A&A 671, A102 (2023)
https://doi.org/10.1051/0004-6361/202245042
© The Authors 2023

Astronomy & Astrophysics

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], H. Domínguez Sánchez[25], P.-A. Duc[26], S. Fotopoulou[27], S. Kruk[28], A. La Marca[22,23], B. Margalef-Bentabol[22], F. R. Marleau[29], C. Tortora[30], N. Aghanim[1], A. Amara[31], N. Auricchio[32], R. Azzollini[33], M. Baldi[34,32,35], R. Bender[28,14], C. Bodendorf[28], E. Branchini[36,37], M. Brescia[38,30], J. Brinchmann[39], S. Camera[40,41,42], V. Capobianco[42], C. Carbone[43], J. Carretero[44,45], F. J. Castander[46,47], S. Cavuoti[30,48,38], A. Cimatti[49,50], R. Cledassou[51,52], G. Congedo[53], L. Conversi[54,55], Y. Copin[56], L. Corcione[42], F. Courbin[57], M. Cropper[33], A. Da Silva[58,59], H. Degaudenzi[15], J. Dinis[58,59], F. Dubath[15], C. A. J. Duncan[60,12], X. Dupac[54], S. Dusini[61], S. Farrens[62], S. Ferriol[56], M. Frailis[63], E. Franceschi[32], M. Fumana[43], S. Galeotta[63], B. Garilli[43], B. Gillis[53], C. Giocoli[64,65], A. Grazian[66], F. Grupp[28,14], S. V. H. Haugan[67], H. Hoekstra[68], W. Holmes[69], F. Hormuth[70], A. Hornstrup[71], P. Hudelot[72], K. Jahnke[73], S. Kermiche[74], A. Kiessling[69], R. Kohley[54], M. Kunz[75], H. Kurki-Suonio[76], S. Ligori[42], P. B. Lilje[67], I. Lloro[77], O. Mansutti[63], O. Marggraf[78], K. Markovic[69], F. Marulli[34,32,35], R. Massey[79], H. J. McCracken[16], E. Medinaceli[32], M. Melchior[80], M. Meneghetti[32,35], G. Meylan[57], M. Moresco[34,32], L. Moscardini[34,32,35], E. Munari[63], S. M. Niemi[81], C. Padilla[44], S. Paltani[15], F. Pasian[63], K. Pedersen[82], W. Percival[83,84,85], V. Pettorino[62], G. Polenta[86], M. Poncet[51], L. Pozzetti[32], F. Raison[28], R. Rebolo[7,87], A. Renzi[88,61], J. Rhodes[69], G. Riccio[30], E. Romelli[63], C. Rosset[2], E. Rossetti[34], R. Saglia[28,14], D. Sapone[89], B. Sartoris[14,63], P. Schneider[78], A. Secroun[74], G. Seidel[73], C. Sirignano[88,61], G. Sirri[35], J. Skottfelt[90], J.-L. Starck[91], P. Tallada-Crespí[92,45], A. N. Taylor[53], I. Tereno[58,11], R. Toledo-Moreo[93], I. Tutusaus[75], E. A. Valentijn[23], L. Valenziano[32,35], T. Vassallo[63], Y. Wang[94], J. Weller[14,28], G. Zamorani[32], J. Zoubian[74], S. Andreon[95], S. Bardelli[32], C. Colodro-Conde[6], D. Di Ferdinando[35], J. Graciá-Carpio[28], V. Lindholm[76], N. Mauri[49,35], S. Mei[2], V. Scottez[72,96], E. Zucca[32], C. Baccigalupi[97,98,63,99], M. Ballardini[100,101,32], F. Bernardeau[102], A. Biviano[63,98], S. Borgani[103,98,63,99], A. S. Borlaff[104], C. Burigana[100,105,106], A. Cappi[107,32], C. S. Carvalho[11], S. Casas[108], G. Castignani[34,32], A. R. Cooray[109], J. Coupon[15], H. M. Courtois[110], S. Davini[111], G. De Lucia[63], G. Desprez[15], J. A. Escartin[28], S. Escoffier[74], M. Fabricius[28,14], M. Farina[112], A. Fontana[8], K. Ganga[2], J. Garcia-Bellido[113], K. George[14], G. Gozaliasl[114], H. Hildebrandt[115], I. Hook[116], O. Ilbert[20], S. Ilić[117,51,24], B. Joachimi[118], V. Kansal[91], E. Keihanen[76], C. C. Kirkpatrick[76], A. Loureiro[53,118,119], J. Macias-Perez[120], M. Magliocchetti[112], R. Maoli[121,8], S. Marcin[80], M. Martinelli[8], N. Martinet[20], M. Maturi[122,123], P. Monaco[103,98,63,99], G. Morgante[32], S. Nadathur[31], A. A. Nucita[124,125,126], L. Patrizii[35], V. Popa[127], C. Porciani[78], D. Potter[128], A. Pourtsidou[53,129], M. Pöntinen[114], P. Reimberg[72], A. G. Sánchez[28], Z. Sakr[24,122,130], M. Schirmer[73], E. Sefusatti[98,63,99], M. Sereno[32,35], J. Stadel[128], R. Teyssier[131], J. Valiviita[132], S. E. van Mierlo[23], A. Veropalumbo[133], M. Viel[103,98,99,97], J. R. Weaver[134,135], and D. Scott[136]

(Affiliations can be found after the references)

Received 22 September 2022 / Accepted 21 October 2022

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 focusses 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

`SourceXtractor++`, 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 official *Euclid* 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. methods: data analysis – galaxies: evolution – galaxies: fundamental parameters – cosmology: observations

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 differentiates 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), 1D 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 (2D 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 (NIR) 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 $0\farcs17$ 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 for determining galaxy morphologies with *Euclid* since no previous large studies have used such a wide filter. While this filter 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 IR bands, $Y_{\rm E}$, $J_{\rm E}$, and $H_{\rm E}$, covering approximately 950 to 2020 nm (Euclid Collaboration 2022a).

