Euclid preparation XXVI: The *Euclid* Morphology Challenge. Towards structural parameters for billions of galaxies

Euclid Collaboration: H. Bretonnière[1,2]⋆, U. Kuchner[3], M. Huertas-Company[4,5,6,7], E. Merlin[8], M. Castellano[8], D. Tuccillo[9], F. Buitrago[10,11], C.J. Conselice[12], A. Boucaud[2], B. Häußler[13], M. Kümmel[14], W. G. Hartley[15], A. Alvarez Ayllon[15], E. Bertin[16,17], F. Ferrari[18], L. Ferreira[19], R. Gavazzi[20,16], D. Hernández-Lang[14], G. Lucatelli[12], A. S. G. Robotham[21], M. Schefer[15], L. Wang[22,23], R. Cabanac[24], P.-A. Duc[25], S. Fotopoulou[26], S. Kruk[27], A. La Marca[22,23], B. Margalef-Bentabol[22], F.R. Marleau[28], C. Tortora[29], N. Aghanim[1], A. Amara[30], N. Auricchio[31], R. Azzollini[32], M. Baldi[33,31,34], R. Bender[27,14], C. Bodendorf[27], E. Branchini[35,36], M. Brescia[37,29], J. Brinchmann[38], S. Camera[39,40,41], V. Capobianco[41], C. Carbone[42], J. Carretero[43,44], F.J. Castander[45,46], S. Cavuoti[29,47,37], A. Cimatti[48,49], R. Cledassou[50,51], G. Congedo[52], L. Conversi[53,54], Y. Copin[55], L. Corcione[41], F. Courbin[56], M. Cropper[32], A. Da Silva[57,58], H. Degaudenzi[15], J. Dinis[57,58], F. Dubath[15], C.A.J. Duncan[59,12], X. Dupac[53], S. Dusini[60], S. Farrens[61], S. Ferriol[55], M. Frailis[62], E. Franceschi[31], M. Fumana[42], S. Galeotta[62], B. Garilli[42], B. Gillis[52], C. Giocoli[63,64], A. Grazian[65], F. Grupp[27,14], S.V.H. Haugan[66], H. Hoekstra[67], W. Holmes[68], F. Hormuth[69], A. Hornstrup[70], P. Hudelot[71], K. Jahnke[72], S. Kermiche[73], A. Kiessling[68], R. Kohley[53], M. Kunz[74], H. Kurki-Suonio[75], S. Ligori[41], P. B. Lilje[66], I. Lloro[76], O. Mansutti[62], O. Marggraf[77], K. Markovic[68], F. Marulli[33,31,34], R. Massey[78], H. J. McCracken[16], E. Medinaceli[31], M. Melchior[79], M. Meneghetti[31,34], G. Meylan[56], M. Moresco[33,31], L. Moscardini[33,31,34], E. Munari[62], S.M. Niemi[80], C. Padilla[43], S. Paltani[15], F. Pasian[62], K. Pedersen[81], W. Percival[82,83,84], V. Pettorino[61], G. Polenta[85], M. Poncet[50], L. Pozzetti[31], F. Raison[27], R. Rebolo[7,86], A. Renzi[87,60], J. Rhodes[68], G. Riccio[29], E. Romelli[62], C. Rosset[2], E. Rossetti[33], R. Saglia[27,14], D. Sapone[88], B. Sartoris[14,62], P. Schneider[77], A. Secroun[73], G. Seidel[72], C. Sirignano[87,60], G. Sirri[34], J. Skottfelt[89], J.-L. Starck[90], P. Tallada-Crespí[91,44], A.N. Taylor[52], I. Tereno[57,11], R. Toledo-Moreo[92], I. Tutusaus[74], E.A. Valentijn[23], L. Valenziano[31,34], T. Vassallo[62], Y. Wang[93], J. Weller[14,27], G. Zamorani[31], J. Zoubian[73], S. Andreon[94], S. Bardelli[31], C. Colodro-Conde[6], D. Di Ferdinando[34], J. Graciá-Carpio[27], V. Lindholm[75], N. Mauri[48,34], S. Mei[2], V. Scottez[71,95], E. Zucca[31], C. Baccigalupi[96,97,62,98], M. Ballardini[99,100,31], F. Bernardeau[101], A. Biviano[62,97], S. Borgani[102,97,62,98], A.S. Borlaff[103], C. Burigana[99,104,105], A. Cappi[106,31], C.S. Carvalho[11], S. Casas[107], G. Castignani[33,31], A.R. Cooray[108], J. Coupon[15], H.M. Courtois[109], S. Davini[110], G. De Lucia[62], G. Desprez[15], J.A. Escartin[27], S. Escoffier[73], M. Fabricius[27,14], M. Farina[111], A. Fontana[8], K. Ganga[2], J. Garcia-Bellido[112], K. George[14], G. Gozaliasl[113], H. Hildebrandt[114], I. Hook[115], O. Ilbert[20], S. Ilić[116,50,24], B. Joachimi[117], V. Kansal[90], E. Keihanen[75], C.C. Kirkpatrick[75], A. Loureiro[52,117,118], J. Macias-Perez[119], M. Magliocchetti[111], R. Maoli[120,8], S. Marcin[79], M. Martinelli[8], N. Martinet[20], M. Maturi[121,122], P. Monaco[102,97,62,98], G. Morgante[31], S. Nadathur[30], A.A. Nucita[123,124,125], L. Patrizii[34], V. Popa[126], C. Porciani[77], D. Potter[127], A. Pourtsidou[52,128], M. Pöntinen[113], P. Reimberg[71], A.G. Sánchez[27], Z. Sakr[24,121,129], M. Schirmer[72], E. Sefusatti[97,62,98], M. Sereno[31,34], J. Stadel[127], R. Teyssier[130], J. Valiviita[131], S.E. van Mierlo[23], A. Veropalumbo[132], M. Viel[102,97,98,96], J. R. Weaver[133,134], D. Scott[135]

(Affiliations can be found after the references)

ABSTRACT

The various *Euclid* imaging surveys will become a reference for studies of galaxy morphology by delivering imaging over an unprecedented area of 15 000 square degrees with high spatial resolution. In order to understand the capabilities of measuring morphologies from *Euclid*-detected galaxies and to help implement measurements in the pipeline of the Organisational Unit MER of the Euclid Science Ground Segment, we have conducted the Euclid Morphology Challenge, which we present in two papers. While the companion paper by Merlin et al. focuses on the analysis of photometry, this paper assesses the accuracy of the parametric galaxy morphology measurements in imaging predicted from within the Euclid Wide Survey. We evaluate the performance of five state-of-the-art surface-brightness-fitting codes `DeepLeGATo`, `Galapagos-2`, `Morfometryka`, `ProFit` and `SourceExtractor++` on a sample of about 1.5 million simulated galaxies (350 000 above 5σ) resembling reduced observations with the *Euclid* VIS and NIR instruments. The simulations include analytic Sérsic profiles with one and two components, as well as more realistic galaxies generated with neural networks. We find that, despite some code-specific differences, all methods tend to achieve reliable structural measurements ($< 10\%$ scatter on ideal Sérsic simulations) down to an apparent magnitude of about $I_{\rm E} = 23$ in one component and $I_{\rm E} = 21$ in two components, which correspond to a signal-to-noise ratio of approximately 1 and 5 respectively. We also show that when tested on non-analytic profiles, the results are typically degraded by a factor of 3, driven by systematics. We conclude that the *Euclid* official Data Releases will deliver robust structural parameters for at least 400 million galaxies in the Euclid Wide Survey by the end of the mission. We find that a key factor for explaining the different behaviour of the codes at the faint end is the set of adopted priors for the various structural parameters.

Key words. Galaxies: structure – Galaxies: evolution – Cosmology: observations

---

⋆ e-mail: hubert.bretonniere@ias.u-psud.fr

1. Introduction

Measurements of galaxy morphology offer easily accessible information for constraining physical processes that regulate galaxy growth and evolution. Galaxy morphologies are therefore among the most important observables available from extragalactic imaging campaigns and continues to be so throughout the era of big data astronomy. This is because the distribution of the stellar light emitted by a galaxy can be correlated to its stellar populations, angular momentum, and the star-formation and merger histories (e.g., Cole et al. 2000; Conselice et al. 2003; Kormendy & Kennicutt 2004; Förster Schreiber et al. 2009; Brennan et al. 2017).

A fundamental goal of extragalactic astronomy is understanding how the diversity of galaxy morphologies is established across time. This is predicated on earlier observations, which already revealed that galaxies come in various types (e.g. Hubble 1926). The most fundamental distinction draws a line between disc-dominated structures that often appear with bright spiral arms, and bulge-dominated galaxies with smooth light distributions. Most galaxies are in fact a combination of both shapes, featuring both a bulge and a disc with varying weights. This simple scheme describes the essential building blocks of nearby galaxies. However a descriptive classification for grouping galaxies into two rough classes is a simplification, and in reality the visible part of most galaxies result from a combination of multiple components.

Characterising and classifying galaxies based on their optical morphologies is not straightforward. A number of different approaches for quantifying galaxy structure and morphology have been developed, documented, and tested in the last few decades, each designed with specific applications in mind. The general goal of all of these methods is to obtain a quantitative measurement, and an error budget, of the morphological properties of galaxies that are easy to understand, use, quantify, and replicate. Contemporary examples include visual classifications (e.g., Lintott et al. 2008; Mortlock et al. 2013; Bait et al. 2017), non-parametric morphologies (Conselice 2003; Lotz et al. 2004; Pawlik et al. 2016), one-dimensional intensity profile fitting of a galaxy's light distribution, either treating each galaxy as a whole (e.g., Sérsic 1968; Peng et al. 2002; Buitrago et al. 2008, 2013), or decomposing them into two separable components (two-dimensional surface brightness fitting, e.g., Simard et al. 2011; Lang et al. 2014), machine learning techniques (e.g., Huertas-Company et al. 2008, 2011; Vega-Ferrero et al. 2021), and structural kinematics (Förster Schreiber et al. 2009; Falcón-Barroso et al. 2017; van de Sande et al. 2017). The increasingly challenging nature of observations of fainter and more distant galaxies makes defining and distinguishing between different structures a non-trivial task. Traditional visual classifications also become ambiguous for many objects, especially for early-type galaxies. In addition, techniques need to be able to efficiently deal with the ever increasing sample sizes of galaxies in contemporary and future all-sky surveys, with an increased statistical accuracy. Light profile fitting is a quantitative, generally automatic, or semi-automatic, and often a faster approach, compared to the qualitative visual classification process. This is especially important for statistical approaches using the very large datasets we are expecting with missions such as *Euclid* in the near future.

Euclid is a European Space Agency 1.2 m space-based telescope mission, primarily designed to investigate dark energy and dark matter by mapping a large fraction of the visible sky (Laureijs et al. 2011). In order to achieve this goal, *Euclid* will conduct a Wide Survey of around 1.5 billion galaxies out to $z \sim 3$ with relatively high spatial resolution wide-field optical and near-infrared imaging, as well as low-resolution grism spectroscopy ($R \sim 250$). These data will be provided by the VIS instrument, which features one broad optical band called $I_{\rm E}$, covering approximately 540 nm to 900 nm (i.e. covering most of the usual *r*, *i* and *z* bands), and a mean image quality of $0\rlap{.}''17$ FWHM (Cropper et al. 2010). The Euclid Wide Survey will therefore provide a unique combination of high spatial resolution and wide area coverage, enabling studies of galaxy morphology and structure with unprecedented statistics. The uncommonly large wavelength range of the VIS filter provides unknown effects on determining galaxy morphologies with *Euclid*, since no previous large studies have used such a wide filter. While this filters was especially designed with *Euclid* core cosmological science in mind, it is essential to fully characterise the use of this filter for the measurement of galaxy morphologies. *Euclid*'s other instrument is the Near Infrared Spectrometer and Photometer (NISP), which will observe in three infrared bands $Y_{\rm E}$, $J_{\rm E}$ and $H_{\rm E}$, covering approximately 950 to 2020 nm (Euclid Collaboration: Schirmer et al. 2022).

Euclid's nominal requirements are to image 15 000 deg^2 or 35 % of the accessible sky down to at least a 10σ depth of magnitude $I_{\rm E} = 24.5$ in the optical and down to a 5σ depth of magnitude 24.3 at near-IR wavelengths ($Y_{\rm E} = 24.3$, $J_{\rm E} = 24.5$, and $H_{\rm E} = 24.4$). Observing strategies and initial tests of the instrument forecast higher sensitivity than the nominal requirements. In addition, the Euclid Deep Survey will provide images two magnitudes deeper in a smaller area of 40 deg^2, as part of the deep fields. *Euclid* will thus provide an unprecedented number of high spatial resolution images for morphological measurements, which will be an extraordinary database for a range of legacy science questions including galaxy formation and evolution, as well as a plethora of follow-up projects.

The Sérsic law (Sérsic 1968) is a commonly used parametric model to describe galaxy radial profiles, which can describe a variety of shapes, from a disc or underlying smooth component of spiral galaxies (Freeman 1970; Kormendy 1977), to elliptical galaxies and bulges (de Vaucouleurs 1948). The practice of fitting the Sérsic law to astronomical images of objects has become widely used. Its aim is to measure and quantify the shapes of galaxy profiles (i.e. the surface brightness profile). The success of Sérsic profiling for morphology measurements has been repeatedly shown. For example, massive elliptical galaxies are well described by one-component Sérsic profiles (Graham & Guzmán 2003; Trujillo et al. 2001) out to around 8 effective radii (Tal & van Dokkum 2011). Deep imaging of large samples of face-on late type galaxies confirm that this type is well represented by an exponential profile (Sérsic profile of $n = 1$) down to faint limits of $\mu = 27$ mag arcsec^{-2} (Pohlen & Trujillo 2006) out to at least 17 effective radii (Bland-Hawthorn et al. 2005).

Given the large number of galaxies that will be observed by *Euclid*, it is essential to obtain a fast and reliable way of measuring morphological parameters of galaxies from images. In order to understand the capabilities of measuring morphologies and structures from *Euclid*-detected galaxies, we have created the Euclid Morphology Challenge, to test, quantify, and evaluate the performance of galaxy morphology measurements by existing parameteric fitting codes on simulated *Euclid* data. The Challenge comprises a simulated dataset of five fields, each realised with single-Sérsic, double-Sérsic, and neural-network-generated galaxies in the $I_{\rm E}$ band. In addition, one of the fields has been simulated in the NIR ($Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$) bands, and in the five *u, g, r, i* and *z* Rubin Observatory bands to test the accu-

racy of multi-band-based model fitting with ancillary data. While Rubin will only cover the southern hemisphere, other facilities such as CFHT (MegaCam) or DES will also cover the northern hemisphere in similar bands. The companion paper (Merlin et al. 2022; hereafter EMC2022a), provides a visualisation of the bandwidth and wavelengths (see their Fig. 1).

In this work, we focus on quantifying galaxy structures through analytic functions that describe the shape of the surface brightness profile of each galaxy. The outcome is a set of parameters that allow the reconstruction of the 2D photometric shape of a galaxy, and thus provides for the statistical study of galaxy evolution. To carry out this challenge we have invited a number of developers of widely used software packages to retrieve morphologies and structures from our large dataset of simulated galaxies. Five teams participated in the challenge. Each team tested the performance of their codes on a common set of simulated *Euclid* galaxies that was provided to them. The codes are (in alphabetical order), `DeepLeGATo` (Tuccillo et al. 2018), `Galapagos-2`[1] (Häußler et al. 2022), `Morfometryka` (Ferrari et al. 2015), `ProFit`[2] (Robotham et al. 2017) and `SourceXtractor++`[3] (Bertin et al. 2020; Kümmel et al. 2020). At their cores, all the software packages describe the morphology or structure of each galaxy from its surface brightness distribution. The five participating model-fitting software packages are described in detail in EMC2022a, and the individual software publications referenced in each section. All but one (`DeepLeGATo`) make use of parametric methods, which use functional forms to fit the light distributions from imaging data. `DeepLeGATo` bases its photometric galaxy profile modeling on convolutional neural networks. All of them fitted at least a single profile to each galaxy in the I_E band, and some teams/codes have extended the challenge to include the simultaneous fitting of multiple images at different wavelengths.

We present the comparison analysis based on the Euclid Morphology Challenge in this paper. We investigate the outcomes from the five participating codes on simulated *Euclid* galaxies. Each software package incorporates its own preferred scheme for dealing with the data and was run by the developers or developing teams themselves. Each participant was free to choose setup parameters and criteria according to their best practice and experience, with the hope that this would ensure the best possible outcomes. This could include independent tests or cross-checks from comparing their software to a subset of the 'true' parameters of the simulated data, which we made available to the developers. Therefore we can expect that each code developer's knowledge contributes to the best possible performance of each code. No further specifics, e.g., in relation to the way of preparing or handling the data, was given to the participants. Each code has different ways of identifying unreliable fits, and we refer the reader to the publications describing each code for additional information. Our goal in this paper is to probe the robustness and accuracy of the most optimal outcome of each software package, to examine the code-to-code scatter, and investigate the known bias toward over-estimating the fitting accuracy. This paper presents a tabulated score of the performance of each code with the ultimate goal of using the optimal code for future *Euclid* observations. Ultimately, one such code will be implemented in the official *Euclid* pipeline to retrieve galaxy morphology parameters for Euclid Legacy Science.

In the rest of this paper, we first describe the data that formed the base of the challenge (Sect. 2): these are single-Sérsic simulations, double-Sérsic simulations, and what we call 'realistic' simulations that use a variational auto-encoder (VAE) trained on observed COSMOS galaxies. We then describe the metric we designed to quantify the comparison between codes (Sect. 3). In our results section (Sect. 4), we discuss each parameter separately and include a comparison of the recovery statistics, for both single-Sérsic and double-Sérsic runs. In Sect. 4.2.6, we briefly summarise multi-band fits for the four codes that provided multi-band results. This is an in-depth investigation that was briefly introduced in the companion paper using the same challenge data, but devoted to comparing results for photometry. The first sub-sections of each 'result section' detail an in-depth analysis. Readers interested in the summary only will find overview comparisons in the summary figures (Figures 6, 13, and 19) and in the 'global score' sub-sections (Sections 4.1.4, 4.2.8, and 4.3.4). Section 4.4 focuses on quantifying the uncertainty predictions that were requested as part of the challenge. We conclude our analysis with a global score in Sect. 5. One goal of this challenge is to find elements that will help to make an appropriate choice for the task of measuring morphological parameters for galaxies observed with *Euclid*. The score we developed here is, however, not able to represent all science objectives, for which individual choices will be required. Information about the reproducibility of the results can be found in the appendix.

2. Data

The Euclid Morphology Challenge addressed the robustness of structural measurements by comparing 'True' input parameters of simulated *Euclid* galaxies to outcomes/fitted 'predicted' values that are output from the software packages (often referred to as 'codes' for simplicity) we test. Simulated galaxies with known input parameters provide full control over measurement errors while minimising systematic errors. In this section, we briefly introduce the data used in the challenge. For more information, we refer the reader to the companion paper, EMC2022a.

We created five fields of 25000 × 25000 pixel each, at $0\rlap{.}''1\,\text{pixel}^{-1}$ scale, corresponding to a field of view (FoV) of about $0.482\,\text{deg}^2$. The fields were made available to the Challenge participants through an online repository, which included a description, lists of source positions and true values of one field that included single-Sérsic and double-Sérsic information for internal consistency checks and for training purposes. Each field was realised in three versions that are described in more detail below: single-Sérsic profiles, double-Sérsic profiles; and simulations with realistic morphologies for the I_E band. In one of the five fields we also provide double-Sérsic simulations in eight different imaging bands, simulating the three NIR Y_E, J_E, H_E filters and ancillary data from the five optical Rubin bands *u, g, r, i,* and *z* to test multi-band capabilities.

We simulated roughly 314 000 galaxies in each field, ranging from $I_E \simeq 15$ to $I_E \simeq 30$ magnitudes. For each field we provided five lists of objects in the format: ID, *x, y* (pixel space) to the participants. Four lists were created which included the simulated objects brighter than a given VIS nominal signal to noise ratio (SNR) thresholds for 100σ ($I_E \simeq 22$), 10σ ($I_E \simeq 24.6$), 5σ ($I_E \simeq 25.25$), and 1σ ($I_E \simeq 27.1$). The fifth list contains all simulated sources, including objects below SNR = 1. We asked the participants to fit those galaxies with at least an SNR over 5σ, where we defined the SNR of a source as the SNR of a point-source in a circular aperture with a diameter of $2''$, and thus this

---

value corresponds to galaxies brighter than $I_E = 25.3$. Note that this definition of the SNR does not consider each galaxy's relative profile, and could impact the completeness in less concentrated profiles (lower Sérsic index or larger effective radius). The vast majority (more than 99%) of galaxies have a magnitude I_E fainter than 20 (Fig. 1), which should be kept in mind when examining the results.

The input catalogues were created using the EGG simulator (version v1.3.1, Schreiber et al. 2017), which outputs a double-Sérsic components catalogue. The single-Sérsic catalogues are derived from the double-Sérsic with empirical formulae to match observations such as the one by the *Hubble Space Telescope* (HST). Figure 1 gives an overview of the distributions of the parameters we analyse in this paper for all galaxies with an SNR greater than 5σ: I_E, effective radius r_e (plotted as logarithmic, $\log_{10} r_e$); axis length ratio q; Sérsic index n for all simulated single component galaxies; and bulge-to-total ratio b/t, which is also shown for double component galaxies. The 5σ limit is defined based on the total flux of the galaxy, and roughly corresponds to $I_E = 25.3$ (see EMC2022a for more details). We describe in more detail the generation of these galaxies in the following sections. Note that the fitted Sérsic indices only range from 0.3 to 6, which are Galsim-related limitations. The same is true for q, where restrictions prevent the simulation of galaxies with an ellipticity larger than 0.9 (q smaller than 0.1).

