{27,26}, E. Zucca², A. Biviano^{11,21}, A. Boucaud⁷⁵, C. Burigana^{7,95}, M. Calabrese^{98,34}, J. A. Escartin Vigo²⁶, J. Gracia-Carpio²⁶, N. Mauri^{40,25}, A. Pezzotta²⁶, M. Pöntinen⁶⁷, C. Porciani⁷⁰, V. Scottez^{77,99}, M. Tenti²⁵, M. Viel^{21,11,23,22,100}, M. Wiesmann⁵⁶, Y. Akrami^{101,102}, V. Alleinato⁹, S. Anselmi^{51,8,103}, M. Ballardini^{104,2,105}, P. Bergamini^{106,2}, M. Bethermin^{107,72}, A. Blanchard⁸⁷, L. Blot^{108,103}, S. Borgani^{109,21,11,22}, S. Bruton¹¹⁰, R. Cabanac⁸⁷, A. Calabro³⁹, G. Canas-Herrera^{79,111}, A. Cappi^{2,112}, C. S. Carvalho⁹³, T. Castro^{11,22,21,100}, K. C. Chambers¹¹³, S. Contarini²⁶, T. Contini⁸⁷, A. R. Cooray¹¹⁴, O. Cucciati², S. Davini³⁰, B. De Caro³⁴, G. Desprez¹¹⁵, A. Díaz-Sánchez¹¹⁶, S. Di Domizio^{29,30}, H. Dole¹⁷, S. Escoffier⁶⁵, A. G. Ferrari^{40,25}, P. G. Ferreira¹¹⁷, I. Ferrero⁵⁶, A. Finoguenov⁶⁷, F. Fornari⁹⁵, L. Gabarra¹¹⁷, K. Ganga⁷⁵, J. García-Bellido¹⁰¹, V. Gautard¹¹⁸, E. Gaztanaga^{16,38,119}, F. Giacomini²⁵, F. Gianotti², G. Gozalias^{120,67}, A. Hall⁴², S. Hemmati¹²¹, H. Hildebrandt¹²², J. Hjorth¹²³, A. Jimenez Muñoz¹²⁴, S. Joudaki¹¹⁹, J. J. E. Kajava^{125,126}, V. Kansal^{127,128}, D. Karagiannis^{129,130}, C. C. Kirkpatrick⁶⁴, J. Le Graet⁶⁵, L. Legrand¹³¹, A. Loureiro^{132,133}, J. Macias-Perez¹²⁴, G. Maggio¹¹, M. Magliocchetti⁵⁰, C. Mancini³⁴, F. Mannucci¹³, R. Maoli^{134,39}, C. J. A. P. Martins^{135,3}, S. Matthew⁴², L. Maurin¹⁷, R. B. Metcalf^{1,2}, P. Monaco^{109,11,22,21}, C. Moretti^{23,100,11,21,22}, G. Morgante², Nicholas A. Walton¹³⁶, L. Patrizii²⁵, V. Popa⁸⁴, D. Potter¹³⁷, I. Risso⁹⁷, P.-F. Rocci¹⁷, M. Sahlén¹³⁸, A. Schneider¹³⁷, M. Schultheis¹¹², M. Sereno^{2,25}, P. Simon⁷⁰, A. Spurio Mancini^{139,140}, S. A. Stanford¹⁴¹, K. Tanidis¹¹⁷, C. Tao⁶⁵, G. Testera³⁰, R. Teyssier¹⁴², S. Toft^{61,143}, S. Tosi^{29,30}, A. Troja^{8,51}, M. Tucci⁶, C. Valieri²⁵, J. Valiviita^{67,68}, D. Vergani², G. Verza^{144,145}, I. A. Zinchenko²⁷, G. Rodighiero^{8,14}, and M. Talia^{1,2}

(Affiliations can be found after the references)

Received 8 July 2024 / Accepted 12 September 2024

* Corresponding author; andrea.enia@unibo.it

ABSTRACT

Euclid will collect an enormous amount of data during the mission’s lifetime, observing billions of galaxies in the extragalactic sky. Along with traditional template-fitting methods, numerous machine learning (ML) algorithms have been presented for computing their photometric redshifts and physical parameters (PPs), requiring significantly less computing effort while producing equivalent performance measures. However, their performance is limited by the quality and amount of input information entering the model (the features), to a level where the recovery of some well-established physical relationships between parameters might not be guaranteed – for example, the star-forming main sequence (SFMS). To forecast the reliability of *Euclid* photo-*z*s and PPs calculations, we produced two mock catalogs simulating the photometry with the UNIONS *ugriz* and *Euclid* filters. We simulated the Euclid Wide Survey (EWS) and Euclid Deep Fields (EDF), alongside two auxiliary fields. We tested the performance of a template-fitting algorithm (Phosphoros) and four ML methods in recovering photo-*z*s, PPs (stellar masses and star formation rates), and the SFMS on the simulated *Euclid* fields. To mimic the *Euclid* processing as closely as possible, the models were trained with Phosphoros-recovered labels and tested on the simulated ground truth. For the EWS, we found that the best results are achieved with a mixed labels approach, training the models with wide survey features and labels from the Phosphoros results on deeper photometry, that is, with the best possible set of labels for a given photometry. This imposes a prior to the input features, helping the models to better discern cases in degenerate regions of feature space, that is, when galaxies have similar magnitudes and colors but different redshifts and PPs, with performance metrics even better than those found with Phosphoros. We found no more than 3% performance degradation using a COSMOS-like reference sample or removing *u* band data, which will not be available until after data release DR1. The best results are obtained for the EDF, with appropriate recovery of photo-*z*, PPs, and the SFMS.

Key words. methods: data analysis – surveys – galaxies: evolution – galaxies: fundamental parameters – galaxies: general

1. Introduction

*Euclid*¹ is an European Space Agency mission whose primary objective is to reveal the geometry of the Universe by measuring precise distances and shapes of $\sim 10^9$ galaxies up to $z \sim 3$, while it is also predicted to observe millions of galaxies at $3 < z < 6$ (Euclid Collaboration 2024f). *Euclid* will observe the extragalactic sky in four optical and near-infrared (NIR) filters: I_E , corresponding to r , i , and z filters (Euclid Collaboration 2024b); and Y_E , J_E , and H_E on the Near Infrared Spectrometer and Photometer (NISF; Euclid Collaboration 2024c). Such a wealth of data will dramatically improve our knowledge of the evolution of galaxies throughout cosmic time.

The Euclid Wide Survey (EWS) will cover $13\,345\,\text{deg}^2$ of the sky up to a 5σ point-like source depth of 26.2 mag in I_E and 24.5 mag in Y_E , J_E , and H_E (Euclid Collaboration 2024h, 2022b). The Euclid Deep Fields (EDF) will probe a smaller ($\sim 53\,\text{deg}^2$) area to a targeted 5σ point-like source depth of 28.2 in I_E and 26.5 in Y_E , J_E , and H_E . In total, *Euclid* is expected to detect approximately ten billion sources and determine roughly 30 million spectroscopic redshifts (e.g., Laureijs et al. 2011). The *Euclid* observations will be complemented with ground-based data from the Ultraviolet Near-Infrared Optical Northern Survey (UNIONS, e.g., Ibata et al. 2017), the Legacy Survey of Space and Time (LSST, Ivezić et al. 2008; LSST Science Collaboration 2009), and the Dark Energy Survey (DES, Flaugher et al. 2015; Dark Energy Survey Collaboration 2016), in order to have a complete wavelength coverage between $0.3\,\mu\text{m}$ and $1.8\,\mu\text{m}$.

Such a vast amount of data are out of computational reach for traditional template-fitting algorithms, which aim to model the observed spectral energy distribution (SED) with a set of synthetic templates searching for the best fit parameters (i.e., photometric redshifts, stellar masses, and star formation rates) with computational times scaling linearly with the number of objects involved. For this reason, a wide set of machine learning (ML) techniques have been proposed, developed, tested, and used to extract the maximum scientific information from such a huge amount of data, especially for the photo-*z*s (Euclid Collaboration 2020, requiring a precision of $\sigma_z < 0.05$ and $< 10\%$ outlier fraction), with the intention of speeding up the computational

efforts while yielding comparable (or even better) performance in recovering the quantities of interest.

The past decade has seen an incredible surge in the use of ML methods for astrophysical data analysis in virtually every possible subfield, from identification and modeling of strong lensing systems (Hezaveh et al. 2017; Gentile et al. 2022, 2023; Euclid Collaboration 2024d), to classification tasks aiming to automatically identifying objects in images and catalogs, or to measure morphologies (Huertas-Company et al. 2015; Dieleman et al. 2015; Tuccillo et al. 2018; Bowles et al. 2021; Guarneri et al. 2021; Cunha & Humphrey 2022; Li et al. 2022a; Euclid Collaboration 2024a; Signor et al. 2024), to regression tasks, for example in finding the relationship between the photometric redshifts and/or physical properties from the observed photometry (Tagliaferri et al. 2003; Collister & Lahav 2004; Brescia et al. 2013; Cavuoti et al. 2017; D’Isanto & Polsterer 2018; Ucci et al. 2018; Bonjean et al. 2019; Delli Veneri et al. 2019; Pasquet et al. 2019; Surana et al. 2020; Mucesh et al. 2021; Razim et al. 2021; Simet et al. 2021; Davidzon et al. 2022; Li et al. 2022b; Carvajal et al. 2023; Euclid Collaboration 2023; Alsing et al. 2023; Leistedt et al. 2023; Alsing et al. 2024; Thorp et al. 2024). Astrophysics has entered the big data era, and the potential of ML methods has been revealed to the whole community.