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 NIR 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 eight 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 $\mu = 27\,\mathrm{mag\,arcsec^{-2}}$ (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 parametric fitting codes on simulated *Euclid* data. The structural measurements evaluated in this work are not tailored to a specific science case. Rather, we provide a comparison of the measurements of parameters (*Euclid* data products) that will enable astronomers to investigate a range of research questions related to galaxy evolution and morphologies or structures with *Euclid*. For example, as Sérsic indices are an approximation to statistically distinguish early- from late-type galaxies, probing those indices in a large range of redshift can help us understand morphology evolution. They will also be combined with other parameters of interest such as colour or stellar mass to scrutinise current models. The depth and volume of Euclid will constrain these relations and open a variety of investigations needed to make progress in galaxy evolution science.

The challenge comprises a simulated dataset of five fields, each realised with single-Sérsic, double-Sérsic, and neural-network-generated galaxies in the I_E band. In addition, one of the fields has been simulated in the NIR (Y_E, J_E, and H_E) bands, and in the five *u, g, r, i,* and *z Rubin* Observatory bands to test the accuracy 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 (Euclid Collaboration 2023; hereafter EMC2023) 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 important information 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 of 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 EMC2023 and in 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 modelling on convolutional neural networks. All of them fitted at least a single profile to each galaxy in the I_E band, and some teams and 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, for example 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, examine the code-to-code scatter, and investigate the known bias towards 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 (Figs. 6, 13, and 19) and in the 'global score' sub-sections (Sects. 4.1.4, 4.2.8, and 4.3.4). Section 4.4 focusses 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.

[1] https://github.com/MegaMorph/galapagos

[2] https://github.com/asgr/ProFound

[3] https://github.com/astrorama/SourceXtractorPlusPlus

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, EMC2023.

We created five fields of 25 000 × 25 000 pixel each, at $0\farcs1$ pixel^{-1} scale, corresponding to a field of view (FoV) of about 0.482 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 $I_\mathrm{E} \simeq 15$ to $I_\mathrm{E} \simeq 30$ 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 (S/N) thresholds for 100σ ($I_\mathrm{E} \simeq 22$), 10σ ($I_\mathrm{E} \simeq 24.6$), 5σ ($I_\mathrm{E} \simeq 25.25$), and 1σ ($I_\mathrm{E} \simeq 27.1$). The fifth list contains all simulated sources, including objects below $S/N = 1$. We asked the participants to fit those galaxies with at least an S/N over 5σ, where we defined the S/N of a source as the S/N of a point-source in a circular aperture with a diameter of $2''$, and thus this value corresponds to galaxies brighter than $I_\mathrm{E} = 25.3$. It is important to note that this definition of the S/N 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 S/N greater than 5σ: I_E, effective radius r_e (plotted as logarithmic, $\log_{10} r_\mathrm{e}$); 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 $I_\mathrm{E} = 25.3$ (see EMC2023 for more details). We describe in more detail the generation of these galaxies in the following sections. We 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 2022b). 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 $0\farcs17$. To convolve the images, the PSF was over-sampled to different degrees: 6 times in VIS; 6 times in NIR at $0\farcs3$ 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 EMC2023.

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. We explain our decisions and production steps for the common catalogues in more detail in EMC2023.

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:

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

where r_e is the effective or half-light radius, 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. We 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.

[4] Euclid's Scientific Challenges are benchmark tests organised inside the Euclid Consortium in preparation for the launch of the satellite.

