

Simulation-based cosmological inference from optically selected galaxy clusters with Capish

Constantin Payerne^{1,★}, Calum Murray^{1,2}, and Hugo Simon¹

¹ Université Paris-Saclay, CEA, IRFU, 91191 Gif-sur-Yvette, France

² Université Paris Cité, CNRS-IN2P3, APC, 75013 Paris, France

Received 5 February 2026 / Accepted 15 May 2026

ABSTRACT

Galaxy clusters are powerful probes of the growth of cosmic structure through measurements of their abundance as a function of mass and redshift. Extracting precise cosmological constraints from cluster surveys is challenging, as we must contend with nontrivial correlations between lensing mass and optical richness, as well as the complex relationship between richness and the underlying halo mass. These difficulties are compounded by systematic effects such as selection function biases, super-sample covariance, and correlated measurement noise between mass proxies. As upcoming photometric surveys are expected to detect tens to hundreds of thousands of galaxy clusters, controlling these systematics becomes essential. In this paper, we present a forward-modeling approach using simulation-based inference (SBI), which provides a natural framework for jointly modeling cluster abundance and lensing mass observables while capturing systematic uncertainties at higher fidelity than analytic likelihood methods – which rely on simplifying assumptions such as fixed covariances and Gaussianity – without requiring an explicit likelihood formulation. We introduce *Capish*, a Python code for generating forward-modeled galaxy cluster catalogs using halo mass functions and incorporating observational effects. We perform SBI using neural density estimation with normalizing flows, trained on abundance and mean lensing mass measurements in observed redshift–richness bins. Key cluster-related summary statistics measured on *Capish* simulations faithfully reproduce their corresponding analytical predictions, and we perform several Bayesian robustness tests of posterior modeling. Our forward model accounts for realistic noise, redshift uncertainties, selection functions, and correlated scatter between lensing mass and observed richness. We find good agreement with explicit-likelihood analyses, with broader SBI posteriors reflecting the increased realism of the forward model. We also test *Capish* on cluster catalogs built from a large cosmological simulation, finding a good fit to the cosmological parameters.

Key words. methods: statistical – galaxies: clusters: general – cosmological parameters

1. Introduction

Galaxy clusters form through the gravitational collapse of large matter density fluctuations and emerge as the largest gravitationally bound objects at the intersection of the cosmic web filaments. Therefore, their formation history, mass, and spatial distribution are highly connected to the fluctuations of the matter density field, the expansion rate of the Universe, and the nature of gravity (e.g., Bartlett 1997; Allen et al. 2011; Kravtsov & Borgani 2012).

Constraining cosmological parameters through the abundance of galaxy clusters relies on connecting the observed cluster count to the underlying mass and redshift distribution of the most massive dark matter halos. Cluster masses are not directly observable. With optical surveys, clusters are detected through the density of their member galaxies (Rykoff et al. 2014), and each cluster is assigned a richness, which is linked to the number of the cluster’s galaxy members. The constraining power of cluster number counts is currently limited by our understanding of the cluster scaling relations (Pratt et al. 2019), linking the underlying cluster mass with what we observe.

Weak gravitational lensing (Bartelmann & Schneider 2001) has become a robust tool for constraining cluster masses (e.g., McClintock et al. 2019; Umetsu 2020; Murray et al. 2022; Mistele & Durakovic 2024; Grandis et al. 2024; Murray et al.

2025), through the coherent distortion of the shapes of background galaxy images, caused by the bending of the light path due to the cluster’s gravitational field. When available, cluster lensing information has been crucial in tightening the constraints on the cluster mass-proxy relation (see, e.g., von der Linden et al. 2014; Penna-Lima et al. 2017).

Over the past two decades, the measurement of the abundance of galaxy clusters has provided competitive constraints on Ω_m and σ_8 , from optically detected clusters by the Dark Energy Survey¹ (DES; Abbott et al. 2020, 2025; To et al. 2021; Costanzi et al. 2021), the Kilo Degree Survey (KiDS; Lesci et al. 2022, 2025), the Sloan Digital Sky Survey (SDSS; Fumagalli et al. 2024; Park et al. 2023; Sunayama et al. 2024), from X-ray-detected clusters by ROSAT (Mantz et al. 2015), eROSITA (Ghirardini et al. 2024) and from clusters detected through the Sunyaev-Zeldovich (SZ) effect at millimeter wavelengths by the *Planck* satellite (Ade et al. 2016; Zubeldia & Challinor 2019; Salvati et al. 2022; Lee et al. 2025; Aymerich et al. 2024), the South Pole Telescope (SPT; Chaubal et al. 2022; Bocquet et al. 2024, 2025) or the Atacama Cosmology Telescope (ACT; Sehgal et al. 2011; Hasselfield et al. 2013). Large cluster surveys such as the Legacy Survey of Space and Time of the Vera Rubin Observatory (LSST Science Collaboration 2009) and the *Euclid* mission

¹ Although it should be noted that the Abbott et al. (2020) found a significant tension with the other DES analysis.

★ Corresponding author: constantin.payerne@gmail.com

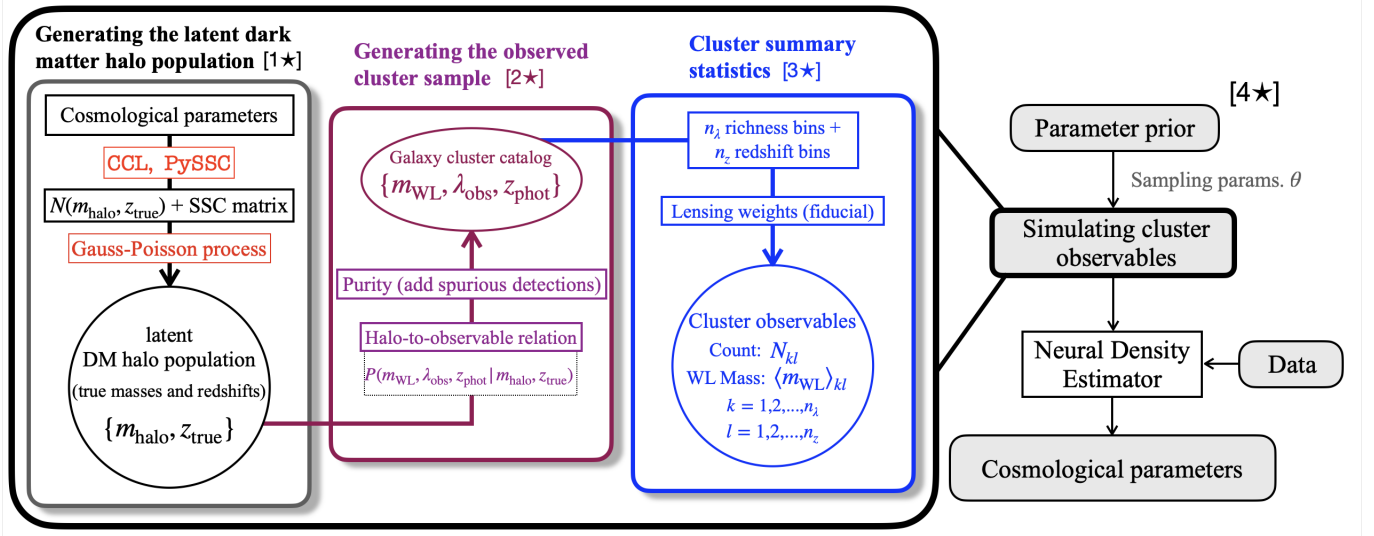


Fig. 1. Organization of Capish, detailed in Sect. 2. The code is organized in three separate blocks: The first block [1★] is dedicated to generate halo masses and true redshift from an underlying halo mass function, accounting for the effect of SSC (presented in Sect. 2.1) The second block [2★] is dedicated to compute the observed richness and lensing mass for halos, along with photometric redshift (presented in Sect. 2.2). The third block [3★] computes the summary statistics (presented in Sect. 2.3). [4★] denotes the overall SBI structure.

(Laureijs et al. 2011) will detect tens to hundreds of thousands of clusters and provide even stronger cosmological constraints. The cosmological constraining power of number counts of optically selected clusters is limited by a range of systematic effects that must be understood and controlled. On the theoretical side, these include uncertainties in the modeling of the dark matter density profile in the cluster field (Becker & Kravtsov 2011; Lee et al. 2018) and in the calibration of the halo mass function (Artis et al. 2021). Cluster number counts also deviate from a simple Poisson process (Poisson 1837), as fluctuations in the underlying matter density field introduce additional correlated scatter known as super-sample covariance (SSC; Hu & Kravtsov 2003; Gouyou Beauchamps et al. 2022), which constitutes an important source of variance for cluster count measurements (Fumagalli et al. 2021; Payerne et al. 2023, 2024). On the observational side, further systematics arise from the calibration of source galaxy shapes (Hernández-Martín et al. 2020), photometric redshift uncertainties in the background galaxy sample (Wright et al. 2020), contamination of the source sample by cluster member galaxies (Varga et al. 2019), and miscentering – offsets between the detected cluster center and the true dark matter halo center (Becker & Kravtsov 2011; Lee et al. 2018; Zhang et al. 2019; Sommer et al. 2022, Murray et al., in prep.). Additionally, selection effects from the cluster detection process can introduce correlations between lensing mass and observed richness that can significantly bias cluster abundance and lensing analyzes (Wu et al. 2022; Zhang et al. 2024; Salcedo et al. 2024).

As the volume and precision of cosmological datasets improve, simplified likelihood assumptions tend to break down², potentially compromising the robustness of inference. Working with next-generation survey data, which are not only high-volume but also high-precision, requires finer resolutions in parameter space and more nuanced modeling of systematics. To address these limitations, the community is increasingly turning to simulation-based inference (SBI), also known as implicit-

likelihood inference. From a methodological standpoint, SBI relies on (i) the generation of a synthetic dataset spanning a relevant range of cosmological and nuisance parameters and (ii) learning the probabilistic relationship between the simulated observables and the latent parameters.

Simulation-based inference has already been successfully applied to real datasets, for example, DES (Jeffrey et al. 2021; Gatti et al. 2025; Prat et al. 2026), KiDS (von Wietersheim-Kramsta et al. 2025), SDSS (Lemos et al. 2023b), and *Planck* cluster number counts (Zubeldia et al. 2025), showing good agreement with explicit-likelihood approaches while naturally accounting for stochastic effects beyond the limitations of Gaussian likelihood assumptions. Several proof-of-concept studies, validated on mock datasets (Reza et al. 2022, 2024; Kosiba et al. 2025; Cerardi et al. 2025; Regamey et al. 2026; Sáez-Casares et al. 2026), have explored the potential of SBI from X-ray, SZ, and optical cluster samples, paving the way for the broader adoption of likelihood-free inference techniques in the analysis of current and upcoming cluster surveys.

In this paper, we present Capish³ (Cluster abundance posterior inference from simulated halos), a Python code dedicated to performing the cosmological analysis of optically selected cluster catalogs in the era of the Vera C. Rubin LSST and the *Euclid* mission, using simulations. In Sect. 2 we present the different functionalities of the Capish code. Its diagram is shown in Fig. 1. We evaluate in Sect. 3 the internal calibration of Capish simulations and self-applied cosmological analyzes, as well as by using the *Euclid* Flagship dark matter halo catalog. We conclude in Sect. 4.

2. The Capish code

Capish is dedicated to (i) simulate analytically galaxy cluster catalogs, given cosmological parameters, survey characteristics, and some observational systematics, (ii) training neural density estimators (NDE), and (iii) performing the cosmological

² The underlying likelihood function is unknown, unavailable, or too expensive computationally.

³ <https://github.com/calumhrmurray/capish>

analyzes of cluster catalogs from the trained NDE. The code structure of Capish is presented in Fig. 1.

2.1. Generating dark matter halo catalogs

Capish generates a dark matter halo catalog from a halo mass function (e.g., [Tinker et al. 2008, 2010](#); [Despali et al. 2016](#)). We first compute the halo count prediction N_{ij} in narrow mass bins (between $\log_{10} m = 12$ to $\log_{10} m = 16.5$) with width $\Delta \log_{10} m_i \ll 1$ and redshift bins (between $z = 0$ to $z = 1.2$) with width $\Delta z_j \ll 1$, given as

$$N_{ij}^{\text{grid}} = \Delta \log_{10}(m_i) \Delta z_j n_h(\log_{10} m_i^{\text{center}}, z_j^{\text{center}}), \quad (1)$$

where $\log_{10} m_i^{\text{center}}$ (respectively, z_j^{center}) is the center of the i -th log-mass (respectively, j -th redshift) bin. In the above equation, $n_h(m, z)$ is the predicted total halo number density per mass and redshift range given by

$$n_h(\log_{10} m, z) = \Omega_S \frac{dn(\log_{10} m, z)}{d \log_{10} m} \frac{d^2 V(z)}{dz d\Omega}, \quad (2)$$

where $dn(m, z)/dm$ is the halo mass function of objects at redshift z and mass m , Ω_S is the survey solid angle and $V(z)$ the comoving volume.

