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,

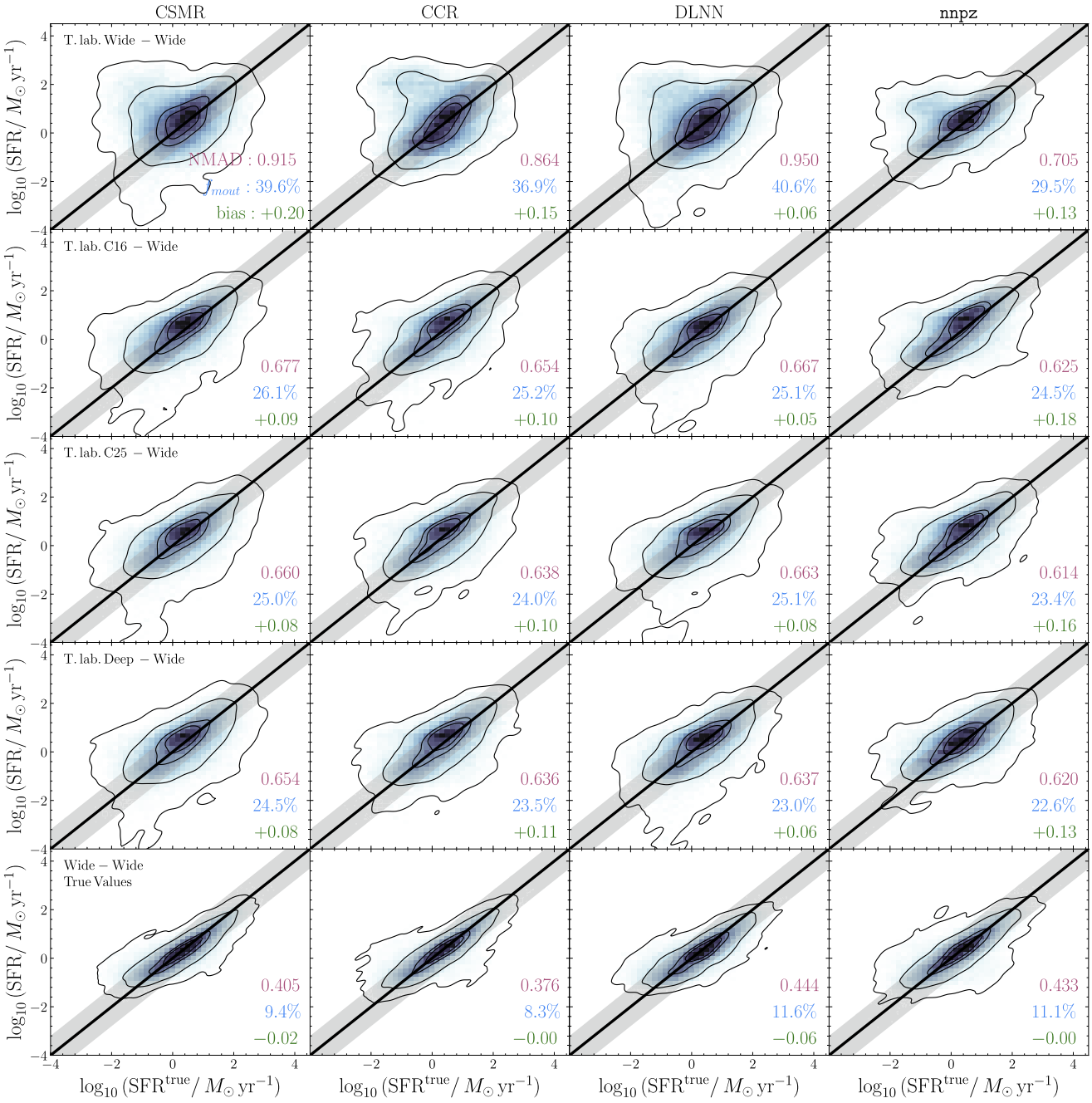

Fig. 7. Same as in Fig. 5, but for star formation rates.

and shown in Figs. 5–7. Notice also that with this approach we reduce the number of galaxies in the reference/training samples, as only those detected in the EWS will have `Phosphoros` recovered labels, thus the number of training galaxies passes from the ∼ one million for the Deep and calibration fields to ∼500 k (in Sect. 4.4.3 we reduce it to ∼230 k to simulate a more realistic COSMOS-alike reference sample).