However, as powerful as they can be, ML techniques are not flawless. The goodness of the predicted quantities is inevitably limited by the quality (and size) of the input information used to train the model. Noisy features hamper a plain association between them and the desired outputs, degrading the final performance to a level where the optimal recovery of the most important quantities to place an observational constraint on galaxy evolution models might not be guaranteed at all. Some kind of agnostic analysis on the performance of ML methods is necessary, as it is determining how those benchmark against classical methods (i.e., template-fitting).

Therefore, it is crucial to evaluate the *Euclid* (and complementary data) capability to recover photometric redshifts, physical parameters (PPs), and the relationships between those, such as the star forming main sequence (SFMS, Daddi et al. 2007; Rodighiero et al. 2014), and doing so in the most realistic way possible. This will help put the forthcoming EWS and EDF results into a more stable context and could act as a benchmark

¹ <https://sci.esa.int/euclid/>

for those that will be obtained by the forthcoming large-area surveys of the next decade, LSST with the *Vera C. Rubin* Observatory (Ivezic et al. 2008), and the *Nancy Roman* Space Telescope (Akeson et al. 2019).

Euclid was successfully launched on July 1, 2023, reaching its observing orbit around the second Lagrange point (L2) the following month. The first public Data Release (DR1), covering $\sim 2500 \text{ deg}^2$, is expected to be in June 2026. In the meantime, in order to estimate the performance of the survey's retrieved physical parameters (and relations), we make use of mock catalogs built from simulations, for which the ground truth (i.e., the real value of the physical parameters) is known.

This paper is outlined as follows: In Sect. 2, we describe the simulations from which we built *Euclid* and ground-based photometry as inputs to the ML models and test their performance. In Sect. 3, we describe the template-fitting and ML methods used. In Sect. 4, we report the results, focusing in particular on the EWS and EDF and what can be done to improve the recovery of photo-zs and physical parameters with the *Euclid* data products. In Sect. 5, we present our conclusions and perspectives on other upcoming wide-area surveys.

In this work, we adopt a flat Lambda cold dark matter (Λ CDM) cosmology with $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.3$, and $\Omega_\Lambda = 0.7$, and assume a Chabrier (2003) initial mass function (IMF). The magnitudes are given in the AB photometric system (Oke & Gunn 1983).

2. Building the mock catalogs

Assessing how good the *Euclid* observations will yield to photometric redshifts and physical parameters necessarily passes through the use of simulated data, for which the ground truth is known. We want these simulations to be as close as possible to the real *Euclid* data, which will not be available until DR1.

2.1. The MAMBO workflow

In this work, we use the *Mocks with Abundance Matching in BOlogna* (MAMBO) workflow (see Girelli 2021, for a thorough description). MAMBO starts from an N-body dark matter simulation to build an empirical mock catalog of galaxies, reproducing their observed physical properties and observables with high accuracy. The cosmological simulation used here is the Millennium dark matter N-body Simulation (Springel et al. 2005), matched to the Planck cosmology following Angulo & White (2010), with a lightcone taken from Henriques et al. (2015), covering 3.14 deg^2 with sub-halo masses $M_{200} > 1.7 \times 10^{10} h^{-1} M_\odot$ up to $z = 6$. Considering a typical stellar-to-halo mass relation (SHMR, Girelli et al. 2020), the corresponding stellar mass at low redshift is on the order of $\log_{10}(M_\star/M_\odot) = 7.5$. In COSMOS2020 (Weaver et al. 2022), galaxies with such a small stellar mass at low redshift are characterized by a H band magnitude of $m_H \sim 25.2$. This is therefore the limit to be considered for the completeness of the MAMBO simulation at very low redshifts $z < 0.2$; however, given that the volume of the simulation is very small at such redshifts, the incompleteness in the case of the simulated EDF is negligible, and the simulation can be considered complete in all the explored regimes. The simulation extends to higher redshifts, in principle, but we cut it at $z = 6$, as it is the default limit of the main *Euclid* pipeline for photometric redshifts.

Starting from the lightcone, the main parameters that we use are the position of each halo in RA and Dec, its redshift z , and the DM sub-halo mass. MAMBO assigns to each galaxy its prop-

erties following empirical prescriptions with a scatter that randomizes the properties. In this way, not only do we ensure a better representation of the observed universe, but we also avoid the possible replication of galaxies that would be caused by a deterministic approach. As for the stellar masses $M_\star$, those come from a SHMR developed using a sub-halo abundance matching technique based on observed stellar mass functions (SMFs) on the SDSS, COSMOS, and CANDELS fields (Girelli et al. 2020). The SMFs are:

- Peng et al. (2010), measured in the SDSS survey and divided into passive and star-forming using the rest-frame ($U - B$) color at $z \sim 0$;
- Ilbert et al. (2013), measured in COSMOS and classified into red or blue using the rest frame color selection ($NUV - r$) vs ($r - J$) at $0.2 < z < 4$;
- Grazian et al. (2015), measured in CANDELS at $z \geq 4$.

Every galaxy is randomly assigned a star-forming or passive and quiescent label based on the ratio of the stellar mass functions (SMFs) for the blue and red populations. Due to the high observational uncertainties of the fraction of SF/Q galaxies at $z > 4$ (Merlin et al. 2018; Girelli et al. 2019), the star-forming fraction f_{SF} was extrapolated from the results at lower redshifts with a limit of $f_{\text{SF}} = 99\%$ up to $z = 6$.

All the other properties, for example, SFR, metallicity, rest-frame, and observed photometry from UV to submillimeter in the desired bands, are extracted with the Empirical Galaxy Generator (EGG, Schreiber et al. 2017), a C++ code that creates a mock catalog of galaxies from a simulated lightcone, whose empirical nature assures that the retrieved physical properties are realistic – as long as the EGG models are. In the configuration of EGG used for MAMBO, each galaxy SED is assigned from a pre-built library of templates from the Bruzual & Charlot (2003) models covering the UVJ -plane (Williams et al. 2009). Models in the library are derived with a Salpeter IMF (Salpeter 1955), but we subsequently converted stellar masses and SFRs to a Chabrier IMF (Chabrier 2003). The physical properties (and type, i.e., star-forming or quiescent) are randomly extracted using empirical relations starting from the stellar mass previously assigned, once again covering the full UVJ -plane.

With MAMBO, we generate a mock catalog of roughly five million galaxies between redshifts zero and six, with the same photometric filters as the ones expected for DR1 in the EWS in the northern hemisphere, where a network of multiple collaborations will obtain data in different bands as part of the Ultraviolet Near-Infrared Optical Northern Survey (UNIONS), whose throughput is shown in Fig. 1. These are the Canada-France Imaging Survey (CFIS; Ibata et al. 2017, on the Canada-France-Hawaii Telescope CFHT) for bands u and r ; Subaru Hyper Suprime-Cam (HSC; Miyazaki et al. 2018) observations for z and g bands as part of the Wide Imaging with Subaru HSC of the *Euclid* Sky (WISHES) and the Waterloo Hawaii IfA G band Survey (WHIGS); PAN-STARRS in band i (Chambers et al. 2016), and the *Euclid* I_E , Y_E , J_E and H_E filters (Cropper et al. 2016; Maciaszek et al. 2016).

The EDF has already been observed with the *Spitzer* Space Telescope's Infrared Array Camera (IRAC, Werner et al. 2004; Fazio et al. 2004) at $3.6 \mu\text{m}$ and $4.5 \mu\text{m}$. These observations are described in detail in Euclid Collaboration (2022a, 2024e). When dealing with the EDF, we also include these two photometric filters, assuming the same observation depth reported in Euclid Collaboration (2024e).

For convenience, the full set of filters is also listed in Table 1, with the corresponding expected 10σ observation depths for a generic extended source (in a $2''$ aperture, i.e., a typical *Euclid*

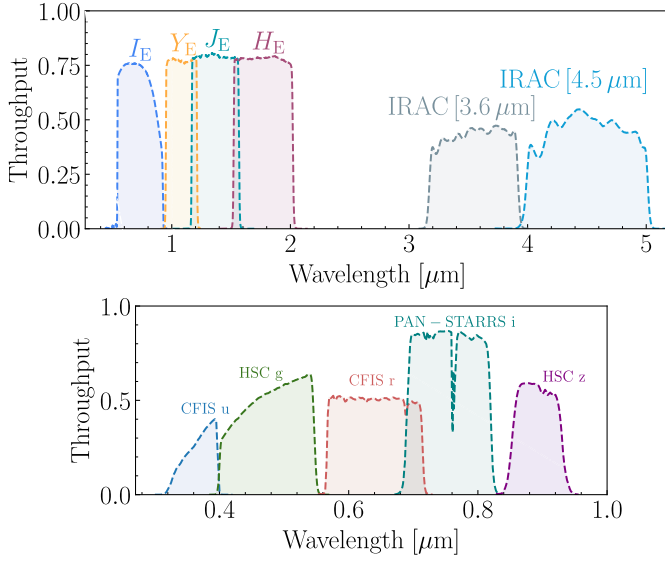


Fig. 1. Throughput of the filters used through this work. On the top panel we show the four *Euclid* filters: I_E , Y_E , J_E , and H_E , along with two IRAC filters at 3.6 μm and 4.5 μm . In the bottom panel, we include the four UNIONS filters that will complement the *Euclid* data in the northern sky: CFIS u and r , HSC g and z , and PAN-STARRS i .