The intrinsic clustering of the underlying matter density field induces a covariance in halo counts. This is often referred to as SSC ([Hu & Kravtsov 2003](#)), and is an important contribution to the variance of large-scale structure probes (particularly galaxy clusters, [Fumagalli et al. 2021](#); [Payerne et al. 2023](#)) as cosmological surveys increase in coverage and depth. The SSC is usually accounted for in the binned count analytic likelihood as an additional covariance term. The effect of SSC is as a correlated scattering of the per-bin halo count prediction in Eq. (1), such as

$$\tilde{N}_{ij}^{\text{grid}} = N_{ij}^{\text{grid}} (1 + \delta_{ij}^{\text{SSC}}), \quad (3)$$

where δ_{ij}^{SSC} is the scattering associated with the SSC, which satisfies $\langle \delta_{ij}^{\text{SSC}} \rangle = 0$ and $\langle \delta_{ij}^{\text{SSC}} \delta_{kl}^{\text{SSC}} \rangle = b_{ij} b_{kl} \sigma_{\text{SSC},jl}^2$, where b_{ij} is the halo bias at mass m_i and redshift z_j , and ([Lacasa et al. 2018, 2023](#))

$$\sigma_{\text{SSC},jl}^2 = \frac{4\pi}{\Omega_S} \int \frac{k^2 dk}{2\pi^2} j_0[kw(z_j)] j_0[kw(z_l)] P_{\text{mm}}(k|z_j, z_l) \quad (4)$$

is the (partial-sky) amplitude of matter density fluctuations⁴. In the equation above, $j_0(x) = \sin(x)/x$ is the zero-th order spherical Bessel function, and $P_{\text{mm}}(k|z_1, z_2)$ is the linear matter power spectrum. In the unbinned regime (given by $\Delta z_j \ll 1$; see, e.g., [Mantz et al. 2010](#); [Penna-Lima et al. 2014](#)), the variance of δ_{ij}^{SSC} becomes maximal, such as the occurrence $\delta_{ij}^{\text{SSC}} < -1$ becomes statistically significant if δ_{ij}^{SSC} is considered to be a Gaussian variable. This could yield to unphysical halo counts through Eq. (3) ([Payerne et al. 2024](#)). In this regime, it is more appropriate to consider δ_{ij}^{SSC} as a log-normal random variable ([Coles & Jones 1991](#); [Wen et al. 2020](#)), such as $\ln(1 + \delta^{\text{SSC}}) \sim \mathcal{N}(\mu^{\text{SSC}}, \Sigma^{\text{SSC}})$ where

$$\mu_{ij}^{\text{SSC}} = -\frac{1}{2} \ln(1 + b_{ij}^2 \sigma_{\text{SSC},jj}^2), \quad (5)$$

$$\Sigma_{ijkl}^{\text{SSC}} = \ln(1 + b_{ij} b_{kl} \sigma_{\text{SSC},jl}^2), \quad (6)$$

⁴ We use PySSC ([Lacasa & Grain 2019](#); [Gouyou Beauchamps et al. 2022](#)) to compute $\sigma_{\text{SSC},jl}$ (we use the PySSC function `sigma2_fullsky`, then rescaled by the sky fraction $\Omega_S/4\pi$). The code is available at <https://github.com/fabienlacasa/PySSC>.

such that the aforementioned conditions for $\langle \delta_{ij}^{\text{SSC}} \rangle$ and $\langle \delta_{ij}^{\text{SSC}} \delta_{kl}^{\text{SSC}} \rangle$ are satisfied. Finally, the number of halos is sampled from a Poisson distribution $\mathcal{P}$ such as

$$\tilde{N}_{ij}^{\text{grid}} \sim \mathcal{P}[\tilde{N}_{ij}^{\text{grid}}]. \quad (7)$$

The resulting dark matter halo catalog is obtained by collecting $\tilde{N}_{ij}^{\text{grid}}$ times samples drawn randomly inside the mass-redshift bin, with centers m_i and z_j .

In the upper panel of Fig. A.1, we show the ratio between (i) the mean of 200 simulated dark matter halo counts with Capish in large mass-redshift bins (for display purposes, to compare with binned prediction) and (ii) the associated theoretical prediction N_{ij}^{true} , computed as the two-dimensional integral of the halo density in Eq. (2) within the same large mass-redshift bins. The ratio is close to one for most mass scales, except at high mass, where the measured halo abundance is smaller than the cosmological prediction because of the finite volume of the simulated catalog. Additionally, we compute the dispersion of halo number counts in the same mass-redshift bins, which we compare to the theoretical variance given by⁵ ([Takada & Spergel 2014](#); [Lacasa & Grain 2019](#))

$$\text{Var}(N_{ij}) = N_{ij}^{\text{true}} + (N_{ij}^{\text{true}})^2 \langle b_{ij} \rangle^2 S_{jj} \quad (8)$$

as the sum of a Poisson term and an SSC term, where $\langle b_{ij} \rangle$ is the mean halo bias in the mass-redshift bin and S_{kl} is given by⁶ ([Lacasa et al. 2018](#))

$$S_{kl} = \int_{z_k}^{z_{k+1}} \int_{z_l}^{z_{l+1}} \frac{dV(z') dV(z'')}{V_k V_l} \sigma_{\text{SSC}}^2(z', z''), \quad (9)$$

where $dV(z)$ is the comoving volume per steradian (V_k is the total volume in the k -th redshift bin). In the bottom panel of Fig. A.1, we see that the measured count dispersion has a good agreement with analytic count variance at most mass scales. We also see that SSC is important for lower mass halos; therefore, from this example, we show that our implementation of SSC in our simulation-based framework is in agreement with the behavior of SSC in the analytical likelihood analysis. The resulting dark matter halo catalog serves as a basis to generate the catalog of observed galaxy clusters.

2.2. Generating the galaxy cluster catalog

2.2.1. From halo properties to cluster observables

The dark matter halo catalogs have to be connected to observational cluster properties, such as the observed richness and the weak gravitational lensing signal. We use a general multivariate Gaussian model ([Evrard et al. 2014](#); [Zhang et al. 2024](#); [Payerne et al. 2025](#)) for the observed richness λ_{obs} and a lensing observable, corresponding to the lensing mass $\log_{10} m_{\text{WL}}$, i.e.,

$$\ln \lambda_{\text{obs}}, \log_{10} m_{\text{WL}} \sim \mathcal{N}[\text{mean}(m, z), \text{Cov}(m, z)] \quad (10)$$

where

$$\text{mean}(m, z) = \{\langle \ln \lambda_{\text{obs}} | m, z \rangle, \langle \log_{10} m_{\text{WL}} | m, z \rangle\} \quad (11)$$

and

$$\text{Cov}(m, z) = \begin{pmatrix} \sigma_{\ln \lambda}^2 & \rho \sigma_{\ln \lambda} \sigma_{\text{WL}} \\ \rho \sigma_{\ln \lambda} \sigma_{\text{WL}} & \sigma_{\text{WL}}^2 \end{pmatrix}. \quad (12)$$

⁵ We use the [Tinker et al. \(2010\)](#) halo bias for the SSC part.

⁶ We use the `Sij_fullsky` function of PySSC, then rescaled to the sky fraction $\Omega_S/4\pi$.

We define the cluster richness λ_{obs} as related to the count of cluster's member galaxies – intrinsically linked to the cluster's formation history – evaluated, for example, within a circular aperture centered on the detected cluster position (à la redMaPPer, Rykoff et al. 2014). The observed richness is linked to the specific selection of these galaxies in photometric surveys; it depends on (i) the cluster detection algorithm to define and measure the cluster's richness⁷ (Rykoff et al. 2014; Bellagamba et al. 2018; Aguena et al. 2021) (ii) observational and/or detection noise. A power-law relation (see, e.g., Mantz et al. 2008; Evrard et al. 2014; Saro et al. 2015; Farahi et al. 2018; Murata et al. 2019; Anbajagane et al. 2020) is commonly adopted to connect the mean cluster's observed (log-) richness $\ln \lambda_{\text{obs}}$ with its halo mass and redshift, via

$$\langle \ln \lambda_{\text{obs}} | m, z \rangle = \mu_0^\lambda + \mu_m^\lambda \log_{10} \left(\frac{m}{m_0} \right) + \mu_z^\lambda \log \left(\frac{1+z}{1+z_0} \right). \quad (13)$$

Two contributions affect the measurement scatter of the cluster's richness. First, the intrinsic scatter ($\sigma_{\ln \lambda, \text{int}}$) denotes the contribution of the intrinsic cluster's formation history. Second, Poisson sampling arises as a statistical process on its own, associated with the observation, since richness is a “count-in-cell” observable. In this work, we use (Farahi et al. 2018; Zhang et al. 2023; To et al. 2025)

$$\sigma_{\ln \lambda}^2 = \sigma_{\ln \lambda, \text{int}}^2 + \exp\{-\langle \ln \lambda_{\text{obs}} | m, z \rangle\}, \quad (14)$$

where the first term denotes the intrinsic scatter and the second the Poisson dispersion⁸ of the richness measurement. This is currently the primary weakness of Capish, as the actual relation between the halo mass and the observed cluster richness is likely much more complex.

Galaxy cluster lensing masses are primarily inferred from the weak gravitational lensing effects of the cluster's gravitational potential, inducing a small distortion in the shapes of background galaxies. This is generally estimated from the excess surface density (Murray et al. 2022), which is taken to be the average in the projected radial bins around the cluster center of the weighted background ellipticities $\widehat{\Delta\Sigma}(R) = \langle \Sigma_{\text{crit}}(z_s, z_l) \epsilon_+^{ls} \rangle(R)$ where $\Sigma_{\text{crit}}(z_s, z_l)$ is the critical surface mass density – a geometrical lensing factor – and ϵ_+^s is the tangential ellipticity of a source at $z_s > z_{\text{cl}}$ within the radial bin. The excess surface density in radial bins is not the unique cluster weak lensing signal estimator that is used in the literature; one can also use binned reduced shear (Becker & Kravtsov 2011), two-dimensional lensing maps (Oguri et al. 2010, Murray et al., in prep.), or weak lensing shear magnification (Murray et al. 2025). The measured excess surface density profile probes $\Sigma(R|m)$ are the projected three-dimensional matter density $\rho_{\text{cl}}(\mathbf{x})$ in the cluster field. The cluster lensing mass m_{WL} (along with other halo parameters such as its concentration) is obtained by fitting the lensing profile with the appropriate model for the matter density

around the cluster center⁹ (McClintock et al. 2019). The lensing mass $m \rightarrow m_{\text{WL}}$ could be biased with respect to its true underlying halo mass, so we use the power-law relation

$$\langle \log_{10} m_{\text{WL}} | m, z \rangle = \mu_0^{\text{WL}} + \mu_m^{\text{WL}} \log_{10} m + \mu_z^{\text{WL}} \ln \left(\frac{1+z}{1+z_0} \right). \quad (15)$$

In the above equation, any deviation from $\mu_0^{\text{WL}} = 0$, $\mu_m^{\text{WL}} = 1$, and $\mu_z^{\text{WL}} = 0$ indicates that the lensing masses are biased with respect to the true halo mass, which may happen if the underlying dark matter density profile is not accurately representing lensing data.