However, despite the reduced training set, this approach improves the overall performance when applying the models to the test sample[9]. In fact, attaching labels obtained with the best possible available photometry acts similar to a prior, which is able to guide the model in better distinguishing the cases in which there are degeneracies in the feature space where two close sets of features yield drastically different solutions (and catastrophic outliers, e.g., two faint galaxies with similar features can be either low-z, low-mass or high-z, high-mass objects), an improvement that totally compensates the loss in sheer number of training examples. This behavior in feature space translates in the photo-z predictions as a vertical strip at $z_{\mathrm{phot}} \sim 1$, which are $z < 1.5$ galaxies mistakenly assumed as being farther away (see upper left panel of Fig. 3 or the top row of Fig. 5), and generating a cloud of higher mass, higher SFR galaxies in their respective plots. The wrong photo-z attribution is dragged onto the stellar masses (top rows of Fig. 6), where the outliers cloud is less prominent as in the photometric redshift case but still present as

[9] In typical ML applications, the relation between the size of the training sample and the quality metrics scales in logarithm scale and saturates after a while; as such, adding (or removing) a factor of two from the training sample could not significantly impact the final metrics.

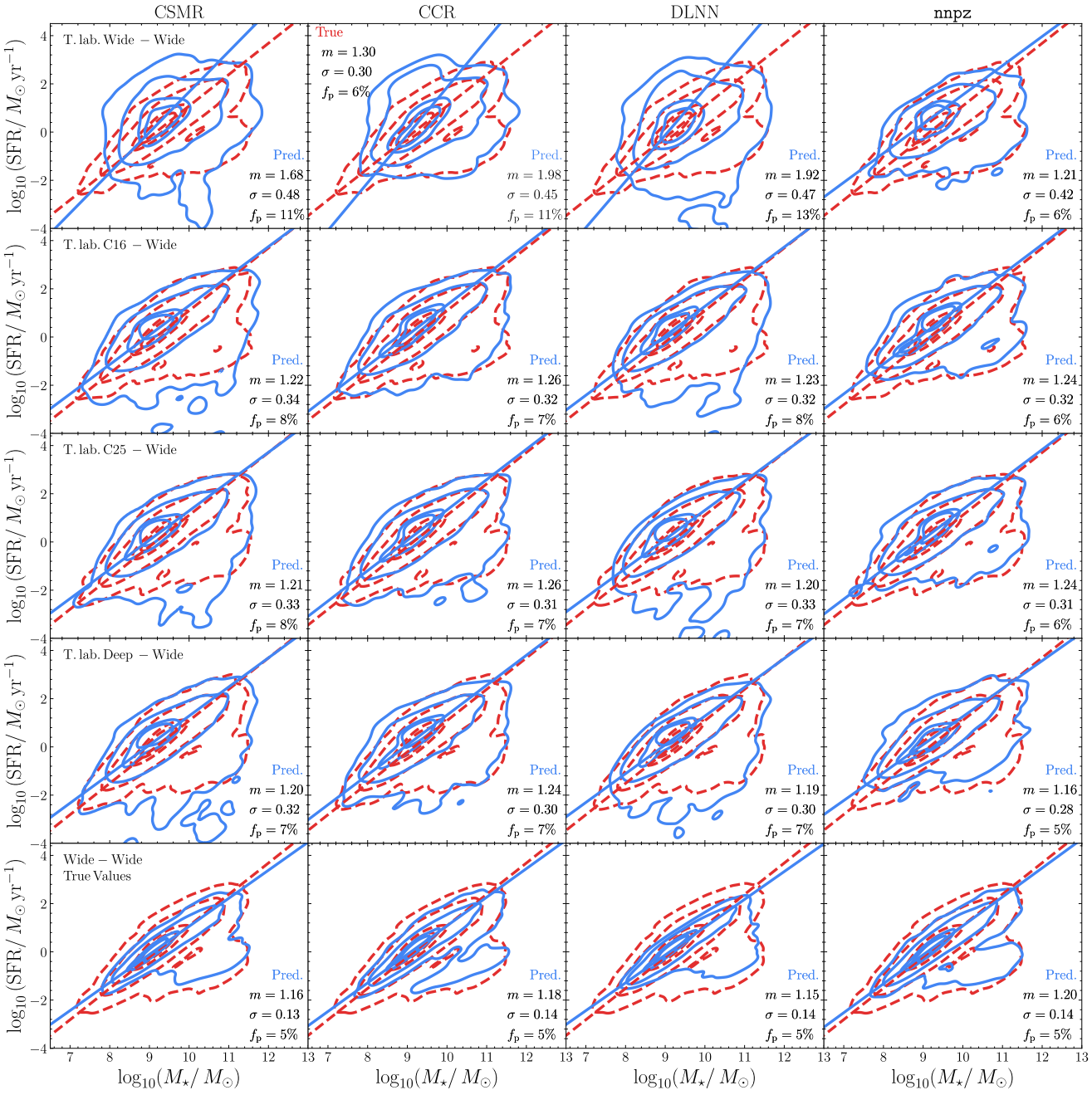

Fig. 8. Same as in Fig. 5, but for the SFMS. Dashed red contours are the test SFMS (i.e., the true values), solid blue is the predicted one. Contour levels are the same as reported in Fig. 5. The lines are the ODR best-fit to the passive-removed distribution (dashed for test SFMS, solid for predicted). The reported metrics are the SFMS slope, scatter, and fraction of passive galaxies, defined in Sect. 3.6.