The galaxy images were then created using the Galsim software. This challenge was designed to mimic the observational depth and conditions of the Euclid Wide Survey (Euclid Collaboration: Scaramella et al. 2021). The point spread function (PSF) models the expected behaviour of the telescope and the VIS instrument. It is more complex than a Gaussian PSF, but has a full width at half maximum (FWHM) equivalent to $0''.17$. To convolve the images, the PSF was over-sampled to different degrees: 6 times in VIS; 6 times in NIR at $0''.3$ pixel scale; and no oversampling in the external bands. Participants received a version of the Euclid PSF before oversampling to use for their measurements. There are no reported temporal or spatial variations in the models, which were taken from Euclid's Scientific Challenge 8[4]. Thus, the PSF is assumed to be constant over the FoV. Rubin's PSFs were simulated with PhoSim (Peterson et al. 2015). We also added noise that matches the Euclid Wide Survey depth, with the noise a sum of two sources, a Gaussian and a Poisson component. The fact that we did not include correlated noise could be a limitation of the simulation. Detailed information about the simulation procedure can be found in EMC2022a.

Our analyses are performed on a common catalogue that consists of 212 000 objects for the single-Sérsic simulations, 207 064 for the double-Sérsic simulations, and 204 229 for the realistic morphologies. Due to a technical issue with one of the contributing software packages that occurred during the measurements of the mono-band single- and double-Sérsic simulations of one of the fields, only four of the five fields were completed by all the participants. As a consequence, we only used the four completed fields for our analysis, and only three fields for the double-Sérsic case because one of the fields was used for the multi-band analysis only. Several codes provide a number of individual quality flags with further information on their fits, including details in relation to reliability. While it is out of the scope of this paper to analyse all the different flags of each code, we test and discuss some important flags in Appendix D.

---

[4] Euclid's Scientific Challenges are benchmark tests organised inside the Euclid Consortium in preparation for the launch of the satellite.
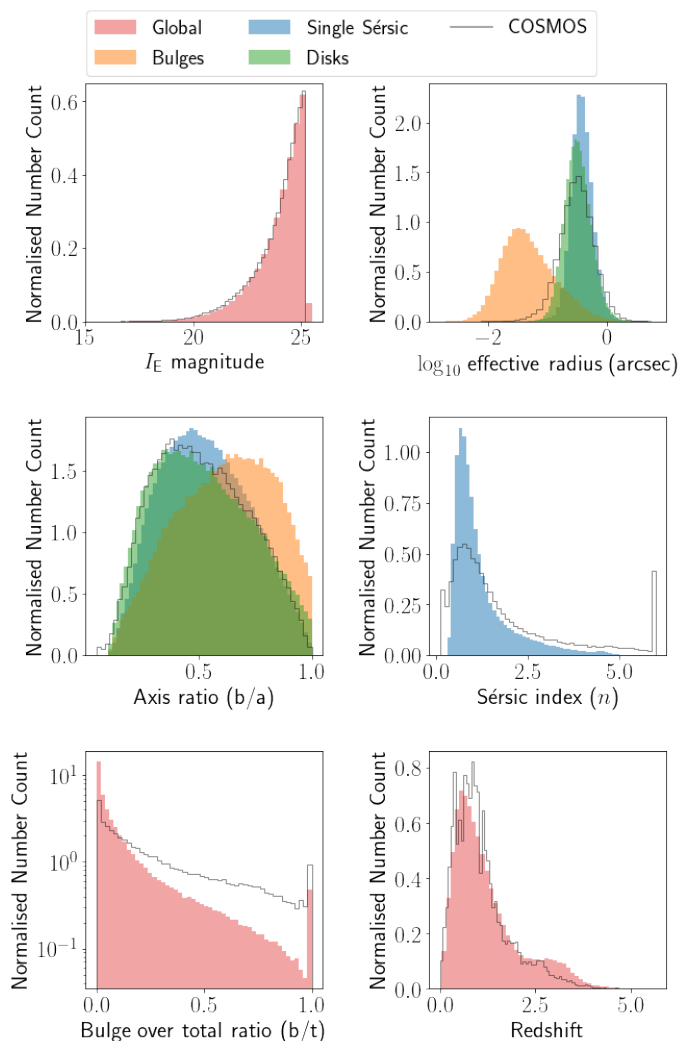
Fig. 1. Distributions of the simulated 'true' galaxy parameters measured in the Euclid Morphology Challenge. Top left: I_E distribution down to 5σ detections. Top right: effective radii for the single component galaxy (blue), and for bulges (orange), and discs (green) separately. Middle Left: Axis ratio distributions. Middle right: Sérsic index distributions for single-component galaxies. Note that Sérsic indices of the bulges are fixed to $n = 4$, and the discs to $n = 1$. Bottom left shows the bulge-to-total ratio distribution. The black solid line shows the COSMOS distribution. Note that for b/t, the y-axis is on a logarithmic scale. The distributions are normalised such that the area is equal to 1. This figure is replicated from EMC2022a.

We explain our decisions and production steps for the common catalogues in more detail in EMC2022a.

2.1. Single-Sérsic simulations

Single-Sérsic profile simulations were created using the Galsim software (version v2.2.1 Rowe et al. 2015) following a Sérsic profile, which is a characterisation of the intensity $I(r)$ of the galaxy as a function of radius. The flux varies with the distance to the centre according to the following relation:

$$I(r) \propto \exp\left[-b_n\left(\frac{r}{r_e}\right)^{1/n}\right],$$ (1)

where r_e is the effective or half-light radius, i.e. the radius in which half of the galaxy's flux is contained. This is usually considered as a proxy for the size of the galaxy and is sometimes abbreviated to 'radius' in this work. The Sérsic index is denoted n, which is a shape parameter describing the curvature of the function. It drives the steepness of the light profile, and thus describes its shape or concentration. Typically, a profile with $n = 4$ fits well to elliptical galaxies, and for $n = 1$, the Sérsic law forms an exponential function, which is often used to describe a disc. Note the presence of b_n, which can be approximated by $b_n = 2n - 1/3$, which links n and r_e (Ciotti 1991). `Galsim` simulates the surface brightness profiles at high spatial resolution, which we then sample at the image pixel scale. This is important to do in order to avoid aliasing effects, especially when the Sérsic index is large.

The galaxy model is then sheared to match the desired ellipticity, or q, which is the semi-minor over semi-major axis of the ellipse shape. The normalisation factor is fixed afterwards to match the total magnitude of the object.

2.2. Double-Sérsic simulations

Galaxy formation and evolution studies gain essential knowledge from tracing the individual galaxy components, i.e. bulges and discs, by fitting two-component models. At the simplest level, light profile decompositions enables the classification of galaxies according to their bulge-dominance. Double-component galaxies are each simulated with `Galsim` as a pixelwise sum of two profiles, one profile for a bulge and one for a disc. The disc is simulated with a Sérsic profile with $n = 1$, which thus simplifies to an exponential profile:

$$I_{\mathrm{disc}}(r) \propto \exp\left[-b_1\left(\frac{r}{r_e}\right)\right] . \tag{2}$$

The bulge profile is fixed with a Sérsic index of $n = 4$, so that the total profile combines to:

$$I(r) \propto (1 - \mathrm{b/t}) \exp\left[-b_1\left(\frac{r}{r_{e,b}}\right)\right] + \mathrm{b/t} \exp\left[-b_4\left(\frac{r}{r_{e,d}}\right)^{1/4}\right] . \tag{3}$$

The two profiles are then sheared to fit the desired ellipticity, q_b and q_d. The flux is first scaled to generate galaxies with suitable b/t, and then the global flux is re-scaled to match the global flux of the galaxy. The two components are always aligned to the same position angle, and the PSF is applied to the global profile. In addition, we created one field with double-Sérsic galaxies that includes images in nine bands, which will be relevant for tests of multi-band fitting routines (Sect. 4.2.6). The structural properties in all bands are kept constant, and therefore our simulations do not model wavelength dependent structural changes.

2.3. Realistic simulation

Simulated galaxy images are inherently difficult to produce realistically, which is why most tests for morphology measurements focus on simulating and fitting smooth analytic profiles. The Euclid Morphology Challenge also provides a dataset with more realistic galaxies learned following a data-driven approach using deep neural networks. This is described in detail in Euclid Collaboration: Bretonnière et al. (2022(a); referred to as B22 from here onwards). Very briefly, we use a deep generative model called the variational auto-encoder (Kingma & Welling 2019),

that compresses and decompresses images to learn a probabilistic latent representation of the training set distribution. Using HST images the model learns how to simulate real 2D noiseless galaxy profiles at a VIS-like resolution. A second generative model, called Normalising Flow (Papamakarios et al. 2021) is then used to condition the latent distribution with the structural parameters. The resulting architecture, called the Flow-Variational AutoEncoder (FVAE), can therefore simulate galaxies directly from a catalogue of parameters, provided that the training set properly covers the range of values. The advantage of the FVAE compared to a classical VAE or other generative network is the ability to constrain the physical parameters of the emulated galaxies.

Given the lack of very large and bright galaxies in the HST data used for training, this dataset does not include galaxies larger than 0.2 arcminutes or brighter than 20.5 mag. This only represents around 1% of the 314 000 simulated galaxies per field. Although this dataset should allow us to quantify the performance of the different codes in more realistic conditions, it is important to emphasise that these simulations are not perfect. Indeed, the conditioning of the latent space with galaxy morphology is not always exact, which can introduce a systematic bias in what we call the 'true' values for these realistic fields; we refer the reader to the discussions in B22. Also note that the model slightly differs from the one used in B22, in the sense that the magnitude is also a parameter conditioned by the Flow, which is then also re-calibrated using `Galsim`. This dependence on the Flow allows us to keep the correlation between morphology and magnitude. The post-processing steps (PSF and noise) are the same as described in the previous sections.

3. Metrics

As in the companion paper EMC2022a, we use four main indicators to evaluate and compare the different codes: completeness (C)[5]; bias $(\mathcal{B})$; dispersion $(\mathcal{D})$; and outlier fraction (O). We also combine these values into a global score, $\mathcal{S}$, to ease the comparison of the different codes. Each of these parameters is computed for each galaxy structural parameter (p), and is plotted in bins of apparent magnitude to quantify the impact of signal-to-noise. In the following, we provide a definition of each of these accuracy estimators, which slightly differs from the ones used in EMC2022a. These differences were necessary to better capture the specifics of our parameter distribution, in particular the large impact of outliers in the dispersion values.

3.1. Bias

The individual bias b_p on a structural parameter p of a galaxy is defined as the difference between the predicted value, Pred_p, and the true simulated value, True_p:

$$b_p = (\mathrm{Pred}_p - \mathrm{True}_p) , \tag{4}$$

where $p = \{r_e, q, n\}$ for single-Sérsic fits and $p = \{\mathrm{b/t}, r_{e,b}, r_{e,d}, q_d, q_b\}$ for double component fits. Sometimes it is appropriate to calculate the relative bias, $\tilde{b}_p$, which is defined as

$$\tilde{b}_p = \frac{\mathrm{Pred}_p - \mathrm{True}_p}{\mathrm{True}_p} . \tag{5}$$

---

[5] The completeness C measures the fraction of objects for which there is a successful fit, see EMC2022a for details.

The use of either the absolute or relative bias depends on the parameter. For example, the same absolute bias has a different meaning in a small galaxy than in a large galaxy: a measurement error of $0\rlap{.}''1$ for a galaxy of $r_{\rm e} = 0\rlap{.}''2$ is more problematic than the same error on a galaxy with $r_{\rm e} = 3\rlap{.}''0$. This is not the case for other parameters, such as q and b/t, which have a constrained dynamical range between 0 and 1. We also chose to use the absolute bias for the Sérsic index, even though this is less straightforward to measure, since the dependence of the profile on n is not linear. For galaxies with $n > 4$, the impact of increasing n on the surface brightness profile is small, which implies that errors on large Sérsic indices are generally less severe than on small values of n. However, since this dependence is not linear, the relative bias does not properly encapsulate this behaviour. In order to make the interpretation easier, we simply use the same absolute definition of b. The choice is also motivated by the fact that the majority of galaxies in our simulations have a low Sérsic index, for which the absolute bias is well suited (see Fig. 1).

We also define the global bias $\mathcal{B}_p$ of a population as the median of all individual biases of the population, $\boldsymbol{b}_p$:

$$\mathcal{B}_p = Q_{0.5}(\boldsymbol{b}_p) , \tag{6}$$

or if we take the relative bias,

$$\widetilde{\mathcal{B}}_p = Q_{0.5}(\tilde{\boldsymbol{b}}_p) , \tag{7}$$

which is the value reported in all subsequent sections. A statistically unbiased measurement thus corresponds to $\mathcal{B}_p = 0$. Notice that $\mathcal{B}_p$ can have positive and negative values if a given parameter is over- or under-estimated, respectively. This metric is computed on all the objects of the common catalogue, without removing the outliers, which are discussed in Sect. 3.3.

3.2. Dispersion

The dispersion of a population, $\mathcal{D}_p$ on a parameter p is defined as the 0.68 quantile ($Q_{0.68}$) of the absolute population biases from which we subtract the median bias:

$$\mathcal{D}_p = Q_{0.68}\left(\left|\boldsymbol{b}_p\right| - Q_{0.5}\left(\boldsymbol{b}_p\right)\right) . \tag{8}$$

Here again, the absolute bias b is used for q, n, and b/t, while the relative bias $\tilde{b}$ is used for the effective radii. The median bias is removed to recentre the distribution around zero, so that the quantile matches the significance of a standard deviation. We use the 0.68 quantile because it is less sensitive to outliers than the standard deviation. Outliers are quantified independently (see Sect. 3.3). Note, however, that for Gaussian distributions both $Q_{0.68}$ and the standard deviation correspond to the same measurement. Figure 2 illustrates the advantage of our dispersion metric compared to a simple standard deviation, comparing the classic standard deviation with our definition in presence of a single outlier. Whenever we use the absolute error $\tilde{b}$, we define the dispersion as $\widetilde{\mathcal{D}}_p$.

3.3. Outlier fit fraction

In addition to bias and dispersion, we also quantify the fraction of 'outliers', which could equally be called 'fraction of bad fits'. We define an outlier on a given structural parameter p when its bias b_p is larger than a given threshold (t_b), which we fix to be $t_b = 0.5$ for all parameters p. The fraction of outliers (O) is thus the number of objects above the threshold divided by the total number of objects in the considered bin. Since the bias b is not always defined in the same way for all parameters (see Sect. 3.1), the meaning of O also differs in the following three cases:
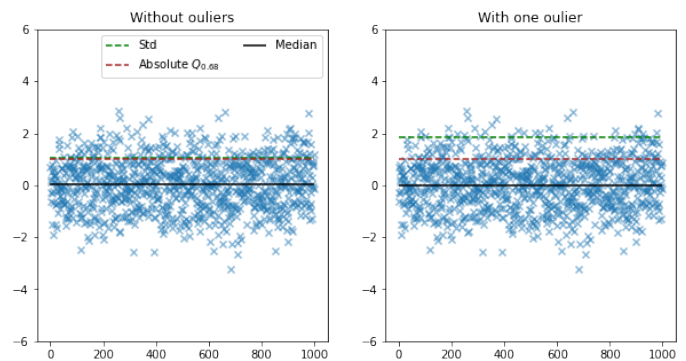
Fig. 2. Illustration of our dispersion metric choice. In both plot, we plot the median, the standard deviation and our definition of the dispersion, defined Eq. 8 for a Normal Gaussian distribution. In the right figure, we add an outlier at $y = 100$. We can see that our definition is not sensible of the presence of an outlier, compared to the standard deviation.

- **Effective radius**: because we use the relative bias $\tilde{b}$, $t_b = 0.5$ means that we consider an outlier if the relative error is larger than 50%.
- **Axis ratio and bulge-to-total ratio**: because the bias is absolute, but the range of possible values is limited to $[0, 1]$, $t_b = 0.5$ means that an outlier is defined when the error is larger than 50% of the maximum possible error.
- **Sérsic index**: since the bias is not relative and the range is not bounded, the outlier definition cannot be seen as a percentage in this case; see the discussion in Sect. 3.1.

We emphasise here that the bias and dispersion metrics are computed including the outliers.

3.4. Global score

Finally, in order to summarise the overall performance of a given code and to compare more easily the codes to one another, we define a global score $\mathcal{S}_p$ on a given parameter p, which encapsulates the four previous measurements, i.e. C, $\mathcal{B}_p$, $\mathcal{D}_p$, O_p:

$$\mathcal{S}_p = (1 - C) + \sum_i w_i \left(k_{\mathcal{B}}\mathcal{B}_{p,i} + k_{\mathcal{D}}\mathcal{D}_{p,i} + k_O O_{p,i}\right) . \tag{9}$$

$k_{\mathcal{B}}$, $k_{\mathcal{D}}$ and k_O are weights applied to each of the different precision indicators. In our case, we set the same relative weight that has been calibrated empirically, so that the order of magnitude of the score, and thus its interpretation, is consistent with the companion paper EMC2022a:

$$k_{\mathcal{B}} = k_{\mathcal{D}} = k_O = 2.1 . \tag{10}$$

With this calibration, scores generally range from 0.2 to 2, the lower the better. The sum is performed over bins of apparent magnitude. The different w_i are therefore factors that weight the score with regard to the SNR of the bin and the fraction of objects in the bin (fewer objects and lower SNR will lead to a smaller weight, and thus smaller impact on $\mathcal{S}$); see EMC2022a for more details, where the definitions of the diagnostics are similar, but not identical, due to different use cases. We emphasise that the score is intended to provide a first-order estimation of the performance of the different codes using a single number, but should not be used on its own to chose a 'best code' appropriate for every scenario. This is due to a number of additional important considerations, like the execution time or user-friendliness, which are left out. We therefore acknowledge that our global

score is a simplification and point out that alternative metrics, which could be adapted for specific science goals, might result in different conclusions. In order to support the user in tailoring the diagnostics to their individual science case, we have created an interactive plotting tool, which is published alongside this paper. It enables the recreation and adaptation of most figures shown in this paper. We describe this tool in Appendix A.

4. Results

Summarising the results in a reasonable number of figures is difficult, since the problem is multi-dimensional with several degeneracies between the different structural parameters. For simplicity, we only show the metrics as a function of apparent I_E magnitude in the main text as taken from the 'true' input values, which is a proxy for SNR. This is a limited representation of the complexity of the problem, but it is a reasonable trade-off between readability and information provided. We also provide an online interactive plotting tool[6] for full exploration of the data. Using this tool it is possible to investigate independently how the fits trend with other parameters, such as Sérsic index or size. In Figs. D.2 and D.3 of the appendix, we show and comment on an example of morphological parameters as a function of the true redshift.

The results are presented as follows. For each type of simulation – single-Sérsic, double-Sérsic and realistic – we measure our three metrics $\mathcal{B}$, $\mathcal{D}$, and $\mathcal{O}$ for each structural parameter and every code on a common dataset containing only galaxies for which all codes provide a valid fit (see also the companion paper EMC2022a). In this way, we ensure a fair comparison between the different codes. These values are summarised in Tables 1 (single-Sérsic and realistic) and 2 (double-Sérsic). Throughout the next sections, we step through our metrics analysis for each of the datasets by discussing two main types of figure. The first figure type is a scatter plot of magnitude versus b or $\tilde{b}$ for individual objects. Because the dispersion increases towards fainter fluxes (high magnitudes), the scatter plots produce a trumpet-like shape, and are therefore referred to as 'trumpet plots'. The two metrics, $\mathcal{B}$ and $\mathcal{D}$ are represented with a running orange line ($\mathcal{D}$ represented as error bars centred on $\mathcal{B}$). In this first type of figure, we also show the distribution of the bias b on the right inset plot, with the reference 0 bias in thick blue lines, and the overall bias in dashed white lines. The outlier threshold t_b is represented by dashed red lines. The second type of plot, which we call the 'summary figure', shows our three metrics $\mathcal{B}$, $\mathcal{D}$, and $\mathcal{O}$ values in 11 bins of magnitude, from magnitude 14 to 26. This allows us to plot in the same figure the five different codes for a direct comparison.

4.1. Single-Sérsic results

In this section, we analyse results from the fitting of single-component Sérsic functions that describe the radial surface brightness profile, fitted on the I_E-band images only. Figure 6 summarises the results, along with Table 1 and Sect.4.1.4. In addition, Figure 22 shows residuals between the simulation and the modelled galaxies. Naturally, single-Sérsic fits are less sensitive to small scale features, since they essentially smooth over the individual components of a galaxy. Despite this drawback, they are generally the fastest and most straightforward measure of the sizes (via the half-light radius, Sect. 4.1.1), axis ratios

---

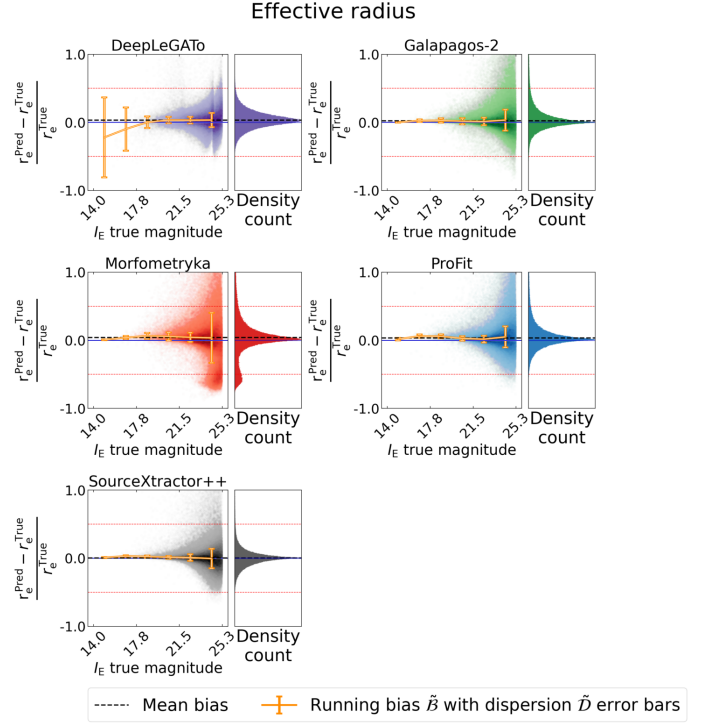[6] https://share.streamlit.io/hbretonniere/euclid_morphology_challenge

---

Fig. 3. Scatter plots showing the recovery of the the half-light radius measured from the single-Sérsic simulation. Each panel shows a different code. The main plot of each panel shows the relative bias per galaxy as a function of apparent I_E magnitude, while we summarise the bias distribution as a histogram on the right. The opacity is proportional to the density; the darker colours mean more points. The blue solid line highlights a zero bias for reference, and the gray dashed line represents the mean value of the bias for all magnitude bins together. The orange points indicate the running mean bias $\mathcal{B}$ in bins of magnitude, with error bars representing the dispersion $\mathcal{D}$ (see Sect. 3).