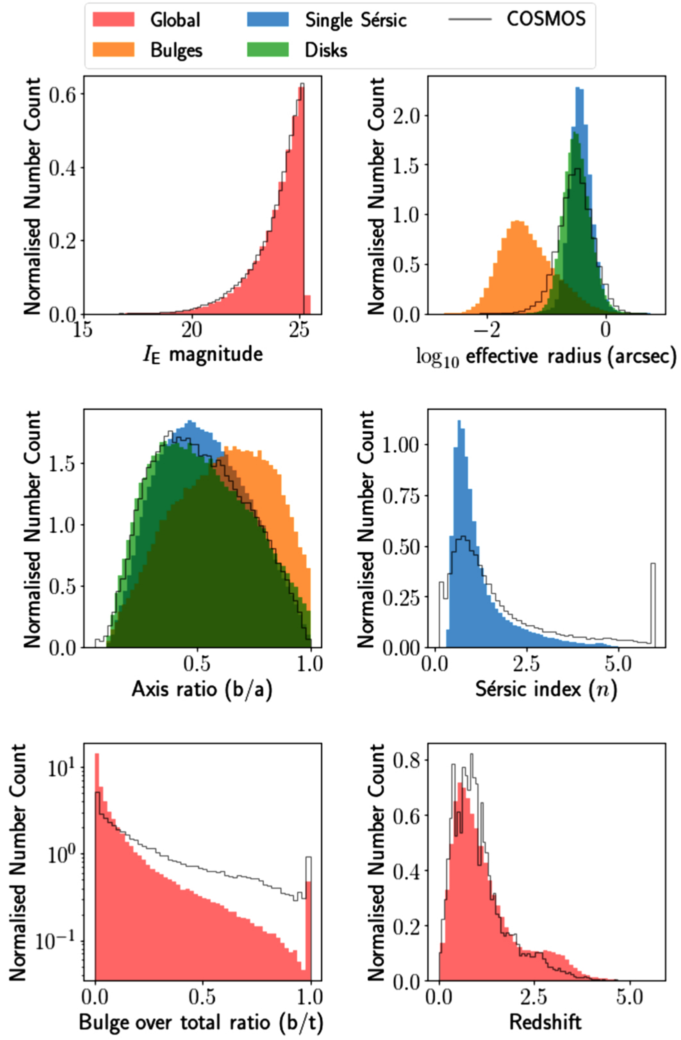

Fig. 1. Distributions of the simulated 'true' galaxy parameters measured in the Euclid Morphology Challenge. Top left: I_E distribution down to 5σ detections. Top right: effective radii for the single component galaxy (blue), and for bulges (orange), and discs (green) separately. Middle Left: Axis ratio distributions. Middle right: Sérsic index distributions for single-component galaxies. We 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. We also 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 EMC2023.

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, that is to say 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 pixel-wise 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:

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

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

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

The two profiles are then sheared to fit the desired ellipticity, $q_{\rm b}$ and $q_{\rm d}$. 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. iven the overall aim of the challenge to probe the capacity of software packages that attempt galaxy model fitting, we chose to test the codes on ideal galaxy simulations with known and fixed Sérsic indices to control for variations across the software packages.

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 (2022c, 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. We 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 EMC2023, we use four main indicators to evaluate and compare the different codes: completeness ($\mathcal{C}$)[5]; bias ($\mathcal{B}$); dispersion ($\mathcal{D}$); and outlier fraction ($\mathcal{O}$). 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 EMC2023. 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 $p = \{r_\mathrm{e}, q, n\}$ for single-Sérsic fits and $p = \{\mathrm{b/t}, r_\mathrm{e,b}, r_\mathrm{e,d}, q_\mathrm{d}, q_\mathrm{b}\}$ for double component fits. Sometimes it is appropriate to calculate the relative bias, $\tilde{b}_p$, which is defined as

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

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.″1 for a galaxy of $r_\mathrm{e} = 0.''2$ is more problematic than the same error on a galaxy with $r_\mathrm{e} = 3.''0$. 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,

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

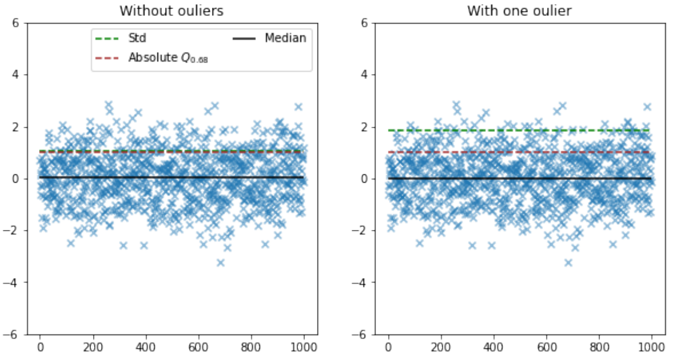

Fig. 2. Illustration of our dispersion metric choice. In both plot, we plot the median, the standard deviation and our definition of the dispersion, defined Eq. (8) for a Normal Gaussian distribution. In the right figure, we add an outlier at $y = 100$. We can see that our definition is not sensible of the presence of an outlier, compared to the standard deviation.

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). We 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 $\tilde{\mathcal{D}}_p$.

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 ($\mathcal{O}$) 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 $\mathcal{O}$ also differs in the following three cases. Firstly, for the 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%. Secondly, for the axis ratio and bulge-to-total ratio: because the bias is absolute, but the range of possible values is

[5] The completeness $\mathcal{C}$ measures the fraction of objects for which there is a successful fit, see EMC2023 for details.

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. Finally, for the 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 C, $\mathcal{B}_p$, $\mathcal{D}_p$, O_p:

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

We note that our three metrics, $k_\mathcal{B}, k_\mathcal{D}$, 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 EMC2023:

$$k_\mathcal{B} = k_\mathcal{D} = k_O = 2.1 . \qquad (10)$$

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 S/N of the bin and the fraction of objects in the bin (fewer objects and lower S/N will lead to a smaller weight, and thus smaller impact on $\mathcal{S}$); see EMC2023 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 S/N. 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, 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 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 EMC2023). 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 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, Fig. 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 (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: 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, that is 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` recognises 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,