a stripe of higher mass galaxies than expected. This also applies to SFRs, which are also heavily affected by the reduced predictive power of the chosen features, as previously described.

To better illustrate what was described in the previous paragraphs, in Fig. 9 we show the distribution of true vs. predicted PPs as a function of the difference between $z_{\rm pred}$ and $z_{\rm true}$. In the true PP – predicted PP plane, $z_{\rm pred} - z_{\rm true}$ is measured as the median of the values falling within each bin. We report the results obtained with `nnpz` on EWS (i.e., the T.lab. Wide-Wide case), but similar things are observed for all the other methods and cases considered.

Both for stellar masses (top panel) and SFRs (bottom panel), the cloud of photo-z catastrophic outliers is visible as a region of blue squares – meaning low-z objects mistakenly placed at high-z – while a minor impact is due to the opposite, in red. These regions are correctly placed outside the defined thresholds for PPs outliers (shaded area), described in Sect. 3.6, thus showing how they better relate with the distribution statistics with respect to a fixed 0.3 dex threshold. Another thing to notice is how the presence of those photo-zs outliers has a limited effect on the PPs distributions scatter (i.e., the NMADs). The majority of the distributions (>80% of points) have $|z_{\rm pred} - z_{\rm true}| < 0.1$. Even removing all the catastrophic photo-z outliers, the NMADs would still be 0.19 (for stellar masses) and 0.58 (for SFRs). This is a consequence of the inherent difficulties of the methods (template-fitting and ML) in recovering PPs with the given set of features,

Table 5. Metrics for the EWS, with the mixed labels approach.