(Sect. 4.1.2), and shapes (via the Sérsic index Sect. 4.1.3) of galaxies. All participants returned results for this analysis, which is why figures in this section have five individual results for comparison.

4.1.1. Half-light radius

Figure 3 shows that the global behaviour of all five software packages is similar, with the expected trumpet shape visible in all plots, i.e., the scatter increases for faint objects. Moreover, the scatter plots generally do not show a significant bias (with the exception of DeepLeGATo for bright objects). Another commonality of all codes is that the trumpet plot is skewed towards positive values, i.e., the majority of outliers (points outside the two red dashed lines) are due to an overestimation of the size.

Beyond this common general behaviour, some peculiarities are notable. This includes the bias in Morfometryka's plot (in red), indicating a bi-modality at the faint end, with around 13% of objects consistently fitted with a lower radius than expected (the relative bias is around -0.5). This is due to convergence problems for objects close to the lower limit, when the fits do not update beyond the first guesses that the software uses, so outputs stall at Sérsic indices between 0.1 and 0.2. Morfometryka recognizes the unreliability of these fits with an internal flag that is given to objects with sizes smaller than the PSF's FWHM. Generally, these objects also have low Sérsic indices. This flag, 'TARGETISSTAR', is designed to flag stars, which these are not,

but their small sizes and low Sérsic indices are recognized internally as such. Such flags were not provided to the authoring team as part of the challenge. They represent around 14% of the common catalogue. We decided to keep these objects in the overlapping catalogue even after the flags were provided. The reason for this is that removing them would bias codes that were generally able to fit these objects, and because of the non-negligible fraction of the catalogue they represent. Nevertheless, even if `Morfometryka` is not able to fit these objects, they are able to recognise the problem and flag them. We show in appendix Fig. D.6 a version of the trumpet plot without those particular objects.

`DeepLeGATo` (in purple) also shows a characteristic behaviour, with a strong negative bias and dispersion for very bright objects ($I_E < 18$), and an apparent discontinuity around 24.5 mag. The first can be explained by the fact that the dataset used to train the model lacks bright objects which are rare in the observations. This is a well known effect of machine learning models, which are sensitive to the distribution of properties apparent in the training dataset. The second distinctive observation of all `DeepLeGATo` plots, the discontinuity around 24.5, is a direct consequence of the training strategy of the neural networks in bins of SNR. The abrupt change corresponds to a change of the deep learning model. Indeed, in an attempt to improve performance on both bright and faint objects, the `DeepLeGATo` algorithm was trained separately for two sets of objects, objects fainter and brighter than magnitude 24.5 (which corresponds to an SNR of 10). This leads to two sets of weights and thus to two models, which can and do behave differently. This behaviour is seen in all structural parameters for which `DeepLeGATo` produced results.

Looking ahead to the 'summary plot' in Fig. 6, the first row of the plot compares the effective radius measurements that we are discussing here. Each column shows one of the three accuracy indicators: bias ($\mathcal{B}$); dispersion ($\mathcal{D}$); and outliers fraction ($\mathcal{O}$). Note that to better highlight the small differences between the codes, the *y*-axis range has been reduced.

The first column, $\mathcal{B}$, reveals that in general all codes slightly overestimate galaxy sizes, which confirms the trend seen in the trumpet plots. Only `DeepLeGATo` dramatically under-estimate the radius of the very bright galaxies, with a decreasing bias from -0.4 (outside the plotted area) at $I_E = 14.5$ to -0.05 at $I_E = 17.5$. In addition to the lack of bright objects in the training set, this can be explained by the fact that `DeepLeGATo` works with a fixed stamp size of 64×64 pixel, which can cut the edges of the galaxy profile and thus lead to an under-estimation of its radius. We can also see that `ProFit` very slightly under-estimates the radius for the first bin (very bright galaxies). However, given that this bin has less than ten galaxies, the statistics may not be large enough to point to a particular trend. We again note that the first four bins only hold around 100 galaxies, which represent less than 1% of the entire catalogue. Importantly though, the absolute value of the bias remains smaller than 7% for all magnitudes and all codes (and for $I_E > 17$ for `DeepLeGATo`, as discussed), which means that despite their different approaches, there are no major differences between the $\mathcal{B}$ values of the different codes. We can see that for the three brightest bins, `Galapagos-2`, `Morfometryka`, and `SourceXtractor++` perform very similarly, with `Galapagos-2` reaching a slightly smaller bias. `ProFit`'s bias is less stable; tt first has a slightly higher bias, which decrease between $I_E = 17$ and $I_E = 23.5$. For those intermediate magnitudes, `Galapagos-2` and `SourceXtractor++` perform very similarly, while `DeepLeGATo` and `Morfometryka` have a higher positive bias. Finally, for the very faint galax-

ies ($I_E > 24$), `SourceXtractor++` has a bias close to zero, followed by `Morfometryka`, `DeepLeGATo`, `Galapagos-2` and `ProFit`.

The second column of the summary figure compares the dispersion $\mathcal{D}$ of all codes. The trends are generally comparable, staying below 0.1 at $I_E < 24$ for all codes except for `DeepLeGATo` for bright objects. Here again, and for the same reasons explained in the previous paragraph, `DeepLeGATo` shows a high dispersion, decreasing from about 0.8 (off the displayed plotting area) at $I_E = 14.5$ to 0.2 at $I_E = 17.5$. We can also see the higher dispersion for `ProFit` in the first magnitude bin. The four codes behave similarly with differences of only a few percent for $I_E < 23.5$, with `SourceXtractor++` having the smaller dispersion, followed by `ProFit` and `Galapagos-2`, `DeepLeGATo` and `Morfometryka`. For fainter objects, `DeepLeGATo`'s dispersion stays below 0.10, while `SourceXtractor++`, `Galapagos-2` and `ProFit` increase to 0.15. `Morfometryka` shows the largest dispersion, up to 0.45 (again, off the plotting area) for the lowest SNR bin. As seen in the trumpet plot, the dispersion at the faint end is dominated by a long tail in the distribution, with a large fraction of objects being estimated to be too large.

Regarding the fraction of outliers (third column), we see that at the bright end, all codes except `DeepLeGATo` have no bad fits (the only bin with a non-zero outlier fraction is `ProFit` and that concerns only one galaxy). For $I_E < 23$, all the codes have less than 10% outliers, with `ProFit` and `Galapagos-2` showing the smallest numbers of bad fits, followed by `SourceXtractor++` and `Morfometryka`. For fainter objects, all measurements except for `DeepLeGATo`, and to some extend `SourceXtractor++`, increase significantly, up to approximately 30% for `Morfometryka` and 20% for `ProFit` and `Galapagos-2`. On the contrary, `DeepLeGATo` has close to zero outliers for $23 \leq I_E \leq 26$ and `SourceXtractor++` also keeps a relatively small fraction of bad fits, with up to 5% for the fainter objects. `Morfometryka`'s outlier fraction for faint objects is due to the accumulation of galaxies around $b = -0.5$, which we have commented on before and are flagged during a regular output catalogue with the flag 'TARGETISSTAR' (see also Fig. D.6).

We remind the reader that even if the individual three metrics in Fig. 6 seem unfavourable for `DeepLeGATo` measurements of bright galaxies, this has little impact on the global score $\mathcal{S}$, affecting only 93 galaxies, i.e less than 1% of the fitted catalogue.

4.1.2. Axis ratio

We now move on to the axis ratio *q*. Recall that *q* has the opposite interpretation compared to ellipticity, a high *q* describing a circular galaxy. We see in the trumpet plot of Fig. 4 an overall good recovery from all codes, with almost zero bias and a reasonably low dispersion. The discontinuities between SNR bins for `DeepLeGATo` is much less noticeable, and the bias for bright objects is also lower. Evidently, also `Morfometryka`'s buildup of unreliable size measurements for small objects (and Sérsic indices as we will see in the next section) are not a problem for providing accurate axis ratios.

The second row of Fig. 6 shows the summary of the three metrics for *q*. Axis ratios are measured remarkably well, with a bias smaller than 3% for $I_E < 26$ for all codes, and for $I_E < 23$ for `Galapagos-2`. `Galapagos-2` has a slightly larger bias than the other codes for the faint objects, with a tendency to estimate more elongated galaxies. However, it remains smaller than 0.07 even in the faintest object bin. We still see a large bias for `DeepLeGATo`, which oscillates between around -0.09 and 0.07 from $I_E = 14$ to $I_E = 17$ (cut by the *y*-axis range in the graph
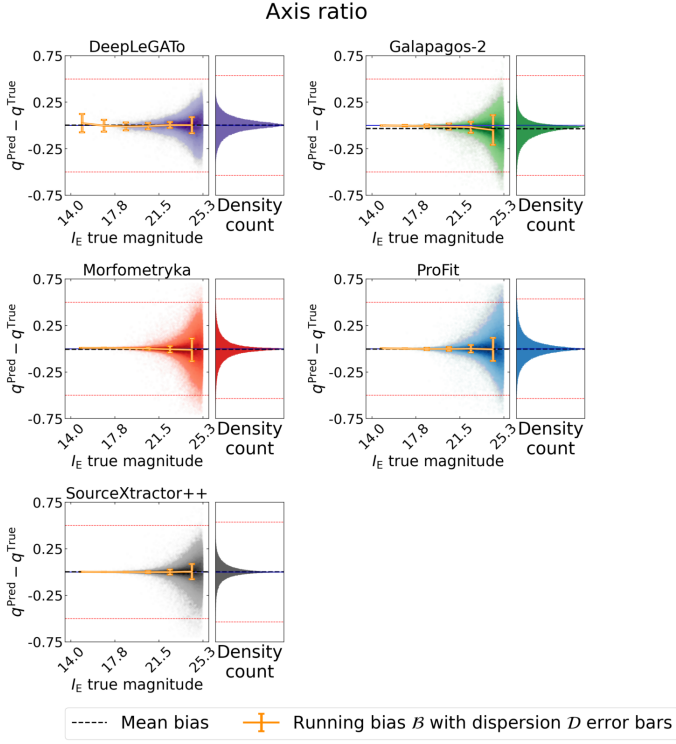
Fig. 4. Fitting results for the axis ratio q of the single-Sérsic simulation. See caption of Fig. 3 and Sect. 4 for further information.

Fig. 5. Fitting results for the Sérsic index of the single-Sérsic simulation. See caption of Fig. 3 and Sect. 4 for further information.

for visualisation purposes). For $I_E < 24$, `SourceXtractor++` and `ProFit` behave similarly well (nearly no bias), followed by a fraction of percent for `Morfometryka` and `Galapagos-2`. `Morfometryka` over-estimates q for faint objects and under-estimates it for bright objects. In the last (faintest) magnitude bin, we can see that `SourceXtractor++` and `DeepLeGATo` slightly over-estimate q, while the other three under-estimate it, which could suggest that the problem comes from the difficulty of the task at very low SNR, rather than a problem linked to the estimation of the PSF.

Regarding the dispersion, all codes except `DeepLeGATo` have a smooth increase with magnitude, from zero up to respectively 0.10 for `SourceXtractor++` and `DeepLeGATo`, 0.15 for `Morfometryka` and `ProFit`, and 0.20 for `Galapagos-2`, and it remains smaller than 0.1 for all codes at $I_E < 24$. For $I_E < 22$, `Morfometryka` and `SourceXtractor++` achieve the smallest dispersion. `DeepLeGATo`'s high dispersion at the bright end relates to issues already expanded on previously.

The outlier fraction (third column in Fig. 6) is overall below 1% for all codes and magnitudes. This is another sign that the ellipticity is one of the parameters which is generally recovered reliably by all software packages, even though an outlier threshold of 0.5 is quite permissive. Indeed, galaxies with a true value of 0.5 cannot be fitted as outliers, but we chose to keep this definition for simplicity of the metric. Furthermore because the metric is the same for all codes, we believe this comparison to be fair. We can see that even though `DeepLeGATo` has the strongest bias and dispersion for bright objects, they are still well below the outlier threshold, and stay very close to zero even for the faintest galaxies. For the other software packages, the fraction of outliers starts to be non-zero for $19 \leq I_E \leq 21$. The interested reader is invited to use the interactive plotting tool released together with this work to investigate the result on the fraction of
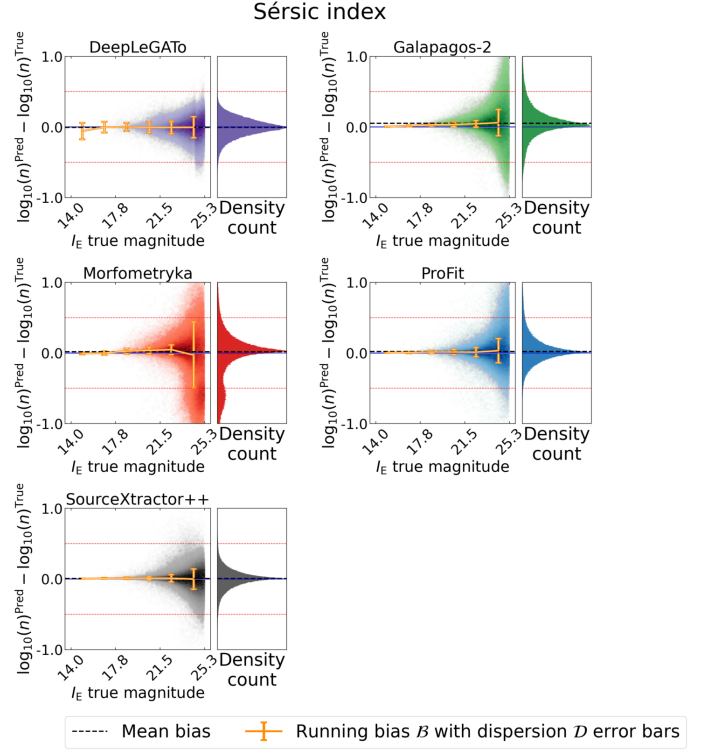
outliers. It allows one change (and therefore to decrease) the outlier threshold.

4.1.3. Sérsic index

In this section we inspect the estimation of the Sérsic index of galaxies (Fig. 5). As a reminder, the Sérsic function is a simplified model that does not capture the entire galaxy, but gives important information about how the intensity varies with radius. Compared to other morphological parameters retrieved from single-Sérsic model fitting, the Sérsic index is regarded as the most challenging parameter to recover (Buitrago et al. 2013; dos Reis et al. 2020). Because the dependence of light profiles on the Sérsic index is exponential, we always analyse $\log_{10}(n)$ instead of n in the following (see e.g. Kelvin et al. 2012 for an extended discussion).

All codes display the familiar trumpet shapes with the known caveats in `DeepLeGATo` and `Morfometryka`. Beyond that, we observe that `DeepLeGATo`, `Morfometryka` and `SourceXtractor++` tend to be skewed towards negative values for faint objects (indicating the prediction of smaller $\log_{10}(n)$ compared to the truth), while `Galapagos-2` and `ProFit` show the opposite trend.

The third row of Fig 6 presents the metrics for the logarithm of the Sérsic index. While `DeepLeGATo`'s performance for fitting bright objects is less biased compared to the previous parameters, it still has the largest negative bias for the smallest magnitude bins, which means it predicts bright galaxies without steep cores (i.e., bulges). Beyond this bright end, `DeepLeGATo` is the only code that does not over-estimate the Sérsic index, i.e., does not predict steeper galaxy profiles in their cores. For fainter galaxies, from $I_E = 17$ to $I_E = 26$, `DeepLeGATo` achieves the most robust bias calibration, mitigated by the fact that it has the highest dispersion. `SourceXtractor++` and `ProFit` have
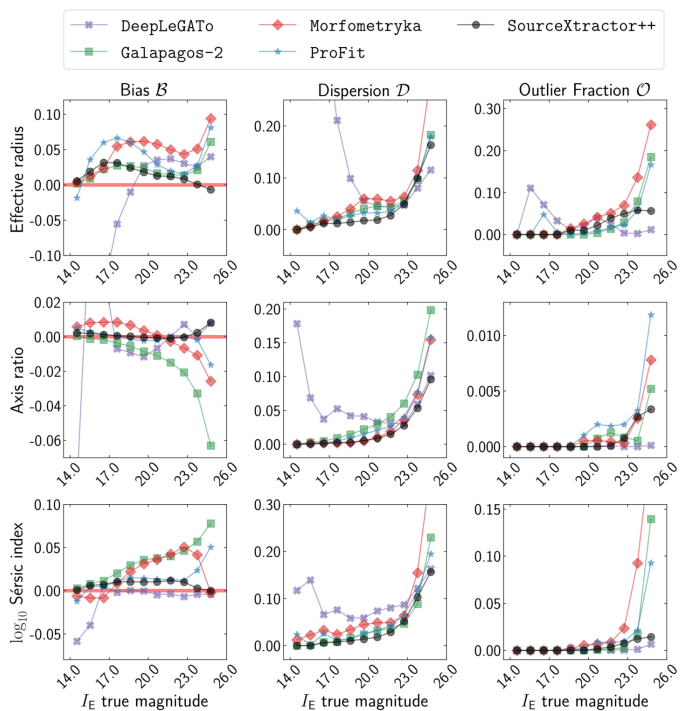
Fig. 6. Summary plot for the single-Sérsic simulation. The different rows show the results for the three different structural parameters: half-light radius r_e (top), axis ratio q (middle) and Sérsic index n (bottom). Columns represent (1) the mean bias $\mathcal{B}$, (2) the dispersion $\mathcal{D}$, and (3) the fraction of outliers $\mathcal{O}$, per bin of I_E magnitude (see text for details). Note that the y-axis is sometimes cut at low values to highlight the small differences between the software packages. Each code is plotted with a different colour as labelled.

Table 1. Comparison of the scores $\mathcal{S}$ obtained by the different software packages in all structural parameters for the single Sérsic simulations. Numbers in blue are results from Galsim simulations (discussed in Sec. 4.1), and numbers in black quote results from measurements of simulations with the deep generative model FVAE (discussed in Sec. 4.3). The last column is the mean of the parameters. A smaller $\mathcal{S}$ means a better fit.

Galsim / FVAE	$\mathcal{S}_{r_e}$	$\mathcal{S}_{b/a}$	$\mathcal{S}_n$	$\mu_\mathcal{S}$
DeepLeGATo	0.37 / $\emptyset$	0.25 / $\emptyset$	0.38 / $\emptyset$	0.33 / $\emptyset$
Galapagos-2	0.58 / 2.05	0.43 / 0.79	0.60 / 1.29	0.54 / 1.38
Morfometryka	1.10 / $\emptyset$	0.37 / $\emptyset$	1.20 / $\emptyset$	0.89 / $\emptyset$
ProFit	0.47 / 1.82	0.21 / 0.65	0.40 / 0.78	0.36 / 1.09
SourceXtractor++	0.38 / 1.84	0.18 / 0.60	0.29 / 0.75	0.28 / 1.06

a similarly small bias (around 0.01) for $I_E < 23$, which then decrease close to zero for `SourceXtractor++` and increases to around 0.5 for `ProFit`. `Morfometryka`'s and `Galapagos-2`'s bias steadily increase for ever fainter galaxies. `Galapagos-2` increase up to 0.07, while `Morfometryka` abruptly falls to -0.1 due to the known accumulation of objects that were not successfully modelled.

The behaviour of the dispersion (second column) is similar for all codes except for `DeepLeGATo` for $I_E < 23$, with a dispersion lower than 0.10. `SourceXtractor++` has the lowest dispersion, followed by `Galapagos-2` and `ProFit`, `Morfometryka`, and `DeepLeGATo`. Here again, the difference between the four first codes is very marginal. The dispersion $\mathcal{D}$ then increases for every code, up to 0.16 for `SourceXtractor++` and `DeepLeGATo`, 0.2 for `ProFit`, 0.25 for `Galapagos-2`, and 0.65 for `Morfometryka`, which can once again be explained by the cluster of points around 0.5. None of the codes suffer from bad fits (third column, $\mathcal{O}$) for $I_E < 19$, and just up to few percents for $I_E < 22$. The fraction then increases steeply at faint magnitudes. The increase is highest for `Morfometryka`, from about 1% at $I_E = 20$ up to 34% at the faintest bin, again related to the discussed failed fits. `Galapagos-2` increases to 15% only in the faintest bin. `DeepLeGATo` achieves the lowest number for all magnitudes, followed by `SourceXtractor++` also at the faint end.

4.1.4. Global scores

The blue numbers in Table 1 summarise the global scores (see Eq. 9) for the three parameters of the single-Sérsic simulations

and for the five codes. An average global score $\mu_\mathcal{S}$ is also provided. They are also plotted in the first part of Fig. 23. The best score is obtained for `SourceXtractor++`, which achieves a value of $\mathcal{S} = 0.28$. In addition, the table also shows that some codes behave better than others for some specific structural parameters. For example, `Morfometryka` is better for the axis ratio than for the effective radius, where it is highly penalised by the large dispersion for faint objects that we discussed. We emphasise again that this score is very sensitive to the different weights on the number of objects, the SNR, and the weights of the metrics. In particular, the weights of the smallest magnitude bins(from $I_E = 14$ to $I_E = 19$) have close to no impact on the score, because of the very small number of objects in those bins. It explains why `DeepLeGATo` has a good global score while performing worse than the other codes for bright objects, while other codes like `Galapagos-2` or `Morfometryka` perform best for certain parameters. By the nature of how we set up the metric, the order of the global score ranking can therefore change if we adjust the different weights to reflect a specific emphasis. We encourage the reader to explore the interactive tool released with this work, to tune this score to their particular science case.

4.2. Double-Sérsic results

We now analyse the measurements from the double-Sérsic simulations. Figure 13 summarises the results, along with Table 2 and Sect. 4.2.8.

As expected, separating the galaxy light into two components is a more degenerate problem than the single-Sérsic model fitting. This is enhanced by the fact that bulges and b/t in our sample are generally small, i.e. the bulge component has a low SNR compared to the disc (see Fig. 1). We also note that `Morfometryka` did not provide results for the bulge-disc decomposition. It is therefore excluded from the comparison in the following sections. Another difference compared to the single-Sérsic dataset is that one of the fields contained multiple bands including *Euclid* NIR and Rubin filters. In the following we only show results for 3/5 fields with VIS-only data. The multi-band dataset is analysed separately in Sect. 4.2.6. Finally, we note that

while the simulations were made with a bulge Sérsic index fixed to $n = 4$, and a disc with a fixed $n = 1$, we asked the participants to also model the galaxies with a free bulge Sérsic index. We compare the results for free and fixed bulge fittings in Sect. 4.2.7. Here, we concentrate on the model using a fixed value of n. Notice that because DeepLeGATo does not fit a model, it does not have those two different versions.