Table 1. Set of filters used in this work.

Band	λ_{eff} [μm]	EWS
CFHT/MegaCam u	0.372	23.6
HSC g	0.480	24.5
CFHT/MegaCam r	0.640	24.1
PAN-STARRS i	0.755	23.2
HSC z	0.891	23.4
VIS/ I_E	0.715	25.0
NISP/ Y_E	1.085	23.5
NISP/ J_E	1.375	23.5
NISP/ H_E	1.773	23.5
IRAC/[3.6 μm]	3.550	24.8
IRAC/[4.5 μm]	4.493	24.7

Notes. Reported magnitudes are the 10σ expected observational depths for an extended source in a $2''$ diameter aperture. IRAC observations will be available only for the EDF; as such, their reported magnitudes are the ones measured in the EDF-N and EDF-F (see Euclid Collaboration 2022a, 2024e).

extended source) per band in the EWS – with attached IRAC observed depths in the same aperture for the EDF.

2.2. The *Euclid* simulated fields

We simulate different versions of *Euclid* observations by adding realistic photometric noise to each band depending on the number of reference observation sequences that are going to be observed (ROS, see Fig. 8 of Euclid Collaboration 2024h) and the expected limiting magnitudes of the survey. A galaxy is kept in the catalog if it is detected either in H_E at $S/N > 5$ or in I_E at $S/N > 10$, given the expected limiting magnitude. Those limits were used because they enable a posteriori selections for other *Euclid* analyses, such as cluster detection and weak lensing analysis.

The four simulated catalogs (see Table 2 and Fig. 2) are:

Table 2. Four simulated *Euclid* catalogs.

	ROS	Sources	I_E lim	H_E lim
EWS	1	512 527	25.00	23.50
C16	16	1 209 598	26.50	25.00
C25	25	1 361 041	26.75	25.25
EDF	40	1 534 023	27.00	25.50

Notes. The limits are the 10σ (I_E) and 5σ (H_E) expected observational depths for an extended source in a $2''$ diameter aperture.

- Wide, a single ROS at limiting magnitudes of $H_{E,\text{lim}} = 23.5$ and $I_{E,\text{lim}} = 25.0$, simulating what is expected from the EWS (Euclid Collaboration 2024h).
- C16, 16 ROS at limiting magnitudes of $H_{E,\text{lim}} = 25.0$ and $I_{E,\text{lim}} = 26.5$, corresponding to a limit 1.5 mags deeper than the EWS. This simulates the so-called *Euclid* auxiliary fields (Euclid Collaboration 2024h), six well-known regions with vast ancillary information, observed for photometric and color calibration; 16 ROS are expected to be observed by the time of DR1.
- C25, 25 ROS at limiting magnitudes of $H_{E,\text{lim}} = 25.25$ and $I_{E,\text{lim}} = 26.75$, corresponding to a limit of 1.75 mags deeper than EWS. This simulates the expected final average number of ROS to the *Euclid* auxiliary fields.
- Deep, 40 ROS reaching limiting magnitudes of $H_{E,\text{lim}} = 25.5$ and $I_{E,\text{lim}} = 27.0$, corresponding to an expected limiting magnitude of 2 mags deeper than EWS, simulating the minimum number of ROS of the different fields composing the EDF (north, south, and Fornax).

We notice that the magnitude limits reported here are different from the ones in Euclid Collaboration (2024h), which refer to point-sources at 5σ . Here instead, we convert those to 10σ limits for an extended source with a $2''$ aperture (as a proxy for a typical *Euclid* extended source).

We are building the calibration fields by improving the EWS photometry on the *Euclid* and UNIONS filters; however, the real auxiliary fields, such as the Cosmic Evolution Survey field (COSMOS; Scoville et al. 2007) will benefit from a wealth of multiwavelength ancillary data (i.e., the COSMOS2020 catalog, Weaver et al. 2022) that will yield better photometric redshifts and physical parameter estimation with respect to what we report in this work.

There are a few caveats about the simulated catalogs. While the photometric noise for all the considered mock catalogs is simulated in the most realistic way possible, we are still dealing with an idealized situation where the photometric procedures are bypassed. Moreover, within the catalogs, we are considering galaxies only, without accounting for any source of contamination that the real *Euclid* data will have to deal with: contaminants such as stars, photometric masks (e.g., from stars) and defects (snowballs, cosmic rays, persistence from solar flares), AGN and QSOs (López-López et al. 2024), under-deblended and over-deblended objects, Local Universe extended objects, and low surface brightness galaxies. While all of those are expected to be reduced to the minimum possible (i.e., by exploiting ML to automatically classify stars and galaxies, see e.g. Cunha & Humphrey 2022, whose reported F1-scores are $\sim 98\%$), some degree of performance degradation will be unavoidable.

Finally, as the absolute best-case scenario, yielding the best possible value for each quality metric, we report the results coming from the unperturbed version of the survey, that is, the MAMBO generated catalog without any photometric noise added,

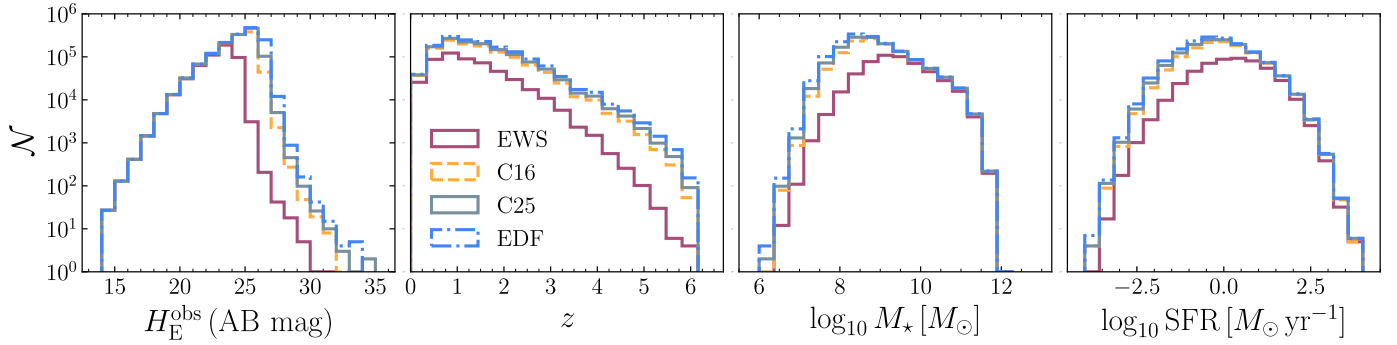


Fig. 2. Four simulated *Euclid* catalogs used in this work (solid purple line, EWS; dashed orange line, C16; solid gray line, C25; dashed-dotted blue line, EDF) shown as the number of sources as a function of the *Euclid* H_E band magnitude (leftmost panel), redshift (center left), stellar mass (center right), and star formation rate (rightmost panel). We notice that the magnitude cut upon which the fields are built is an OR condition on the S/N in H_E , I_E filters; as such, simulated galaxies are found below the nominal limiting magnitude cut for H_E band, as those are detected at S/N > 10 in I_E .

run on the ground-truth magnitudes. Regardless of the flaws that may be inherent in the simulations, whatever uncertainty is generated from this set of photometric values and parameters depends only on the technique used to derive the second from the first, such as badly interpolated holes in the feature space for ML algorithms or a lack of SED models and degeneracies between colors and physical properties for template-fitting algorithms.

3. Methodology

In this section, we describe the algorithms and metrics used to assess the model's performance in recovering the ground truth. In particular, we focus on the recovery of photometric redshifts z and two physical parameters: stellar masses $\log_{10}(M_*/M_\odot)$ and star formation rates $\log_{10}(\text{SFR}/M_\odot \text{ yr}^{-1})$, and the relation between them, the star-forming main-sequence (SFMS).

3.1. Feature, labels, and samples

In line with the standard ML terminology, we now designate the catalogs' photometry and subproducts (i.e., broad-band photometry and colors) as features, and the model output (i.e., redshifts and physical properties) as labels. In this work, we address two versions of the latter:

- the true labels, which are the ground-truth z and physical properties extracted from MAMBO;
- the recovered labels, whose values have been obtained by running a traditional template-fitting code (Phosphoros, see Sect. 3.3) on the (simulated) *Euclid* observed features.