		CSMR			CCR			DLNN			nnpz		
		NMAD	f_{out}	bias	NMAD	f_{out}	bias	NMAD	f_{out}	bias	NMAD	f_{out}	bias
EWS	z	0.08	30%	−0.02	0.06	23%	−0.03	0.08	28%	−0.02	0.06	22%	−0.03
	$M_\star$	0.25	20%	−0.02	0.24	20%	−0.03	0.27	20%	−0.04	0.24	19%	−0.01
	SFR	0.92	40%	0.20	0.86	37%	0.15	0.95	41%	0.06	0.71	30%	0.13
C16	z	0.06	19%	−0.04	0.05	15%	−0.03	0.06	20%	−0.03	0.06	16%	−0.04
	$M_\star$	0.21	14%	−0.10	0.19	12%	−0.08	0.22	14%	−0.10	0.21	13%	−0.06
	SFR	0.68	26%	0.09	0.65	25%	0.10	0.67	25%	0.05	0.62	25%	0.18
C25	z	0.06	19%	−0.04	0.05	14%	−0.03	0.07	20%	−0.03	0.06	16%	−0.04
	$M_\star$	0.21	14%	−0.10	0.19	12%	−0.09	0.21	15%	−0.12	0.20	13%	−0.07
	SFR	0.66	25%	0.08	0.64	24%	0.11	0.66	25%	0.08	0.61	23%	0.16
EDF	z	0.06	18%	−0.03	0.05	14%	−0.03	0.07	21%	−0.05	0.05	15%	−0.03
	$M_\star$	0.21	14%	−0.10	0.19	12%	−0.09	0.22	15%	−0.12	0.19	11%	−0.07
	SFR	0.65	25%	0.08	0.64	24%	0.11	0.64	23%	0.06	0.62	23%	0.13
True	z	0.05	16%	0.00	0.04	13%	−0.00	0.06	18%	−0.01	0.05	15%	−0.01
	$M_\star$	0.15	9%	−0.01	0.14	8%	−0.00	0.18	11%	−0.02	0.16	10%	0.01
	SFR	0.41	9%	−0.02	0.38	8%	−0.00	0.44	12%	−0.06	0.43	11%	−0.01

Notes. Leftmost column refers to the training sample, i.e., EDF means a model trained with EWS features and labels from `Phosphoros` results to the EDF photometry, True means a model trained with EWS features and the ground truth labels. All of those models are then tested on galaxies with features from the EWS survey and ground truth as labels. The reported metrics are the ones presented in Sect. 3.6. $M_\star$ refers to $\log_{10}(M_\star/M_\odot)$, SFR to $\log_{10}(\mathrm{SFR}/M_\odot\,\mathrm{yr}^{-1})$.

Table 6. Metrics for the recovered SFMS in the EWS, with the mixed labels approach.

		CSMR			CCR			DLNN			nnpz		
		m	σ	f_p	m	σ	f_p	m	σ	f_p	m	σ	f_p
EWS		1.68	0.48	0.11	1.98	0.45	0.11	1.92	0.47	0.13	1.22	0.41	0.08
C16		1.22	0.34	0.08	1.26	0.32	0.08	1.23	0.32	0.08	1.24	0.32	0.06
C25	SFMS	1.21	0.33	0.08	1.26	0.31	0.07	1.20	0.33	0.07	1.24	0.31	0.06
EDF		1.20	0.32	0.07	1.24	0.30	0.07	1.19	0.30	0.07	1.16	0.28	0.05
True		1.16	0.13	0.05	1.18	0.14	0.05	1.15	0.14	0.05	1.20	0.14	0.05

Notes. Leftmost column refers to the training sample, i.e., EDF means a model trained with EWS features and labels from `Phosphoros` results to the EDF photometry, True to models trained with EWS features and the ground truth labels. The reported metrics are the ones presented in Sect. 3.6. The SFMS ground truth values, injected in the simulation, are $m = 1.30$, $\sigma = 0.30$, $f_p = 0.06$.

that is, filters chosen to sample the galaxies emissions, even in cases where photo-z is correctly measured.

This fraction (∼9–10%) of low-z galaxies mistakenly assumed to be high-z skew the respective luminosity functions to overestimation. This is also observed in the SFMS, where the overall effect is somehow compensated by the mass-SFR scaling in the same (wrong) direction. The fit to the relation is usually way steeper than the true one when the training sample is at the same depth as the EWS, with the exception of `nnpz`.

With the mixed labels approach, the models will preferentially place fainter objects at lower-z with lower masses instead of the opposite. The other side of the coin is that now a fraction of truly high-z galaxies will be placed at lower redshifts. This is particularly visible with `nnpz` (right column of Fig. 5), and a little less with the other models. However, the net result is an improvement, as only ∼1–2% of these kinds of outliers are present in our results, while the number of low-z objects previously mistaken for high-z ones reduces by a half (from ∼10% to ∼5%).