4.2.1. Bulge-to-total flux ratio

We first inspect how accurately the bulge-to-total flux ratio b/t is recovered. The results are shown in Fig. 7. First, we see that SourceXtractor++ and DeepLeGATo are less impacted by the low SNR at the faint end of the plot than the other two codes, with the trumpet shape highly concentrated towards zero bias (peaked Gaussian distribution in the histograms). Galapagos-2 and ProFit have highly non-Gaussian distributions of biases, with a tendency of over-estimating b/t for faint objects. This is obvious both in the distributions of b/t and of the bulge radius (Fig. 8). This suggests that in cases where the bulges are small and faint, these codes tend to fail to properly disentangle the flux of the bulge from the flux of the disc. As a consequence, a part of the disc's flux gets attributed to the bulge. A possible explanation for the SourceXtractor++ and DeepLeGATo ability to avoid this effect could be the use of favourable priors. Surprisingly, the figure shows that the metrics are better for faint objects, where the constraining power of the data is theoretically the lowest, and therefore the estimation is mostly driven by the prior. SourceXtractor++ uses an explicit prior of 0.022 for b/t, which matches the average b/t in the simulation. It was calibrated by the participants on a sub-sample of the dataset with known ground truth. DeepLeGATo also implicitly learns the prior from the data during training, by maximizing the likelihood. Galapagos-2 uses arbitrary priors and initially places half the light in the bulge and half in the disc. ProFit starts with reasonable initial guesses for the profile solution based on runs of the ProFound software on the cutouts (Robotham et al. 2018), but these initial conditions remain less accurate than the ones used by SourceXtractor++. These trends seem to confirm that the information contained in the images at the faint end is limited and therefore the final results are in most cases driven by the priors.

The summary of the metrics is provided in Fig. 13; the first row detailing b/t. For $I_E < 23$, Galapagos-2 achieves the lowest bias, followed by SourceXtractor++. ProFit has a tendency to over-estimate b/t, even for the brighter objects with increasing bias up to 0.37 for the faintest objects. Galapagos-2 has a similar bias in the faint end, but starts rising at fainter magnitudes ($I_E \simeq 23$ versus $I_E \simeq 19$ for ProFit). DeepLeGATo starts to be competitive around $I_E = 20$, and achieves the lowest bias at the faint end, followed by SourceXtractor++. DeepLeGATo generally under-estimates b/t, which is the opposite trend than the one seen in the other codes. This may be due to DeepLeGATo's learning being driven by the implicitly learned prior rather than by a disentangling of light based on profile fitting. Galapagos-2 has the smallest dispersion (second column) for the brightest objects, but then $\mathcal{D}$ increases to 0.14 for fainter objects. This is comparable to ProFit from $I_E \simeq 17.5$ onward. DeepLeGATo has a high dispersion up to $I_E \simeq 21$, which decreases from 0.5 to 0.05 at the faint end – a similar dispersion to SourceXtractor++. SourceXtractor++ stays relatively stable at all magnitudes, with dispersion between ~ 0.05 (bright) and 0.10 (faint). The trends for the outlier fractions are similar in all codes, with ProFit's outliers starting to
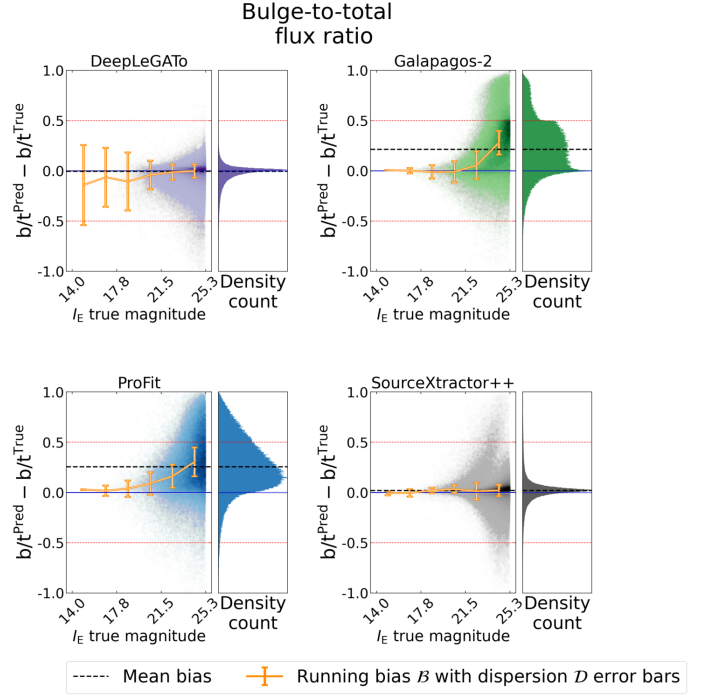

Fig. 7. Fitting results for the bulge-to-total flux ratio using the double-Sérsic simulation. See caption of Fig. 3 for further information.

increase from $I_E \sim 19$ onward and up to a fraction of 30% for the faintest galaxies. Galapagos-2 has close to no outliers up to $I_E \simeq 22$ and a fraction of 0.28 for the faintest galaxies. Compared to Galapagos-2, SourceXtractor++ has a slightly larger outlier fraction, but then keeps outliers to under 5% in the faintest bins. DeepLeGATo retains the lowest number of outliers for $20 < I_E < 26$, but reports some bad fits among the brightest objects.

4.2.2. Bulge half-light radius

We now inspect the estimation of the effective radius of the bulge component. Figure 8 clearly reflects the difficulty in obtaining reliable structural measurements of bulges. First, for all codes, the bias distributions are skewed towards positive values, i.e. an over-estimation of the true size. This can be directly linked to the fact that the bulge-to-total flux ratios are generally over-estimated. Figure 1 shows that bulge radii are small, and because the bulges are generally smaller than the discs, they are submerged inside the disc profiles, making it increasingly challenging to accurately estimate their radii.

The second row of the summary figure (Fig. 13) details this observation. Note that the scale is logarithmic for the bias and the dispersion, to help appreciate the differences for faint objects. We can see that the four codes (except DeepLeGATo) for the first two bins of very bright objects) have a similar value absolute for $I_E < 20$. SourceXtractor++ and ProFit slightly over-estimate the radius while DeepLeGATo and Galapagos-2 slightly under-estimate it. For fainter objects, DeepLeGATo keeps the lowest bias, followed by Galapagos-2 and SourceXtractor++ and then ProFit for $I_E < 22.5$. For the challenging faint galaxies, SourceXtractor++ decreases to close to zero bias, which could be explained by the correct choice of priors, as discussed in the previous subsection. Galapagos-2 and ProFit's $\mathcal{B}$ rise up to approximately 10 and 60, respectively, at the faint end. A similar behaviour is visible in

the dispersion: `ProFit` and `Galapagos-2` increase in similar ways up to 4 at $I_E = 23$, which rises up to 80 for `ProFit`, while `DeepLeGATo` and `SourceXtractor++` keep their dispersions below 1. The challenge of fitting bulges becomes even more obvious when we look at the outlier fraction. Indeed, we can see that for the faintest bins – and always according to our arbitrary definition of outlier – more than half of the galaxies are poorly fit, close to 100% for `ProFit`. For brighter objects ($I_E < 23.5$), `Galapagos-2` maintains the lowest number of outliers, from close to zero to around 10%, while `SourceXtractor++` goes up to $\sim 30\%$, and `ProFit` 50%. Again, this seems to reflect the fact that when the fit can be robustly constrained by the data because it has high SNR, `Galapagos-2` performs well since the prior is not that relevant.

In order to better understand this large bias and fraction of outliers, we show in Fig. 9 the different metrics as a function of the bulge-to-total fraction (*x*-axis) in addition to magnitude (*y*-axis). It is well known that the accuracy of bulge-disc decompositions are correlated with magnitude and bulge-to-total ratios. Understanding the metrics in relation to the true value of a galaxy's b/t can help to disentangled those two effects. In this figure, we want to highlight the absolute magnitude of the bias and dispersion, independent of their sign. The plot therefore shows for which types of objects measurement errors are large versus where they are small. For this, we compute the absolute mean bias per bin of magnitude and b/t ,

$$|\widetilde{\mathcal{B}_{r_e}}| = \overline{|\tilde{\boldsymbol{b}_{r_e}}|} \, ,$$

while the dispersion is the same as for the other cases (see Eq. 8). In this figure, the colour of the square shows the bias $|\widetilde{\mathcal{B}_p}|$ (lighter colours indicate smaller bias), and the coloured discs indicate the dispersion (the redder the point the smaller the dispersion). The first column plots results for the bulge radius, the second for the disc radius, and each line is a different software code. Note that we limit the magnitudes to faint galaxies ($I_E > 18.5$) and that for `ProFit` and `Galapagos-2`, the colour-bars are on a logarithmic scale to accommodate the large values. The expected behaviour is particularly clear for `ProFit` (third row), which we use here to for demonstration. The bias of the bulge radius $\mathcal{B}$ becomes smaller for brighter and more bulge-dominated galaxies (lower right corner of the plot) and the dispersion is low. On the contrary, a faint galaxy with small b/t has a high bias and high dispersion. The opposite is seen for disc radii: biases are highest in faint bulge-dominated galaxies. The figure therefore confirms that most of the catastrophic fits for `Galapagos-2` and `ProFit` correspond to faint galaxies with low b/t. When the bulge component is dominant, the overall accuracy improves significantly. For example, for `Galapagos-2`, the dispersion stays below 1.5 if we remove the extreme bin of b/t (b/t < 0.2), and the bias remains under 2. We can see the same behaviour for `ProFit` if we remove the low b/t (first column), and faint objects (top row), with a dispersion and bias staying lower than 3. The plot also uncovers some unexpected behaviour: `SourceXtractor++` struggles to measure bulge and disc radii for faint bulge dominated objects, but also for brighter objects (disc radius) and bright objects with small bulges. We encourage the reader to go to the online platform and adapt those graphs according to their interests, e.g., removing the extreme cases, for a better visualisation.
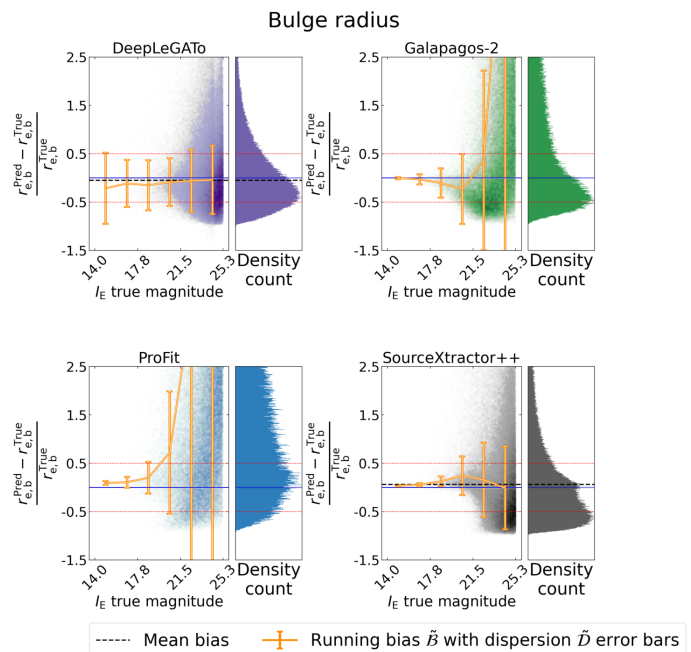

Fig. 8. Fitting results for the bulge radius using the double-Sérsic simulation. Notice that only four codes provided results for the double-Sérsic simulation. From top to bottom and from left to right: `DeepLeGATo`; `Galapagos-2`; `ProFit`; and `SourceXtractor++`. See caption of Fig. 3 for further information.

4.2.3. Disc half-light radius

Figure 10 shows the trumpet plots for the half-light radius measurements of the disc component. Results are noticeably more symmetric than for the bulge component and in fact are similar to the results reported for the single-Sérsic case. One noticeable difference is the bias of `DeepLeGATo`, which is inverted; bright galaxies are estimated with larger discs compared to the truth. As previously discussed this is related to discs generally being larger than bulges and the small bulges contained in the simulations. While being symmetric, the 'trumpets' (and thus the bias distributions) are significantly wider, with prominent wings in the histograms.

The third row of Fig. 13, confirms that the overall reliability of the estimation of the disc structural parameters is comparable to the single-Sérsic r_e fit, with a slightly higher bias and dispersion. Beyond that global view, trends for bright galaxies are opposite to these for the single-Sérsic radius estimation: an over-estimation of the radius for `DeepLeGATo` and an underestimation for the three others. We can see that all codes maintain absolute biases smaller than 0.04 – apart from `DeepLeGATo` , for galaxies brighter than 21, and the fainter bin of `Galapagos-2` (with a bias of 0.15). `SourceXtractor++` retains its disc radius bias close to 0 over all magnitudes except for the last one, where it goes up to 0.04. `ProFit` has a slightly larger negative bias at intermediate magnitudes. Similarly to bulges, the dependence on magnitude is not as obvious as for the single-Sérsic case, because of the additional dependence on b/t. However, the impact is less obvious for discs, given the b/t distribution skewed towards small bulges (Fig. 1). The second column of Fig. 9 again explores bias and dispersion for b/t and magnitude trends. It shows that accuracy increases for bright objects and low b/t. Regarding the dispersion (Fig. 13), we can see a steady increase with magnitude, peaking at 0.19, (`SourceXtractor++`), 0.21 (`DeepLeGATo`), 0.30 (`Galapagos-2`), and 0.47 (`ProFit`). The
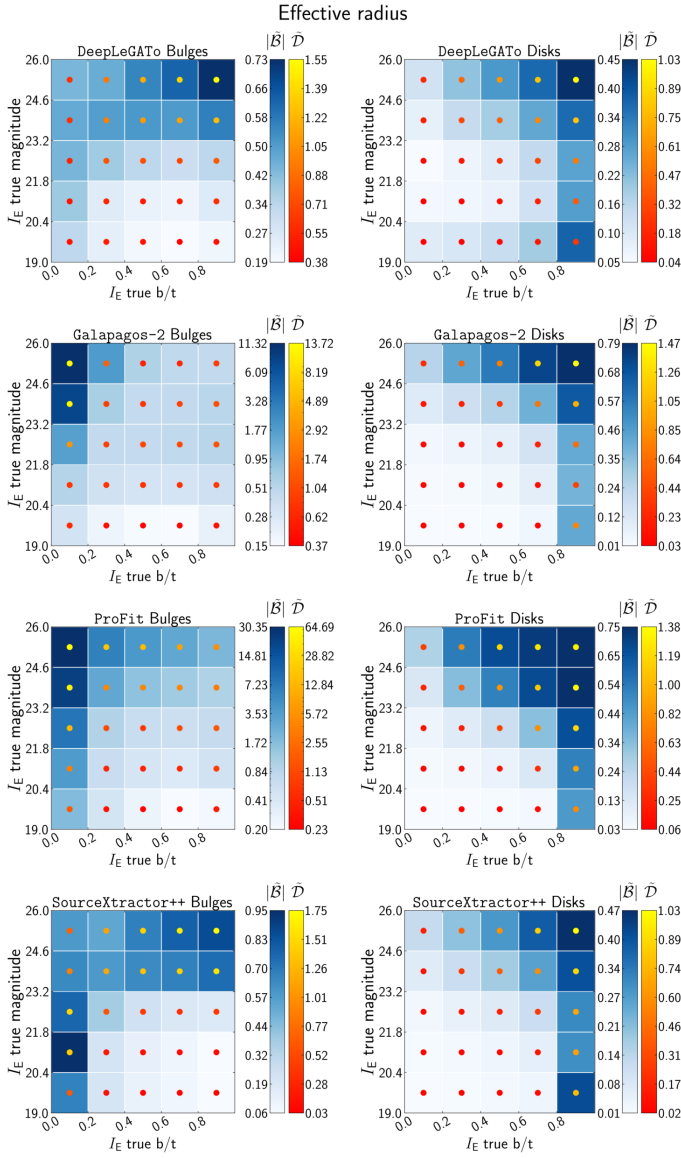
Fig. 9. Absolute bias $|\widetilde{\mathcal{B}}|$ and dispersion $\widetilde{\mathcal{D}}$ for the effective radius of bulge (left column) and disc (right column) components in the double-Sérsic simulation, as a function of bulge-to-total ratio (*x*-axis) and apparent I_E magnitude (*y*-axis). Each row shows a different code. For ProFit and Galapagos-2, the colour scale is logarithmic. In each panel, the colour of the squares is proportional to the mean bias $\mathcal{D}$ (lighter being smaller), and the colour of the dot inside each square indicates the dispersion $\mathcal{D}$ (redder being lower). For most of the codes, we find the expected behaviour: both the bias and the dispersion increase for faint objects, as well as at small b/t for bulges and large b/t for discs.

outlier fraction is less linear but with similar ranking. ProFit has the lowest fraction of bright outliers, but is the highest in the faint bins. For the faint bin, DeepLeGATo has the smallest fraction, followed by SourceXtractor++. The fractions in the last bins are nevertheless higher than for the single-Sérsic fit, with respectively, 5%, 10%, 29%, and 30% for DeepLeGATo, SourceXtractor++, Galapagos-2, and ProFit.

4.2.4. Bulge axis ratio

Figure 11 presents the accuracy in the estimation of the axis ratio q of the bulge components. The characteristic trumpet shape is no longer preserved, and distributions tend to
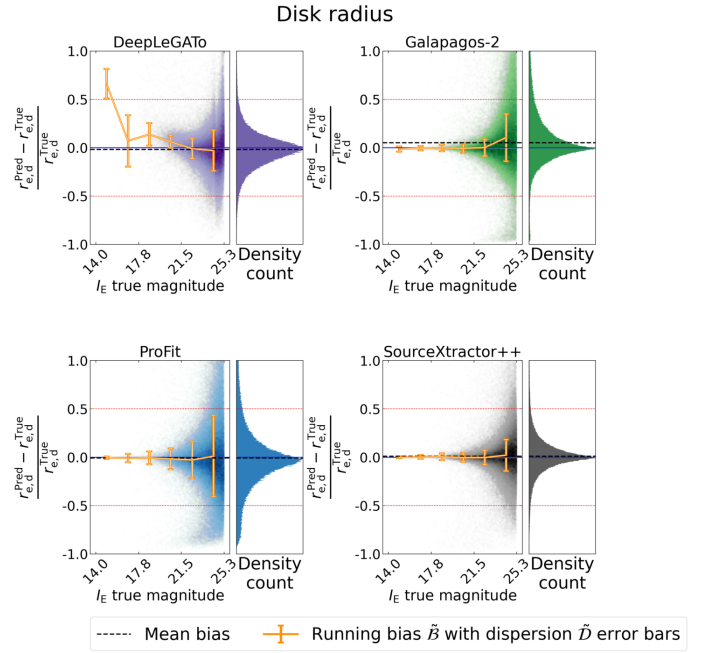
Fig. 10. Fitting results for the disc radius using the double-Sérsic simulation. See caption of Fig. 3 for further information.

be flatter, especially for the faint objects. These results are quantified in the fourth row of Fig. 13. SourceXtractor++, ProFit, and DeepLeGATo maintain an absolute bias smaller than 0.1 for $17 < I_E < 26$. SourceXtractor++ has close to no bias, while ProFit has a tendency to under-estimate the bulges q, i.e predicting galaxies that are too elongated. It is the opposite for DeepLeGATo, which over-estimates q, especially for the brightest galaxies. Galapagos-2 is well calibrated for $I_E < 19$, and then starts to under-estimate q, with a negative bias down to $\mathcal{B} = -0.42$ on the faintest galaxies. For the dispersion $\mathcal{D}$, DeepLeGATo and SourceXtractor++ achieve the lowest values for faint objects, around 0.25. ProFit and Galapagos-2 have a strong increase for $I_E > 20$, up to 0.5 and 1, respectively. For brighter objects, all codes except DeepLeGATo achieve comparable results. Finally, DeepLeGATo and SourceXtractor++ achieve a very low outlier fraction only with few percent. For $I_E < 20$, the three codes (excluding DeepLeGATo) behave similarly, but then O rises for ProFit and Galapagos-2 for $I_E < 22.5$, and ends at 0.3 for ProFit and 0.42 for Galapagos-2. DeepLeGATo's fraction of outliers ranges from approximately 100% (bright) to 1% (faint).

We also investigated the 2D distributions of the metrics as a function of magnitude and b/t, in the same way as we did for the radius in Fig. 9. We found that removing cases with extreme b/t significantly improves the results at all magnitudes. We let the interested readers explore this behaviour with the online tool.

4.2.5. Disc axis ratio

In general, software packages were able to measure the axis ratio q of the disc components (Fig. 12) more accurately than for the bulges. They are comparable to results from the single-Sérsic case, albeit with a higher dispersion and a larger negative bias for faint objects for Galapagos-2 which tends to under-estimate q.

We can make a more in-depth comparison of the metrics by looking at the last row of Fig. 13. Their general behaviour is comparable to the bulge axis ratio, but with smaller val-
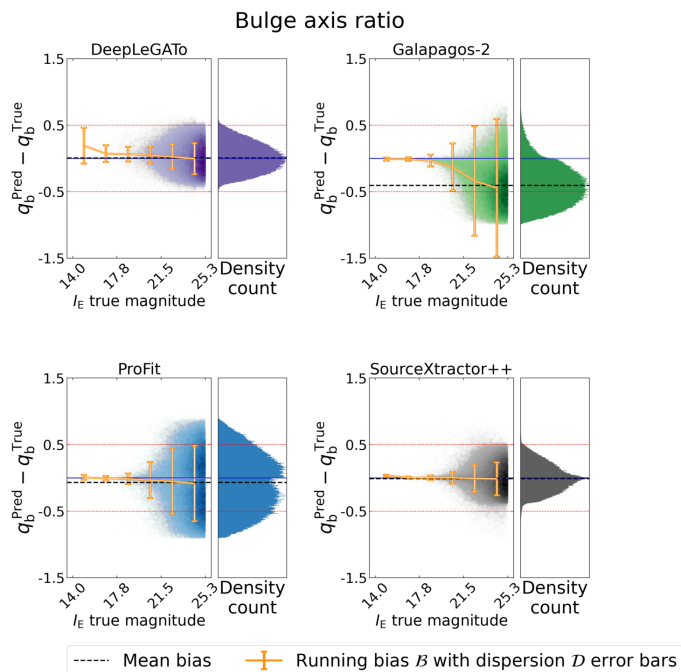
Fig. 11. Fitting results for the bulge axis ratio using the double-Sérsic simulation. See caption of Fig. 3 for further information.