Thus, we can check what the best possible performance for a particular run is (i.e., when the redshifts and physical parameters are perfectly known) to compare with the more realistic ones that will be obtained with *Euclid* data, when the ground truth will be inevitably unknown and recovered labels will be built from the observed features as the model input samples. This information is useful, especially in cases where the EWS performance are evaluated with a reference sample built from the calibration fields (see in the next paragraphs). As reported in Sect. 2.2, the simulated ones have labels recovered from the same set of filters as the EWS, but the real ones will benefit from lots more multiwavelength data, with better recovered photo- z and PPs. The expected real performance of such cases should therefore be in between the recovered and true labels performance.

Every supervised ML application is composed of a training (or reference²) sample from which the relations between features and labels are inferred, and a target (or test³) sample on which the models are applied. As described in Sect. 2, we have four different simulated versions of *Euclid* observations, mimicking the expected outcome of the EWS, EDF, and two calibration (auxiliary) fields. As common practice in ML applications, we split those catalogs, using part of the sample for training (or as references) and the rest equally split for cross-validation and testing. In this work, we used 90% of the samples for training (translating into training sets between 500 k and one million galaxies), ensuring at least ~ 50 k galaxies in the test sample, which is more than enough to evaluate the model's performance.

To understand the actual performance expected from the observed *Euclid* data, we explore the predictive capabilities of models trained on deeper photometry when applied to a shallower one (in this case, the EWS). For instance, this is achieved by training a model using the EDF catalog and subsequently evaluating it on the EWS catalog. In those cases, at test time, we share the same set of train and test sources between the catalogs to be as consistent as possible. The same is done for every ML method used in this work, and we share the same train, reference, and test samples for the same catalog permutation between different models (e.g., when testing the performance of a model trained on the EDF catalog and tested on the EWS one, the training and test source IDs are the same for all the methods considered).

When dealing with recovered labels, in order to simulate a typical application where the ground truth is unknown as *Euclid* will observe photometry from which the photo- z s and physical parameters will be derived, we train the models on those and test on the true labels.

3.2. Features engineering

As reported in Sect. 2, in standard ML terminology, the catalogs observed photometric values are the features of the mod-

² For some ML methods (i.e., the nearest-neighbors algorithms), there is no training phase (or it could be considered instantaneous training); as such, the sample from which the predictions are inferred is not referred to as the training sample but as the reference sample.

³ Similarly, both terms are applied to data samples that will not be utilized for the model's training or as a reference sample. The subtle distinction lies in the utilization of a test sample, specifically employed for evaluating the model's performance.

els. At the base level, each entry in the features space is a single galaxy’s simulated photometry, in magnitudes; that is, the nine *Euclid* + UNIONS bands for the EWS, with the addition of two IRAC bands for the EDF. In order to improve the quality of the models, thus the model inferences, we also include derived features as the colors (pairwise differences of the magnitudes, excluding permutations), increasing the number of total features to 45 (EWS) and 66 (EDF). This is the number of features for each of the previously described methods, with the notable exception of the CatBoost chained regressors (CCR), where the inferred labels are added on top of those as new training features at each iteration, as described in Sect. 3.5.2.

All of the methods presented in this section are not sensitive to the dynamic range and scales of the input features, except for the Deep Learning Neural Network (DLNN). In that case, we scale the features to a similar dynamic range with a standard Z-score normalization.

3.3. Phosphoros

Phosphoros⁴ (Paltani et al., in prep.) is a Bayesian template-fitting tool for galaxies SED developed within the *Euclid* collaboration. In the *Euclid* photo-*z* data challenge, which evaluated the performance metrics of different template-fitting and ML codes in retrieving the photometric redshift of a mock catalog, Phosphoros yielded the best performance along with LePhare (Euclid Collaboration 2020).

Phosphoros can be used to evaluate at the same time the photometric redshift and the physical properties of galaxies that have to be provided as tags for the templates. In the present work, we have used 1254 templates from Bruzual & Charlot (2003, in the 2016 version⁵) with Chabrier IMF (Chabrier 2003), considering exponentially declining (e-folding timescale $\tau = 0.1, 0.3, 1, 2, 3, 5, 10, 15, 30$ Gyr) and delayed (characteristic timescale $\tau = 1, 3$ Gyr) star formation histories, 2 metallicities ($Z = 0.008, 0.02$) and 57 ages between 0.01 and 13.5 Gyr. The internal dust attenuation has been modeled with Calzetti’s law (Calzetti et al. 2000) with $E(B - V)$ values in the range [0.0, 0.5]. We tested whether the IMF choice for the templates might bias the performance by running Phosphoros with templates built with Salpeter IMF (Salpeter 1955), finding identical results in terms of performance metrics (see Sect. 3.6), though almost monolithically shifted by a factor 0.23 dex in logarithm with respect to the Chabrier results, shown throughout the paper.

As a first step, a grid with model photometry is derived for all the templates in the redshift range $z \in [0.0, 6.0]$ with steps of $dz = 0.01$. When comparing the model to the observed photometry, the only factor considered prior is the “volume-prior”, proportional to the redshift-dependent differential comoving volume. Upper limits are treated in a statistical sense, as models with fluxes over the limit in the undetected bands are still considered when looking for the best-fit model; in those cases, the χ^2 evaluation follows the indications in Sawicki (2012). In the version that will be used for *Euclid* data, there will also be the possibility of consistently dereddening the photometry for the Galactic extinction (Galametz et al. 2017) and considering the variability of the filter transmission functions across the field of view (Euclid Collaboration 2024g). A recipe to add emission

lines is also implemented in Phosphoros, but not used in this work.

The final result of the computation is the characterization of the multidimensional posterior with a density sampling of 100 values for each galaxy, as well as the values of the physical properties and redshift from the best posterior model as the mean, the median (used in this work), or the mode of the distribution.

In this work, we use Phosphoros results to benchmark ML methods against a standard template-fitting algorithm and as a necessary step to build the reference sample to use as input for *nnpz*.

3.4. *nnpz*

The Nearest-Neighbors Photometric Redshift (*nnpz*, see Tanaka et al. 2018, for a first application on the HSC-SSP survey) is a supervised-learning technique mapping a given set of features to known labels with an upgraded version of the *k*-nearest neighbors algorithm (*k*-NN). In its most simple form, a *k*-NN algorithm combines an integer number of *k* neighbors in a reference sample closest to the target in feature space with respect to some distance metric (e.g., Euclidean) and predicts a label based on some user-defined combination of the metrics of the *k*-NN reference sample labels (e.g., a mean weighted by the distance in feature space). The same conceptual approach can be employed to provide a posterior distribution function (PDF) for the desired label by combining some a priori known PDFs for the reference sample under the assumption that similar observations with similar uncertainties would naturally produce similar results. Predictions and confidence intervals will naturally follow from the output PDF.

This is the concept behind *nnpz* in a nutshell. The reference sample is built starting from Phosphoros as a set of objects whose full parameters’ PDF has been sampled with 100 randomly extracted points according to the PDF density distribution. The samples of the *k*-NNs are then combined to produce the target PDF, from which a punctual prediction is obtained from the mean, the median, or the mode of the distribution.

In this work, we used the 1.2.2 version of *nnpz* available on the *Euclid Local Development Environment* (LODEEN) version 3.1.0, a virtual machine containing all of the *Euclid* software and pipelines. As for the code hyperparameters, after a first skim in a batch of at least 1000 nearest neighbors in the target space obtained from a space partition with KDTree (necessary to speed up the whole process instead of simply brute-forcing the search), we fix the final *k* from which the target labels are evaluated to 30 nearest neighbors. To generate a prediction, each neighbor is weighted with its χ^2 likelihood, which is the χ^2 distance between the reference neighbor and the target point in the feature space. *nnpz* combines the posterior coming from all the nearest neighbors and produces a PDF for the predicted target galaxy, from which we extract the point prediction as the median value of the distribution. We perform the same tests presented in this work with the mode of the distribution (i.e., the maximum-likelihood estimator) as the point prediction without noticing a significant change in the results.

In fact, returning a source’s multivariate PDF samples as output instead of a single-point prediction is one of the great advantages of *nnpz*. This information is in principle recoverable with other ML algorithms, such as CatBoost (see Mucesh et al. 2021, for an application to a simple random forest) if considering all the training samples in a particular leaf as PDF samples, though this is computationally and memory-wise less feasible than the ~ 100 samples per galaxy of *nnpz*.

⁴ <https://phosphoros.readthedocs.io>

⁵ http://www.bruzual.org/bc03/Updated_version_2016/

3.5. Other machine learning techniques

Apart from `nnpz`, we performed similar tests using previously tested ML techniques that have been shown to be extremely efficient for redshift and galaxy property estimation: Gradient-Boosted Decision-Trees (GBDT) and DLNN.

3.5.1. CatBoost single-model regressor

A GBDT is rooted in decision trees, a building block of widely used and successful techniques for regression and classification tasks. In a decision tree, the data is recursively split into smaller subsets based on the features that best separate the data according to some information gain (for classification) or variance minimization (for regression) criteria, until a stopping criterion is met. The result is a tree-like structure, with each internal node representing a feature, each branch representing a potential value for that feature, and each final node (leaf) representing a class (for classification) or predicted value (for regression). This scheme has been improved by what is called gradient boosting, which decreases the randomness improvement in training by starting with a set of imprecise decision trees (“*weak learners*”) and iteratively improving them, focusing on what these are predicting wrong rather than generating a new random subset of the data.

