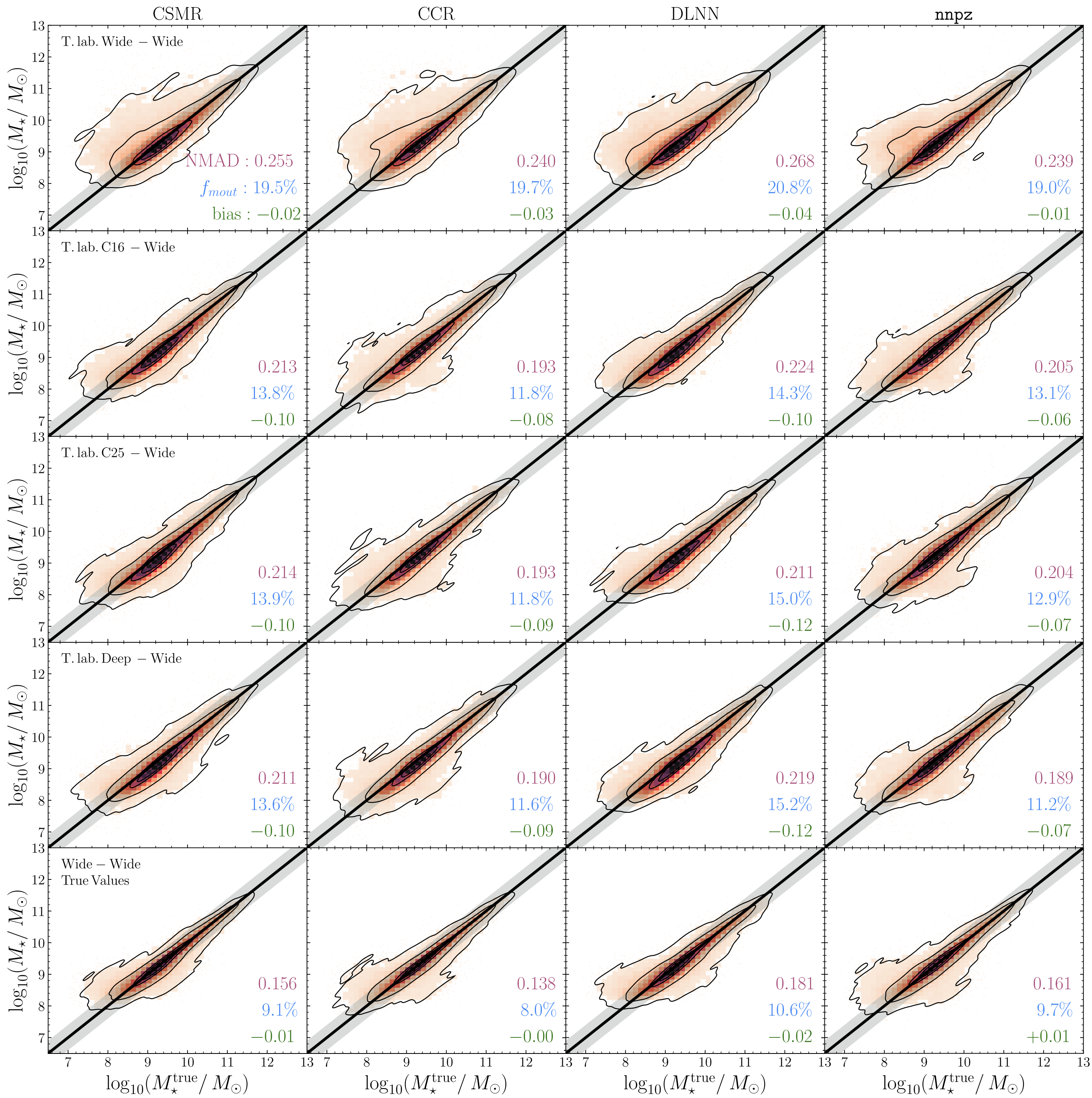

Fig. 6. Same as in Fig. 5, but for stellar masses.

4.4.2. Mixed labels approach

Trying to mitigate the effect of the aforementioned cloud of catastrophic outliers, we tried another approach, rooted in the belief that better performance should arise in training the models with the best possible set of labels for a given set of features. We refer to this one as the mixed labels approach, whose results are reported in Figs. 5–8 and Tables 5–6.

Differently from the previous approach, here we train the models with features (magnitudes and colors) always coming from the EWS catalog. However, for the deeper fields, the training labels are the `Phosphoros` results obtained with the corresponding photometry. This is specified in the plot with the text Training Label (T.lab.) followed by the name of the field. The model is then tested on features and (true) labels of the EWS.