Fig. 12. Fitting results for the disc axis ratio using the double-Sérsic simulation. See caption of Fig. 3 for further information.

ues. The absolute bias remains smaller than 0.3 for all codes for $I_E < 23$ (apart from `DeepLeGATo` which is again unreliable for bright objects). `Galapagos-2`'s and `ProFit`'s biases are well calibrated for $I_E < 22.5$, but then decline to a value of -0.2. For the faintest bins, `SourceXtractor++` has a slight tendency to under-estimate the axis ratio, while `ProFit` and `DeepLeGATo` over-estimate it. For the dispersion, `Galapagos-2` is also the best calibrated for $I_E < 21$, but increases up to 0.45 for the faintest bins, while `ProFit` increases to 0.23, `SourceXtractor++` to 0.18, and `DeepLeGATo` to 0.15. `DeepLeGATo` again starts to be comparable to other codes for $I_E > 20$, and improves to achieve the smallest dispersion for faint objects. `Galapagos-2` is also the best calibrated for $I_E \lesssim 18$ for the fraction of outliers, with less than 4% of outliers, and up to 12% in the faintest bins. It is still the second lowest for intermediate bins, followed by `SourceXtractor++` and `ProFit` by a few percent. From magnitudes 18 to 26, `DeepLeGATo` achieves the lowest number of bad fits, below 3%, followed by a fraction of percent by `SourceXtractor++` in the faintest bins. At intermediate magnitudes ($I_E \simeq 17$), `DeepLeGATo` and `ProFit` increase to around 5%, and `SourceXtractor++` to 7%.

4.2.6. Multi-band fits

Galaxies change appearance with varying wavelengths (Kelvin et al. 2012; Vulcani et al. 2014; Kennedy et al. 2015). As a result, the chosen waveband may influence the classification and the determination of a galaxy structural parameters (see e.g., Häußler et al. 2022 for a detailed discussion). As discussed in the introduction, in addition to the VIS images, which deliver the highest spatial resolution, *Euclid* will also provide NIR images in three filters. In addition, a variety of ground-based surveys such at the LSST will overlap with the *Euclid* footprint. While the main focus of the Euclid Morphology Challenge is on VIS, we included the option to test the capability of software packages to
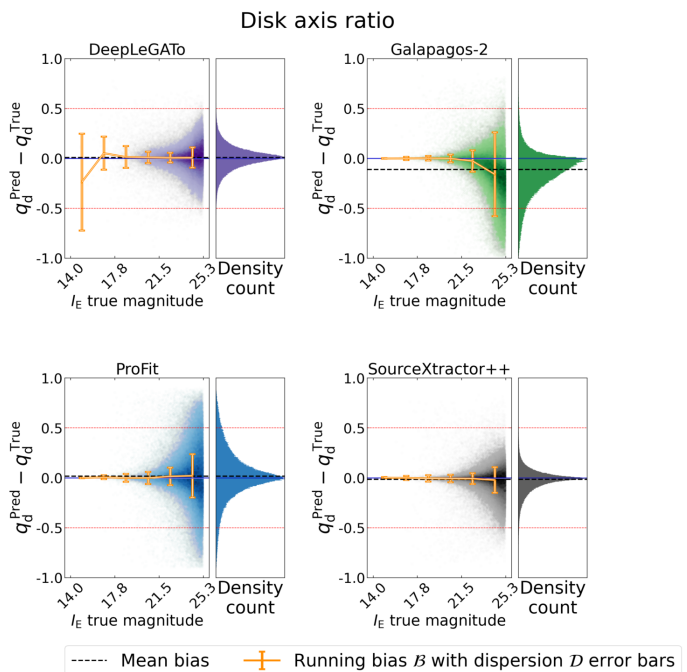
fit images in multiple wavelength ranges. The multi-wavelength simulations we provided are rather simplistic, with only the total magnitude and the bulge-to-total ratio b/t changing with wavelength. While the first is extensively analysed in EMC2022a, we focus here on the results for b/t. We expect that b/t is best recovered in VIS and challenging in other bands due to their lower SNR, lower resolution, noise correlation, and artifacts related to the re-sampling. However, it is interesting to check whether the constraints provided by VIS help to improve the morphology estimated in the lower resolution images.

We received multi-band fitting measurements from `Galapagos-2`, `ProFit`, and `SourceXtractor++`. Not every team interpreted the task to provide multi-band fitting in the same way and thus methods and decisions vary from code to code. The `Galapagos-2` team ran all the bands simultaneously to produce the different parameters. In their bulge-disc decompositions, they fixed all the parameters apart from the magnitude, for which complete freedom to vary with wavelength was ensured. The b/t we compare in Fig. 14 is constructed from these magnitude outputs. Note that we only show results as a function of I_E magnitudes, which is the deepest image by far. The strength of codes like `Galapagos-2` lies in improvements for shallower data, like the NIR images. These can be explored in the online tool. `SourceXtractor++` also fitted all the bands in a joint analysis, with the exception of b/t, which `SourceXtractor++` provides directly. This means that the b/t parameter was fit independently in each band and the overall model amplitude could scale freely. `ProFit` fitted all bands independently, and thus galaxies can have different structural parameters in the different bands. This choice disadvantages the fitting process in the faint or low SNR bands (filters with narrow pass-bands). It did however give us a good indication that $\mathcal{B}$, $\mathcal{D}$, and $\mathcal{O}$ increase for all morphological parameters that we probe, from I_E to NIR y band, typically from a few percent in bright galaxies to 10 and more percent in faint galaxies. Note that `ProFit` has the option for a multi-band joint analysis, but this mode was not used for the challenge.
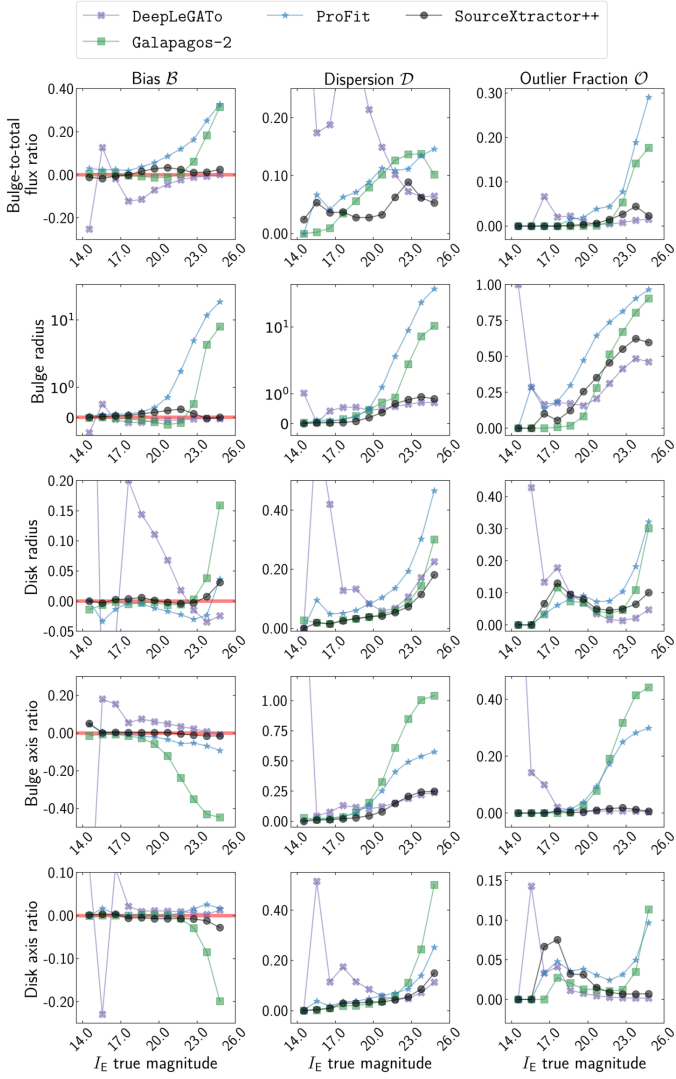
Fig. 13. Summary plot for the double-Sérsic simulations. From top to bottom: bulge effective radius, disc effective radius, bulge axis ratio, disc axis ratio, and bulge over total flux ratio. See caption of Fig. 6 for further information.

Figure 14 summarises the results of fitting b/t across the nine bands, roughly arranged by wavelength[7] (four *Euclid* filters highlighted by blue shading and five Rubin filters highlighted by red shading), and for three classes of galaxies: bright $(14 \leq I_E \leq 20$, top row); intermediate $(20 \leq I_E \leq 23$, middle raw); and faint $(23 \leq I_E \leq 26$, bottom row). The figure uses successful results from galaxies in the one multi-band field, combined in an overlapping catalogue of around 70 700 objects. Bias $\mathcal{B}$, dispersion $\mathcal{D}$ and outlier fraction O increase in the five Rubin compared to the *Euclid* bands. This is expected, since they are narrower and thus have less throughput and being a ground based telescope, Rubin's PSF is larger than *Euclid*'s PSF ($0''7$ vs. $0''16$), which leads to a lower resolution. Second, looking at the brighter galaxies, we see that the difference between *Euclid* and Rubin bands is larger for `ProFit` than for the two other software packages. This is a direct indication of the benefit of fitting images simultaneously. A multi-band approach increases the SNR of measurements in faint bands (e.g., Häußler et al. 2013). On the contrary, for `Galapagos-2` and `SourceXtractor++`, we

---

[7] Recall that I_E is very wide, $550 - 900$ nm, essentially combining Rubin's *r, i,* and *z* bands. See Fig. 1 of EMC2022a for more information.
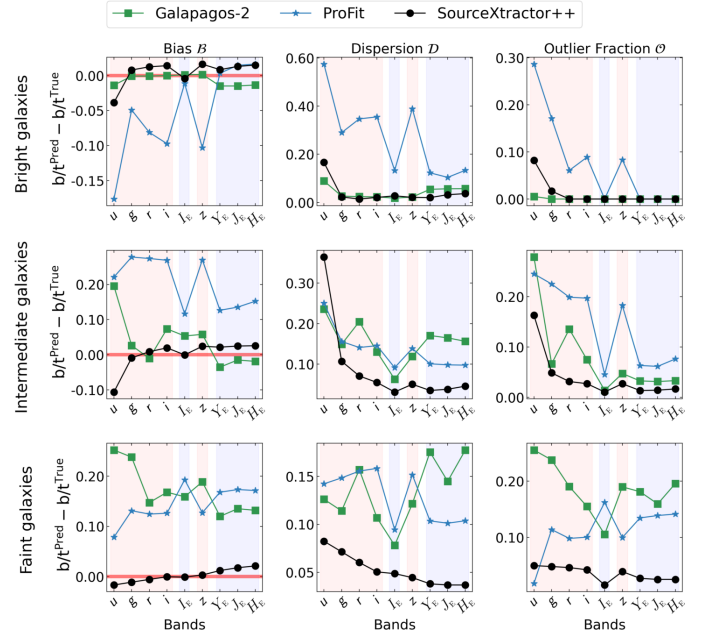

Fig. 14. Results for the fitting of b/t in double-Sérsic multi-band data. The three lines represent the results for three different selections of galaxies. From top to bottom: bright galaxies $(14 < I_E < 20)$; intermediate galaxies $(20 < I_E < 23)$; and faint galaxies $(23 < I_E < 26)$. The three columns represents our three metrics, the bias $\mathcal{B}$, the dispersion $\mathcal{D}$, and the outlier fraction O, which are plotted on the y-axis of the corresponding columns. The x−axis represents the different bands, ordered by increasing wavelength (I_E overlaps with *g, r,* and *i*). The background colours show the *Euclid* bands in blue and the Rubin bands in red.

can see that there is close to no effect of the band width on the three metrics of the bright objects. The effect stays low even for intermediate and faint galaxies (second and third rows). An interesting exception is the extremely faint Rubin *u* band, where results are especially unreliable. This result relates back to the dominance of I_E, which is by far the widest, and thus offers the most information from its deeper image.

In general, results are comparable to the analysis of single-band I_E b/t measurements (see Sect. 4.2.1 and Fig. 13). Differences between `Galapagos-2` and `SourceXtractor++` become apparent for faint galaxies (bottom row), where `SourceXtractor++` results are consistently more stable across all bands. The difference again highlights the importance of the prior or initial guess, especially for faint galaxies. These results suggest that the simultaneous analysis of *Euclid* and Rubin will improve the accuracy of measurements in the narrow bands, mainly by exploiting information from the deep I_E band. It also highlights the important synergies between the two experiments (Guy et al. 2022).

4.2.7. Free versus fixed bulge component

We additionally asked participants to perform the bulge-disc decomposition with two different settings. In the first, participants kept the Sérsic index of the bulge fixed to $n = 4$, which is also the value that was used in the simulations. In the second setting, the Sérsic index was set as a free parameter. A comparison is provided in Fig. 15, where dashed lines indicate the case of a free bulge Sérsic index and solid lines when the bulge is fixed. The disc effective radius and axis ratio measurements (right column) deteriorate in all codes when the bulge Sérsic index is left free,

from a few percent to 10% in extreme cases (fainter objects). The effect is weakest for `SourceXtractor`++, which seems to be less sensitive to the change of model. As discussed before, the favourable results could come from their choice of prior for n. We notice the most dramatic effect of switching from free to fixed bulge Sérsic index in `ProFit`, which increases the dispersion of b/t measurements by more than 10% for bright objects. Changes for the bulge component are less clear, with some codes achieving better results with n free and others worse. For instance, `Galapagos-2` has a higher bias when inferring the bulge effective radius with a free bulge Sérsic index, while `ProFit` obtains a better result for bias and dispersion, but much higher outlier fraction for bright objects. Finally, the change in the bulge components for `SourceXtractor`++ is very small, which could confirm the interpretation about priors, for which we saw in the previous section that it is stronger for bulge components. In some cases, we observe a significant difference between the free and fixed Sérsic index models especially for the very bright objects. We cannot find a straightforward explanation for this behaviour, but we highlight again that this bin represents only a handful of objects. Finally, the last row of Fig. 15 illustrates how well a fixed Sérsic index $n = 4$ is recovered when the parameter is left free in the model. We observe the same type of behaviour as for the other bulge-component parameters: `SourceXtractor`++ achieves the lowest bias and dispersion, driven by an appropriate prior selection, which is in this case perfectly known. `Galapagos-2` achieves comparable results for the very bright objects, but then tends to over-estimate n for faint galaxies, i.e. giving too steep bulge profiles. The dispersion also gets higher, as for `ProFit`. The effect on the bias is the opposite for `ProFit`, which always under-estimates n.

4.2.8. Global scores

Finally, Table 2 presents the scores $\mathcal{S}$ for the double-Sérsic simulations, fitted with a fixed (in red) and a free (in black) bulge Sérsic index. The score reflects and summarises the analysis previously discussed. First, the average score for the bulge component parameters (first number of the last column) is much larger than the values obtained for the disc parameters. It is indeed penalised by the large values of bulge radius fitting, which range from 2.42 to 89.37. For the disc components, we see that the average score is much lower than for bulges, with values smaller than 1.2 and hence comparable to the single-Sérsic simulation. `DeepLeGATo` again scores lowest for the two components, since it maintains accuracy for faint objects, which are the majority of the dataset. This outweighs its poor performance at the bright end. On the contrary, software packages that achieve better results for the bright objects are penalised by their lower performance at the faint end. The table also summarises differences between free and fixed Sérsic index, as discussed in the previous section.

4.3. Realistic simulation results

We call 'realistic' simulations those based on an approach using deep neural networks (Sect. 2.3). Figure 19 summarises the results, along with Table 1 and Sect. 4.3.4.

These are therefore inherently different, and by design more closely resemble real galaxies. In our analysis, each galaxy is characterised by the three parameters effective radius r_e, axis ratio q, and Sérsic index n, comparable to the description of single-Sérsic simulations, but with more complex morphologies.
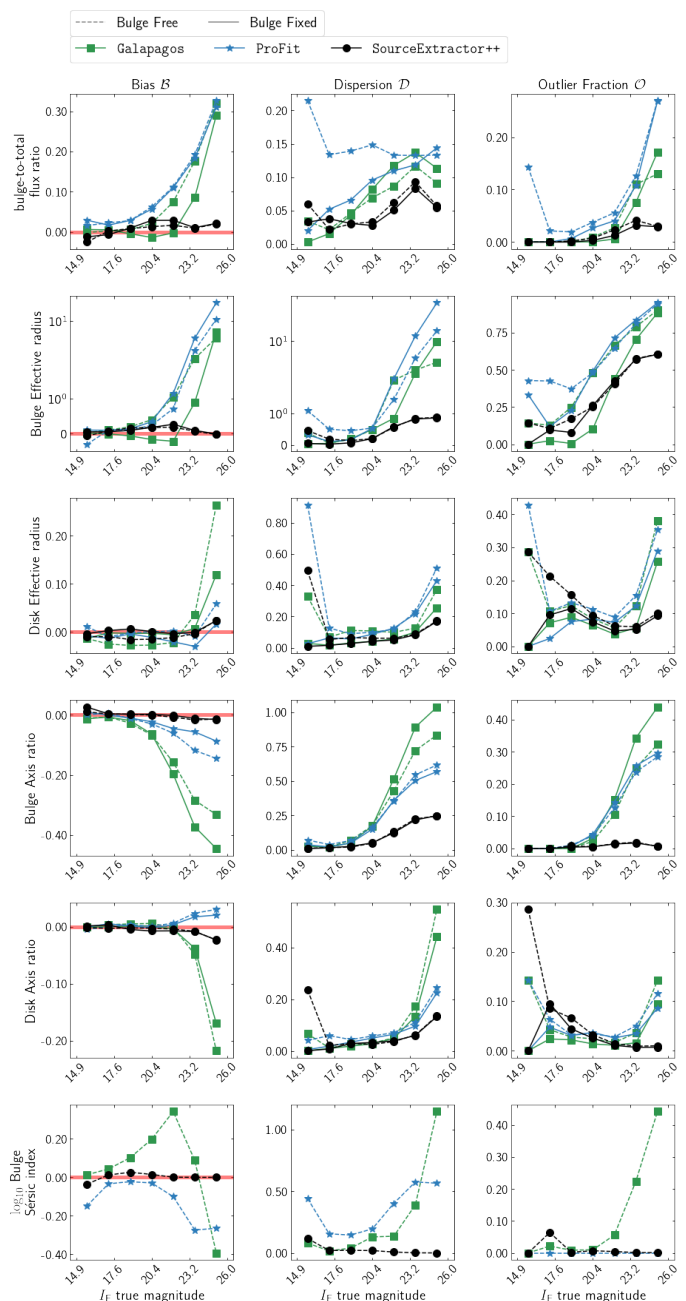

Fig. 15. Comparison of the fitting between the fixed (solid line) and free (dashed line) bulge Sérsic index models. We can see that overall, letting n free worsens the results according to our metrics for the disc component parameters and b/t, while it is less clear for the bulge parameters. In addition to the usual double-Sérsic parameters, we show in the last row the fitting of the bulge Sérsic index for the "bulge free" model.

We also notice that, as explained earlier in Sect. 2.3, the limited control of the simulation input parameter can induce a bias in the true parameters. A comparison with the results of the Sérsic simulations is therefore not straightforward, but it offers valuable insights into what impact complex structures have on galaxy profile fitting. Notice that only `Galapagos-2`, `ProFit`, and `SourceXtractor`++ provided predictions for the realistic simulation, which is why figures are limited to three panels. To the best of our knowledge, a comparison study on the same set of realistic galaxy morphologies has not been conducted before.

Table 2. Comparison of the scores $\mathcal{S}$ obtained by the different codes in all structural parameters for the double Sérsic simulation (with a fixed bulge Sérsic index fit in red, and with with a free bulge Sérsic index fit in black). The last column is the mean of the parameters, first for the bulge components and b/t, second for the disk components and b/t. `DeepLeGATo` is not a model-fitting algorithm, and thus has no fixed or free bulge Sérsic index modes. A smaller $\mathcal{S}$ means a better fit.

Fix bulge fit / Free bulge fit	$\mathcal{S}_{r_{e,b}}$	$\mathcal{S}_{r_{e,d}}$	$\mathcal{S}_{q_b}$	$\mathcal{S}_{q_d}$	$\mathcal{S}_{b/t}$	μ_b, μ_d
`DeepLeGATo`	2.42 / ∅	0.49 / ∅	0.50 / ∅	0.21 / ∅	0.23 / ∅	1.05 ; 0.31 / ∅
`Galapagos-2`	23.13 / 17.45	0.91 / 1.46	3.50 / 2.74	0.51 / 1.29	0.95 / 0.99	9.19 ; 0.96 / 7.06 ; 1.26
`Profit`	89.37 / 37.65	1.14 / 1.37	1.76 / 1.92	0.51 / 0.60	1.17 / 1.22	30.76 ; 0.93 / 13.6 ; 1.06
`SourceXtractor++`	3.00 / 3.00	0.48 / 0.51	0.53 / 0.54	0.29 / 0.29	0.26 / 0.27	1.26 ; 0.34 / 1.26 ; 0.36

4.3.1. Half-light radius

Figure 16 shows the scatter plot for the half-light radius based on the deep learning simulations. The overall performance is degraded compared to the single-Sérsic simulation, meaning that the distributions are wider and the measurements are significantly biased, especially at the faint end. However, these factors are difficult to disentangle. A decreasing accuracy is expected because the Sérsic model is less suited for complex surface-brightness profiles compared to the analytic simulations. However, the process of generating these simulations is not free of biases either. As explained in Sect. 2, galaxies in this case have been generated following a data-driven approach trained on HST observations. The mapping between structural parameters and galaxy images is also learned empirically and it is therefore not perfect. Interestingly, all codes behave similarly, i.e., there is an under-estimation of the radius for faint objects. This might be an indication that the degradation could be partly explained by the fact that the input is also biased. If this is the case, then we can expect performance for all metrics on real *Euclid* observations to be between the analytic and the realistic results.