CatBoost (Prokhorenkova et al. 2018)⁶ is a cutting-edge ML algorithm specifically designed for gradient boosting on decision trees. There are some specific features that help reduce some typical issues in gradient boosting algorithm implementations, such as the ordered boosting to reduce overfitting and the oblivious trees to regularize while increasing speed.

In this work, we use CatBoost in two different ways. With the CatBoost single-model regressor (CSMR), we train a single model to solve a multiregression problem. Each set of features is associated with a pool of labels (z_{phot} , $M_{\star}$, SFR) and not just a single label per time (as in Sect. 3.5.2), finding the best model with a Multivariate Root Mean Square Error (*MultiRMSE*) loss. In each case presented, the final model is trained with 1000 estimators, allowing for a maximum depth of 11.

3.5.2. CatBoost chained regressors

With the CCR we train on a set of features one label per time, and iteratively append the predicted labels to the features up until convergence, allowing the model to naturally learn the correlation between parameters through an iterative approach. A thorough and more detailed description can be found in Humphrey et al. (in prep.). Here, we summarize it in the following paragraph.

We start with a training set (X_{train} , y_{train}) and a test set (X_{test} , y_{test}). The first iteration goes as follows:

1. The model is trained on X_{train} whose features are the full set of colors and magnitudes (with permutations, see in Sect. 3.2), with only z_{phot} as the lone label in y_{train} . From this model, we can predict some z_{phot} and evaluate their performance metrics on the test sample.
2. Now, the model is trained on a new X_{train} that is composed of the previous ones (magnitudes and colors) and the z_{phot} predicted in 1. From this model, we predict $\log_{10}(M_{\star}/M_{\odot})$ and evaluate the $\log_{10}(M_{\star}/M_{\odot})$ performance metrics on the test

Table 3. Architecture of the DLNN used in this work.

Layer	N_{in}	N_{out}
Input	N_{feat}	4096
Dense	4096	2084
Dense	2048	1204
Dense	1024	512
Dense	512	64
Output	64	3

Notes. All layers are dense, fully connected with a ReLU activation function. N_{feat} values are reported in Sect. 3.2.

sample; of course, the X_{test} has been extended to incorporate the new feature from the predicted z_{phot} on the test sample.

3. Then, the model is trained on another X_{train} composed of the previous features plus the predicted $\log_{10}(M_{\star}/M_{\odot})$ in the previous step. With this model, we predict $\log_{10}(\text{SFR}/M_{\odot} \text{ yr}^{-1})$ and evaluate the $\log_{10}(\text{SFR}/M_{\odot} \text{ yr}^{-1})$ performance metrics on the (once again, extended) test sample.

Now the second iteration starts.

4. The model is trained on an X_{train} composed of the previous features – including the z_{phot} and $\log_{10}(M_{\star}/M_{\odot})$ predicted in steps (1) and (2) – plus the predicted SFRs in step (3), and with this model, we re-predict z_{phot} and evaluate their performance metrics.
5. Again, another model is retrained with the previous features plus the new z_{phot} predicted in the previous step. $\log_{10}(M_{\star}/M_{\odot})$ is re-predicted, and the model performance on the label is evaluated.

The whole procedure goes on for four iterations, when we observe a convergence of the evaluated metrics in agreement with Humphrey et al. (in prep.).

As such, the model features are (in square brackets, the step in which they have been evaluated):

1. magnitudes and colors;
2. magnitudes, colors, and z_{phot} ;
3. magnitudes, colors, z_{phot} , and $\log_{10}(M_{\star}/M_{\odot})$;
4. magnitudes, colors, z_{phot} , $\log_{10}(M_{\star}/M_{\odot})$, and $\log_{10}(\text{SFR}/M_{\odot} \text{ yr}^{-1})$;
5. magnitudes, colors, z_{phot} , $\log_{10}(M_{\star}/M_{\odot})$, $\log_{10}(\text{SFR}/M_{\odot} \text{ yr}^{-1})$, and z_{phot} .

And so on, for four iterations.

Finally, we store the final set of label predictions for the test set on which we evaluate the performance metrics (see Sect. 3.6). In running CatBoost we use the same set of hyperparameters as in CSMR and in Humphrey et al. (in prep.).

3.5.3. Deep learning neural network

As we only deal with structured (i.e., tabular) data, we also test the performance of a simple, multilayered DLNN. Here we adopt a typical architecture that has been widely used in the literature in searching for photometric redshifts and physical parameters (e.g. Firth et al. 2003; Collister & Lahav 2004; Euclid Collaboration 2023).

The DLNN inputs are the training features (magnitudes and colors, with permutations, see Sect. 3.2), and the output is a set of three labels (z_{phot} , $M_{\star}$, SFR). The DLNN architecture (described in Table 3) consists of five fully connected layers

⁶ <https://catboost.ai/>

with a decreasing power of two hidden units for each layer. The adopted activation function for each layer is a Rectified Linear Unit (ReLU, [Nair & Hinton 2010](#)), Mean Squared Error (MSE) for the loss function with L_2 regularization to avoid overfitting, and the model is optimized with the Adaptive Moment estimator (Adam, [Kingma & Ba 2014](#)).

The DLNN for each model are trained and tested on the same train and test samples as for all the other methods. We run the training on mini-batches of size 512.

3.6. Metrics and quality assessment

We use standard metrics to quantify the model's performance. Those are defined differently when referring to redshifts or PPs.

The first is the normalized median absolute deviation, defined as:

$$\text{NMAD} = 1.48 \times \text{median} \begin{cases} \frac{|z_{\text{pred}} - z_{\text{test}}|}{1 + z_{\text{test}}} - b & \text{for redshifts;} \\ |y_{\text{pred}} - y_{\text{test}}| - b & \text{for PPs,} \end{cases} \quad (1)$$

with b being the model bias (see below).

The outlier fraction f_{out} is defined as the fraction of catastrophic outliers ([Hildebrandt et al. 2010](#)) over a certain threshold (in log space for physical parameters, linear for redshifts):

$$f_{\text{out}} : \begin{cases} \frac{|z_{\text{pred}} - z_{\text{test}}|}{1 + z_{\text{test}}} > 0.15 & \text{for redshifts;} \\ |y_{\text{pred}} - y_{\text{test}}| > t_{\text{out}} & \text{for PPs.} \end{cases} \quad (2)$$

These thresholds have been evaluated looking at the standard deviation of the PPs distribution, considering only sources with a good photo- z recovery – that is, below the $0.15(1+z)$ threshold – for all the methods considered, trained with true labels (see Sect. 3.1). As a consequence, the PPs thresholds are not the same for stellar masses or SFRs. The chosen thresholds are two times the mean standard deviation between the prediction and the true values found for all the considered methods, rounded to the nearest decimal. The values are $t_{\text{out}} = 0.4$ dex for stellar masses and $t_{\text{out}} = 0.8$ dex for SFRs.

Defining the catastrophic outliers in this way is different than assuming a plain 0.3 dex difference between the prediction and the true values of the physical parameters (corresponding to a factor of two) that is found in recent literature (e.g. [Euclid Collaboration 2023](#)). Fixing the same value in dex space for every PP is a penalizing choice, especially for SFRs, where a 0.3 dex difference might even be the order of (or even below) 1σ of the distribution, which is too little to define a galaxy as a catastrophic outlier. In this way, we adopt a more robust definition from a statistical sense, which actually returns an informative quantitative description of what a catastrophic outlier is for a stellar mass estimate or a star formation rate.

Finally, a model's overall bias b is:

$$b = \text{median} \begin{cases} \left(\frac{z_{\text{pred}} - z_{\text{test}}}{1 + z_{\text{test}}} \right) & \text{for redshifts;} \\ (y_{\text{pred}} - y_{\text{test}}) & \text{for PPs.} \end{cases} \quad (3)$$

In all three cases, the closer to zero, the better the predicted values resemble the test ones. Of all three, only the bias can take either positive or negative values.

We notice that those metrics are different from the *Euclid* requirements, that depend on the redshift probability distribution functions (PDZ, see their definition in Sect. 4.2 of

[Euclid Collaboration 2020](#)). We use photo- z and PPs point estimates instead.

As for the SFMS, we evaluate the performance of the recovered relation by evaluating three parameters:

- the relation slope m , measured with an orthogonal distance regression (ODR);
- the fraction of passive galaxies f_p , defined as the fraction of objects with specific-SFR $\log_{10}(\text{sSFR}/\text{Gyr}^{-1}) < -1$. This limit has been determined looking at the divide between passive and nonpassive galaxies in MAMBO and is in accordance with values found in the literature (e.g. [Pozzetti et al. 2010](#); [Ilbert et al. 2013](#));
- the relation scatter σ , measured only on nonpassive sources.