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

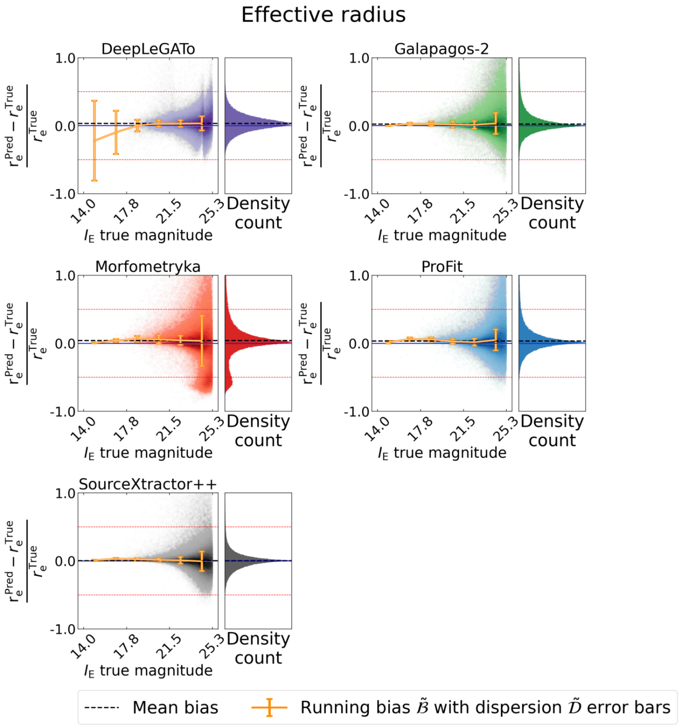

Fig. 3. Scatter plots showing the recovery of the half-light radius measured from the single-Sérsic simulation. Each panel shows a different code. The main plot of each panel shows the relative bias per galaxy as a function of apparent I_E magnitude, while we summarise the bias distribution as a histogram 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 grey 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).

'TARGETISSTAR', is designed to flag stars, which these are not, but their small sizes and low Sérsic indices are recognised 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 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 ($I_\mathrm{E} < 18$), 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 S/N. 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 S/N 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 ($\mathcal{O}$). We 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 $I_\mathrm{E} = 14.5$ to -0.05 at $I_\mathrm{E} = 17.5$. 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 $I_\mathrm{E} > 17$ 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 $I_\mathrm{E} = 17$ and $I_\mathrm{E} = 23.5$. 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 galaxies ($I_\mathrm{E} > 24$), 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 $I_\mathrm{E} < 24$ 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 $I_\mathrm{E} = 14.5$ to 0.2 at $I_\mathrm{E} = 17.5$. 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 $I_\mathrm{E} < 23.5$, 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 S/N 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 $I_\mathrm{E} < 23$, all the codes have less than 10% outliers, with ProFit and Galapagos-2 showing the smallest numbers of bad fits,

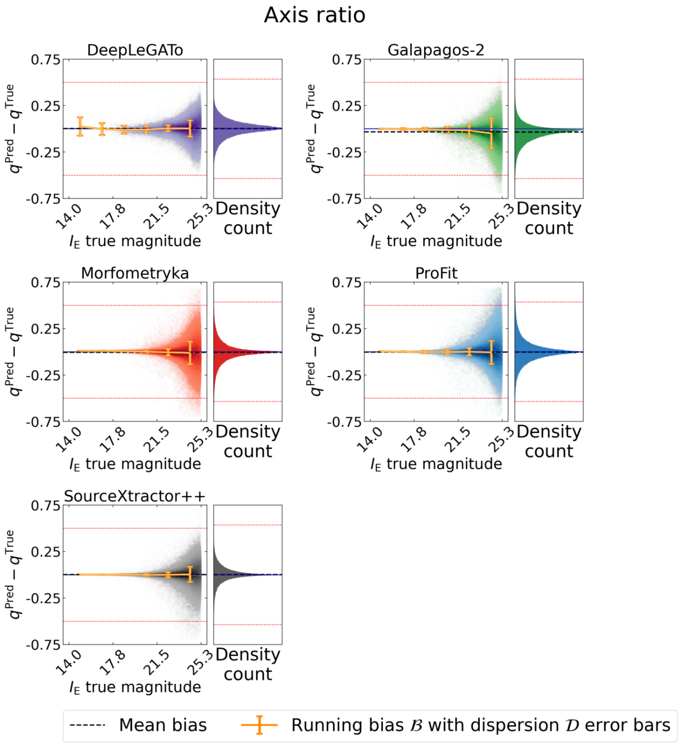

Fig. 4. Fitting results for the axis ratio q of the single-Sérsic simulation. See caption of Fig. 3 and Sect. 4 for further information.

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 $23 \leq I_{\rm E} \leq 26$ 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, 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 S/N 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 subsequently 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 $I_{\rm E} < 26$ for all codes, and for $I_{\rm E} < 23$ 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 $I_{\rm E} = 14$ to $I_{\rm E} = 17$ (cut by the y-axis range in the graph for visualisation purposes). For $I_{\rm E} < 24$, 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 S/N, 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 $I_{\rm E} < 24$. For $I_{\rm E} < 22$, 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 $19 \leq I_{\rm E} \leq 21$. 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.

We highlight that the error in the axis ratio measurement is the sum of at least two procedures: the prediction of the two semi-axis lengths (impacted by the S/N and the PSF) but also of the position angle (PA) of the galaxy, necessary to define the two semi-axis. We note that the PA is not part of our current comparison.

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.