The top row of Fig. 19 (note that because bright galaxies were not simulated for this type of simulation, we reduce the *x*-axis range from $I_E = 20$ to $I_E = 26$) shows that the bias is similar for the three codes for $I_E < 23$, but remains more stable for `Galapagos-2` ($\mathcal{B} = 14$), followed by `ProFit` ($\mathcal{B} = 19$) and `SourceXtractor++` ($\mathcal{B} = 24$). The dispersion ranges from around 0.3 (0.39 for `ProFit`) at the bright end to 0.6 at the faint end, with few percent of differences. This is a factor of approximately 3 increase compared to the analytic Sérsic simulation. The fraction of outliers also increases to 15% (`SourceXtractor++`), 23% (`ProFit`), and 34% (`Galapagos-2`), which is slightly higher compared to the single-Sérsic case. For the three codes, $\mathcal{O}$ follows a U-shape, increasing at the bright and faint ends. The increase for bright objects might again reflect a simulation bias since the model was trained with a small number of very bright galaxies, but also the fact that brighter galaxies have more structures, and thus a larger departure from a Sérsic profile.
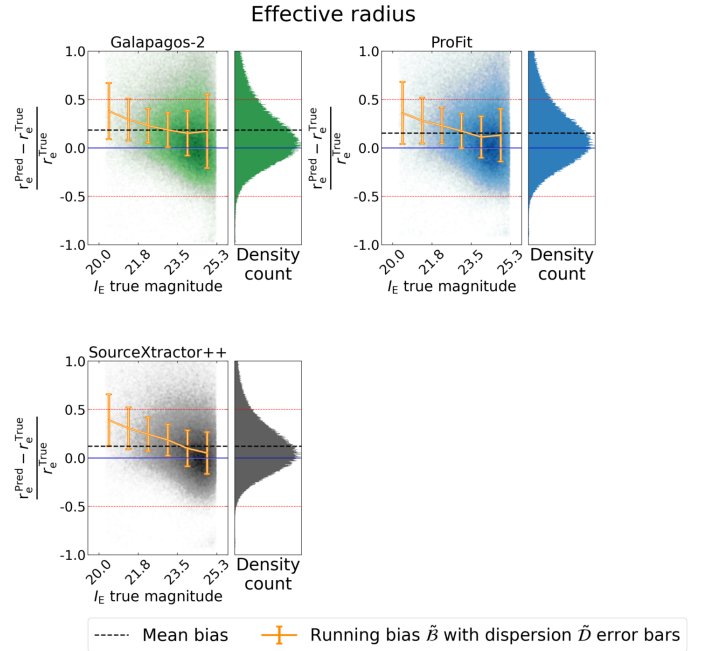

Fig. 16. Fitting results for the effective radius of the realistic simulation. Notice that only three codes provided results for the double-Sérsic simulation. From top to bottom and from left to right: `Galapagos-2`, `ProFit`, `SourceXtractor++`. See caption of Fig. 3 for further information.

4.3.2. Axis ratio

Figure 17 shows the results for the galaxy axis ratio *q*. The first impression is similar to that for the half-light radius: the wings of the trumpet shape are wider compared to the single-Sérsic case; however, the bias is no longer an issue as it was for the radius. The three codes are overall very consistent with each other.

Looking at the second row of Fig. 19, we see a small bias for all codes, from less than a 0.01 for `Galapagos-2` (0.02 for `SourceXtractor++` and 0.035 for `ProFit`) up to 0.04 for `ProFit` and `SourceXtractor++` for faint objects. As for the single-Sérsic case, the bias for `Galapagos-2` and `ProFit` decreases in the fainter magnitude bins, down to −0.04 for `Galapagos-2`, but are low in general ($|\mathcal{B}| < 0.04$). This might
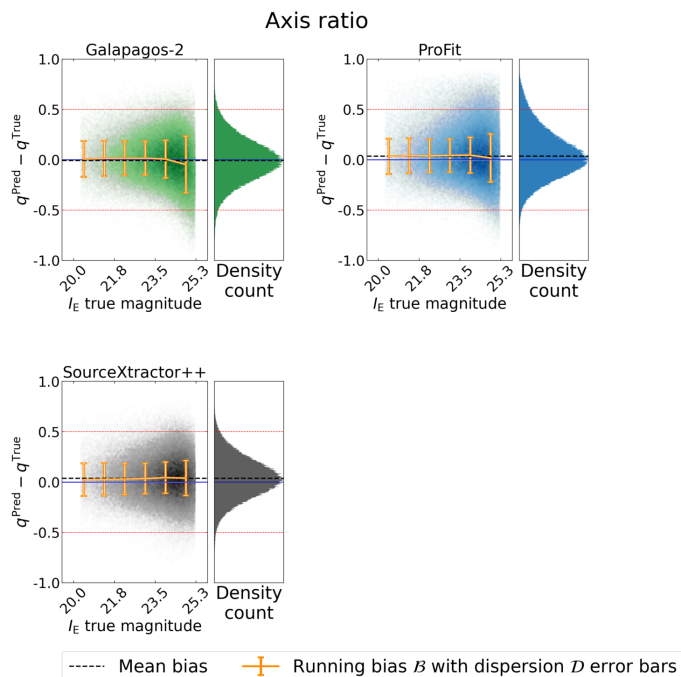
Fig. 17. Fitting results for the axis ratio of the Realistic simulation. See caption of Fig. 3 for further information.

Fig. 18. Fitting results for the Sérsic index of the realistic simulation. See caption of Fig. 3 for further information.

be an indication that the input simulations are less affected by systematics for this structural parameter in particular. The overall ellipticity is indeed less dependent on the galaxy surface-brightness profile. The dispersions are somewhat higher than for the single-Sérsic case, but stable for `SourceXtractor++` around 0.16, and from 0.17 up to 0.28 for the other two software packages. For the outliers, `SourceXtractor++` and `Galapagos-2` achieve similar results, with a fraction around 2% (4% for `Galapagos-2` at the fainter end), while `ProFit` seems more affected by the features in the simulation, with 4.5% to 7% of outliers, which is approximately 10 times higher than for the single-Sérsic case (compared to 5 times larger for the other software codes).

4.3.3. Sérsic index

We finally focus on the estimation of the Sérsic index n. As explained in Sect. 4.1.3, we analyse $\log_{10}(n)$ instead of n. Figure 18 is very similar to that of the axis ratio: the trumpet shape is wider than for the single-Sérsic case, but with no major bias. In the last row of Fig. 19, we see that indeed, `SourceXtractor++` has close to no bias for the all ranges of magnitude, while `ProFit` and `Galapagos-2` retain a low bias (< 0.05), with an increase up to 0.14 for `Galapagos-2` in the faintest magnitude bin. The dispersion of `Galapagos-2` and `ProFit` are very similar (and the lowest, compared to `SourceXtractor++`) around 0.18, here again until the faintest magnitudes. In the last bin, `SourceXtractor++` achieves a reasonable bias (0.25), while `ProFit` goes up to 0.30 and `Galapagos-2` to 0.43. The dispersion and biases are roughly 2 times larger than in the single-Sérsic case. The same behaviour can be seen in the outlier fraction, with a small fraction (5%) for $I_{\mathrm{E}} < 23$. The outlier fraction of `SourceXtractor++` does not increase significantly for fainter objects, but it $\mathcal{O}$ reaches 20% in `ProFit`, and 40% in `Galapagos-2`, which is, for all codes, much larger than the out-
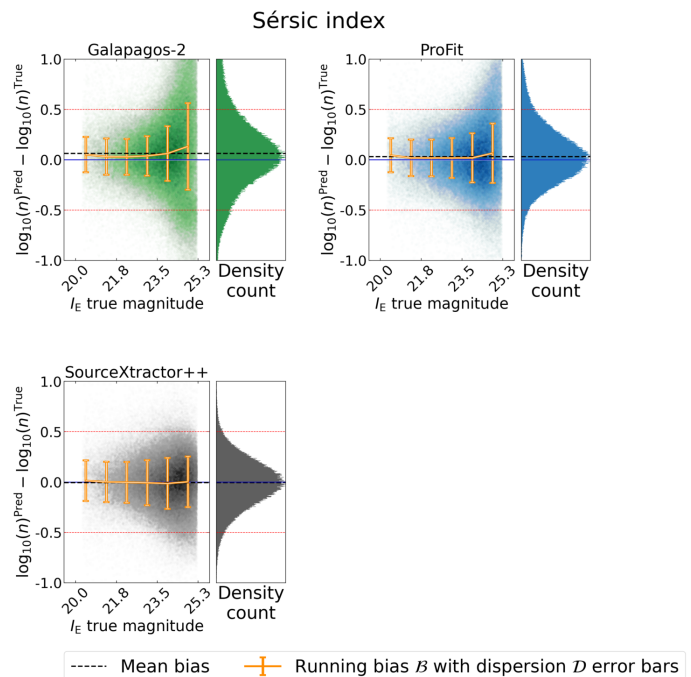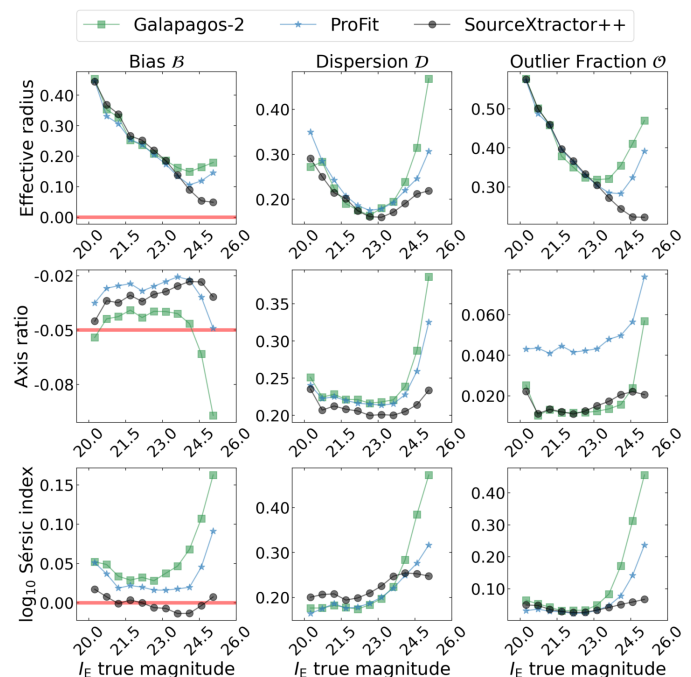

Fig. 19. Summary plot for the realistic simulation. From top to bottom: effective radius; axis ratio; and Sérsic index. See caption of Fig. 6 for further information.

lier fractions we reported from simulations created with analytical profiles.

4.3.4. Global scores

Table 1, which we already discussed in Sect. 4.1.4, also shows in black the results for the more realistic simulations. Overall, measurements are less accurate in the deep learning generated images, with a degradation between 2 times to 5 times, based on

our metrics. The most stable parameter compared to the analytic simulation is q, followed by n. The hardest to capture in realistic simulations is the radius r_e. Overall, `SourceXtractor++` and `ProFit` achieve similar results, where `ProFit` benefits from its high completeness, and `Galapagos-2`, though less complete, achieves robust results in more complex morphologies. Average scores μ_S, range from 1.8 (`Galapagos-2`), to 2.5 (`ProFit`), and 3.8 (`SourceXtractor++`).

4.4. Uncertainty quantification

Another goal of this challenge is to review the ability of software packages to predict the uncertainty related to each measure. Each participant was asked to provide the 1σ uncertainty for every parameter and every galaxy, called σ_p. To quantify the quality of this prediction, we tested whether each parameter prediction P_p of each galaxy falls inside the predicted error interval,

$$T_p \in [P_p - \sigma_p, P_p + \sigma_p]. \tag{11}$$

If it does, then we attach a flag '1'. If it does not fall within this interval, then the prediction is flagged '0'. Following a frequentist approach[8], if σ_p is indeed the 1σ uncertainty, then 68% of the galaxies should be flagged 1. `DeepLeGATo` does not feature in this section because the network only provides point estimates and no reliable uncertainty estimation was available in the version that was used to obtain results for this challenge.

4.4.1. Single-Sérsic simulation

Figure 20 assesses how well uncertainties are predicted in each code for the single-Sérsic -simulations per bin of I_E magnitude. The red dashed line indicates the fraction of objects that should be obtained if the uncertainty was well calibrated. We also show the overall calibration of the entire sample, which we present on the same plot on the right, highlighted by blue shading.

It becomes immediately obvious that the uncertainty is always under-estimated for all codes and parameters. `Galapagos-2`, `Morfometryka`, and `SourceXtractor++` under-estimate the uncertainties of their effective radius measurements in similar ways: around 58% of object are well-calibrated (instead of 68%). Size uncertainty estimations are worse for bright objects, which is common for all codes. This suggests that the uncertainty is mostly related to the flux of the objects. This under-estimation is most substantial in `Morfometryka`, which has almost no well-calibrated bright objects, but over-estimates uncertainties in the faint end. This behaviour is similar, but less striking for `SourceXtractor++`. The middle and right panels show uncertainty calibrations for axis ratio and Sérsic index measurements. The performance we just described is roughly comparable in all parameters. Nevertheless, reported uncertainties for the axis ratio seem to be best calibrated by `SourceXtractor++` (the bar is close to 68%). They are also less affected by the magnitude. `ProFit`'s uncertainty estimation of axis ratios is also much better calibrated than for the radius, with a global score of 60%, similar to `Morfometryka`. Overall, `Galapagos-2` has only 45% of its axis ratio measurements well calibrated. This is also the average estimation for uncertainties of the Sérsic index (between 45% and 50% are correctly estimated), worse in bright objects than in faint objects.

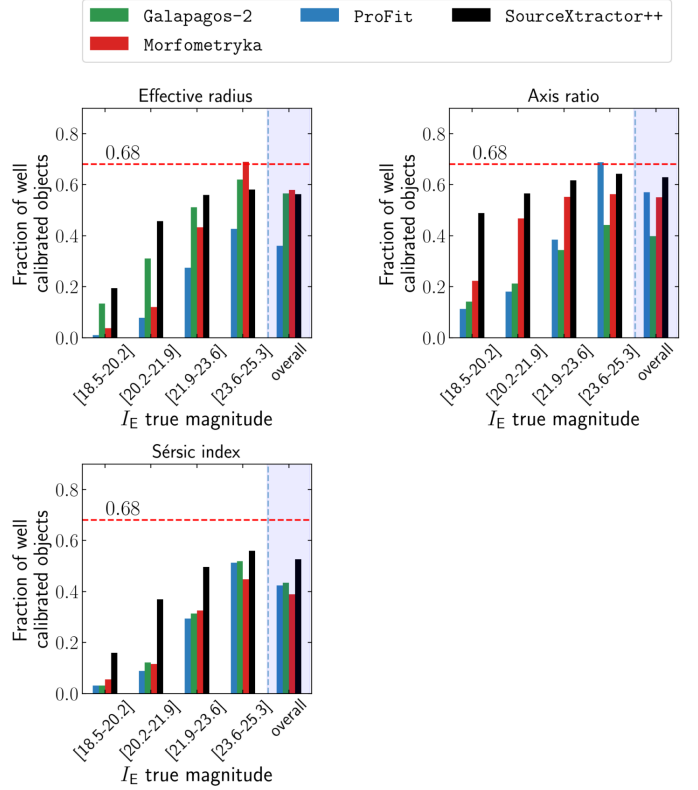[8] We only have one prediction for each object, thus we are not able to perform a proper Bayesian analysis.

Fig. 20. Uncertainty calibration for the single-Sérsic simulation. The *x*-axis represents four bins of I_E magnitude. The *y*-axis shows the fraction of object per bin of magnitude for which the True value of a parameter falls in an interval of the predicted 1σ uncertainty centred on the predicted value. The final bin is for all objects, regardless of their magnitude.
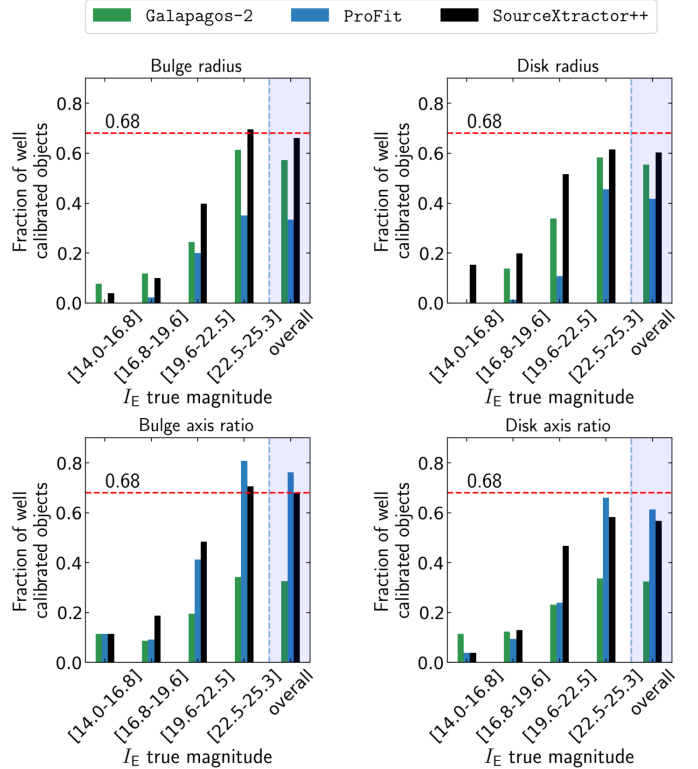

Fig. 21. Uncertainty calibration for the double-Sérsic simulation. See caption of Fig. 20 for further information.

4.4.2. Double-Sérsic simulation

In Fig. 21 we evaluate uncertainty estimations of measurements from the double-Sérsic simulations. We collect and compare uncertainty estimates for three codes `SourceXtractor`++, `Galapagos-2`, and `ProFit`. While we see similar trends to single-Sérsic estimates–better calibrations for faint than for bright objects, we also notice some improvements. e.g. `SourceXtractor`++ is overall well calibrated for the bulge radius and for the bulge q. Importantly though, the all codes still under-estimate their uncertainties, especially in bright objects. For the disc q, `ProFit` is better calibrated, with a score of 60%, while `Galapagos-2` only scores 30%, and the same for the bulge q. Some of these may be alleviated by employing `Galapagos-2`-specific flags, which were not used in this analysis but introduced in EMC2022a, and investigated in detail in Häußler et al. (2022). For the bulge q, `ProFit` over-estimates the uncertainty. For the disc radius, all the codes are under calibrated, with a score of about 60% for `SourceXtractor`++ and `Galapagos-2`, and 40% for `ProFit`.

5. Summary and conclusions

The Euclid Surveys will become a reference dataset for studies involving galaxy structure and morphology due to the unique combination of wide area and high spatial resolution. In this paper, we have described the results of the Euclid Morphology Challenge to quantify the performance of five state-of-the-art surface-brightness-fitting software packages on simulated *Euclid* imaging data. This paper is the second part of the two paper discussing the EMC. While the companion paper (Euclid Collaboration: Merlin et al. 2022) focuses on results related to the photometry, we have focused on the results concerning the morphological parameters only. We compare the results after measuring structural properties of simulated galaxies in fields that mimic *Euclid* observations with the VIS instrument from `DeepLeGATo`, `Galapagos-2`, `Morfometryka`, `ProFit`, and `SourceXtractor`++. The simulations we use include single- and double-component Sérsic profiles, as well as more realistic data-driven generated galaxies. In addition, one field was provided in multiple bands that included *Euclid* NIR Y_E, J_E, and H_E filters and ancillary data from the five optical Rubin bands *u, g, r, i,* and *z,* to test possible multi-band fitting routines.

Figure 22 visually summarises some of the main results of this work. This figure shows one example of a bright, intermediate, and a faint galaxy as fitted by each software package and for each type of simulation. We also show the residual images obtained after subtracting the best fit model from the original image. The figure illustrates several of the key trends raised in this work.

Single-Sérsic simulations (first three rows) yield the best results, with the lowest residuals, always remaining below 4σ of the noise level. All codes show very similar results, confirming that one component fits are robust down to an SNR of 5 for simple simulations.

Bulge-disk decompositions (middle three rows) are more challenging to perform, as evident from the residuals of subtraction, with clear features remaining. However, the results vary for each code. `DeepLeGATo` and `SourceXtractor`++ in particular appear to show smaller residuals at the faint end. We argue that this result is mainly due to assuming a prior distribution that is close to the true distribution, as opposed to the other codes.

The fit at low SNR is therefore highly dependent on the assumed prior, reflecting the limited constraining power of the data.

The fitting of realistic simulations (last three rows) results in systematic uncertainties that arise from using simple Sérsic models to describe the surface brightness distribution of galaxies at high spatial resolution. As a consequence, the residuals disclose the more complex structures that the basic model is unable to capture. For example, the brightest example galaxy (first row of the block), reveals a complex morphology with spiral arms and clumps. `Galapagos-2` and `SourceXtractor`++ tend to fit the bright bulge of the galaxy, but not the disc. However, in order to capture the global flux distribution of the galaxy, the bulge flux is over-estimated, leading to a very strong positive residual in the bulge, and a negative residual in the disc. `ProFit` fits a galaxy with a lower Sérsic index, which leads to a lower residual, where only the non-linear complex features of the galaxy remains. For the intermediate and faint galaxies (two last rows), we see here again some complex structures in the residual, which are very similar in all the codes. To our knowledge, this is the first time that several fitting codes have been tested on the same non-analytic simulations.

Figure 23 summarises the main trends through the global score $\mathcal{S}$ for every parameter, as well as for each code, and type of simulation. We emphasise that while this score is informative, it is not intended as a universal ranking of the different codes. For this reason, we provide an online interactive tool along with this publication, which allows the user to plot different quantities to accommodate and support choices of each unique science case.

We summarize below the main conclusions of our analysis below.