4. Results

In this section, we present the results of the methods presented in Sect. 3 on the simulated *Euclid* fields described in Sect. 2. Our primary objective is to evaluate the performance of the methods described in Sect. 3 and find the optimal strategies to extract the maximum amount of information available in the *Euclid* survey, in particular the EWS and EDF. We ought to do so by dealing with realistic information, that is, what the survey will actually deliver as a scientific product. As such, we present results for the EWS and for the EDF, obtained by combining the Phosphoros results (photometric redshifts, physical parameters) in the fields with the available photometry. As described in Sect. 2, we deal with two kinds of training labels: the true ones (i.e., the ground-truth simulated parameters), which of course are unknown in a real-life application, and as such, we only use them to assess what the best-case scenario for a particular field could be, and the recovered labels, whose values are the Phosphoros outputs resulting from the observed photometry, which is what an actual application to real *Euclid* data will have to deal with. However, it is worth noticing that spectroscopic redshifts will be available for a smaller sample (still around the order of millions of sources), which is the closest thing to true labels that the EWS and EDF will yield. A similar argument (with lots of attached caveats) could apply to those sources with $H\alpha$ -derived SFRs, though the numbers in this case are sensibly reduced with respect to the spectroscopic redshifts. For each trained model, we carefully checked that the performance metrics evaluated on the training set do not differ significantly from the ones obtained by applying the model to the test set, thus excluding any kind of overfitting to the training set.

4.1. Computational performance

Most of the runs presented here – specifically, the CSMR, CCR and DLNN results – are performed on Galileo100, a high-performance computing (HPC) system located at Cineca, within the *Italian SuperComputing Resource Allocation* (ISCRA⁷) Class C program, as part of the PPRESCIA-HP10CBAZOH program (PI, Enia). Galileo100⁸ is a DUAL-Socket Dell PowerEdge cluster, hosting 636 computing nodes each with two x86 Intel(R) Xeon(R) Platinum 8276-8276L, with 24 cores each. In fact, the main advantage of ML methods over classic template-fitting templates is the dramatic speed-up to the inference problem, at least when dealing with point value prediction for the parameters, coupled with the improved computational performance in training these models in HPC systems.

⁷ <https://iscra.cineca.it/>

⁸ <https://www.hpc.cineca.it/systems/hardware/galileo100/>

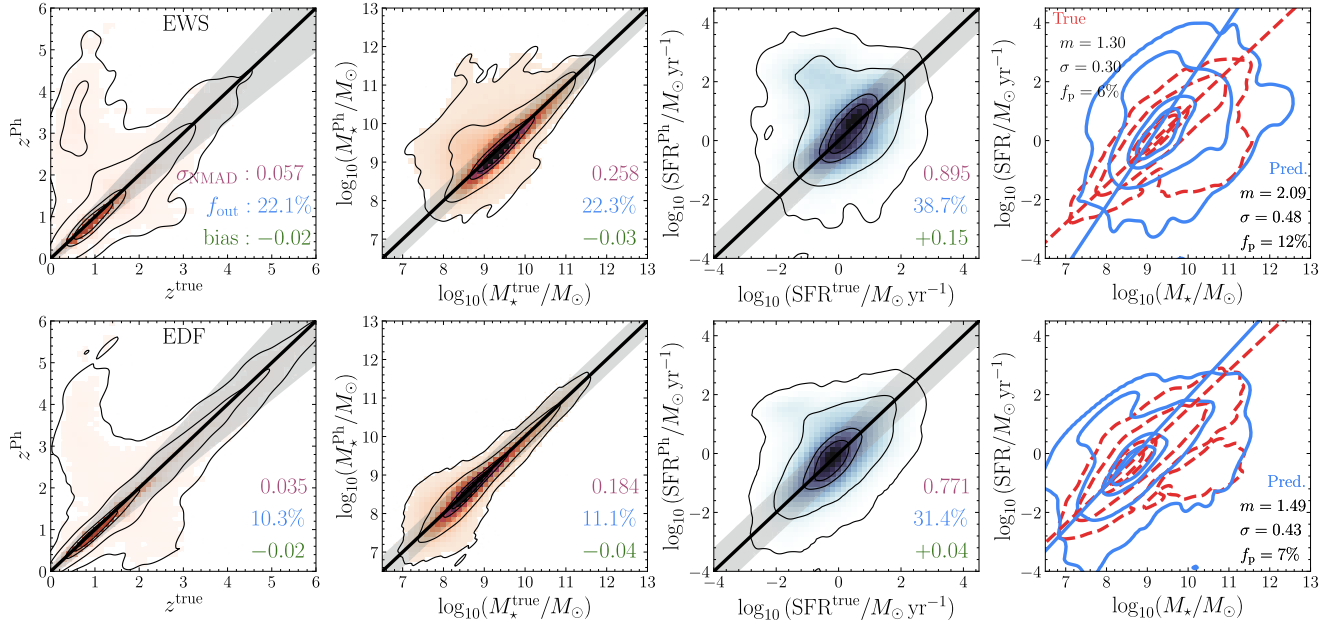


Fig. 3. Phosphoros results on two simulated *Euclid* catalogs, EWS (top panels) and EDF (bottom panels), with true values plotted against the Phosphoros recovered ones. The black line is the 1:1 relation; the shaded area is the region beyond which a prediction is an outlier. In every plot, the four contours are the area containing 98%, 86%, 39% (corresponding to the 3σ , 2σ and 1σ levels for a 2D histogram) and 20% of the sample. For SFMS the true distribution is reported in red (dashed), the predicted one in blue (solid). The lines are the ODR best-fit to the (passive-removed) distribution. The reported metrics are NMAD (purple), the outlier fraction f_{out} (blue) and the bias (green) for the photometric redshifts and physical parameters, and the slope m , scatter σ and fraction of passive galaxies f_p for the SFMS, all defined in Sect. 3.6.

Phosphoros and *nnpz* are run on a PowerEdge T640 machine with an Intel(R) Xeon(R) Silver 4116 CPU @ 2.10 GHz processor, and 24 available cores. A typical run of Phosphoros requires ~ 0.8 seconds per galaxy; for a number of galaxies around a million (as in our cases), it translates into uninterrupted runs of a couple of weeks on the 24 available cores of our workstation.

With *nnpz*, which technically does not need training as the whole computational load is on the shoulders of the neighbor search and PDF combination, a typical run on a target sample of ~ 50 k galaxies requires ~ 8 minutes of time, or 0.006 seconds per galaxy, a speed-up of a factor 100 with respect to Phosphoros.

CatBoost-based runs and DLNN require training instead, after which the inference is almost instantaneous. How long those methods will run depends on the size of the training set, how complex the model is allowed to be and the number of training epochs for DLNN; for a typical training set size of a million galaxies, it translates into training runs of ~ 15 minutes for CSMR on 16 cores, ~ 0.002 seconds of training time per galaxy. For CCR the training time per galaxy is similar, though the final run is of course longer since a model is trained at every iteration for each label. We run the CCR on Galileo100, asking for a single node of 48 cores, whose overall run lasted for ~ 1 hour time.

Finally, on the same HPC system, we trained the DLNN for 300 epochs, translating into ~ 7 hours of training time (~ 80 seconds per epoch) for a $\sim$ million galaxies in the training sample, a training time of ~ 0.003 seconds per galaxy.

4.2. Phosphoros results

The first results we present are the template-fitting runs with Phosphoros on all the galaxies present in the training (or reference) samples. We refer the reader to Sect. 3.3 for further details

on how Phosphoros has been run. The results are shown in Fig. 3, for the simulated EWS and EDF. In Appendix B we also show the results for the two auxiliary fields at 16 and 25 ROS. In each plot, the true values are plotted against the recovered ones, and the performance metrics are reported in the bottom right of each plot.

These results are sort of the blueprint for all the others found in this work. The first thing that jumps out is the difficulty in recovering the correct SFR, as both the EWS and EDF simulations display high NMADs (0.90–0.77, respectively) and fractions of outliers ($>30\%$). The recovered SFRs for the EWS are also biased toward higher values by a factor ~ 1.4 (a bias of 0.13–0.15 in the logarithm).

Optimal recovery is obtained for photometric redshifts instead, with NMADs that improve from 0.057 to 0.035 passing from Wide to Deep photometry – and the addition of the two IRAC bands – and f_{out} reducing from 22% to 10%, with half of this reduction the consequence of an improvement in correctly distinguishing faint low- z , low-mass objects from high- z , high-mass ones. For the EWS, worse results are obtained for the stellar masses, with higher NMADs (0.258) and fractions of outliers (22%). The combined effect of deeper photometry plus the two IRAC bands sensibly improve the recovered stellar masses in the EDF, with NMADS decreasing to 0.18 and f_{out} to 11%. Both the recovered photometric redshifts and stellar masses show low biases (absolute values smaller than 0.04) with respect to the ones found in the SFRs.

These are not unexpected findings, given the specific set of filters used as input. As reported in Sect. 2 (see also Table 1 and Fig. 1), for the EWS we use 9 filters with rest-frame λ_{eff} between $0.37 \mu\text{m}$ and $1.77 \mu\text{m}$. As the photometric redshifts are more sensitive to colors in the ultraviolet (UV)-to-NIR part of the spectrum, these are well recovered with the given wavelength range and the number of filters. Moreover, dropouts in

Table 4. Metrics for the unperturbed simulation.