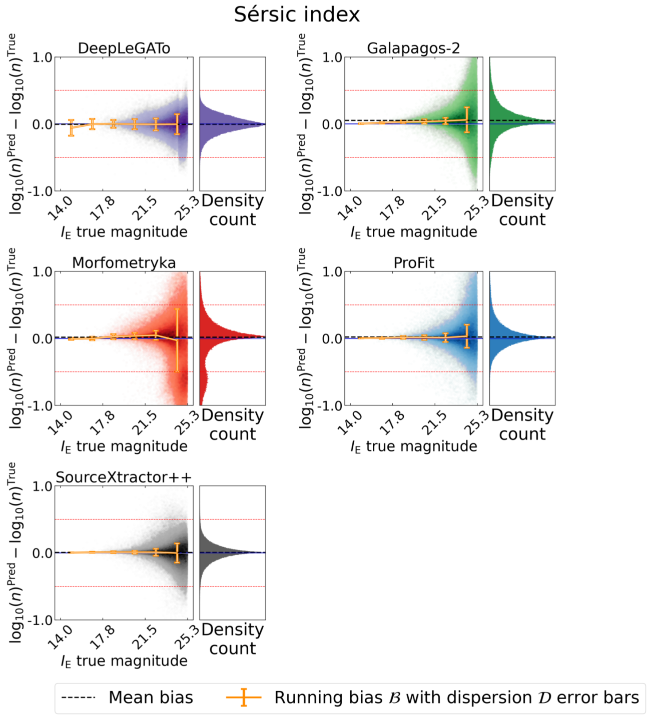

Fig. 5. Fitting results for the Sérsic index of the single-Sérsic simulation. See caption of Fig. 3 and Sect. 4 for further information.

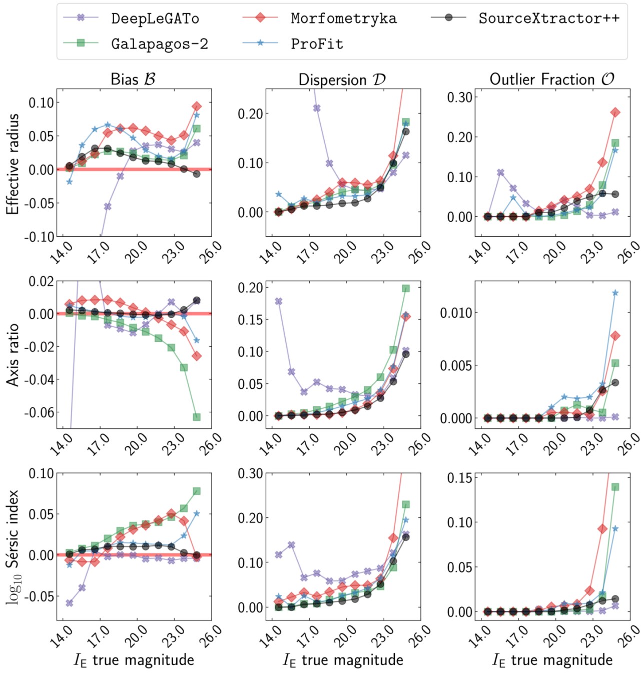

Fig. 6. Summary plot for the single-Sérsic simulation. The different rows show the results for the three different structural parameters: half-light radius $r_{\rm e}$ (top), axis ratio q (middle) and Sérsic index n (bottom). Columns represent (1) the mean bias $\mathcal{B}$, (2) the dispersion $\mathcal{D}$, and (3) the fraction of outliers $\mathcal{O}$, per bin of $I_{\rm E}$ magnitude (see text for details). We 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.

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, which means it does not predict steeper galaxy profiles in their cores. For fainter galaxies, from $I_{\rm E} = 17$ to $I_{\rm E} = 26$, DeepLeGATo achieves the most robust bias calibration, mitigated by the fact that it has the highest dispersion. SourceXtractor++ and ProFit have a similarly small bias (around 0.01) for $I_{\rm E} < 23$, 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 $I_{\rm E} < 23$, 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, O) for $I_{\rm E} < 19$, and just up to few percents for $I_{\rm E} < 22$. The fraction then increases steeply at faint magnitudes. The increase is highest for Morfometryka, from about 1% at $I_{\rm E} = 20$ 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 S/N, and the weights of the metrics. In particular, the weights of the smallest magnitude bins(from $I_{\rm E} = 14$ to $I_{\rm E} = 19$) 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.

Table 1. Comparison of the scores $\mathcal{S}$ obtained by the different software packages in all structural parameters for the single Sérsic simulations.

Galsim FVAE	$\mathcal{S}_{r_e}$	$\mathcal{S}_{b/a}$	$\mathcal{S}_n$	$\mu_{\mathcal{S}}$
DeepLeGATo	0.37 ∅	0.25 ∅	0.38 ∅	0.33 ∅
Galapagos-2	0.58 2.05	0.43 0.79	0.60 1.29	0.54 1.38
Morfometryka	1.10 ∅	0.37 ∅	1.20 ∅	0.89 ∅
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

Notes. Numbers in blue are results from GALSIM simulations (discussed in Sect. 4.1), and numbers in black quote results from measurements of simulations with the deep generative model FVAE (discussed in Sect. 4.3). The last column is the mean of the parameters. A smaller $\mathcal{S}$ means a better fit.