- Single-Sérsic parameters (r_e, q, n) of *Euclid* VIS galaxies can be estimated with a bias and a dispersion lower than 10% down to a SNR per pixel of approximately 1. This will include imaging for 400 million objects by the end of the mission. Despite some differences between the different software codes tools, they all achieve consistent results.
- We find that bulge-disc decompositions are often unstable. In particular the properties of internal structural components (i.e. bulges and discs) can only be recovered in a reliable way ($\mathcal{B} < 0.1$, $\mathcal{D} < 0.1$) when they dominate light distributions (b/t > 0.8 for bulges and b/t < 0.3 for discs), even for bright objects. Interestingly, we find that the bulge-to-total ratio is well recovered with less than a 10% dispersion, especially by `DeepLeGATo` and `SourceXtractor`++. We argue that this is related to the priors assumed by these two codes, which are closer to the true distribution.
- The different software packages are also tested on galaxies generated with more realistic morphologies. We find the the overall performance degrades by a factor between 2 and 5 on average, depending on the structural parameter in all the considered metrics, i.e., bias, dispersion and number of outliers. We discuss, however, that this factor might be enhanced by some intrinsic biases in the simulations.
- All codes tend to underestimate the uncertainties in the measurement by a factor of around 2. This is even true for Bayesian codes like `ProFit`.
- A simultaneous fit of *Euclid* and Rubin-like images results in an improvement of the structural parameters obtained from ground based data, highlighting the significant synergies between the two projects.
- Deriving good initial values or priors is important for all fitting parameters, especially during a bulge/disc decomposition (e.g., Lange et al. 2016) where the constraining power of data is limited.

Fig. 22. Illustration of residuals from the galaxy modeled with the predicted parameters by the different software packages. The plot has the same structure for the three types of simulations (single-Sérsic in three first rows, double-Sérsic in the three middle rows, and realistic galaxies in the three last rows). For each type, the first row presents a bright galaxy, the second an intermediate galaxy, and the third a faint one). In the two first columns, we show the actual image of the galaxy, with and without noise. Then, each pair of columns presents: (1) the noiseless model constructed from the predicted parameters of each code, (2) the residual between the noisy true galaxy and the noise-less model. All the colour bars are related to the Euclid Wide Survey noise level. For this reason, a perfect model would lead to a colour-bar ranging from -3σ and 3σ. Finally, to better capture the details of the residuals, we clipped the value to a certain σ value, which is fixed for all the codes and magnitudes in a certain type of simulation. For the residuals, the color-bar is symmetric, centred on zero.
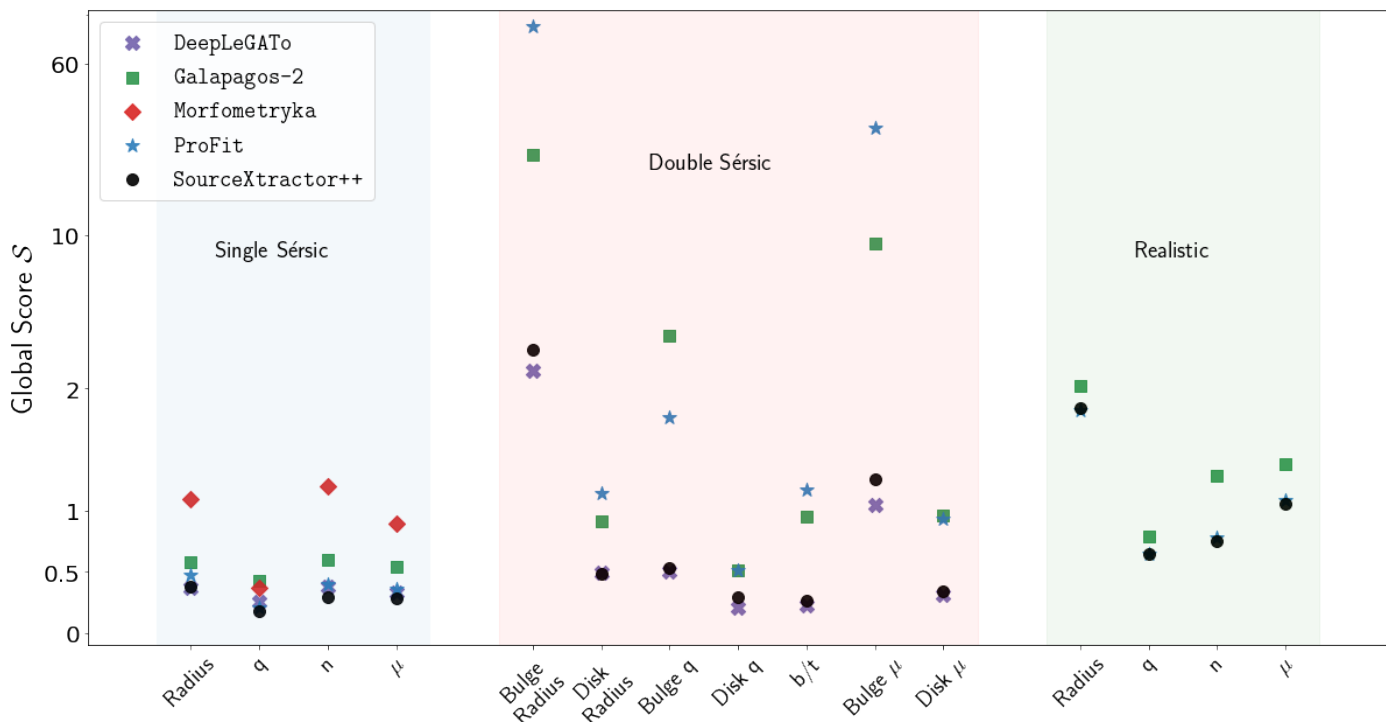
Fig. 23. Summary of the global scores $\mathcal{S}_p$ for the single-Sérsic, double-Sérsic , and realistic simulations. The *x*-axis shows the different parameters, of the different simulations (single-Sérsic, double-Sérsic , and realistic in shaded blue, red and green, respectively). The global score $\mu_\mathcal{S}$ corresponds to the mean of the $\mathcal{S}_p$. The *y*-axis represents the corresponding global score $\mathcal{S}$, for each parameter and code, the lower the score the better.

- An improved model of the PSF that takes into account the under-sampling can play an important role in improving the accuracy of the shape measurements.
- While we do not perform an in-depth analysis of the computational time and resources, we can stress that there are non-negligible differences between the various software packages, with, e.g. `DeepLeGATo` being very fast in GPUs, while Bayesian analysis such as `ProFit` requires longer runs. See the appendix of EMC2022a for a more detailed analysis.

In the following, we review and repeat general conclusions and global results of the EMC, which are also discussed in the companion paper, EMC2022a.

"We did not attempt to investigate the performance of the model-fitting algorithms in the context of the *Euclid* Deep Survey (Laureijs et al. 2011). It is reasonable to expect that the results should be similar to those obtained for the Wide Survey at equivalent S/N, although the increased level of contamination and blending might worsen the performance. Dedicated simulations would be required to quantitatively confirm this hypothesis, but this is beyond the scope of the present study.

The results of the EMC set the baseline to decide which algorithm is to be implemented in the *Euclid* pipeline. One possible way to proceed might be to use the computational power of `DeepLeGATo` to provide fast and reliable priors to `SourceXtractor`++; we will investigate this option in the next future. However, given the many different approaches and techniques adopted by the participants (different use of priors, different pre-processing strategies, different approaches for multiband fitting, etc.) and the other parameters not included in the metric that should be taken into account (such as the computing time, required resources, compatibility with the current pipeline, accuracy of uncertainty budget estimates etc.), it is important to stress that the results presented in this work and in EMC2022[a]

must be interpreted with caution. We followed one possible approach to analyse in a compact way a very complex and multi-layered dataset; and we emphasise that some of the software packages might be better suited than the ones obtaining the best score here, for other specific science cases. We invite the interested reader to check the full set of results using the online tool.

It is worth stressing that the results obtained for the EMC [...] were used by the developers to improve and in some cases significantly upgrade the software packages, which is *per se* a relevant outcome of this work.

Future work will include testing at least some of the software packages in more realistic environments using *Euclid* official simulations, and finally the implementation in the *Euclid* pipeline of the chosen algorithms."

References

Bait, O., Barway, S., & Wadadekar, Y. 2017, MNRAS, 471, 2687

Bertin, E., Schefer, M., Apostolakos, N., et al. 2020, in Astronomical Society of the Pacific Conference Series, Vol. 527, Astronomical Data Analysis Software and Systems XXIX, ed. R. Pizzo, E. R. Deul, J. D. Mol, J. de Plaa, & H. Verkouter, 461

Bland-Hawthorn, J., Vlajić, M., Freeman, K. C., & Draine, B. T. 2005, ApJ, 629, 239

Brennan, R., Pandya, V., Somerville, R. S., et al. 2017, MNRAS, 465, 619

Buitrago, F., Trujillo, I., Conselice, C. J., et al. 2008, ApJ, 687, L61

Buitrago, F., Trujillo, I., Conselice, C. J., & Häußler, B. 2013, MNRAS, 428, 1460

Ciotti, L. 1991, A&A, 249, 99

Cole, S., Lacey, C. G., Baugh, C. M., & Frenk, C. S. 2000, MNRAS, 319, 168

Conselice, C. J. 2003, ApJS, 147, 1

Conselice, C. J., Gallagher, John S., I., & Wyse, R. F. G. 2003, AJ, 125, 66

Cropper, M., Refregier, A., Guttridge, P., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7731, Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave, ed. J. Oschmann, Jacobus M., M. C. Clampin, & H. A. MacEwen, 77311J

de Vaucouleurs, G. 1948, Annales d'Astrophysique, 11, 247

dos Reis, S. N., Buitrago, F., Papaderos, P., et al. 2020, A&A, 634, A11

Euclid Collaboration: Bretonnière, H., Huertas-Company, M., Boucaud, A., et al. 2022, A&A, 657, A90

Euclid Collaboration: Scaramella, R., Amiaux, J., Mellier, Y., et al. 2021, arXiv:2108.01201

Euclid Collaboration: Schirmer, M., Jahnke, K., Seidel, G., et al. 2022, A&A, 662, A92

Falcón-Barroso, J., Lyubenova, M., van de Ven, G., et al. 2017, A&A, 597, A48

Ferrari, F., de Carvalho, R. R., & Trevisan, M. 2015, ApJ, 814, 55

Förster Schreiber, N. M., Genzel, R., Bouché, N., et al. 2009, ApJ, 706, 1364

Freeman, K. C. 1970, ApJ, 160, 811

Graham, A. W. & Guzmán, R. 2003, AJ, 125, 2936

Guy, L. P., Cuillandre, J.-C., Bachelet, E., et al. 2022, in Zenodo id. 5836022, Vol. 58

Häußler, B., Bamford, S. P., Vika, M., et al. 2013, MNRAS, 430, 330

Häußler, B., Vika, M., Bamford, S. P., et al. 2022, arXiv, arXiv:2204.05907

Hubble, E. P. 1926, ApJ, 64, 321

Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., & Sánchez Almeida, J. 2011, A&A, 525, A157

Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971

Kelvin, L. S., Driver, S. P., Robotham, A. S. G., et al. 2012, MNRAS, 421, 1007

Kennedy, R., Bamford, S. P., Baldry, I., et al. 2015, MNRAS, 454, 806

Kingma, D. P. & Welling, M. 2019, Foundations and Trends® in Machine Learning, 12, 307–392

Kormendy, J. 1977, ApJ, 218, 333

Kormendy, J. & Kennicutt, Robert C., J. 2004, ARA&A, 42, 603

Kümmel, M., Bertin, E., Schefer, M., et al. 2020, in Astronomical Society of the Pacific Conference Series, Vol. 527, Astronomical Data Analysis Software and Systems XXIX, ed. R. Pizzo, E. R. Deul, J. D. Mol, J. de Plaa, & H. Verkouter, 29

Lang, P., Wuyts, S., Somerville, R. S., et al. 2014, ApJ, 788, 11

Lange, R., Moffett, A. J., Driver, S. P., et al. 2016, MNRAS, 462, 1470

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv, 1110.3193

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179

Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163

Mortlock, A., Conselice, C. J., Hartley, W. G., et al. 2013, MNRAS, 433, 1185

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. 2021, Journal of Machine Learning Research, 22, 1

Pawlik, M. M., Wild, V., Walcher, C. J., et al. 2016, MNRAS, 456, 3032

Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266

Peterson, J. R., Jernigan, J. G., Kahn, S. M., et al. 2015, ApJS, 218, 14

Pohlen, M. & Trujillo, I. 2006, A&A, 454, 759

Robotham, A. S. G., Davies, L. J. M., Driver, S. P., et al. 2018, MNRAS, 476, 3137

Robotham, A. S. G., Taranu, D. S., Tobar, R., Moffett, A., & Driver, S. P. 2017, MNRAS, 466, 1513

Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, Astronomy and Computing, 10, 121

Schreiber, C., Elbaz, D., Pannella, M., et al. 2017, A&A, 602, A96

Sérsic, J. L. 1968, Atlas de galaxias australes (Cordoba, Argentina: Observatorio Astronomico, 1968)

Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, ApJS, 196, 11

Tal, T. & van Dokkum, P. G. 2011, ApJ, 731, 89

Trujillo, I., Graham, A. W., & Caon, N. 2001, MNRAS, 326, 869

Tuccillo, D., Huertas-Company, M., Decencière, E., et al. 2018, MNRAS, 475, 894

van de Sande, J., Bland-Hawthorn, J., Fogarty, L. M. R., et al. 2017, ApJ, 835, 104

Vega-Ferrero, J., Domínguez Sánchez, H., Bernardi, M., et al. 2021, MNRAS, 506, 1927

Vulcani, B., Bamford, S. P., Häußler, B., et al. 2014, MNRAS, 441, 1340

[1] Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, 91405, Orsay, France

[2] Université Paris Cité, CNRS, Astroparticule et Cosmologie, F-75013 Paris, France

[3] School of Physics and Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, UK

[4] Université Paris-Cité, 5 Rue Thomas Mann, 75013, Paris, France

[5] Université PSL, Observatoire de Paris, Sorbonne Université, CNRS, LERMA, F-75014, Paris, France

[6] Instituto de Astrofísica de Canarias (IAC); Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38200, La Laguna, Tenerife, Spain

[7] Instituto de Astrofísica de Canarias, Calle Vía Láctea s/n, E-38204, San Cristóbal de La Laguna, Tenerife, Spain

[8] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, I-00078 Monteporzio Catone, Italy

[9] Instituto de Física de Cantabria, Edificio Juan Jordá, Avenida de los Castros, E-39005 Santander, Spain

[10] Departamento de Física Teórica, Atómica y Óptica, Universidad de Valladolid, 47011 Valladolid, Spain

[11] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, PT-1349-018 Lisboa, Portugal

[12] Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK

[13] European Southern Observatory, Alonso de Cordova 3107, Casilla 19001, Santiago, Chile

[14] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679 München, Germany

[15] Department of Astronomy, University of Geneva, ch. dÉcogia 16, CH-1290 Versoix, Switzerland

[16] Institut d'Astrophysique de Paris, UMR 7095, CNRS, and Sorbonne Université, 98 bis boulevard Arago, 75014 Paris, France

[17] Canada-France-Hawaii Telescope, 65-1238 Mamalahoa Hwy, Kamuela, HI 96743, USA

[18] Instituto de Matemática Estatística e Física, Universidade Federal do Rio Grande, 96203-900, Rio Grande, RS, Brazil

[19] University of Nottingham, University Park, Nottingham NG7 2RD, UK

[20] Aix-Marseille Univ, CNRS, CNES, LAM, Marseille, France

[21] ICRAR, M468, University of Western Australia, Crawley, WA 6009, Australia

[22] SRON Netherlands Institute for Space Research, Landleven 12, 9747 AD, Groningen, The Netherlands

[23] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands

[24] Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, F-31400 Toulouse, France

[25] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, F-67000 Strasbourg, France

[26] School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL, UK

[27] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, D-85748 Garching, Germany

[28] Institut für Astro- und Teilchenphysik, Universität Innsbruck, Technikerstr. 25/8, 6020 Innsbruck, Austria

[29] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, I-80131 Napoli, Italy

[30] Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK

[31] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna,

Via Piero Gobetti 93/3, I-40129 Bologna, Italy

[32] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK

[33] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Università di Bologna, via Piero Gobetti 93/2, I-40129 Bologna, Italy

[34] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, I-40127 Bologna, Italy

[35] Dipartimento di Fisica, Università di Genova, Via Dodecaneso 33, I-16146, Genova, Italy

[36] INFN-Sezione di Roma Tre, Via della Vasca Navale 84, I-00146, Roma, Italy

[37] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, I-80126, Napoli, Italy

[38] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal

[39] Dipartimento di Fisica, Universitá degli Studi di Torino, Via P. Giuria 1, I-10125 Torino, Italy

[40] INFN-Sezione di Torino, Via P. Giuria 1, I-10125 Torino, Italy

[41] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, I-10025 Pino Torinese (TO), Italy

[42] INAF-IASF Milano, Via Alfonso Corti 12, I-20133 Milano, Italy

[43] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain

[44] Port d'Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain

[45] Institut d'Estudis Espacials de Catalunya (IEEC), Carrer Gran Capitá 2-4, 08034 Barcelona, Spain

[46] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain

[47] INFN section of Naples, Via Cinthia 6, I-80126, Napoli, Italy

[48] Dipartimento di Fisica e Astronomia "Augusto Righi" - Alma Mater Studiorum Universitá di Bologna, Viale Berti Pichat 6/2, I-40127 Bologna, Italy

[49] INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, I-50125, Firenze, Italy

[50] Centre National d'Etudes Spatiales, Toulouse, France

[51] Institut national de physique nucléaire et de physique des particules, 3 rue Michel-Ange, 75794 Paris Cédex 16, France

[52] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK

[53] ESAC/ESA, Camino Bajo del Castillo, s/n., Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain

[54] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy

[55] Univ Lyon, Univ Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, F-69622, Villeurbanne, France

[56] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland

[57] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisboa, Portugal

[58] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, PT-1749-016 Lisboa, Portugal

[59] Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK

[60] INFN-Padova, Via Marzolo 8, I-35131 Padova, Italy

[61] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, Astrophysique, Instrumentation et Modélisation Paris-Saclay, 91191 Gif-sur-Yvette, France

[62] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, I-34143 Trieste, Italy

[63] Istituto Nazionale di Astrofisica (INAF) - Osservatorio di Astrofisica e Scienza dello Spazio (OAS), Via Gobetti 93/3, I-40127 Bologna, Italy

[64] Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Via Irnerio 46, I-40126 Bologna, Italy

[65] INAF-Osservatorio Astronomico di Padova, Via dell'Osservatorio 5, I-35122 Padova, Italy

[66] Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, N-0315 Oslo, Norway

[67] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands

[68] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109, USA

[69] von Hoerner & Sulger GmbH, SchloßPlatz 8, D-68723 Schwetzingen, Germany

[70] Technical University of Denmark, Elektrovej 327, 2800 Kgs. Lyngby, Denmark

[71] Institut d'Astrophysique de Paris, 98bis Boulevard Arago, F-75014, Paris, France

[72] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

[73] Aix-Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France

[74] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland

[75] Department of Physics and Helsinki Institute of Physics, Gustaf Hällströmin katu 2, 00014 University of Helsinki, Finland

[76] NOVA optical infrared instrumentation group at ASTRON, Oude Hoogeveensedijk 4, 7991PD, Dwingeloo, The Netherlands

[77] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany

[78] Department of Physics, Institute for Computational Cosmology, Durham University, South Road, DH1 3LE, UK

[79] University of Applied Sciences and Arts of Northwestern Switzerland, School of Engineering, 5210 Windisch, Switzerland

[80] European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands

[81] Department of Physics and Astronomy, University of Aarhus, Ny Munkegade 120, DK-8000 Aarhus C, Denmark

[82] Centre for Astrophysics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

[83] Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

[84] Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada

[85] Space Science Data Center, Italian Space Agency, via del Politecnico snc, 00133 Roma, Italy

[86] Departamento de Astrofísica, Universidad de La Laguna, E-38206, La Laguna, Tenerife, Spain

[87] Dipartimento di Fisica e Astronomia "G.Galilei", Universitá di Padova, Via Marzolo 8, I-35131 Padova, Italy

[88] Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile

[89] Centre for Electronic Imaging, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK

[90] AIM, CEA, CNRS, Université Paris-Saclay, Université de Paris, F-91191 Gif-sur-Yvette, France

[91] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain

[92] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, 30202 Cartagena, Spain

[93] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA

[94] INAF-Osservatorio Astronomico di Brera, Via Brera 28, I-20122 Milano, Italy

[95] Junia, EPA department, F 59000 Lille, France

[96] SISSA, International School for Advanced Studies, Via Bonomea 265, I-34136 Trieste TS, Italy

[97] IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy

[98] INFN, Sezione di Trieste, Via Valerio 2, I-34127 Trieste TS, Italy

[99] Dipartimento di Fisica e Scienze della Terra, Universitá degli Studi di Ferrara, Via Giuseppe Saragat 1, I-44122 Ferrara, Italy

[100] Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara, Via Giuseppe Saragat 1, I-44122 Ferrara, Italy

[101] Institut de Physique Théorique, CEA, CNRS, Université Paris-Saclay F-91191 Gif-sur-Yvette Cedex, France

[102] Dipartimento di Fisica - Sezione di Astronomia, Universitá di Tri-

este, Via Tiepolo 11, I-34131 Trieste, Italy

[103] NASA Ames Research Center, Moffett Field, CA 94035, USA

[104] INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, I-40129 Bologna, Italy

[105] INFN-Bologna, Via Irnerio 46, I-40126 Bologna, Italy

[106] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice cedex 4, France

[107] Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, D-52056 Aachen, Germany

[108] Department of Physics & Astronomy, University of California Irvine, Irvine CA 92697, USA

[109] University of Lyon, UCB Lyon 1, CNRS/IN2P3, IUF, IP2I Lyon, France

[110] INFN-Sezione di Genova, Via Dodecaneso 33, I-16146, Genova, Italy

[111] INAF-Istituto di Astrofisica e Planetologia Spaziali, via del Fosso del Cavaliere, 100, I-00100 Roma, Italy

[112] Instituto de Física Teórica UAM-CSIC, Campus de Cantoblanco, E-28049 Madrid, Spain

[113] Department of Physics, P.O. Box 64, 00014 University of Helsinki, Finland

[114] Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing (GCCL), 44780 Bochum, Germany

[115] Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK

[116] Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France

[117] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

[118] Astrophysics Group, Blackett Laboratory, Imperial College London, London SW7 2AZ, UK

[119] Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, 53, Avenue des Martyrs, 38000, Grenoble, France

[120] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 2, I-00185 Roma, Italy

[121] Institut für Theoretische Physik, University of Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany

[122] Zentrum für Astronomie, Universität Heidelberg, Philosophenweg 12, D- 69120 Heidelberg, Germany

[123] Department of Mathematics and Physics E. De Giorgi, University of Salento, Via per Arnesano, CP-I93, I-73100, Lecce, Italy

[124] INAF-Sezione di Lecce, c/o Dipartimento Matematica e Fisica, Via per Arnesano, I-73100, Lecce, Italy

[125] INFN, Sezione di Lecce, Via per Arnesano, CP-193, I-73100, Lecce, Italy

[126] Institute of Space Science, Bucharest, Ro-077125, Romania

[127] Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[128] Higgs Centre for Theoretical Physics, School of Physics and Astronomy, The University of Edinburgh, Edinburgh EH9 3FD, UK

[129] Université St Joseph; Faculty of Sciences, Beirut, Lebanon

[130] Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA

[131] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland

[132] Department of Mathematics and Physics, Roma Tre University, Via della Vasca Navale 84, I-00146 Rome, Italy

[133] Cosmic Dawn Center (DAWN)

[134] Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen, Denmark

[135] Departement of Physics and Astronomy, University of British Columbia, Vancouver, BC V6T 1Z1, Canada

Appendix A: Interactive plot website

We describe here the interactive web page we have created to accompany this paper: https://share.streamlit.io/ hbretonniere/euclid_morphology_challenge. It will allow the reader to create most of the plots we present in this paper but with control over some choices we made for the representation. The tool was constructed using the `Streamlit` python package. All plots can be computed on the fly, with the common catalogues for each type of simulation (i.e., using the results of the Euclid Morphology Challenge, which were the basis for the analysis presented in the paper). The following is a brief introduction to the tool. We recommend the use of Firefox, Opera, or Chrome browsers, and found that the platform is unstable in Safari. `JavaScript` has to be authorised for the platform to work. In some cases, the data needs to be download on the server, and it can take several minutes. For any additional information, contact H.Bretonnière.