	CSMR			CCR			DLNN			nnpz		
	NMAD	f_{out}	bias	NMAD	f_{out}	bias	NMAD	f_{out}	bias	NMAD	f_{out}	bias
z	0.005	0.1%	$<10^{-3}$	0.002	0.0%	$<10^{-3}$	0.017	1.4%	0.002	0.001	0.1%	$<10^{-3}$
$M_{\star}$	0.032	0.1%	$<10^{-3}$	0.022	0.1%	$<10^{-3}$	0.083	1.5%	$<10^{-3}$	0.019	0.1%	$<10^{-3}$
SFR	0.178	0.2%	0.005	0.168	0.2%	$<10^{-3}$	0.275	1.5%	0.004	0.132	0.2%	$<10^{-3}$

Notes. Here, the methods try to find the mapping between the true labels and the noise-free features. As such, these should be intended as the absolute lower limits for each parameter. $M_{\star}$ refers to $\log_{10}(M_{\star}/M_{\odot})$, SFR to $\log_{10}(\text{SFR}/M_{\odot} \text{ yr}^{-1})$.

different filters are an excellent proxy for high- z galaxies. Stellar masses correlate well with rest-frame NIR photometry, in particular the H_E band, and most of our simulated sample ($>60\%$) reside between $0 < z < 1.5$ where NIR is still sampled by *Euclid* filters. The addition of the first two IRAC channels helps significantly in improving the stellar masses recovery. Things are harder for SFRs, as they correlate the most with mid-IR to far-IR photometry (Kennicutt & Evans 2012), tracing obscured star formation, and secondly with UV rest-frame monochromatic fluxes at 1550 Å (FUV, Bell & Kennicutt 2001) and 2800 Å (NUV, Bell et al. 2005), tracing unobscured star formation. The former, stronger proxy is inaccessible with the chosen set of filters, while the latter is a weaker one. This makes the recovery of SFRs difficult even in an ideal, pristine situation (see Sect. 4.3 and Table 4) and extremely complicated when more sources of uncertainty are added. These could be improved by imposing some SFR-related priors to the template-fitting algorithm, something that will be carefully considered when dealing with real *Euclid* data.

The main fraction of photo- z catastrophic outliers (around 10% for the EWS, 5% for the EDF) is composed of faint low-redshift ($z^{\text{true}} < 1$), low-mass [$\log_{10}(M_{\star}^{\text{true}}/M_{\odot}) < 9$] and low-SFR galaxies [$\log_{10}(\text{SFR}^{\text{true}}/M_{\odot} \text{ yr}^{-1}) < 0$] that are instead misplaced at higher redshifts ($z > 2$) with at least one order of magnitude higher masses [$\log_{10}(M_{\star}/M_{\odot}) > 10$] and SFRs [$\log_{10}(\text{SFR}/M_{\odot} \text{ yr}^{-1}) > 1$]. This is reflected in the SFMS. In the EWS case, the higher SFR overestimation with respect to stellar masses yields a fitted relation with a sensibly higher slope ($m = 2.1$) with respect to the true one ($m = 1.3$). The uncertainties on the recovered parameters translate also into a higher scatter of $\sigma = 0.48$ (ground truth of 0.30) and a fraction of passive galaxies higher ($f_p = 12\%$ instead of 6%). Things get better for the EDF, with metrics still distant from the true ones though.

4.3. The unperturbed simulation

One might wonder what the absolute best-case scenario is in terms of performance when applying the methods described in Sect. 3 to a pristine, unperturbed set of features mapping to the true labels. This is the same as asking what order of magnitude the irremovable inherent uncertainty of those methods is, which will always affect the measured metrics, even in a more realistic scenario where the noise affecting the features (and labels) will dominate.

To answer this question, we run the methods defined in Sect. 3 on an unperturbed, noise-free features version of the MAMBO catalog with true labels (see Sect. 2 for definitions). Any uncertainty depends only on the specifics of the technique used to map features to labels and, from a broader perspective, on how well those specific features (magnitudes and colors) are able to recover those particular labels (photometric redshifts, stellar masses, and star formation rates).

The results are reported in Table 4. What stands out is the perfect recovery of the photometric redshifts, up to a 0.1% fraction of outliers. Those nine filters and the associated colors are able to correctly put a galaxy in its right place in the cosmic picture (see Appendix A for a quantification on the feature importance). This is similar for stellar masses, though with metrics degraded by an order of magnitude, NMADs of $\sim 10^{-2}$ vs $\sim 10^{-3}$ for photo- z s – and comparable outlier fractions – as expected once considering that the rest-frame H band is a well-known tracer for correctly identifying the galaxy mass content, which is still true as the majority of the simulated sources are at $z < 1.5$.

Star formation rates are harder to recover, though. Even in an ideal, perfect scenario, it is impossible to go below NMADs of ~ 0.13 for that particular set of features. Of course, there is room for improvement if adding other features more sensitive to the star formation processes when available, for instance, mid-IR or far-IR photometry, or spectral features, such as the $H\alpha$ emission line. A more detailed dissertation is beyond the scope of this work, which focuses mainly on the EWS and EDF, without considering other ancillary (or new) filters. However, the complete exploitation of the full spectrophotometric (and morphologic) information in *Euclid* will be explored in a forthcoming work.

4.4. Results for the Euclid Wide Survey

The unperturbed case gives back an extremely optimistic best-case scenario. In reality, all the observed photometry in *Euclid* will be affected by some degree of uncertainty, whose effect is to make the feature space noisier, mixing together sources with different labels. At times, even with extremely different ones, in degenerate regions of the feature space (i.e., fainter and less massive or brighter and more massive), making it hard – or even impossible – to correctly understand which label is associated with that particular set of features. This unavoidably degrades the quality of the model and the performance metrics when applied to a sizable sample of data.

As reported in Sect. 2, we simulate four different versions of *Euclid* observed catalogs: the EWS and EDF, and two calibration fields with 16 and 25 ROS, respectively, mimicking the *Euclid* auxiliary fields for photometric and color gradient calibration (Euclid Collaboration 2024h). In this section, we focus on the EWS, and the performance observed when training the models on deeper samples.

We present two possible approaches for this task. In the flowchart shown in Fig. 4, we summarize what has been done in obtaining the reported results for the EWS (top panel) and the EDF (bottom panel). The flowchart describes the different approaches employed when dealing with simulations at different depths of the same field.

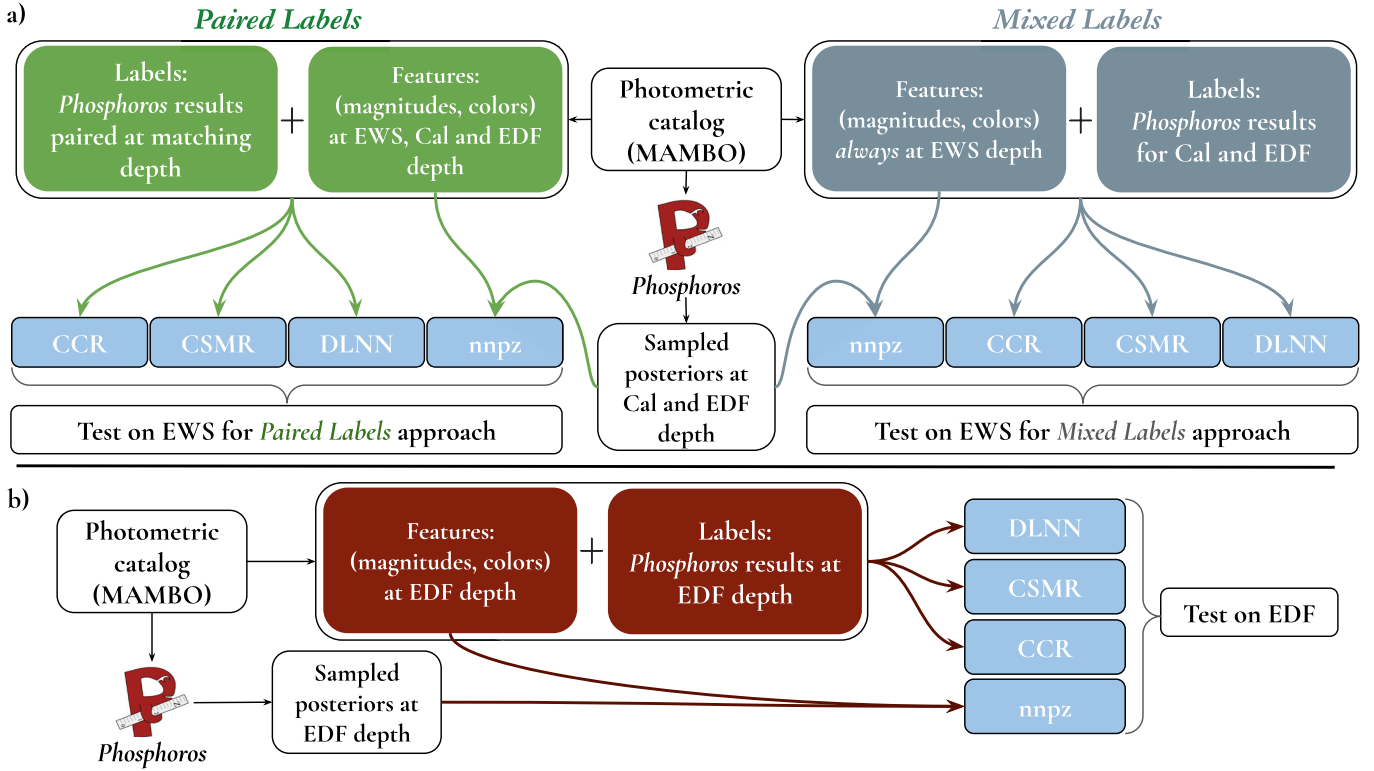


Fig. 4. Flowchart followed for the reported results on the EWS and EDF. Panel (a) summarizes what has been done for the EWS. In this case, we employed two different approaches: pairing features to labels coming from *Phosphoros* results at the corresponding depth (paired labels), or with features always from the Wide simulated catalog and labels coming from *Phosphoros* results at the Calibration and Deep fields depth for the corresponding sources (mixed labels). These pairs of (features, labels) – or (features, posteriors for *nnpz* (see Sect. 3.4) – are thus given as input for the ML models described in Sect. 3. Panel (b) illustrates the straighter flowchart for the EDF, where the pairs (features, labels) or (features, posteriors) always come from the simulated Deep field.