The lensing profile $\widehat{\Delta\Sigma}(R)$ is a scattered measurement, arising from several independent contributions (Hoekstra 2001, 2003; Becker & Kravtsov 2011; Gruen et al. 2015; Wu et al. 2019). Its dominant noise term ($\sigma_{\text{WLgal}}^{\Delta\Sigma}$) arises from intrinsic galaxy shape dispersion and limited source galaxy samples for the lensing profile estimation, scaling as $\sigma_\epsilon / \sqrt{\bar{n}_{\text{gal}}(z)}$, where $\sigma_\epsilon^2 = \sigma_{\epsilon, \text{meas.}}^2 + \sigma_{\epsilon, \text{SN}}^2$ being the dispersion of galaxy shapes due to (i) intrinsic galaxy shape noise ($\sigma_{\epsilon, \text{SN}} \approx 0.25$, Chang et al. 2013), (ii) errors in shape measurement algorithms ($\sigma_{\epsilon, \text{meas.}}$), and $\bar{n}_{\text{gal}}(z)$ the surface number density of lensed background galaxies for a cluster at redshift z . Additional noise-independent of source density – comes from the intrinsic dispersion in halo morphology¹⁰ ($\sigma_{\text{WLint}}^{\Delta\Sigma}$) and the impact of correlated and/or uncorrelated large-scale structure ($\sigma_{\text{WLCSS}}^{\Delta\Sigma}$ and $\sigma_{\text{WLuLSS}}^{\Delta\Sigma}$). On the observational side, further systematics arise from shape measurement errors (Hernández-Martín et al. 2020), photometric-redshift uncertainties (Wright et al. 2020), cluster-member contamination (Varga et al. 2019), and miscentering (Zhang et al. 2019; Sommer et al. 2022). In summary, the total variance $\sigma_{\text{WL}}^{\Delta\Sigma}$ of the measured excess surface density profile $\Delta\Sigma(R)$ is given by the quadratic sum of these terms. The lensing profile variance depends on the underlying halo mass, its redshift, the cluster's local environment, and finally, the lensing survey characteristics. In this work, we rely on the lensing mass; its variance – as for the lensing profile – can be decomposed as

$$\sigma_{\text{WL}}^2 = \sigma_{\text{WLgal}}^2 + \sigma_{\text{WLint}}^2 + \sigma_{\text{WLCSS}}^2 + \sigma_{\text{WLuLSS}}^2. \quad (16)$$

The above depends on the number of parameters (and correlations) that are fitted jointly with the lensing mass (see Appendix B). For simplicity, in this work we only consider the σ_{WLgal}^2 component, and leave the implementation of the other terms for future works, which can be obtained separately from the theory (Shirasaki & Takada 2018) of dedicated simulations (Becker & Kravtsov 2011; Wu et al. 2019, Murray et al. in prep.). As the number density of sources approaches that expected from space-based observations or LSST conditions (e.g., DES provided $\bar{n}_{\text{gal}} = 10 \text{ arcmin}^{-2}$, where LSST will provide $\bar{n}_{\text{gal}} = 40 \text{ arcmin}^{-2}$), the contribution to the scatter in the weak lensing masses from galaxy shape noise becomes comparable or even subdominant to the intrinsic scatter terms and large-scale structures (Hoekstra 2001; Becker & Kravtsov 2011). A fixed parameter can be used, but Capish also provides an information matrix-based analytical model $\sigma_{\text{WLgal}}(m, z)$ as explained in Appendix B, for improved realism, as shown in Fig. B.1, accounting for (i) the cluster's redshift-dependent number density of source galaxies, (ii) the theoretical modeling for

⁷ Generally scaling with, but not equal to, the number counts of cluster member galaxies. For instance, Rykoff et al. (2014) defines the richness as the sum of membership probabilities for red-sequence selected galaxies in the vicinity of a cluster, where Aguena et al. (2021) defines clusters as the overlap of multiple galaxies' photometric redshift distributions happening at the same position on the sky.

⁸ Most papers use $(\exp(\langle \ln \lambda_{\text{obs}} | m, z \rangle) - 1) / \exp(2 \langle \ln \lambda_{\text{obs}} | m, z \rangle)$, which is a good approximation for large richness values. Since we want our code to be stable over the full richness range, we use Eq. (14).

⁹ We generally consider that the lensing signal is dominated by the 1-halo term at scales below 3 Mpc, originating from the cluster itself, and the 2-halo term becomes increasingly important at larger scales, originating from neighboring halos.

¹⁰ Because most of the time we use a simplified spherical model to constrain the mass of nonspherical clusters.

$\Delta\Sigma(R)$, and (iii) the considered radial fitting range. In this error model, the scatter is larger for low-mass halos (whose weak lensing signals have a lower signal-to-noise ratio) or high-redshift halos (which have fewer source galaxies to measure their lensing profiles).

Within *Capish*, it is also possible instead to add lensing mass scatter at the summary statistics level (the detailed description of the summary statistics we used in this study – evaluated in bins of richness and redshift – is discussed later in Sect. 2.3). First, for each i -th halo in the halo catalog, we computed its lensing mass $\log_{10} m_{\text{WL},i} = \langle \log_{10} m_{\text{WL}} | m_i, z_i \rangle$ by using Eq. (15) (i.e., no individual lensing mass dispersion). Then, for an ensemble (or stack) of N_{cl} galaxy clusters within a given richness-redshift bin, the mean halo mass within the stack is scattered accordingly with $\sigma_{\text{WL}}^{\text{stack}} = \sigma_{\text{WL}} / \sqrt{N_{\text{cl}}}$. This is closer to the modeling used in recent cluster cosmology analyzes and weak lensing mass-relation calibration techniques (Abbott et al. 2020; Lesci et al. 2025; Payerne et al. 2025), where the mean mass of clusters is inferred from a clusters’ stacked weak lensing profiles¹¹.

For optically selected clusters, an important systematic is the covariance between the lensing observable and the observed richness that is measured by cluster finders. This correlation coefficient is effective, with no distinction between intrinsic and extrinsic scatters (Wu et al. 2019; Zhou et al. 2024); the intrinsic part of this covariance reflects galaxy assembly bias tied to secondary halo properties such as concentration, mass accretion rate, or dynamical state (Zhang et al. 2024). The extrinsic part arises from selection biases within cluster-finding algorithms, shaped by observational systematics, including photometric-redshift errors (Graham et al. 2018), shear shape noise (Wu et al. 2019), projection and percolation effects (Costanzi et al. 2019), triaxiality bias (Zhang et al. 2023), and additional complexities associated with realistic cluster finders such as redMaPPer (Rykoff et al. 2014). This covariance induces additive biases that cannot be mitigated by increasing the source density or reducing shape noise (Nord et al. 2008; Evrard et al. 2014; Farahi et al. 2018; Wu et al. 2019; Nde et al. 2026), but must be accurately quantified to achieve a percent-level mass calibration (Rozo et al. 2014; Zhou et al. 2024). Salcedo et al. (2025) showed that projection effects that impact weak-lensing measurements of DES Y1 clusters can, if not properly accounted for, induce significant biases in inferred cosmological parameters (Abbott et al. 2020). These effects can be efficiently modeled using a simulation-based forward approach to stacked lensing observables. In *Capish*, we adopt a single parameter, $\rho = \text{Corr}(\log_{10} m_{\text{WL}}, \ln \lambda_{\text{obs}} | m, z)$, to describe the overall correlation between the lensing observable and the richness observed assigned by the cluster finder at a fixed true mass and redshift. We further allow ρ to vary with mass and redshift, as selection biases–arising from the mechanisms discussed above–are expected to depend on local cluster properties and environment and therefore lead to a mass-dependent correlation between cluster lensing and cluster richness.

2.2.2. Cluster photometric redshifts

Individual observed clusters’ redshifts are usually derived from their member galaxies (see, e.g., Rykoff et al. 2014). It is linked to the redshift of its underlying dark matter halo through

$$z_{\text{phot}} \sim \mathcal{N}(z + b(z), \sigma_{\text{PZ}}(z)^2), \quad (17)$$

¹¹ In this case, there is no difference in considering $\log_{10} m_{\text{WL}}$ or m_{WL} as the random Gaussian lensing variable.

where $b(z)$ represents a possible systematic bias, and $\mathcal{N}(0, \sigma_{\text{PZ}}^2)$ denotes a Gaussian random variable with zero mean and redshift-dependent scatter $\sigma_{\text{PZ}}(z) = \sigma_{\text{PZ},0}(1 + z)$, and depends on the optical survey characteristics, depth, and so on, setting $\sigma_{\text{PZ},0}$ at $z = 0$. This functional form captures the typical increase in uncertainty at a higher redshift due to observational limitations. It is, however, simplistic (although sufficient for this proof-of-concept study) since in practice cluster redshift estimates are affected by additional systematic effects, including filter transitions of key spectral features used in redshift reconstruction. For example, the 4000 Å break shifting between the g and r bands at $z \sim 0.4$ can impact both the inferred cluster redshift and its uncertainty (see, e.g., Kluge et al. 2024). Moreover, the uncertainty model depends on the number and wavelength coverage of the available filters. In particular, low-redshift reconstruction ($z < 0.05$) could be significantly improved with LSST u -band observations (Ivezić et al. 2019), which extend to shorter wavelengths than the current blue limit set by the g band.

2.2.3. Cluster selection function

To account for the cluster selection function arising from the performance of the cluster finder, the ideal framework would forward model the full pipeline: galaxy catalog, cluster detection algorithm, and richness assignment. This is beyond our current scope. We aim, then, to include cluster finding characteristics in the simulated cluster catalog. The first effect in cluster finding is the presence of spurious detection, encoded in the cluster finder purity; optical surveys suffer from artifacts–unmasked stellar diffraction spikes, globular clusters misidentified as galaxy overdensities, line-of-sight projections–that inject spurious detections into the catalog. We therefore include a purity model in *Capish* to account for these false positives. We define purity $p(\lambda_{\text{obs}}, z_{\text{phot}})$ as the fraction of detected objects that correspond to genuine clusters. Following Agüena & Lima (2018), we adopt the form

$$p(\lambda_{\text{obs}}, z_{\text{obs}}) = \frac{(\lambda_{\text{obs}}/\lambda_p)^{n_{\text{pur}}(z_{\text{obs}})}}{1 + (\lambda_{\text{obs}}/\lambda_p)^{n_{\text{pur}}(z_{\text{obs}})}}, \quad (18)$$

where n_{pur} and λ_p set the shape and scale of the purity function. This parameterization ensures purity approaches unity at high richness while allowing significant contamination at low richness, consistent with the behavior of real cluster finders. For each richness-redshift bin, we first compute the histogram of true detections $\hat{N}_{\text{true}}(\lambda_{k,\text{obs}}, z_{l,\text{phot}})$, then draw the number of spurious detections from

$$\hat{N}_{kl,\text{fake}} \sim \mathcal{P}\left(\hat{N}_{\text{true}}(\lambda_{k,\text{obs}}, z_{l,\text{phot}}) \times \frac{1 - p(\lambda_{k,\text{obs}}, z_{l,\text{phot}})}{p(\lambda_{k,\text{obs}}, z_{l,\text{phot}})}\right). \quad (19)$$

Each spurious detection is assigned a richness and redshift drawn uniformly within its bin. We set the lensing mass of these objects to zero, reflecting the assumption that they produce a negligible weak lensing signal¹².

Another feature in the cluster selection function is the cluster catalog completeness, denoting the underlying halo population that is effectively probed by the detected galaxy clusters.

¹² Any identified structure would inevitably produce a positive lensing signal, as it may consist of a low-mass population or result from the cumulative contribution of projected structures along the line of sight; however, this has not been explored extensively in the literature, and this is not the purpose of this work, so we keep using the aforementioned assumption.