We find that `nnpz` and CCR return the best performance of all the ML methods described in Sect. 3, followed by CSMR and the DLNN. In particular, `nnpz` returns the best results for the stellar masses and SFRs for the EWS, reducing the outlier fraction from the ∼28% of the paired labels approach to the (still high) ∼20–30% in the mixed labels one, and NMADs down to 0.61 from 0.67.

`nnpz` is also the best method for recovering the SFMS, as shown in Fig. 8. In the T.lab Deep – Wide case, all the methods get close to the true values, with `nnpz` being the closest. In fact, even in the worst case where Wide features and labels are employed in training, `nnpz` results are better than the ones obtained with the other methods, with a close to correct recovery of the slope and fraction of passive galaxies, despite a higher scatter ($\sigma = 0.42$ instead of $\sigma = 0.30$) and a very small parallel displacement of the relation due to an overall overestimation of the SFRs.

The whole EDF will not be finalized until the end of the mission. The EWS results will be largely inferred from the auxiliary fields. These results show that this will not significantly affect the EWS scientific outcomes, as the performance of the auxiliary fields at 16 or 25 ROS is only slightly (a few percentage points) worse than the ones with the full EDF photometry available.

We find photo-z metrics slightly outside the mission requirements ($\sigma_z < 0.05$, $f_{out} < 10\%$) As reported in Sect. 3.6, the metrics reported here are the ones measured on point predictions, while the requirements in Euclid Collaboration (2020)

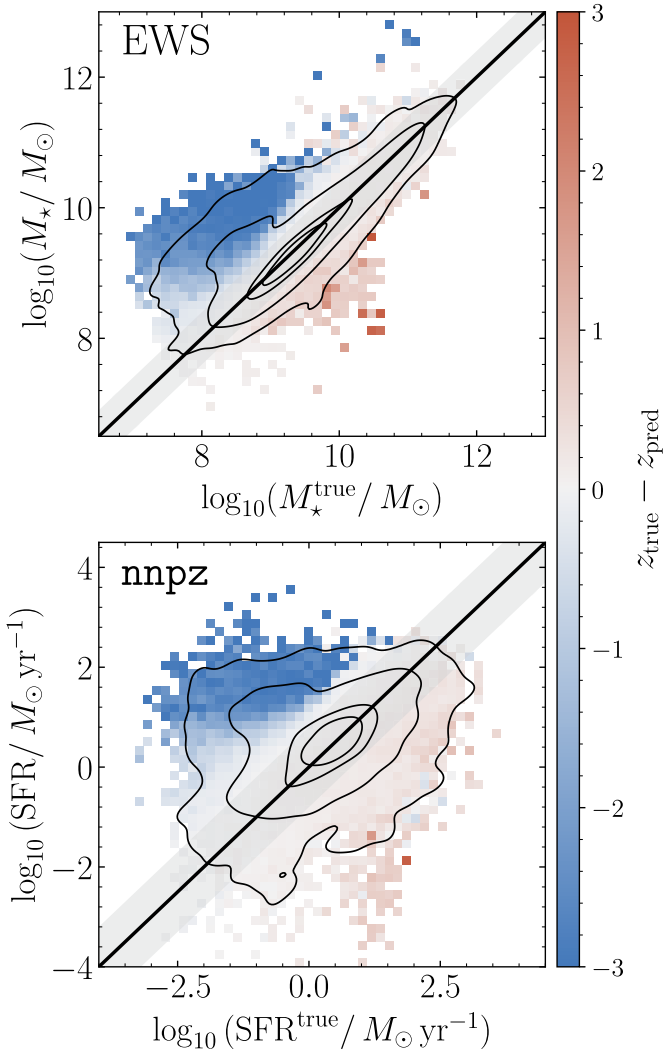

Fig. 9. Stellar masses (top panel) and SFRs (bottom panel), color coded with the difference between the true and predicted redshifts, for the nnpz results in the EWS trained with wide features (the upper-right plots in Figs. 6–7). Similar results are obtained for all the other methods considered. The black line is the 1:1 relation. Shaded area is the region beyond which a prediction is an outlier (0.4 dex for stellar masses, 0.8 dex for SFRs, see Sect. 3.6). 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. The $z_{\rm pred} - z_{\rm true}$ are measured as the median of the values falling in each true vs predicted bin. It is clearly visible how the wrong photo-z attribution generates the main bulk of catastrophic outliers. However, the distribution scatter along the 1:1 relation – i.e., the NMAD – is mainly due to the inherent difficulties of the methods in assigning the correct PPs given the set of input features independently from the inferred photo-z. This is particularly true for SFRs. This plot also illustrates how penalizing a 0.3 dex definition for catastrophic outliers would be with respect with the ones chosen for this work, which actually follows better the distribution of true and predicted PPs.