To give an example, when referring to T.lab Deep - Wide we mean:

- training features from the EWS;
- training labels from the `Phosphoros` results obtained with the *Euclid* Deep photometry;
- test with features from the EWS and as labels the true values for $z_{\rm phot}$, $\log_{10}(M_\star/M_\odot)$ and $\log_{10}\,({\rm SFR}/M_\odot\,{\rm yr}^{-1})$.

The mixed labels T.lab. Wide-Wide case is exactly the same as the paired labels Wide-Wide case, so the first rows of Table C.1 report the same values as the first rows of Table 5

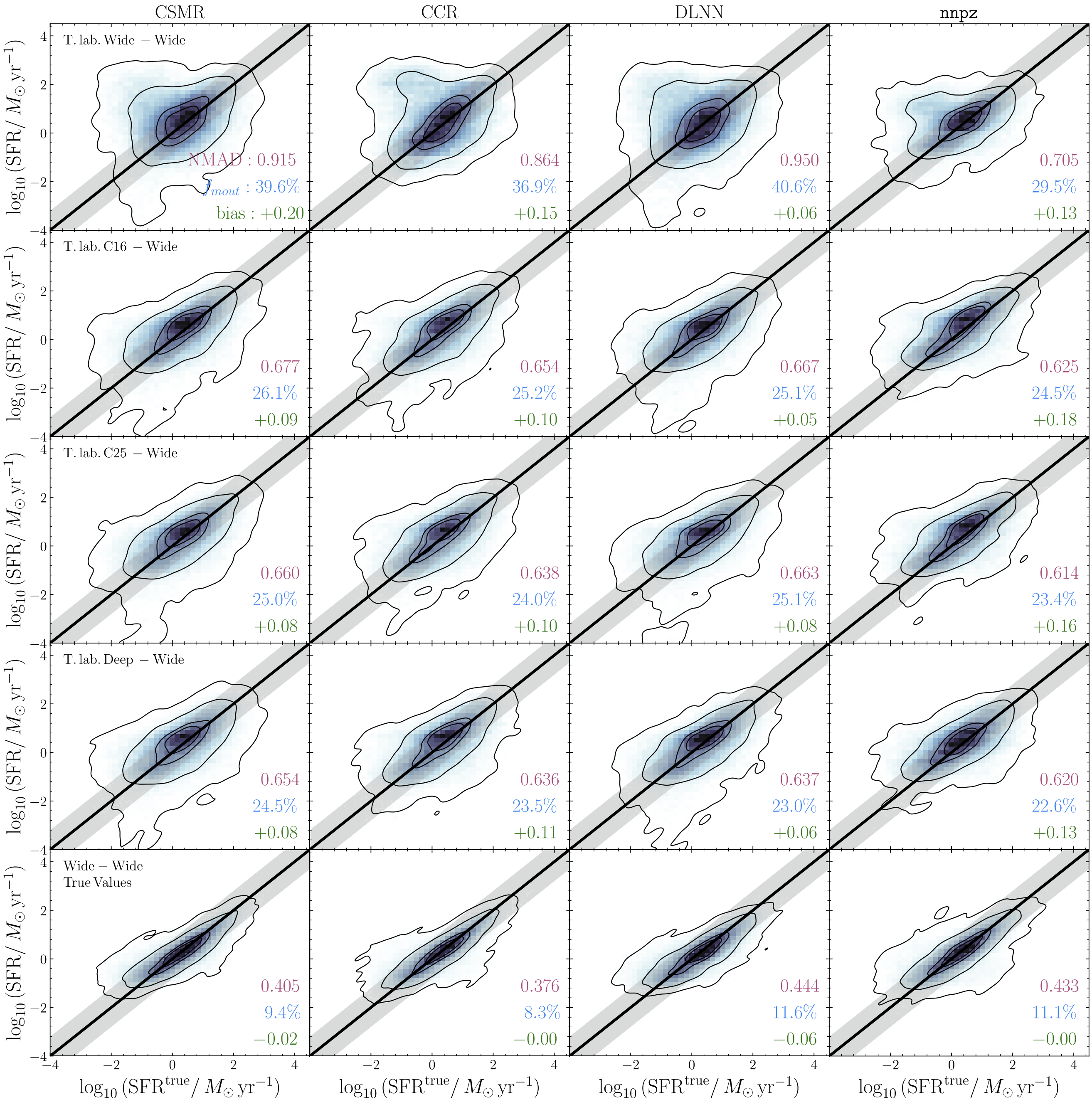

Fig. 7. Same as in Fig. 5, but for star formation rates.

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.