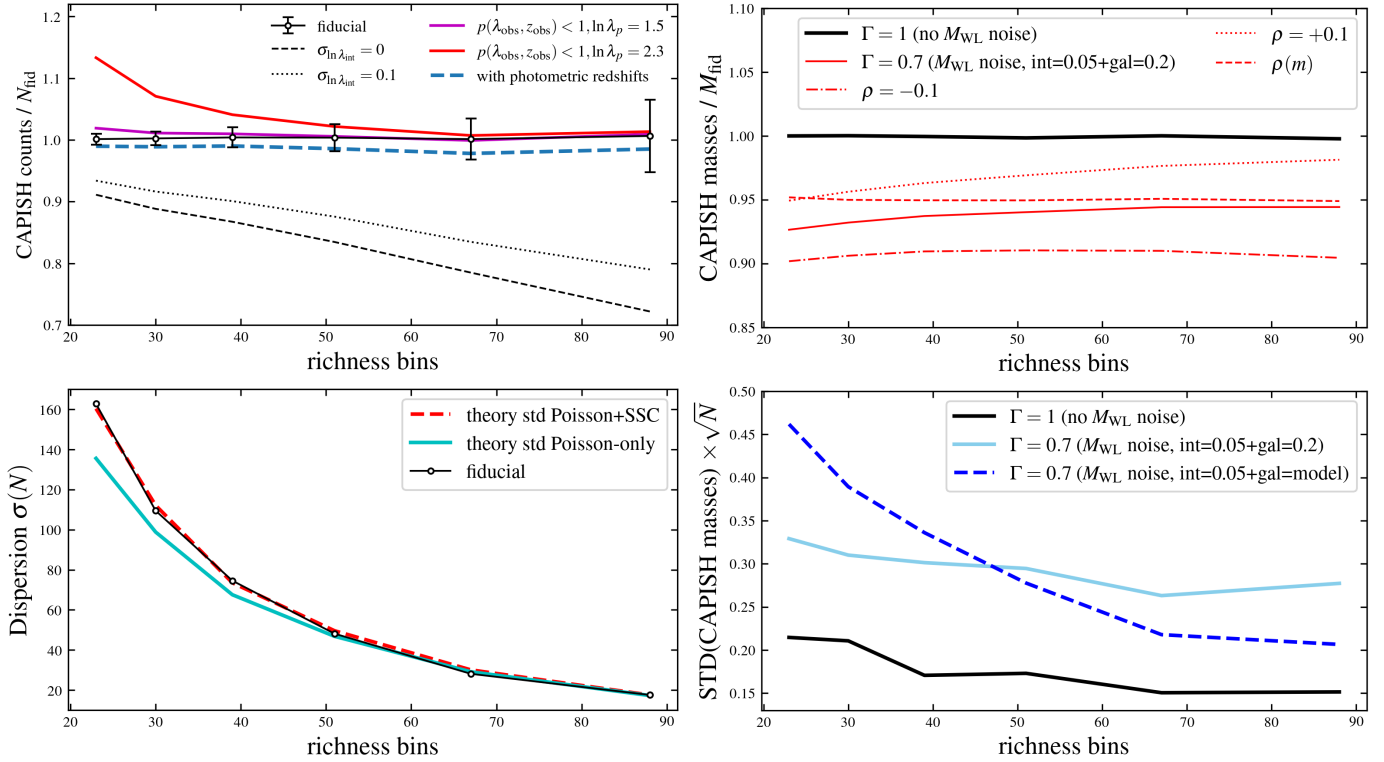


Fig. 2. *Top left:* CapiSH mean count outputs, averaged over 100 simulations, compared to the theoretical prediction for the mean cluster count in redshift richness bins. *Bottom left:* Standard deviation of the same counts over the 100 simulations. *Top right:* CapiSH mean mass outputs, averaged over 100 simulations, compared to the theoretical prediction for the mean cluster masses in redshift-richness bins. *Bottom right:* Standard deviation of the same mean masses over the 100 simulations.

Working at the catalog level, imposing a low richness cut (e.g., $\lambda_{\text{obs}} > 20$) restricts the probed halo masses above a certain value, determined by the richness-mass relation (and its scatter) in Eq. (13). As an illustration, Fig. C.1 shows how completeness (given by the ratio between the detected clusters that are true halos and the true halo population) varies with halo mass for different richness thresholds and intrinsic scatter values. We see that lower thresholds and smaller scatter both steepen the selection function and push it to lower masses. As a result, the effect of completeness is inherent in the presence of a statistical mass-richness relation of the cluster catalog.

Although not implemented here, Appendix D discusses an alternative completeness definition. Following Rozo et al. (2010), Aguena & Lima (2018), Euclid Collaboration: Adam et al. (2019), Lesci et al. (2022, 2025), Payerne et al. (2025), completeness can be modeled as a mass- and redshift-dependent function $c(m, z)$, independent of the mass-richness relation in Eq. (13). It quantifies the fraction of halos missed by the cluster finder due to intrinsic detection limitations, rather than the effective mass threshold of the richness cut (see Fig. C.1). Incorporating $c(m, z) < 1$ would reduce the predicted cluster counts, in contrast to purity, which increases them.

2.3. Cluster count and cluster lensing summary statistics

CapiSH returns two summary statistics:

- Count in observed richness-redshift bins: The cluster abundance summary statistic is given by

$$\widehat{N}_{kl} = \sum_{cl} \mathbb{1}\{cl \in (k, l)\}, \quad (20)$$

which gives the measured counts in bins of observed richness and photometric redshifts. In the above equation, $\mathbb{1}\{cl \in (k, l)\} = 1$ if the cluster belongs to the (k, l) richness-redshift bin, and 0 otherwise.

- Mean mass in observed richness-redshift bins: The mean mass in a given richness-redshift bin is given as the mean of individual lensing masses. However, the usual way to recover the mean lensing mass of clusters within the kl -th observed richness-observed redshift bins is to fit the corresponding stacked lensing profile $\widehat{\Delta\Sigma}_{kl}$ in the same kl richness-redshift bin (see, e.g., Murata et al. 2019). From this, the link between the “true” mean mass of a stack and the mean mass inferred from the weak lensing stacked profile $\widehat{\Delta\Sigma}_{kl}$ may not be trivial and may be affected by the dispersion of individual masses within the stack (McClintock et al. 2019). Melchior et al. (2017) proposed that since $\Delta\Sigma(R|M) \propto M^{\Gamma(R|M)}$ (at first order) where $\Gamma(R|M)$ is the logarithmic slope of the excess surface density¹³, we use $\Gamma = 0.7$ in this work; the mean weak lensing mass inferred from $\widehat{\Delta\Sigma}_{kl}$ (Abbott et al. 2020; Lesci et al. 2022; Fumagalli et al. 2024) can be modeled and/or interpreted as

$$\widehat{M}_{kl} = \left(\frac{1}{\sum_{cl} W(z_{cl}) \mathbb{1}\{cl \in (k, l)\}} \sum_{cl} W(z_{cl}) m_{cl, \text{WL}}^{\Gamma} \mathbb{1}\{cl \in (k, l)\} \right)^{1/\Gamma} \quad (21)$$

¹³ For instance, Melchior et al. (2017) found a typical value of $\Gamma = 0.74$ for a Navarro-Frenk-White (NFW; Navarro et al. 1997) profile within the one-halo regime.

where, considering $\sigma_\epsilon^2(z_s) = \sigma_{\epsilon, \text{meas.}}^2(z_s) + \sigma_{\epsilon, \text{SN}}^2(z_s)$ (in this work, we set $\sigma_{\epsilon, \text{meas.}} = 0$), the weights $W(z_{\text{cl}})$ are the maximum signal-to-noise ratio weights (Sheldon et al. 2004) and are given by

$$W(z_{\text{cl}}) = \frac{1}{\int_{z_{\text{cl}}}^{\infty} dz'_s n_{\text{gal}}(z'_s)} \int_{z_{\text{cl}}}^{\infty} dz'_s n_{\text{gal}}(z'_s) \frac{\Sigma_{\text{crit}}^{-2}(z'_s, z_{\text{cl}})}{\sigma_\epsilon^2(z'_s)}, \quad (22)$$

where $n_{\text{gal}}(z)$ is the source galaxy redshift distribution, taken from Chang et al. (2013).

As stated in Sect. 2.2, a choice is to consider that (i) individual masses are not scattered in Eq. (12), and (ii) that the weak lensing mass scatter is instead applied at the stacked level, such as

$$\log_{10} \widehat{M}_{kl} \rightarrow \log_{10} \widehat{M}_{kl} + \widehat{\epsilon}_{kl}, \quad (23)$$

where $\widehat{\epsilon}_{kl} \sim \mathcal{N}(0, (\sigma_{\text{WL}, kl}^{\text{stack}})^2)$. The stack scatter is given

by $\sigma_{\text{WL}, kl}^{\text{stack}} = \sigma_{\text{WL}, kl} / \sqrt{\widehat{N}_{kl}}$, with $\sigma_{\text{WL}, kl}$ denoting the weak lensing mass scatter for a single cluster lying in the (k, l) richness-redshift bin. This quantity $\sigma_{\text{WL}, kl}$ is either fixed or evaluated using the error model for the mean mass of the stack (see Appendix B). This is useful if we want to match the lensing mass error bars that are obtained in a given analysis by just setting them manually.

The simulated cluster count may become zero in one or more richness-redshift bins for a given simulation. In that case, it is not possible to compute the appropriate mean cluster mass for the corresponding bins. As an output of Capish and as an input of the neural density estimators (see the next section), we use the “mass” summary statistics

$$\widehat{N} \widehat{M}_{kl} = \begin{cases} \widehat{N}_{kl} \times \widehat{M}_{kl} & \text{if } \widehat{N}_{kl} \neq 0 \\ 0 & \text{if } \widehat{N}_{kl} = 0, \end{cases} \quad (24)$$

where $\widehat{N}_{kl}$ is given in Eq. (20), and $\widehat{M}_{kl}$ is given in Eq. (21). So, Eq. (24) is set to 0 when the bin is unfilled, instead of being not defined. This enables us to smooth the summary statistics for extreme cases, which will facilitate the training of neural density estimators.

3. Validation

3.1. Validating Capish outputs

Before any cosmological analysis, we can explore how Capish outputs compare to the theoretical prediction of cluster counts and cluster masses (see Appendix H). For the redshift bin $0.2 < z < 0.5$ and in seven richness bins, we show in Fig. 2 (left panel, upper plot) the bias between the mean Capish-counts (averaged over 100 simulations) and an analytical count prediction in Eq. (H.1), which is computed without accounting for selection function with fiducial parameters listed in Table 1 (the underlying halo mass definition used throughout this work is $M_{200\text{m}}$, and we used the Tinker et al. (2010) halo mass function). For the simplest case (no selection function) Capish counts are in good agreement with the theoretical prediction. Accounting for the selection function¹⁴, we see that adding purity to Capish simulations affects the recovered counts at low richness, by adding

Table 1. Fiducial values used in this work.

Parameters	Default values in Capish	Equations
(Ω_m, Ω_b)	(0.319, 0.048)	Eq. (2)(4)
(σ_8, n_s)	(0.813, 0.96)	–
h	0.7	–
(w_0, w_a)	(−1, 0)	–
Ω_s	$\pi/2$ ($f_{\text{sky}} = 1/8$)	–
$(\mu_0^\lambda, \mu_m^\lambda, \mu_z^\lambda)$	(3.5, 1.72, 0.0)	Eq. (13)
$(\log_{10} m_0, z_0)$	(14.5, 0.5)	–
$\sigma_{\ln \lambda, \text{int}}$	0.2	Eq. (14)
ρ	0	Eq. (12)
$(\mu_0^{\text{WL}}, \mu_m^{\text{WL}}, \mu_z^{\text{WL}})$	(0, 1, 0)	Eq. (15)
$(\sigma_{\text{WL}, \text{gal}}, \sigma_{\text{WL}, \text{int}})$	(0.2 or theory-based, 0)	Eq. (16)
$(\sigma_{\text{PZ}, 0}, b(z_{\text{true}}))$	(0, 0)	Eq. (17)
$(n_{\text{pur}}, \ln \lambda_p)$	(2.5, 1.5)	Eq. (18)
$(\alpha_{\text{gal}}, \beta_{\text{gal}}, z_{\text{gal}}^0)$	(2.0, 1.5, 0.5)	Eq. (B.4)
Γ	0.7	Eq. (21)

Notes. The halo mass function is taken from Despali et al. (2016) and the halo bias from Tinker et al. (2010). The first block corresponds to the fiducial cosmological parameters. The second block is for generating the cluster catalog from the underlying halo catalog. The last block is for computing summary statistics from the cluster catalog. The parameters $\mu_0^\lambda, \mu_m^\lambda$, and μ_z^λ are chosen to mimic the DES Y1 best fit of the cluster mass-richness relation (Abbott et al. 2020), the latter following a different parameterization.

fake clusters. We also see that adding photometric redshift (with $\sigma_{\text{PZ}}(z) = 0.02(1+z)$, and $b(z) = 0$ in Eq. (17)) induces a 2–3% bias with respect to the true redshift case. In Figure 2 (left panel, lower plot), we see that the variance of Capish counts is in good agreement with the analytical prediction of the cluster count variance, accounting for Poisson noise and SSC in Eq. (H.3).

Figure 2 (right panel, upper plot) shows the bias between the mean simulated Capish masses and a mean-mass prediction (computed using Eq. (H.5)), again with the simplest assumptions (i) no selection function (ii) no scatter for the lensing mass and (iii) $\Gamma = 1$. When Capish is run under these simplest assumptions, the bias is zero. Introducing a slope $\Gamma = 0.7$ produces a 5% bias, and adding noise – from galaxy shape and/or shot noise and intrinsic scatter (0.2 and 0.05, respectively) – to the lensing mass (assuming the $\log_{10} m_{\text{WL}}$ parameterization) increases this bias to 10%. Furthermore, varying the constant correlation parameter $\rho \in \{-0.1, +0.1\}$ causes the mean lensing masses to be biased low and high, respectively. More realistically, Capish allows the use of a mass-dependent function $\rho(m) = 0.3 \times \exp(-2(\log_{10} m - 13.3))$, decreasing with mass, whose effect on mean mass is more pronounced at lower richness (or lower mass) values. This is more realistic with regard to the mass-dependence of ρ (Sunayama et al. 2020; Wu et al. 2022; Zhou et al. 2024).