The main plotting parameters are located on the top of the left panel of the page. First, choose a type of simulation in the dataset option, single-Sérsic, double-Sérsic, realistic, or multi-band. Depending on the type of simulation, additional options appear, for example the fitting made with a free Sérsic index for the bulge component of the double-Sérsic simulations. Then, just below the dataset panel, the different types of plots will appear: the summary plots (e.g. Fig. 6), the trumpet plots (e.g. Fig. 3), the summary score plot (Fig. 23), or the error prediction (e.g. Fig20). The options vary depending on the selected dataset (e.g., 2D plots are only available for the double-Sérsic dataset). Next, choose the parameters to analyse and select the different software packages. For each, the user can select or deselect the parameter/software, or click the 'all parameters' or 'all softwares' button to select them all. Finally, for the summary and trumpet plots, the user can change the *x*-axis, i.e. the true parameter in addition to the number of bins. The available options are the I_E magnitude, redshift, half-light radius or Sérsic index. According to the options selected in the panel on the left, the figure will change in the panel on the right. Above the plot, there is a check box called `Demo version`. When selected, only 1% of the catalogue is used to draw the plots. This option should only be used to play with the parameters until the right plot is selected. Then, the box must be un-checked to produce plots including all measurements of the combined catalogue, as was used in the plots of the paper. The slider button allows easy control over the value of the outlier threshold (see Sect. 3). Here again, the plot will change accordingly, for example changing the third columns of the summary type plots. Another slider control changes the *x*-axis range, and the *y*-axis if we are working on the 2D summary plots. By default, the *y*-axis range is fixed to hard-coded values. The user can undo this limitation by un-checking the `limit y axis range` button. If only one parameter is selected, three other sliders will appear to control the *y*-axis ranges.

Finally, below the figure, a checkbox allows one to save the results used to produce the summary plot, as a python dictionary. The dictionary *D* is structured as follows:

$$D = \mathrm{dict}[i][p] \,, \tag{A.1}$$

and contains all the bins of the metric *i*, on a parameter *p* and software name *s* where $i = 0, 1,$ or 2 for the bias, dispersion and outlier fraction, respectively.

The different buttons and options change depending on the dataset or plot type. For example, when plotting the global score, four additional sliders appear to control the weights of the metrics (see Eq. 9). The user will also be able to remove the bin weighting, to have an equal weight for all magnitudes. Some of EMC2022a's plots can also be plotted by changing the "select EMC paper" option on the top button of the left panel.

Appendix B: Effect of `PSFEx` on the `SourceXtractor++` fitting

The challenge instructions to the participants did not restrict or predict procedure steps, but left maximal freedom to allow for best choices by the experts. This lead to individual decisions and interpretations of the Challenge requirements. One such choice by `SourceXtractor++` was to employ the full pipeline, including some important pre-processing steps. In preparation for model fitting with `SourceXtractor++`, the team made modifications to the data that was provided by the challenge team, in order to treat them in the same way they would treat real data. In particular, the Rubin-like and *Euclid* NIR images were re-sampled back to their original pixel scale (0″2 and 0″3, respectively). The team also extracted oversampled PSF images using the PSFEx software. They thus used the provided PSF images in the given sampling and have oversampled PSF images in all bands. This choice was made based on the team's assumption that it was an essential condition to deliver satisfactory results. The outcome of the `SourceXtractor++` team's internal checks found that PSF images were sufficiently well oversampled (two times for Rubin-like data, and three times for the *Euclid* NIR bands), but not for the I_E band. They concluded that their own PSF, created by PSFEx, could improve results for morphology and stellar photometry with a 1/0.45 oversampling factor. Figure D.1 compares outputs that do and do not use this step of the `SourceXtractor++` pipeline. Dashed lines (without PSFEx) use the PSF we provided, while the solid lines use the re-sampled PSF after PSFEx was employed. The performance of measuring the axis ratio and Sérsic index particularly improve, but also the effective radius.

Appendix C: Performance of metrics at different redshifts

The behaviour of all parameters with varying redshifts can be explored in the online tool. Here, we detail the general trend by looking at selected examples for *q* in the single-Sérsic simulations. In Figs. D.2 and D.3, we show examples of the scatter- and summary-plots as a function of the true redshift of galaxies (rather than I_E magnitude, which was shown throughout the paper). We first see that most galaxies are at a redshift smaller than 2 (see also Fig. 1). In the scatter plot, the dependence on the quality of (the exemplary) *q* prediction and the redshift is less clear than it was with magnitude, without the characteristic trumpet shape. Nevertheless, both the running bias and dispersion demonstrate that the dispersion is higher at higher redshifts. We detect clear

trends in all codes in the three metrics, bias, dispersion and outlier fraction (Fig. D.3). We can also see that in general, the metrics get worse from redshift $z \simeq 0$ to $z \simeq 2$, and then reach a plateau.

In EMC2022a, we detail in figure 2 the distribution of redshifts as a function of $I_{\rm E}$. This further helps to link the papers presented here with redshift.

Appendix D: Inspection of selected quality flags

Most participating software packages also include flags with additional information that allow the user to choose to reject objects based on certain criteria. In addition, some codes give more flags than others and therefore lead to more or less restrictive common catalogues. Since we are interested in relative fit qualities, we have decided to use a common catalogue that is as inclusive as possible, i.e., that maximizes the number of objects to test how the individual codes deal with the same set of difficult objects. This naturally affects the result on a code-by-code basis to some extent. In order to investigate the seriousness of this affect, we here test a number of selected flags.

`Galapagos-2` returns five quality flags that link back to whether fits (of one or more components) encountered fitting constraints or indicate the relative brightness of the double-Sérsic components (see EMC2022a for a summary of all flags). These are thoroughly described and tested in Häußler et al. (2022). Here, we test the impact of the constraint flags (USE_FLAG_SS, USE_FLAG_BULGE_CONSTR and USE_FLAG_DISK_CONSTR) on the summary figures. For the single-Sérsic case, we remove all galaxies that are not constrained, which represent 4% of the common catalogue. These are mainly faint galaxies (Fig. D.4), improving the results by few percent in the last magnitude bin. Removing those galaxies also improves the results for `Morfometryka` and `ProFit`, which indicates that all three codes returned unreliable measurements for these galaxies, but only `Galapagos-2` specified this with a flag. For the double-Sérsic simulation, presented in Fig. D.5, we remove galaxies where the bulge or the disk fits were not constrained. This is the case for 13% of the catalogue, mainly because the bulges were not constrained. This is a non negligible impact for the bulge components, with the impact greater for `Galapagos-2` than for other codes. This indicates that the same galaxies are successfully fit by some codes, and not by others; some of the codes (like `Galapagos-2` and `Morfometryka`) have additional quality flags in place to inform the user.

As mentioned in the text, we also studied the impact of the objects flagged as 'TARGETISSTAR' by `Morfometryka`. We show in Fig. D.6 the trumpet plot of the fitting of the single-Sérsic effective radius fitting by `Morfometryka`, and conclude that taking into account this flag will remove the bi-modality, as explained Sect. 4.1.
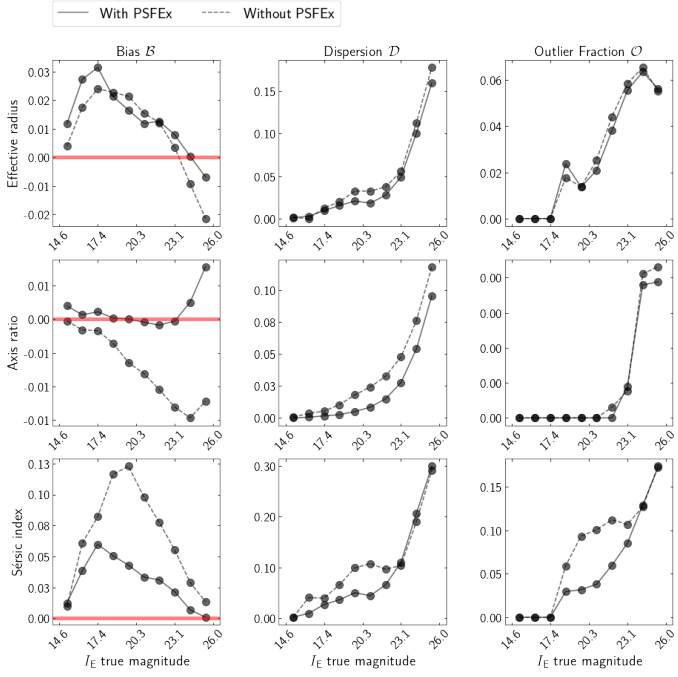
Fig. D.1. Summary plot with the comparison of the fitting with (solid line) or without (dashed line) the use of PSFEx. With PSFEx (as in the rest of the paper), the results are still slightly different than the one presented earlier because only one field is analysed.
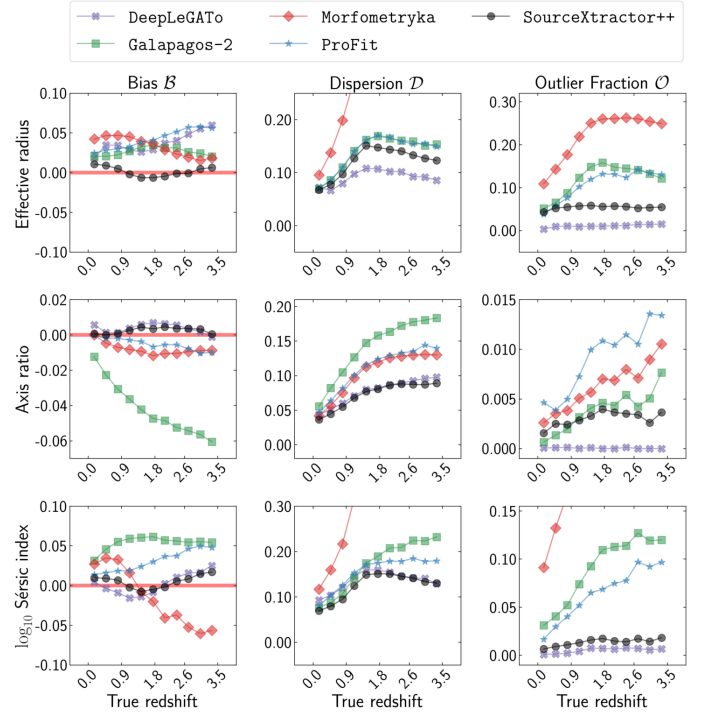

Fig. D.3. Summary plot for the fitting of the single Sérsic simulations regarding the true redshift.
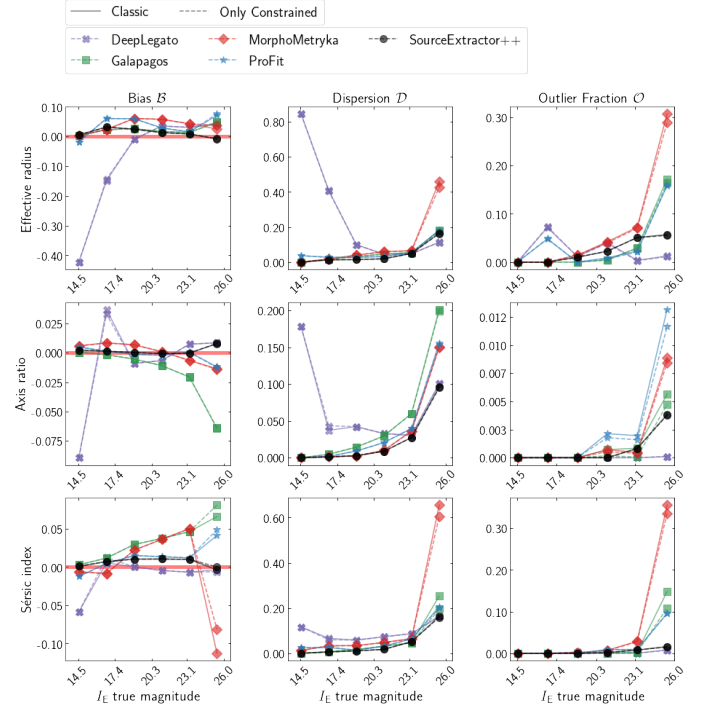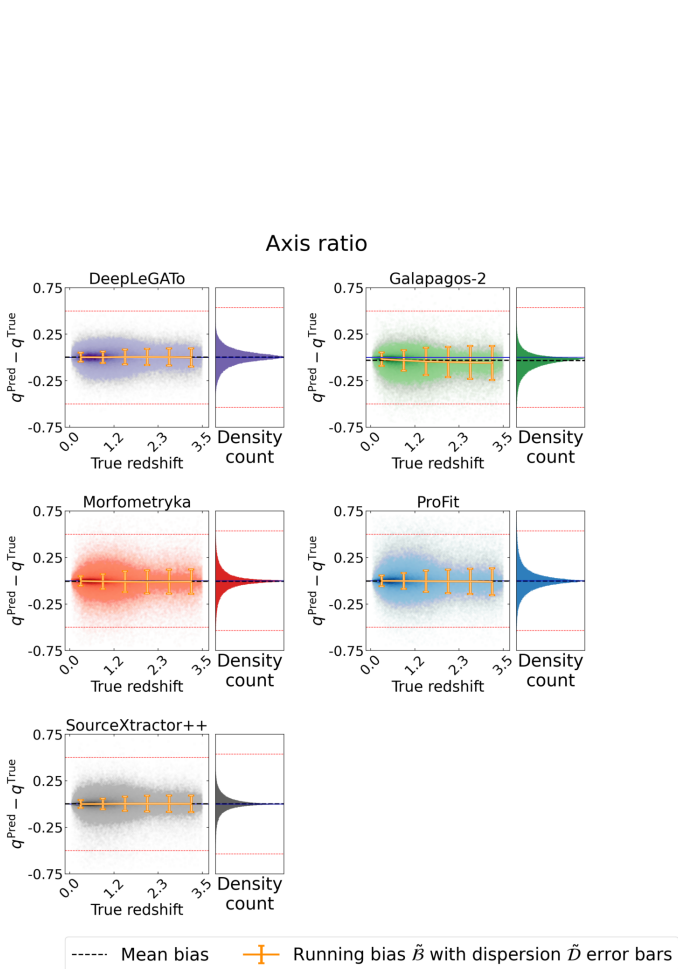

Fig. D.2. Trumpet plot for the single-Sérsic q fitting as a function of the true redshift. See Fig. 3 for further information.

Fig. D.4. Summary plot comparing the fit on the single-Sérsic simulations with or without considering `Galapagos-2` constraint flags.
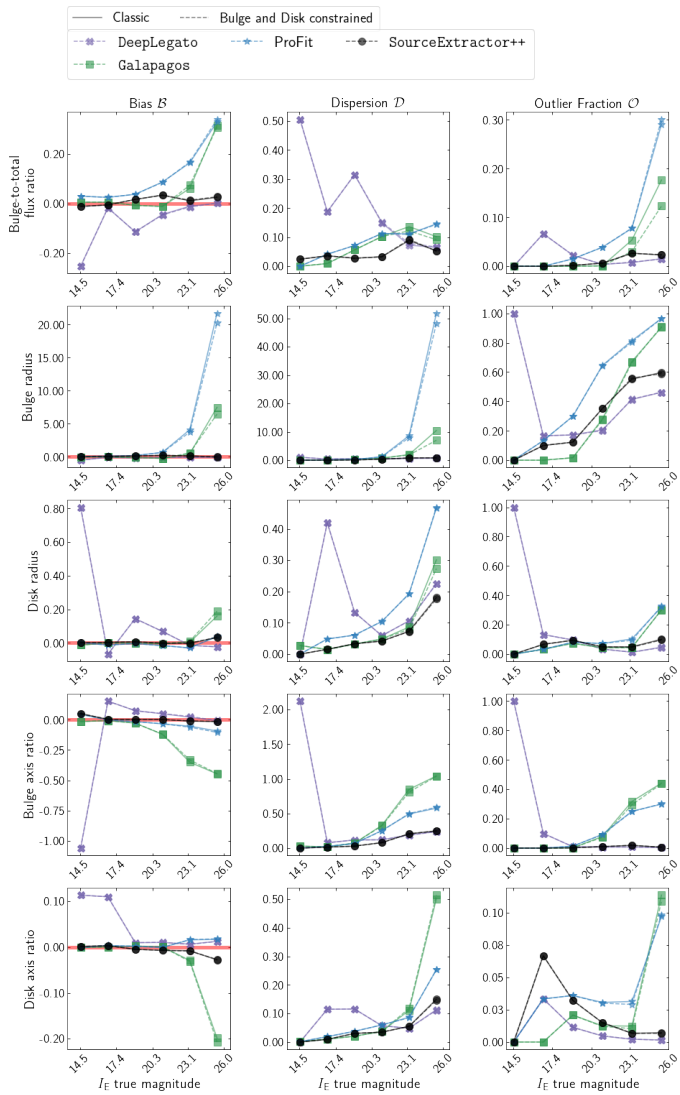
Fig. D.5. Summary plot comparing the fit on the double-Sérsic simulations with or without considering `Galapagos-2` constraint flags.
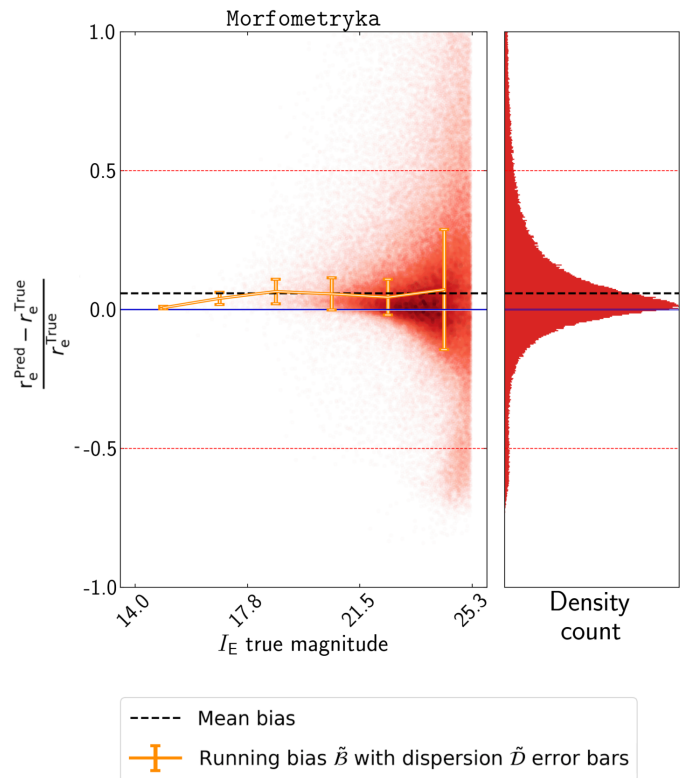

Fig. D.6. Trumpet plot for the single-Sérsic effective radius fitting, removing the objects flagged as "stars" by `Morfometryka`. Wee see that the bi-modality saw in Fig. 3 vanishes here.

Appendix E: Contact for reproducibility

We provide information about the precise version of the software packages used for the EMC, along with the configuration files and `README`s, where available. These can be found in a folder in the same repository that we provide for the interactive plot website (https://github.com/Hbretonniere/Euclid_Morphology_Challenge).

We also detail here the role of the different authors in the production of the paper to give the reader the possibility to contact for further details.

The simulations were produced in a joint effort by E. Merlin, M. Castellano, D. Tucillo, M. Huertas-Company and H. Bretonnière. Contact E. Merlin for information about the single- and double-Sérsic simulations, and H. Bretonnière for the FVAE simulation.

Regarding the use of the different codes presented in the challenge:

1. `DeepLeGATo` was run by D. Tucillo. Contact M. Huertas-Company for further information.
2. `Galapagos-2` was run by B. Häußler.
3. `Morfometryka` was run by L. Ferreira, F.Ferrari and F.G. Lucatelli, with C. Conselice. Contact L. Ferreira for further information.
4. `ProFit` was run by A.S.G. Robotham.
5. `SourceXtractor++` was run in a joint effort by A. Álvarez-Ayllón, E. Bertin, P. Dubath, R. Gavazzi, W. Hartley, D. Hernandez Lang, M.Kümmel, and M. Schefer. Contact W. Hartley for further information.

The analysis was conducted and the paper was written by H. Bretonnière, U. Kuchner and M. Huertas-Company with input from members of the Euclid morphology working group and challenge participants.