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, that is the bulge component has a low S/N 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 S/N 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 maximising 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 $I_E \simeq 17.5$ onwards. `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 increase from $I_E \sim 19$ onwards 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, that is 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. We note that the scale is logarithmic for the bias and the dispersion, to help appreciate the differences for faint

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).

Fix bulge fit Free bulge fit	$\mathcal{S}_{r_{\rm e,b}}$	$\mathcal{S}_{r_{\rm e,d}}$	$\mathcal{S}_{q_{\rm b}}$	$\mathcal{S}_{q_{\rm d}}$	$\mathcal{S}_{\rm b/t}$	$\mu_{\rm b}, \mu_{\rm d}$
`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

Notes. The last column is the mean of the parameters, first for the bulge components and b/t, second for the disc 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.

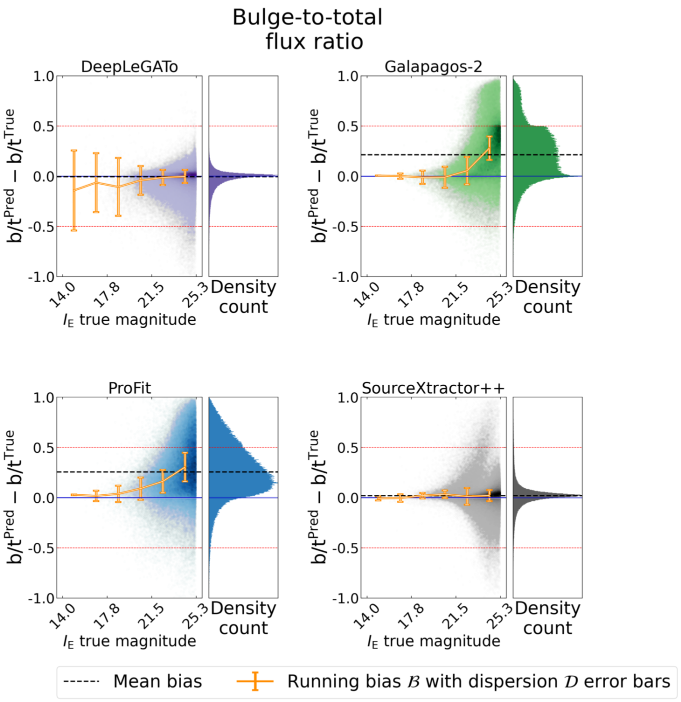

Fig. 7. Fitting results for the bulge-to-total flux ratio using the double-Sérsic simulation. See caption of Fig. 3 for further information.

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 $I_{\rm E} < 20$. `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 $I_{\rm E} < 22.5$. 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 $I_{\rm E} = 23$, 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 ($I_{\rm E} < 23.5$), `Galapagos-2` maintains the lowest number of outliers, from close to zero to around 10%, while `SourceXtractor++` goes up to ∼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 S/N, `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,

$$|\tilde{\mathcal{B}}_{r_{\rm e}}| = \overline{|\tilde{\boldsymbol{b}}_{r_{\rm e}}|}\ ,$$

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 $|\mathcal{B}_p|$ (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. We note that we limit the magnitudes to faint galaxies ($I_{\rm E} > 18.5$) and that for `ProFit` and `Galapagos-2`, the colourbars 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

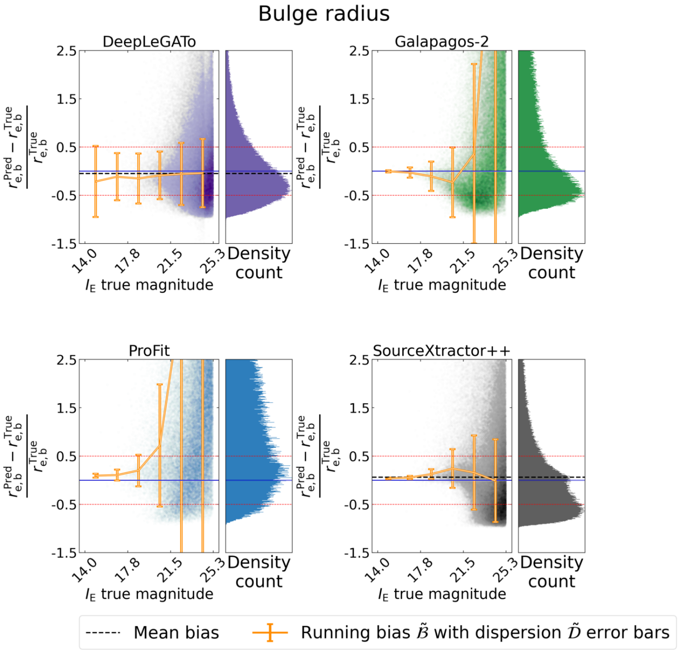

Fig. 8. Fitting results for the bulge radius using the double-Sérsic simulation. Notice that only four codes provided results for the double-Sérsic simulation. From top to bottom and from left to right: `DeepLeGATo`; `Galapagos-2`; `ProFit`; and `SourceXtractor++`. See caption of Fig. 3 for further information.

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, for example removing the extreme cases, for a better visualisation.