The right panel of Fig. 2 (lower plot) shows the dispersion of the Capish mean masses. In the absence of lensing-mass noise, the dispersion is at its lowest level, originating from Poisson and SSC sampling of the halo mass function. When lensing-mass noise is included, the dispersion increases accordingly. We also see that using the error model described in Appendix B (with $\bar{n}_{\text{gal}} = 25 \text{ arcmin}^{-2}$, and considering the fitting radius between 1 Mpc and 5 Mpc) increases the errors at low richness and decreases them at high richness, compared to adopting a fixed value for σ_{WL} . All of these preliminary tests of the Capish

¹⁴ The fiducial selection function parameters in Table 1 are chosen such that the cluster catalog is increasingly purer at higher richnesses, with 50% purity at richness $\lambda_{\text{obs}} = 5$.

simulator ensure that the summary statistics behave as expected, as for their estimators and their covariances.

3.2. Probability coverage

The posterior prediction is built on top of the `sbi`¹⁵ Python package (Tejero-Cantero et al. 2020), a flexible toolkit for SBI. The required inputs for `sbi` Python package are the summary statistics obtained from the Capish simulator. Then, `sbi` trains neural networks, typically normalizing flows, to learn the posterior $\mathbb{P}(\hat{\theta}_{\text{true},k} \mid \hat{D}_k)$ using N_{sim} simulated pairs $\{\hat{\theta}_{\text{true},k}, \hat{D}_k\}$ for $k = 1, 2, \dots, N_{\text{sim}}$, where $\hat{D}_k \sim \text{Simulator}(\hat{\theta}_{\text{true},k})$.

We assess the performance of the posterior generator across the free parameter space $\theta = \{\Omega_m, \sigma_8, \mu_0^\lambda, \mu_m^\lambda, \mu_z^\lambda, \sigma_{\ln \lambda, \text{int}}\}$. For the Capish simulator, we consider: (i) a sky area of $\Omega_S = \pi/2$ ($f_{\text{sky}} = 1/8$); (ii) the Tinker et al. (2010) halo mass function with the 200*m* mass definition; (iii) pure cluster samples; (iv) the scaling functional form of the scaling relation in Eq. (13); (v) no photometric redshifts; and (vi) we consider the stacked approach for the scatter of weak lensing mass, as explained in Sect. 2.3, and applied to the lensing mass through Eq. (23) with a theory model for σ_{WL} (with $\bar{n}_{\text{gal}} = 25 \text{ arcmin}^{-2}$; see Appendix B); (vii) using the alternative mass summary statistics in Eq. (24). Summary statistics (counts and modified lensing masses) are computed using the richness bin edges $\lambda_{\text{obs}} = \{20, 30, 40, 60, 100, 200\}$ and redshift bin edges $z = \{0.2, 0.35, 0.5, 0.7, 1.0\}$. We use the lower richness cut (respectively, redshift) $\lambda_{\text{obs}} > 20$ (respectively, $z > 0.2$), to be consistent with DES-Y1 analyzes McClintock et al. (2019), Abbott et al. (2020); the $\lambda_{\text{obs}} > 20$ cut is generally used so that it ensures a high-purity cluster sample (Costanzi et al. 2019). Restricting to $z > 0.2$ mimics the conservative cut used in DES cluster-based analyzes (Abbott et al. 2020, 2025), to prevent the degradation of redMaPPer performance at low redshifts, where the red-sequence galaxy population becomes harder to isolate due to the lack of *u*-band data in the DES analyzes¹⁶. We note that robust detection of low-redshift clusters below $z = 0.2$ will be feasible with the Rubin LSST, since *u*-band imaging will be available over 18 000 square degrees of the Southern sky. For the fiducial parameters, the total number of clusters within the considered richness and redshift bins is $\sim 77\,000$. For the first redshift bin, $N_{z_{1,\lambda}} \sim \{6200, 1900, 1100, 340, 45\}$. For the second redshift bin, $N_{z_{2,\lambda}} \sim \{10^4, 2900, 1500, 400, 45\}$. For the third redshift bin, $N_{z_{3,\lambda}} \sim \{1.6 \times 10^4, 4200, 1900, 430, 34\}$ and for the fourth redshift bin $N_{z_{4,\lambda}} \sim \{2.3 \times 10^4, 4900, 1900, 320, 17\}$. The typical binned $\langle \log_{10}(M_{200\text{m}}/M_\odot) \rangle$ is rather stable with redshift and goes from 14.2 (low richness bin) to 15 (high richness bin). Our upper richness cut, $\lambda_{\text{obs}} < 200$, ensures that all massive simulated clusters are included in the counts¹⁷. While our analysis focuses on the sensitivity of cluster abundance to growth-of-structure parameters (such as Ω_m and σ_8) within the standard Lambda cold dark matter (Λ CDM) paradigm, galaxy clusters also provide a powerful probe of extensions beyond Λ CDM. These include scenarios with massive neutrinos (Böhlinger & Chon 2016), modified gravity (Cataneo et al. 2015), and primordial nonGaussianity (Robinson et al. 2000). Testing such extensions typically requires probing the abundance of rarer systems, i.e., very massive and/or high-redshift clusters. Since the last richness

bin remains relatively broad (spanning $\lambda_{\text{obs}} = 100$ to $\lambda_{\text{obs}} = 200$), it could be refined in future work to better capture the high-mass tail of the halo mass function. This would allow us to more fully exploit SBI for constraining extended cosmological models.

We generate $N_{\text{sim}} = 60\,000$ simulations that cover the parameter priors $\Omega_m \sim \mathcal{U}(0.2, 0.45)$, $\sigma_8 \sim \mathcal{U}(0.6, 0.95)$, $\mu_0^\lambda \sim \mathcal{U}(3, 4)$, $\mu_m^\lambda \sim \mathcal{U}(1.3, 2.1)$, $\mu_z^\lambda \sim \mathcal{U}(-0.7, 0.7)$, and $\sigma_{\ln \lambda, \text{int}} \sim \mathcal{U}(0.1, 0.5)$, which encompass the fiducial values listed in Table 1. This stage took approximately two hours of CPU time.

We train the posterior generator using the neural posterior estimator (NPE) method, using alternative cluster counts in the dedicated richness-redshift bins (this configuration is labeled count; see Eq. (20)), mean lensing masses multiplied by the cluster counts (labeled Nm; see Eq. (24)), and the combination of the two (labeled count_Nm). The SBI training for each setup took approximately 25 minutes of CPU time. After training, the `sbi` package returns a prediction for the posterior distribution for the six parameters $\theta = \{\Omega_m, \sigma_8, \mu_0^\lambda, \mu_m^\lambda, \mu_z^\lambda, \sigma_{\ln \lambda, \text{int}}\}$, given an observed data vector $\hat{D}$ (being cluster counts, cluster masses, or a combination of the two). For a given data vector $\hat{D}$, approximately 500 000 samples of the posterior $\hat{\theta} \sim \mathbb{P}(\cdot \mid \hat{D})$ can be drawn within a second.

Validating the accuracy of the posterior generator is critical. To this end, we assess its performance using the coverage calibration test, a key diagnostic of the posterior that tests if the reported posterior credible sets have their intended probabilistic meaning.

Given a dataset $\hat{D}$ simulated at a parameter point $\hat{\theta}_{\text{true}}$ drawn from the prior (i) we compute the corresponding posterior estimate $\mathbb{P}(\hat{\theta} \mid \hat{D})$ from the trained NPE, and (ii) we extract a γ -credible set¹⁸ $\text{Cred}_\gamma(\hat{D})$ for this estimated posterior. For a nominal coverage probability γ , the Bayesian coverage expresses as

$$p_\gamma(\hat{D}) := \mathbb{P}(\hat{\theta}_{\text{true}} \in \text{Cred}_\gamma(\hat{D}) \mid \hat{D}) \quad (25)$$

$$= \mathbb{E}_{\hat{\theta}_{\text{true}}} [\mathbb{1}\{\hat{\theta}_{\text{true}} \in \text{Cred}_\gamma(\hat{D})\} \mid \hat{D}]. \quad (26)$$

A posterior estimate is calibrated in the Bayesian sense if $p_\gamma(\hat{D}) = \gamma$ for any $\hat{D}$, that is, the true parameter value $\hat{\theta}_{\text{true}}$ should fall within the estimated posterior γ -credible set a fraction γ of the time, and so for any observed data $\hat{D}$. Under mild assumptions, this calibration is equivalent to the estimated posterior being equal to the true posterior. However, we do not have access to the true posterior by construction, and therefore the computation of p_γ from simulated pairs would require some binning of the observations. The common method is to average $p_\gamma(\hat{D})$ also on the data, i.e., over the simulated pairs $\{\hat{\theta}_{\text{true},k}, \hat{D}_k\}$, giving

$$p_\gamma := \mathbb{P}(\hat{\theta}_{\text{true}} \in \text{Cred}_\gamma(\hat{D})) = \mathbb{E}_{\hat{\theta}_{\text{true}}, \hat{D}} [\mathbb{1}\{\hat{\theta}_{\text{true}} \in \text{Cred}_\gamma(\hat{D})\}], \quad (27)$$

called expected coverage probability (ECP). We say that the posterior is calibrated with respect to ECP when $p_\gamma = \gamma$. For information, a sub-identity $p_\gamma < \gamma$ (respectively, super-identity $p_\gamma > \gamma$) calibration curve reveals an underestimation (respectively, overestimation) of the uncertainty.

This calibration test allows us to detect some potential failures in the posterior estimation. For example, due to the continuous nature of the normalizing flows employed in the neural density estimation (NDE), the posterior estimate can be highly

¹⁵ <https://sbi-dev.github.io/sbi/latest/>

¹⁶ The redMaPPer detection in the different DES cluster analyzes used the *g*, *r*, *i*, and *z* bands. For $z < 0.2$ galaxies, the distinctive break features at $\sim 4000 \text{ \AA}$ fall in the *u*-band.

¹⁷ This holds for simulations run with the fiducial parameters in Table 1.

¹⁸ A γ -credible set is a set that contains the parameter value with probability γ , e.g., the interval between the $\gamma/2$ -quantile and the $1 - \gamma/2$ -quantile in 1D.

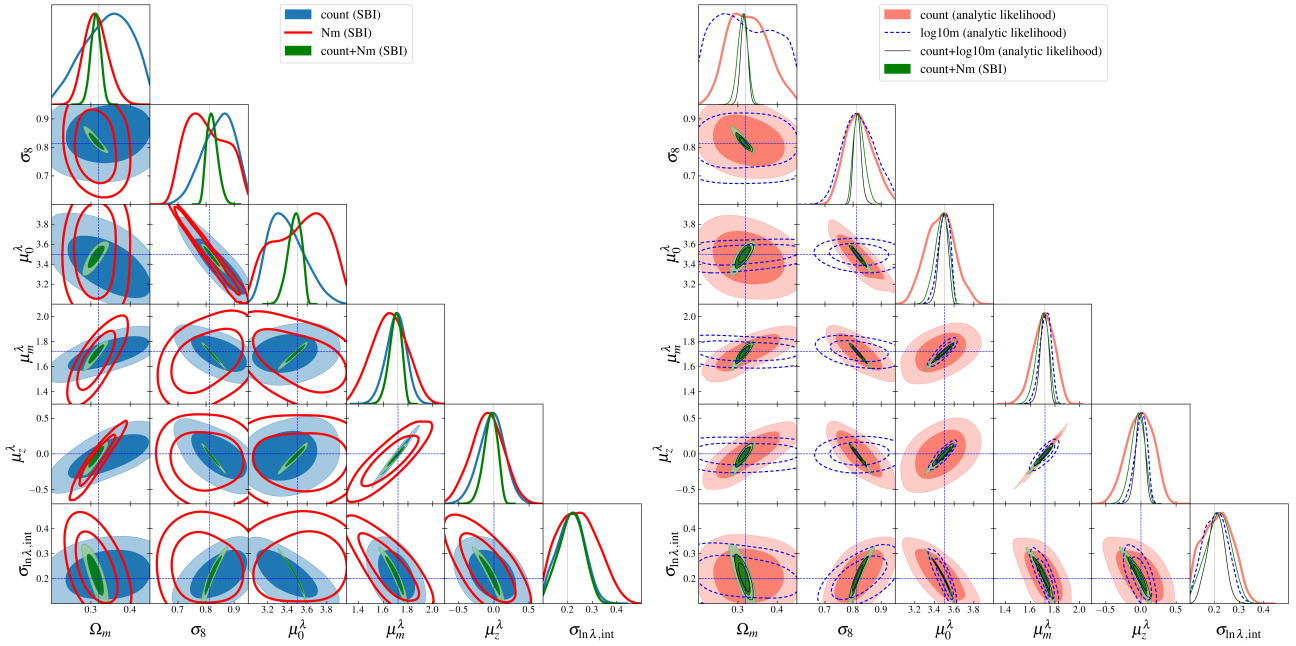


Fig. 3. *Left:* Posterior distributions obtained from the trained posterior generator for the three configurations count, Nm, and count_Nm. *Right:* Corresponding posteriors from the analytic likelihood approach, sampled with MCMC for count, log10m, and count_log10m (the SBI count_Nm results are overplotted for comparison).