4.4.1. Paired labels approach

The first one is the paired labels approach. Here, we train each model (or build a reference sample) with features and labels coming both from a particular field (EWS, EDF, or the two calibration fields), and test on the EWS. The labels are the recovered ones (see Sect. 3.1), that is, the *Phosphoros* results for photo- z and physical parameters on the field-correspondent photometry. The results are summarized in Table C.1, where for each pair of training/reference – test field we report the performance metrics for all the considered labels, and Table C.2, where we report the same for the SFMS results.

The photometric redshifts performance is good, in line with the template-fitting results in Sect. 4.2 (see top panels of Fig. 5). There is a slight improvement in training the model with photometry and labels coming from deeper fields, with NMADs reducing by ~ 0.01 and outliers by $\sim 5\%$ at best. *nnpz* has the best results overall (NMAD ~ 0.06 , $f_{\text{out}} \sim 18\%$), for every possible case of training field involved. The vast majority of outliers – raising the NMAD too – are $z < 1.5$ galaxies mistakenly assumed to be higher redshift ones at $z > 2$ (more on that in the next paragraphs). When looking at their magnitudes, these objects are revealed to be faint galaxies, with a distribution peaking close to the magnitude limits for each band.

This wrong distance attribution is carried over to the stellar mass prediction. A part of the degradation in the NMAD and most of the one in f_{out} is a consequence of those lower- z , lower-mass galaxies mistakenly assumed to be as high- z , high-mass ones. At best, with the given features and true labels in the

training sample, no less than NMAD ~ 0.14 and $f_{\text{out}} \sim 13\%$ is expected (with the CCR, see bottom panel of Fig. 6). For stellar masses, no improvement is observed when using deeper calibration fields for training but rather a degradation (see Table C.1, with the exception of EDF field, with the two IRAC channels). This is not unexpected, as it is common in ML applications to see cases where training and testing on noisier data altogether yields better results than training with better features and testing on the noisy ones.

The most worrisome metrics are the ones associated with SFRs. The outlier fraction, defined as points with predicted SFR above or below a certain threshold to the true value (0.8 difference in log space, Sect. 3.6), is over 30% for every method with the notable exception of *nnpz*, where it stays between 26% and 30%. We already showed in Sect. 4.3 how recovering SFR with the given set of features is harder than photo- z s or stellar masses even in the ideal, unperturbed case. In a more realistic scenario, the results of the EWS are far from ideal, even when the true values for SFR are used in the training process (no less than an NMAD of 0.38 and 10% of outliers, see bottom panel of Fig. 7). The template-fitting algorithm finds it hard to recover SFR indeed, as reported in Sect. 4.2 ($\sim 39\%$). Differently than stellar masses, this is not just a matter of the wrong photo- z s attribution affecting the SFRs (i.e., closer and less star-forming vs farther and more star-forming; more on that in the following paragraphs), but an inherent degeneracy due to the filters and colors used in the inference process.

The occurrence of simultaneous wrong predictions for stellar masses and SFRs (both overestimated or underestimated)

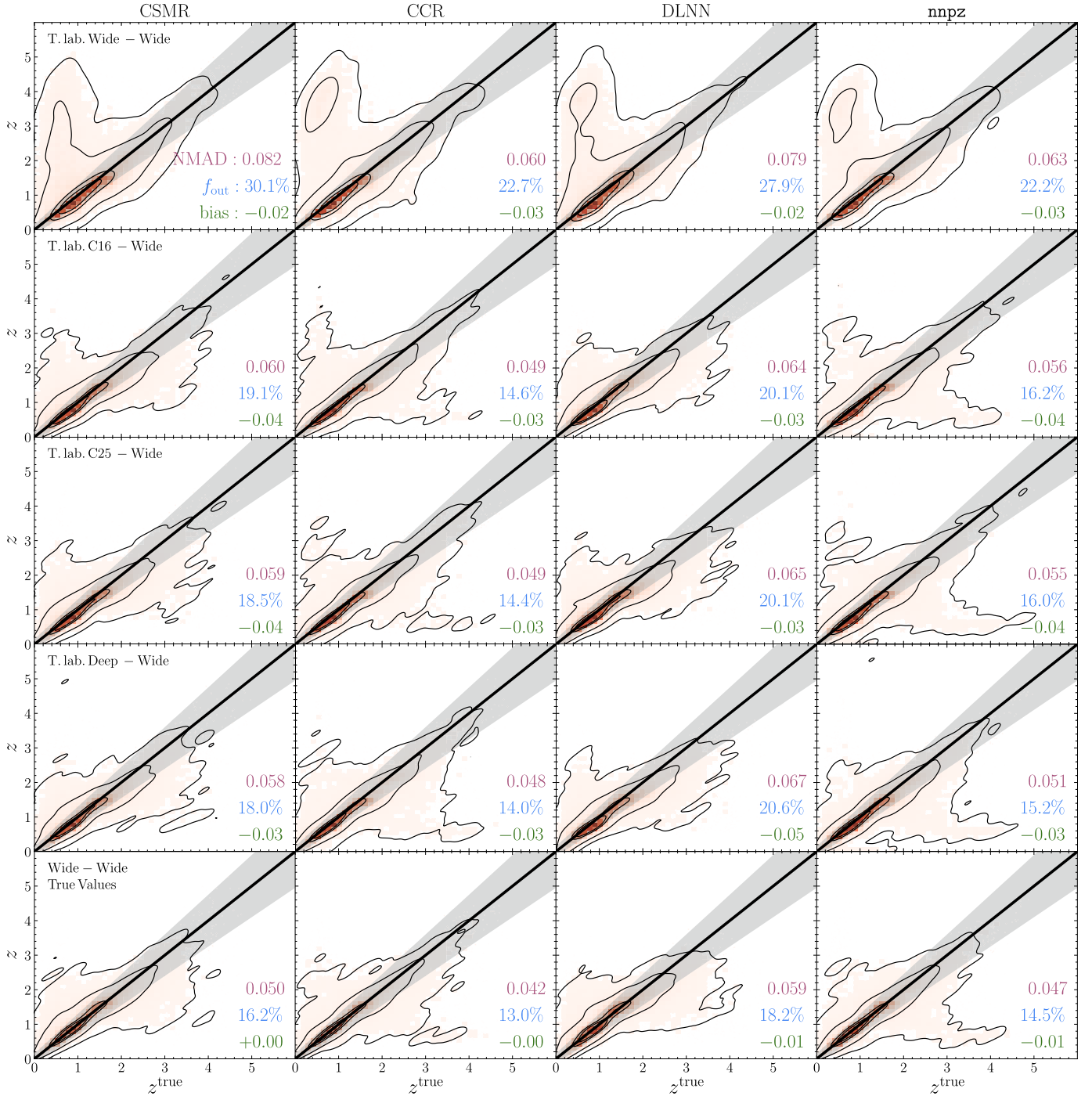


Fig. 5. Results for the EWS with the mixed labels approach. The true values on the x -axis are plotted against the predicted values on y . The black line is the 1:1 relation; the shaded area is the region beyond which a prediction is an outlier. Contours are the area containing 98%, 86%, 39% (corresponding to the 3σ , 2σ and 1σ levels for a 2D histogram) and 20% of the sample. Each column represents the results for the methods described in Sect. 3. In the first four rows, the training labels are the recovered ones, coming from Phosphoros results to the mock photometry at the same depth of the field reported in the leftmost plot legend and tested on the EWS (see Sect. 2 for further details). The T.lab Wide-Wide case is exactly the same as the Wide-Wide case in Table C.1. In the fifth row, we show the results of the EWS training the models with their true labels as the best-case scenario for that particular field. The reported metrics are NMAD (purple), the outlier fraction f_{out} (blue) and the bias (green) for the photometric redshifts and physical parameters, as well as the slope m , scatter σ and fraction of passive galaxies f_p for the SFMS, all defined in Sect. 3.6).

mitigates the impact on the recovered SFMS, at least regarding the relation slope m , when training with deeper photometry (Table C.2). However, with the notable exception of *nnpz*, which yields the best performance in terms of SFRs, the recov-

ered fraction of passive galaxies [$\log_{10}(\text{sSFR}/\text{Gyr}^{-1}) < -1$] is usually well overestimated by a factor of two, the true one being 6%. No method at whatever training depth is able to recover the correct relation scatter ($\sigma = 0.24$).