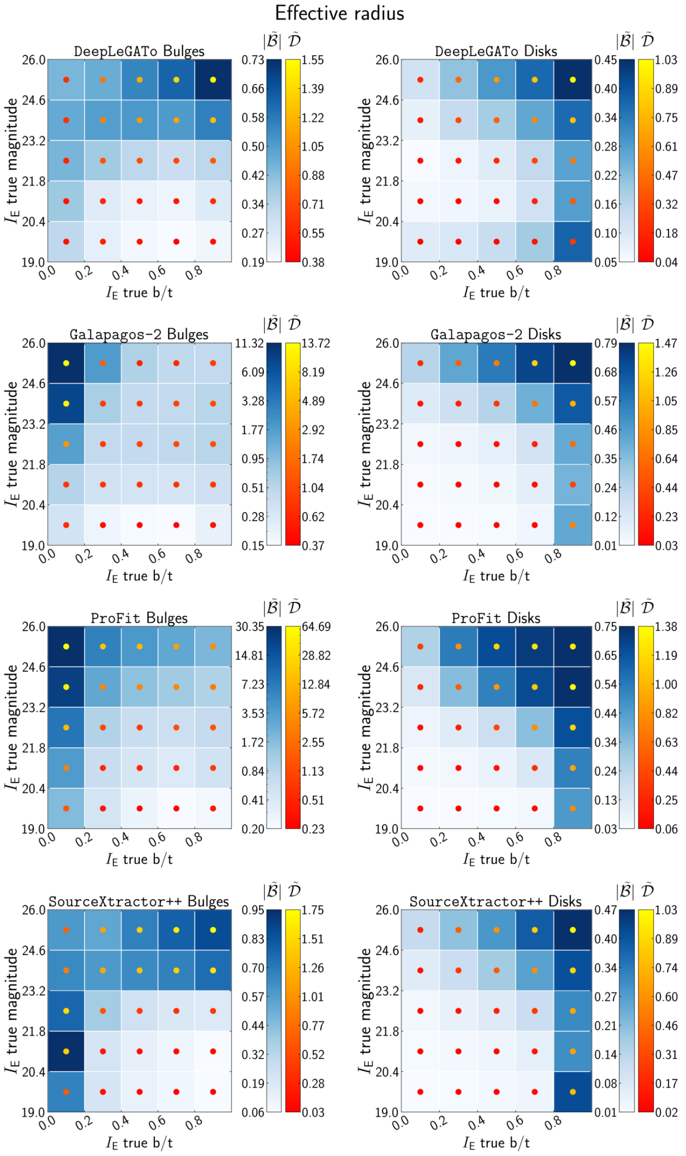

Fig. 9. Absolute bias $|\tilde{\mathcal{B}}|$ and dispersion $\tilde{\mathcal{D}}$ for the effective radius of bulge (left column) and disc (right column) components in the double-Sérsic simulation, as a function of bulge-to-total ratio (x-axis) and apparent I_{E} magnitude (y-axis). Each row shows a 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.

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 under-estimation 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

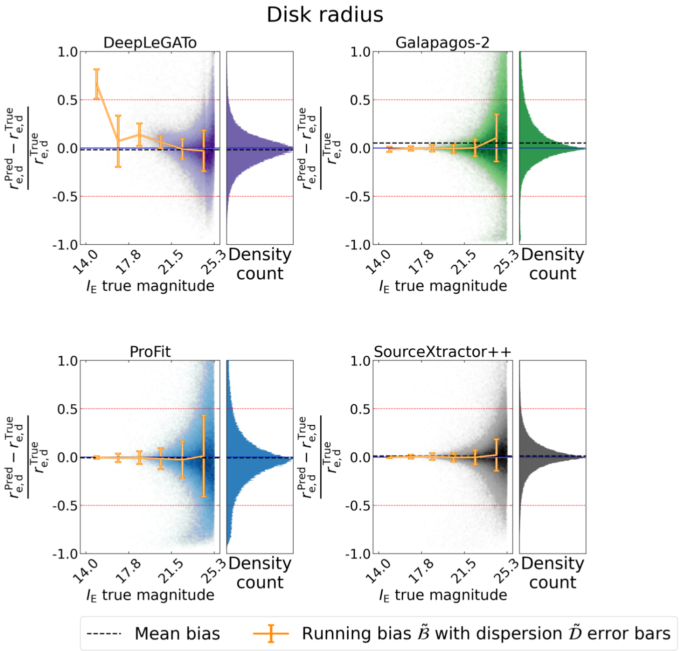

Fig. 10. Fitting results for the disc radius using the double-Sérsic simulation. See caption of Fig. 3 for further information.

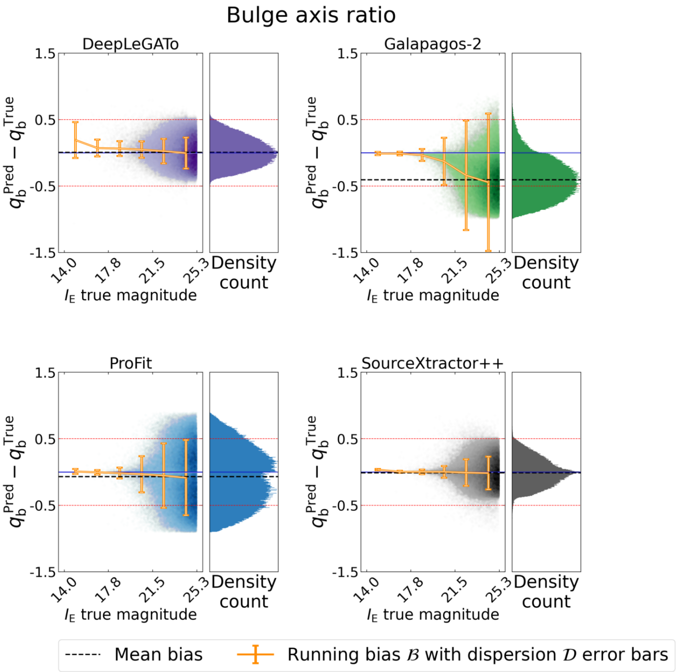

Fig. 11. Fitting results for the bulge axis ratio using the double-Sérsic simulation. See caption of Fig. 3 for further information.