inaccurate near sharp posterior features such as prior edges (for example, Reza et al. 2024). This leads to an underestimation of the probability density near the prior boundaries such that a probability coverage test would exhibit a mild tendency toward overconfidence. First, we do not account for this effect and compute the ECP over the full prior range, for the three configurations (counts, mean masses, and their combination), as shown in Fig. F.1, where it appears that $p_\gamma < \gamma$, i.e., the uncertainties are indeed underestimated. We therefore recompute the same probability coverage for the count_Nm configuration (combination of counts and mean masses, the most relevant), but restricting the evaluation to fewer points $\hat{\theta}_{\text{true}} \in [\theta_{\min} + \Delta\theta/2, \theta_{\max} - \Delta\theta/2]$, where $\theta_{\min}$ (respectively, $\theta_{\max}$) is the lower (respectively, upper) bound of the parameter prior, $\Delta\theta$ represents $\sim 5\%$ of the prior size, and we “mask” posterior samples accordingly by removing samples outside this updated prior range. As shown in the left panel of Fig. F.2 (for the coverage plot, we consider 1000 pairs $\{\hat{\theta}_{\text{true},k}, \hat{D}_k\}$ with $\hat{\theta}_{\text{true},k}$ within the restricted prior region, and posteriors are estimated with 50 000 samples), the resulting calibration curve is closer to identity, emphasizing that this prior problem can be solved by (i) training on a broader prior, (ii) restricting the posterior to a tighter prior range, chosen to match the analysis requirement.

Additionally, to test not only the marginalized but also the joint posterior, we compute an alternative ECP quantity by using the TARP method (Tests of Accuracy with Random Points; Lemos et al. 2023a), shown in the left panel of Fig. F.2, also close to identity, ensuring good calibration with respect to this metric. Although they do not guarantee the full validity of the posterior estimation, these tests are effective in detecting potential miscalibration, and all pass successfully here.

3.3. Cosmological analysis: Internal validation

Given the configuration count_Nm, and considering 500 pairs $\{\hat{\theta}_{\text{true},k}, \hat{D}_k\}$, we compute the estimated posterior mean $E[\hat{\theta}_k | \hat{D}_k]$

that we compare to the true values $\hat{\theta}_{\text{true},k}$. The comparison is shown in Fig. G.1. This consistency test shows that the recovered cosmological parameters are efficiently recovered from the posterior, with some scatter around the true value, with no strong apparent bias.

We study now the shape of the posterior, for which we generate 500 simulations at the same fiducial parameter values (in Table 1). By averaging the data vector over these realizations, we obtain an effectively noiseless data vector, which we use to perform consistency tests of the Bayesian inference. Figure 3 shows the resulting posterior distributions¹⁹ (using count, Nm, or count_Nm) for the six parameters (the best fits are reported in Table J.1). In all cases, we find good agreement between the inferred posteriors and the fiducial parameter values, the latter lying within the 1σ region of each posterior distribution, as expected from the test in Fig. G.1.

As mentioned before, it is possible that cluster number counts of Capish simulations display one or more empty bins, making the computation of the mean mass impossible (that is why we choose the alternative lensing mass statistics). In Appendix E we explore the effect of removing the simulations that contain at least one empty cluster-count bin in the training. In this case, $\log_{10} M_{kl}$ can be computed and then used in the NDE training. In Fig. E.1 we show the fraction of nonremoved objects after masking over empty count bins, displaying an inhomogeneous shape over the prior space (with $\sim 40\%$ remaining simulations in the bottom left corner of the prior region). This masking induces a complex, implicit prior on the posterior and breaks the desired transparency of the pipeline to empty bins, and shifts the posterior (as shown in Fig. E.2). This confirms that the alternative summary statistics Nm in Eq. (24) are particularly reliable for this task.

We test how the selection-bias parameter ρ in Eq. (12) – which quantifies the correlation between observed weak-lensing masses and optical richness – impacts cosmological parameter

¹⁹ Posteriors are displayed with GetDist (Lewis 2025).

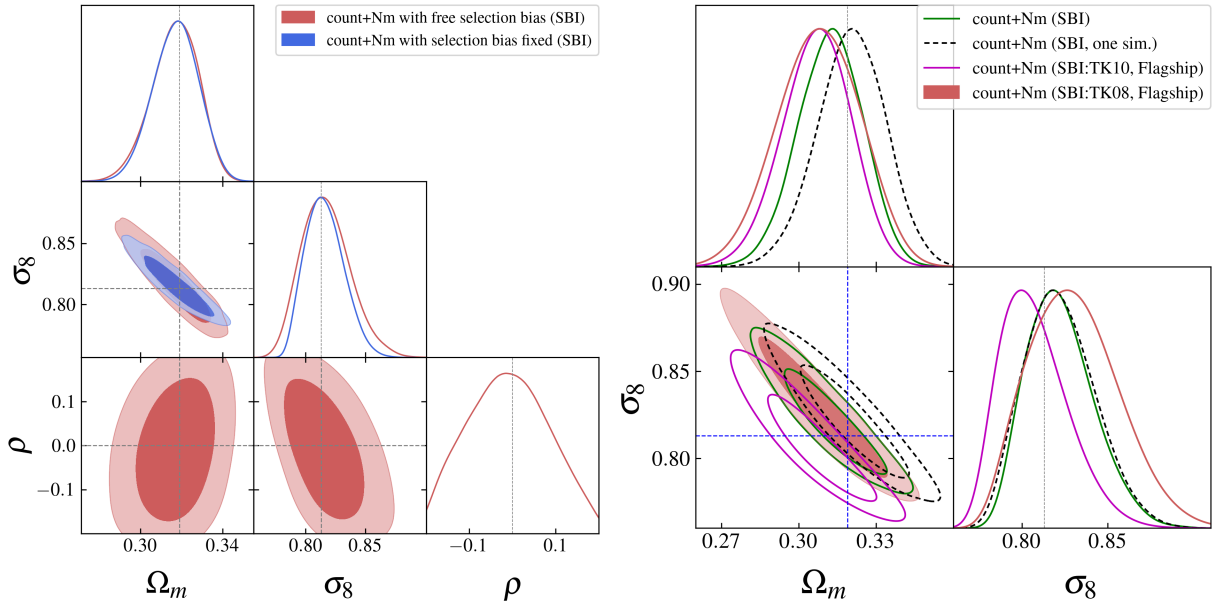


Fig. 4. *Left:* Posterior distributions obtained on a noiseless data vector with the count_Nm setup; the blue posterior is obtained by training on only 6 parameters, and the red is trained adding the lensing mass-richness correlation parameter ρ . *Right:* Posterior distribution over the Ω_m and σ_8 parameters. (i) count_Nm applied to the noiseless data vector, (ii) count_Nm applied to a single fiducial simulation, (iii) count_Nm applied to the Flagship simulation with richness drawn from Capish, trained on Tinker et al. (2010) halo mass function and (iv) same as (iii), but trained on Tinker et al. (2008) halo mass function.

constraints. To this end, we adopt a different noise model than in previous analyzes: instead of applying lensing-mass scatter to the stacked mass, the scatter is applied to individual lensing mass measurements. As a baseline, we first train a new count_Nm NDE with ρ being fixed. We then train a second count_Nm NDE that includes $\rho \sim \mathcal{U}(-0.2, 0.2)$. The resulting posterior samples, applied to the noiseless data vector (computed with $\rho = 0$), are shown in Fig. 4 (left panel), marginalized over the parameters μ_0^λ , μ_m^λ , μ_z^λ , and $\sigma_{\ln \lambda, \text{int}}$. Allowing ρ to vary leads to slightly broader constraints on σ_8 , revealing a negative correlation between the selection-bias parameter and σ_8 , and keeping the constraints on Ω_m unchanged. This example illustrates that Capish incorporates the effect of lensing mass-richness correlation, and that, for the considered prior and current parametrization for ρ , including this mass-dependent correlation does not modify significantly the inference of cosmological parameters.

3.4. Cosmological analysis: Comparison to the explicit-likelihood approach

In this subsection we show how the SBI posteriors compare to the posteriors obtained from an explicit-likelihood approach. The elements for the explicit-likelihood approach are detailed in Appendix H; we define a Gaussian cluster abundance likelihood $\mathcal{L}_{\text{BLC}}^{\text{Gauss-SN+SSC}}$ in Eq. (H.2) accounting for the Poisson noise (Poisson 1837) and SSC (Hu & Kravtsov 2003). For the mean lensing mass likelihood $\mathcal{L}_{M_{\text{WL}}}$, we adopt a Gaussian model in Eq. (H.4) for the logarithm of stacked cluster masses, where the covariance is taken to be diagonal. Its diagonal components (i.e., the dispersion of each stacked mass) are computed assuming a Fisher-like approach on an NFW (Navarro et al. 1997) profile model, accounting for the shape noise and shot noise of source galaxies, and the total number of clusters within the stack.

We draw samples from the parameter posterior distribution with the emcee package (Foreman-Mackey et al. 2013) given by the Bayes theorem $\mathbb{P}(\theta | \text{data}) \propto \mathcal{L}_{\text{tot}}(\text{data} | \theta) \times \pi(\theta)$, where $\pi(\theta)$

is the prior distribution mentioned earlier. Each Markov chain Monte Carlo (MCMC) run took about 4 hours of CPU time. The results are shown in the right panel of Fig. 3; the explicit-likelihood constraints are consistent with the fiducial parameters, showcasing that the simulation-based noiseless Capish data vectors are not biased between the two methods. The combined constraints and count-only show a comparable shape to the SBI results (left panel of Fig. 3), whereas the explicit-likelihood log10m constraints are very different from the Nm approach, since not the same information is encoded between the two statistics. When combined with counts, however, the discrepancy between SBI and the explicit-likelihood approaches decreases, since the combined constraints show roughly similar posterior widths and correlations as those inferred by SBI. However, we see that explicit-likelihood error bars are slightly smaller, which we expect is from a series of effects; in the explicit-likelihood approach, we adopt a fixed covariance matrix²⁰ for the lensing masses and cluster counts (to speed up the calculations), whereas in the SBI approach, the covariance can vary across simulations. Our implementation of the explicit-likelihood approach neglects correlation between probes, which will tighten parameter posterior and Capish accounts for a sampling dispersion (the intrinsic variation of the mean mass in each richness-redshift bin due to the variation of the number of clusters), which is absent from the explicit-likelihood approach²¹. Moreover, the explicit-likelihood approach assumes variables are Gaussian, whereas Capish will account for any nonGaussian noise in our forward model. Given these effects, it is expected that Capish has slightly broader posteriors than the explicit-likelihood approach.

It is important to note that posterior coverage remains a relevant diagnostic for explicit-likelihood approaches (see, e.g., Payerne et al. 2023), as it provides a direct test of the accuracy

²⁰ Computed at the fiducial cosmology and scaling relation parameters.

²¹ For which the error bars are computed as $\sigma_{\text{WL,gal}} / \sqrt{N}$, where N is the cluster count at fiducial values.

of the recovered posterior, similarly to SBI methods. However, we do not conduct this test in this study, given the high computational cost of evaluating a single explicit-likelihood posterior (4 hours for a single posterior).