depend on the PDZ. Moreover, the calibration fields used here have recovered labels from applying Phosphoros to the 9 bands described in Sect. 2, while the real calibration fields will benefit from more multiwavelength observations. The net effect will be more reliable labels for training and, thus, improved metrics. While our findings lie just outside the mission requirements for photo-z, we can likely assume that real ones will meet them.

4.4.3. A COSMOS-like reference sample

The photometric and color gradient calibration in *Euclid* will be performed observing six of the most observed fields in the sky (the so-called *Euclid* auxiliary fields; see Sect. 6.2 in Euclid Collaboration 2024h, for further details). The COSMOS field (Scoville et al. 2007) will be one of the first to be observed, and the widest, covering ∼2 deg^2.

When applying the same expected depth cuts of the *Euclid* auxiliary fields to the COSMOS2020 catalog (Weaver et al. 2022), we are left with ∼230 k galaxies. As such, we try to quantify if and how much the performance degrades with such a reduced number of training galaxies. Therefore, we run all the previously reported tests with this smaller sample of ∼230 k galaxies for all the reference fields with the mixed labels approach and test on the EWS. This corresponds to a ∼50% cut to the training samples, though, as previously reported (Sect. 4.4), this does not automatically translate into a catastrophic reduction in performance metrics. We observe a reduction in NMAD and $f_{\rm out}$ between less than 1% and 2–3% indeed, a sign that a ∼230 k COSMOS-like reference sample is enough to reach close to the saturation limit for performance metrics given the specifics of the surveys.

4.4.4. Removing the u band in the target sample

The final design (and timing) for the EWS are still under redefinition with respect to what was reported in Euclid Collaboration (2024h). However, we already know that for DR1 we will have different observations in the northern and southern sky, with the latter lacking a u band filter in the complementary ground-based observations.

Therefore, we perform another test to quantify the performance degradation once we remove the u filter from the target sample. As reported in Appendix A, except for the $u - g$ color, which is the second most important one, the u band is typically absent from the first ∼30 features in terms of importance and usually appears with less than 0.5% importance. As such, we observe a small reduction in the metrics performance, of the order of ∼3%, for all the methods and fields considered.

4.5. Results for the Euclid Deep Fields

For the EDF, *Euclid* will observe 53 deg^2 with at least 40 ROS, pushing the expected magnitude limits two magnitudes deeper in all the bands. Moreover, the EDF will benefit from two additional bands at 3.6 µm and 4.5 µm. This will not mean just deeper data but also more robust photometry with smaller uncertainties, translating into a more reliable estimation of photometric redshifts and physical parameters, especially stellar masses.

To quantify how more reliable the EDF will be with respect to the EWS, we perform the same tests in Sect. 4.4 on a set of training (reference) and test (target) samples coming both from the MAMBO simulated EDF, with 40 ROS, the minimum expected for the Deep fields (see Sect. 2 for further details). It is useful to point out that these deeper data push significantly toward higher-z, lower stellar masses, and lower SFRs. As shown in Fig. 2, the number of $z > 4$ galaxies increases from a few 10^3 to 10^4 (in the simulated 3.14 deg^2 of the lightcone), one order of magnitude higher. A similar increase is observed for galaxies with $\log_{10}(M_\star/M_\odot) < 8$, $\log_{10}({\rm SFR}/M_\odot\,{\rm yr}^{-1}) < 0$. Correctly predicting those values will become increasingly difficult as they become more distant and less massive or star-forming, as a consequence of their inherently lower S/N and the shifting of