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 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 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, that is 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 values. 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

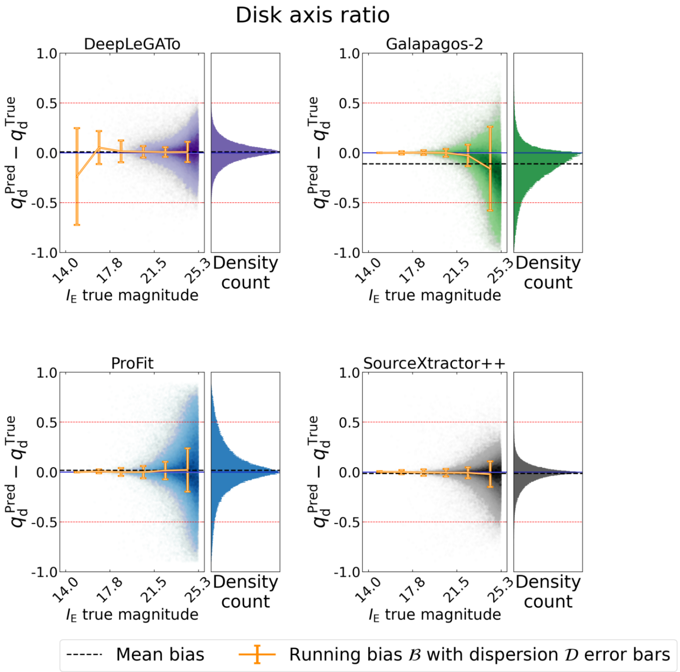

Fig. 12. Fitting results for the disc axis ratio using the double-Sérsic simulation. See caption of Fig. 3 for further information.

few percent. From magnitudes18 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 EMC2023, 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 S/N, lower resolution, noise correlation, and artefacts 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. We note that the results are only shown 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 S/N 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. We note that ProFit has the option for a multi-band joint analysis, but this mode was not used for the challenge.

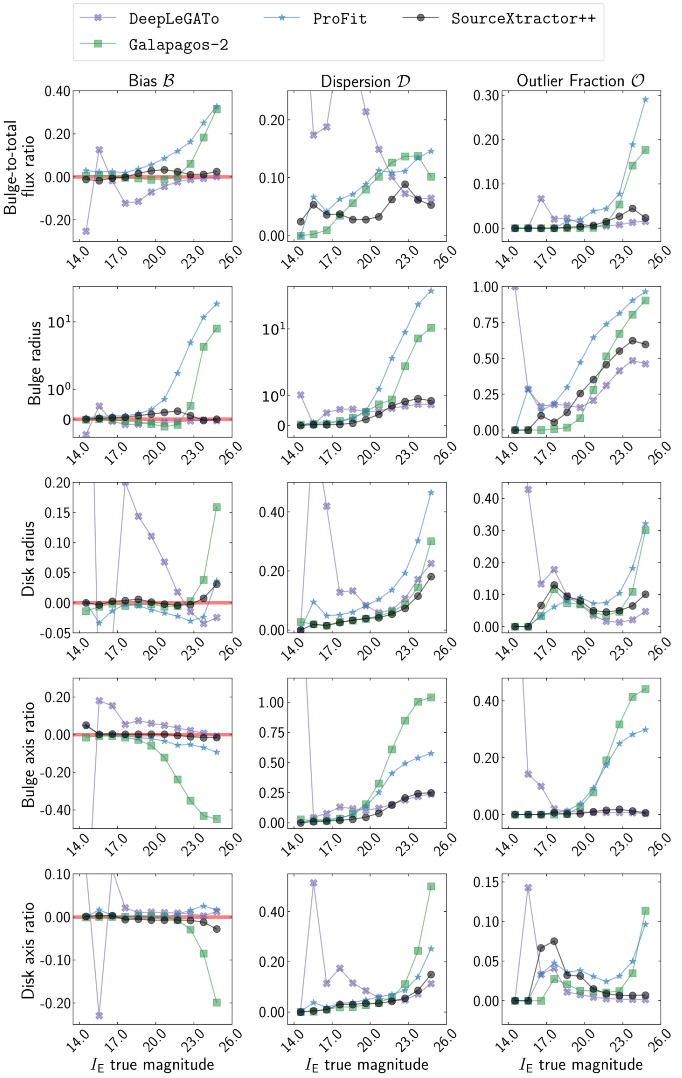

Fig. 13. Summary plot for the double-Sérsic simulations. From top to bottom: bulge effective radius, disc effective radius, bulge axis ratio, disc axis ratio, and bulge over total flux ratio. See caption of Fig. 6 for further information.