3.5. Cosmological analysis with external simulated datasets

As a final validation, we apply our trained posterior generator to the *Euclid* Flagship simulation (Euclid Collaboration: Castander et al. 2025)²². Flagship is a large-scale mock galaxy catalog designed to support the scientific exploitation of ESA's *Euclid* mission. The simulation was run using the PKDGRAV3 code (Potter et al. 2017) under a flat- Λ CDM cosmology, with cosmological parameters close to those inferred by Planck Collaboration XXIV (2016). A light cone extending to $z = 3$ was generated on the fly, covering one octant of the sky (approximately 5157 deg^2). Dark matter halos were identified using the rockstar algorithm (Behroozi et al. 2013). From these halos with true masses m_{halo} and true redshift z_{halo} , we construct a cluster catalog by assigning an observed richness λ_{obs} to each Flagship halo using Eq. (13) and the fiducial parameters listed in Table 1. Individual lensing masses m_{WL} are taken to be equal to the halo masses m_{halo} , with stochasticity applied only at the stacked level through Eq. (23), when constructing the stacked lensing mass matrices. Figure 4 (right panel) shows the posterior distributions for the cosmological parameters (Ω_m, σ_8) obtained with our SBI framework using three different data vectors: (i) a noiseless data vector, (ii) a single Capish simulation, and (iii) the Flagship simulation with Capish-like richnesses. In the case of the single Capish realization, the posterior is slightly shifted, primarily due to statistical noise in the data. For the SBI applied to the Flagship catalog when using the Tinker et al. (2010) [TK10], we find a good agreement with the fiducial cosmology for the marginalized posteriors, but still with a $\sim 2\sigma$ tension in the Ω_m, σ_8 plane. This residual tension can be attributed to the fact that we performed our training assuming the Tinker et al. (2010) halo mass function, which is different from the effective mass function measured in the Flagship simulation (Euclid Collaboration: Castander et al. 2025), whose underlying halo mass function was found to be closer to the Tinker et al. (2008) [TK08] implementation (see Appendix I and the right panel of Fig. F.2). We repeat the training this time using the TK08 halo mass function, which enables us to recover the good cosmological parameters with no significant tension (below 1σ ; see the right panel of Fig. 4). In practice, the underlying Flagship halo mass function is unknown and thus may differ from TK08 or TK10. Incorporating additional nuisance parameters in the training (denoting any possible deviation from TK08 or TK10) with appropriated priors could help mitigate this discrepancy (Wu et al. 2010; Cunha & Evrard 2010; Abbott et al. 2020; Artis et al. 2021).

4. Conclusions

In this work, we have presented Capish, a simulation-based inference framework for the cosmological analysis of galaxy cluster datasets. Rather than relying on explicit likelihood modeling, Capish learns posterior distributions directly from forward-modeled cluster observables – specifically cluster number counts and mean weak-lensing masses in richness–redshift bins – naturally capturing nonGaussian features of the data that analytic likelihood approaches cannot. The forward model

incorporates Poisson sampling noise and super-sample covariance using log-normal variables, extending the approach of Payerne et al. (2024), as well as lensing mass scatter, selection effects through purity and selection biases, and an alternative lensing mass statistic – given by the product of count and mean lensing mass – to regularize the neural density estimator training. Capish performs posterior inference using neural posterior estimation within the sbi package (Tejero-Cantero et al. 2020).

The code has been validated by comparing its outputs (at fiducial values in Table 1) against analytical predictions, confirming that the simulated cluster counts accurately reproduce the expected SSC variance. The flexibility of the selection function implementation has been tested, and the impact of lensing mass–richness covariance on recovered masses has been assessed. Bayesian coverage tests indicate that the posteriors are unbiased and provide robust uncertainty estimates for both cosmological and scaling-relation parameters. We have also examined how the effective lensing mass–richness correlation parameter (Wu et al. 2022; Zhang et al. 2024), one of the major systematic effects in cosmology with optically selected clusters, propagates into the cosmological posteriors. Comparisons with explicit-likelihood analyzes (Payerne et al. 2023) show good overall agreement, with the broader SBI posteriors reflecting the increased realism of the forward model. Another advantage of Capish for simulation-based inference is the computational time, as Capish generates simulations and trains a posterior generator within 2.5 hours (which samples a posterior within a second), whereas one MCMC posterior sampling based on an explicit likelihood takes 4 hours. As a final validation, Capish has been applied to the *Euclid* Flagship dataset (Euclid Collaboration: Castander et al. 2025) with Capish-assigned richness, recovering the Flagship cosmological parameters within 1σ . This framework offers a flexible and robust alternative for analyzes of surveys such as DES, *Euclid* (Laureijs et al. 2011), and LSST (LSST Science Collaboration 2009), and we propose it as a test bench for validating explicit-likelihood cosmological pipelines and assessing their internal consistency.

Acknowledgements. The authors thank the anonymous reviewer for their insightful comments and suggestions. We gratefully acknowledge support from the CNRS/IN2P3 Computing Center (Lyon – France) for providing computing and data-processing resources needed for this work. We thank the developers and maintainers of the following software tools used in this work: sbi (Tejero-Cantero et al. 2020), NumPy (van der Walt et al. 2011), SciPy (Virtanen et al. 2020), Matplotlib (Hunter 2007), GetDist (Lewis 2025), emcee (Foreman-Mackey et al. 2013), Jupyter (Kluyver et al. 2016).

References

- Abbott, T., Aguena, M., Alarcon, A., et al. 2020, *Phys. Rev. D*, **102**, 023509
- Abbott, T. M. C., Aguena, M., Alarcon, A., et al. 2025, *Phys. Rev. D*, **112**, 083535
- Ade, P., Aghanim, N., Armitage-Caplan, C., et al. 2016, *A&A*, **594**, A24
- Aguena, M., & Lima, M. 2018, *Phys. Rev. D*, **98**, 123529
- Aguena, M., Benoist, C., da Costa, L. N., et al. 2021, *MNRAS*, **502**, 4435
- Allen, S. W., Evrard, A. E., & Mantz, A. B. 2011, *ARA&A*, **49**, 409
- Anbajagane, D., Evrard, A. E., Farahi, A., et al. 2020, *MNRAS*, **495**, 686
- Artis, E., Melin, J.-B., Bartlett, J. G., & Murray, C. 2021, *A&A*, **649**, A47
- Aymerich, G., Douspis, M., Pratt, G. W., et al. 2024, *A&A*, **690**, A238
- Bartelmann, M., & Schneider, P. 2001, *Phys. Rep.*, **340**, 291
- Bartlett, J. G. 1997, *ASP Conf. Ser.*, **126**, 365
- Becker, M. R., & Kravtsov, A. V. 2011, *ApJ*, **740**, 25
- Behroozi, P. S., Wechsler, R. H., & Wu, H.-Y. 2013, *ApJ*, **762**, 109
- Bellagamba, F., Roncarelli, M., Maturi, M., & Moscardini, L. 2018, *MNRAS*, **473**, 5221
- Bocquet, S., Grandis, S., Bleem, L. E., et al. 2024, *Phys. Rev. D*, **110**, 083510
- Bocquet, S., Grandis, S., Krause, E., et al. 2025, *Phys. Rev. D*, **111**, 063533

²² Publicly available at <https://cosmohub.pic.es>

- Böhringer, H., & Chon, G. 2016, *Mod. Phys. Lett. A*, **31**, 1640008
- Cataneo, M., Rapetti, D., Schmidt, F., et al. 2015, *Phys. Rev. D*, **92**, 044009
- Cerardi, N., Pierre, M., Lanusse, F., & Corap, X. 2025, *A&A*, **701**, A110
- Chang, C., Jarvis, M., Jain, B., et al. 2013, *MNRAS*, **434**, 2121
- Chaubal, P. S., Reichardt, C. L., Gupta, N., et al. 2022, *ApJ*, **931**, 139
- Coles, P., & Jones, B. 1991, *MNRAS*, **248**, 1
- Costanzi, M., Rozo, E., Rykoff, E. S., et al. 2019, *MNRAS*, **482**, 490
- Costanzi, M., Rozo, E., Simet, M., et al. 2019, *MNRAS*, **488**, 4779
- Costanzi, M., Saro, A., Bocquet, S., et al. 2021, *Phys. Rev. D*, **103**, 043522
- Cunha, C. E., & Evrard, A. E. 2010, *Phys. Rev. D*, **81**, 083509
- Despali, G., Giocoli, C., Angulo, R. E., et al. 2016, *MNRAS*, **456**, 2486
- Duffy, A. R., Schaye, J., Kay, S. T., & Dalla Vecchia, C. 2008, *MNRAS*, **390**, L64
- Euclid Collaboration (Adam, R., et al.) 2019, *A&A*, **627**, A23
- Euclid Collaboration (Castander, F. J., et al.) 2025, *A&A*, **697**, A5
- Evrard, A. E., Arnault, P., Huterer, D., & Farahi, A. 2014, *MNRAS*, **441**, 3562
- Farahi, A., Evrard, A. E., McCarthy, I., Barnes, D. J., & Kay, S. T. 2018, *MNRAS*, **478**, 2618
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, **125**, 306
- Fumagalli, A., Saro, A., Borgani, S., et al. 2021, *A&A*, **652**, A21
- Fumagalli, A., Costanzi, M., Saro, A., Castro, T., & Borgani, S. 2024, *A&A*, **682**, A148
- Gatti, M., Campailla, G., Jeffrey, N., et al. 2025, *Phys. Rev. D*, **111**, 063504
- Ghirardini, V., Bulbul, E., Artis, E., et al. 2024, *A&A*, **689**, A298
- Gouyou Beauchamps, S., Lacasa, F., Tutusaus, I., et al. 2022, *A&A*, **659**, A128
- Graham, M. L., Connolly, A. J., Ivezić, Ž., et al. 2018, *AJ*, **155**, 1
- Grandis, S., Ghirardini, V., Bocquet, S., et al. 2024, *A&A*, **687**, A178
- Gruen, D., Seitz, S., Becker, M. R., Friedrich, O., & Mana, A. 2015, *MNRAS*, **449**, 4264
- Hasselfield, M., Hilton, M., Marriage, T. A., et al. 2013, *JCAP*, **2013**, 008
- Hernández-Martín, B., Schrabback, T., Hoekstra, H., et al. 2020, *A&A*, **640**, A117
- Hoekstra, H. 2001, *A&A*, **370**, 743
- Hoekstra, H. 2003, *MNRAS*, **339**, 1155
- Hu, W., & Kravtsov, A. V. 2003, *ApJ*, **584**, 702
- Hunter, J. D. 2007, *Comput. Sci. Eng.*, **9**, 90
- Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, *ApJ*, **873**, 111
- Jeffrey, N., Alsing, J., & Lanusse, F. 2021, *MNRAS*, **501**, 954
- Kluge, M., Comparat, J., Liu, A., et al. 2024, *A&A*, **688**, A210
- Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, *Jupyter Notebooks - A Publishing Format for Reproducible Computational Workflows* (IOS Press), 87
- Kosiba, M., Cerardi, N., Pierre, M., et al. 2025, *A&A*, **693**, A46
- Kravtsov, A. V., & Borgani, S. 2012, *ARA&A*, **50**, 353
- Lacasa, F., & Grain, J. 2019, *A&A*, **624**, A61
- Lacasa, F., Lima, M., & Agüena, M. 2018, *A&A*, **611**, A83
- Lacasa, F., Aubert, M., Baratta, P., et al. 2023, *A&A*, **671**, A115
- Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints [arXiv:1110.3193]
- Lee, B. E., Le Brun, A. M. C., Haq, M. E., et al. 2018, *MNRAS*, **479**, 890
- Lee, E., Battye, R., & Bolliet, B. 2025, *MNRAS*, **538**, 2768
- Lemos, P., Coogan, A., Hezaveh, Y., & Perreault-Levasseur, L. 2023a, in *40th International Conference on Machine Learning*, 202, 19256
- Lemos, P., Parker, L. H., Hahn, C., et al. 2023b, *Mach. Learn. Astrophys.*, **18**
- Lesci, G. F., Marulli, F., Moscardini, L., et al. 2022, *A&A*, **659**, A88
- Lesci, G. F., Marulli, F., Moscardini, L., et al. 2025, *A&A*, **703**, A25
- Lewis, A. 2025, *JCAP*, **2025**, 025
- LSST Science Collaboration (Abell, P. A., et al.) 2009, arXiv e-prints [arXiv:0912.0201]
- Mantz, A. B. 2019, *MNRAS*, **485**, 4863
- Mantz, A., Allen, S. W., Ebeling, H., & Rapetti, D. 2008, *MNRAS*, **387**, 1179
- Mantz, A., Allen, S. W., Rapetti, D., & Ebeling, H. 2010, *MNRAS*, **406**, 1759
- Mantz, A. B., von der Linden, A., Allen, S. W., et al. 2015, *MNRAS*, **446**, 2205
- McClintock, T., Varga, T. N., Gruen, D., et al. 2019, *MNRAS*, **482**, 1352
- Melchior, P., Gruen, D., McClintock, T., et al. 2017, *MNRAS*, **469**, 4899
- Mistele, T., & Durakovic, A. 2024, *OJAp*, **7**, 120
- Murata, R., Oguri, M., Nishimichi, T., et al. 2019, *PASJ*, **71**, 107
- Murray, C., Bartlett, J. G., Artis, E., & Melin, J.-B. 2022, *MNRAS*, **512**, 4785
- Murray, C., Combet, C., Payerne, C., & Ricci, M. 2025, *A&A*, **697**, A141
- Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, *ApJ*, **490**, 493
- Nde, T. N., Wu, H.-Y., Cao, S., et al. 2026, *Phys. Rev. D*, **113**, 063559
- Nord, B., Stanek, R., Rasia, E., & Evrard, A. E. 2008, *MNRAS*, **383**, L10
- Oguri, M., Takada, M., Okabe, N., & Smith, G. P. 2010, *MNRAS*, **405**, 2215
- Park, Y., Sunayama, T., Takada, M., et al. 2023, *MNRAS*, **518**, 5171
- Payerne, C., Murray, C., Combet, C., et al. 2023, *MNRAS*, **520**, 6223
- Payerne, C., Murray, C., Combet, C., & Penna-Lima, M. 2024, *MNRAS*, **532**, 381
- Payerne, C., Zhang, Z., Agüena, M., et al. 2025, *A&A*, **700**, A34
- Penna-Lima, M., Makler, M., & Wuensche, C. A. 2014, *JCAP*, **2014**, 039
- Penna-Lima, M., Bartlett, J. G., Rozo, E., et al. 2017, *A&A*, **604**, A89
- Planck Collaboration XXIV. 2016, *A&A*, **594**, A24
- Poisson, S. D. 1837, Bachelier, Imprimeur-Librairie, 510
- Potter, D., Stadel, J., & Teyssier, R. 2017, *COMPAC*, **4**, 2
- Prat, J., Gatti, M., Doux, C., et al. 2026, *MNRAS*, **545**, staf2152
- Pratt, G. W., Arnaud, M., Biviano, A., et al. 2019, *Space Sci. Rev.*, **215**, 25
- Regamey, M., Eckert, D., Seppi, R., et al. 2026, *A&A*, **708**, A260
- Reza, M., Zhang, Y., Nord, B., et al. 2022, *Mach. Learn. Astrophys.*, **20**
- Reza, M., Zhang, Y., Avestruz, C., et al. 2024, arXiv e-prints [arXiv:2409.20507]
- Robinson, J., Gawiser, E., & Silk, J. 2000, *ApJ*, **532**, 1
- Rozo, E., Wechsler, R. H., Rykoff, E. S., et al. 2010, *ApJ*, **708**, 645
- Rozo, E., Evrard, A. E., Rykoff, E. S., & Bartlett, J. G. 2014, *MNRAS*, **438**, 62
- Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, *ApJ*, **785**, 104
- Sáez-Casares, I., Calabrese, M., Bianchi, D., et al. 2026, *Astron. Comput.*, **55**, 101067
- Salcedo, A. N., Wu, H.-Y., Rozo, E., et al. 2024, *Phys. Rev. Lett.*, **133**, 221002
- Salcedo, A. N., Rozo, E., Wu, H. Y., et al. 2025, arXiv e-prints [arXiv:2510.25706]
- Salvati, L., Saro, A., Bocquet, S., et al. 2022, *ApJ*, **934**, 129
- Saro, A., Bocquet, S., Rozo, E., et al. 2015, *MNRAS*, **454**, 2305
- Sehgal, N., Trac, H., Acquaviva, V., et al. 2011, *ApJ*, **732**, 44
- Sheldon, E. S., Johnston, D. E., Frieman, J. A., et al. 2004, *AJ*, **127**, 2544
- Shirasaki, M., & Takada, M. 2018, *MNRAS*, **478**, 4277
- Sommer, M. W., Schrabback, T., Applegate, D. E., et al. 2022, *MNRAS*, **509**, 1127
- Sunayama, T., Park, Y., Takada, M., et al. 2020, *MNRAS*, **496**, 4468
- Sunayama, T., Miyatake, H., Sugiyama, S., et al. 2024, *Phys. Rev. D*, **110**, 083511
- Takada, M., & Spergel, D. N. 2014, *MNRAS*, **441**, 2456
- Tejero-Cantero, A., Boelts, J., Deistler, M., et al. 2020, *J. Open Source Software*, **5**, 2505
- Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, *ApJ*, **688**, 709
- Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, *ApJ*, **724**, 878
- To, C., Krause, E., Rozo, E., et al. 2021, *Phys. Rev. Lett.*, **126**, 141301
- To, C.-H., Krause, E., Chang, C., et al. 2025, *Phys. Rev. D*, **112**, 063537
- Umetsu, K. 2020, *A&A Rev.*, **28**, 7
- van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *Comput. Sci. Eng.*, **13**, 22
- Varga, T. N., DeRose, J., Gruen, D., et al. 2019, *MNRAS*, **489**, 2511
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nature Methods*, **17**, 261
- von der Linden, A., Mantz, A., Allen, S. W., et al. 2014, *MNRAS*, **443**, 1973
- von Wietersheim-Kramsta, M., Lin, K., Tessore, N., et al. 2025, *A&A*, **694**, A223
- Wen, D., Kembell, A. J., & Saslaw, W. C. 2020, *ApJ*, **890**, 160
- Wright, A. H., Hildebrandt, H., van den Busch, J. L., & Heymans, C. 2020, *A&A*, **637**, A100
- Wu, H.-Y., Zentner, A. R., & Wechsler, R. H. 2010, *ApJ*, **713**, 856
- Wu, H.-Y., Weinberg, D. H., Salcedo, A. N., Winking, B. D., & Zu, Y. 2019, *MNRAS*, **490**, 2606
- Wu, H.-Y., Costanzi, M., To, C.-H., et al. 2022, *MNRAS*, **515**, 4471
- Zhang, Y., Jeltrema, T., Hollowood, D. L., et al. 2019, *MNRAS*, **487**, 2578
- Zhang, Z., Wu, H.-Y., Zhang, Y., et al. 2023, *MNRAS*, **523**, 1994
- Zhang, Z., Farahi, A., Nagai, D., et al. 2024, *MNRAS*, **530**, 3127
- Zhou, C., Wu, H.-Y., Salcedo, A. N., et al. 2024, *Phys. Rev. D*, **110**, 103508
- Zubeldia, Í., & Challinor, A. 2019, *MNRAS*, **489**, 401
- Zubeldia, Í., Bolliet, B., Challinor, A., & Handley, W. 2025, *Phys. Rev. D*, **112**, 083536

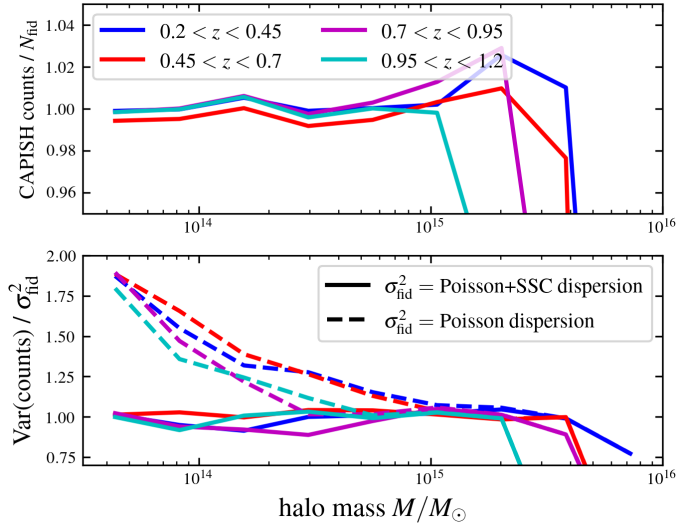


Fig. A.1. *Top:* Mean Capish halo count compared to analytical prediction. *Bottom:* Variance of Capish halo counts compared to Poisson-only or Poisson+SSC variance analytical prediction.

Appendix A: SSC validation at the level of the halo mass distribution

We show in the top panel of Fig. A.1 the comparison between the mean Capish halo count in bins of true mass and true redshift, and the standard prediction based on Tinker et al. (2010) halo mass function. The lower panel compares the variance of Capish counts in the same mass and redshift bins, to the Poisson-only variance prediction, and the full Poisson+SSC variance prediction.

Appendix B: Error model for lensing mass

For a single halo with mass m and redshift z , the Fisher information on the weak lensing mass is given by

$$F_{MM} = \int_{R_{\min}}^{R_{\max}} \sigma_{\Delta\Sigma}^{-2}(R) \left(\frac{\partial \Delta\Sigma(R)}{\partial M} \right)^2 dR. \quad (\text{B.1})$$

In the above equation, the integral runs over the available radial range that is used for the weak lensing mass fitting. The quantity $\Delta\Sigma(R) = \Delta\Sigma_{\text{NFW}}(R|M, c)$ denotes the predicted excess surface density for an NFW (Navarro et al. 1997) halo of halo mass m , concentration $c(m, z)$, taken to follow the (Duffy et al. 2008) concentration mass-relation. The per-radius dispersion of the excess surface density profile entering Eq. (B.1) has several sources listed in Sect. 2.2. We focus on the shape and shot noise of the source galaxy sample, given by

$$\sigma_{\Delta\Sigma}^2(R) = \frac{\Sigma_{\text{crit}}^2(z_{\text{cl}}) \sigma_{\epsilon}^2}{n_{\text{gal}}^{\text{bgd}}(z_{\text{cl}}) 2\pi R}, \quad (\text{B.2})$$

where σ_{ϵ} is the intrinsic galaxy shape noise and $n_{\text{gal}}^{\text{bgd}}(z)$ is the surface density of background galaxies behind a cluster at redshift z , given by

$$n_{\text{gal}}^{\text{bgd}}(z_{\text{cl}}) = \bar{n}_{\text{gal}} \frac{\int_{z_{\text{cl}}+0.2}^{\infty} n_{\text{gal}}(z_s) dz_s}{\int_0^{\infty} n_{\text{gal}}(z_s) dz_s}. \quad (\text{B.3})$$

where the Chang et al. (2013) normalized redshift distribution is given by

$$n_{\text{gal}}(z_s) \propto z_s^{\alpha_{\text{gal}}} \exp\left(-\left(\frac{z_s}{z_{\text{gal}}^0}\right)^{\beta_{\text{gal}}}\right), \quad (\text{B.4})$$

where $\alpha_{\text{gal}}, \beta_{\text{gal}}$ and z_{gal}^0 are given in Table 1. The effective critical surface density is averaged over the source redshift distribution $n(z_s)$, such as

$$\Sigma_{\text{crit}}^{-2}(z_{\text{cl}}) = \frac{\int_{z_{\text{cl}}+0.2}^{\infty} n_{\text{gal}}(z_s) \Sigma_{\text{crit}}^{-2}(z_{\text{cl}}, z_s) dz_s}{\int_{z_{\text{cl}}+0.2}^{\infty} n_{\text{gal}}(z_s) dz_s}. \quad (\text{B.5})$$

Then, we get that

$$\sigma_{\text{WL}}(m, z) = \frac{\sqrt{F_{MM}^{-1}}}{\ln(10)m}, \quad (\text{B.6})$$

represented in Fig. B.1 as a function of mass and redshift.

Appendix C: Detected/full halo population after minimal richness cut

Figure C.1 illustrates the fraction of detected Capish halos for 3 different minimal cuts in cluster's richness, and for 3 different values of the cluster richness-mass relation intrinsic scatter $\sigma_{\ln \lambda_{\text{int}}}$ in Eq. (14).

Appendix D: Alternative interpretation of completeness

In this appendix we discuss an alternative implementation of the survey selection function. In addition to purity, denoting the contamination level of the cluster catalog by spurious detections (implemented in Capish), we discuss here the possibility of including an independent completeness function, representing the fraction of the underlying dark matter halo population that is systematically missing from our dataset (Mantz et al. 2010; Mantz 2019). Following the formalism in Agüena & Lima (2018), Payerne et al. (2025), Lesci et al. (2022, 2025), the expected cluster number count per interval of halo mass, true redshift, observed redshift, and observed richness is given by

$$\frac{\partial^4 N(m, z, \lambda_{\text{obs}}, z_{\text{obs}})}{\partial \lambda_{\text{obs}} \partial m \partial z_{\text{obs}} \partial z} = \frac{\partial^2 N(m, z)}{\partial m \partial z} \frac{c(m, z)}{p(\lambda_{\text{obs}}, z_{\text{obs}})} \times \frac{P(\lambda_{\text{obs}} | m, z) P(z_{\text{obs}} | z)}{P(\lambda_{\text{obs}} | m, z) P(z_{\text{obs}} | z)}, \quad (\text{D.1})$$

where $\frac{\partial^2 N(m, z)}{\partial m \partial z}$ denotes for the mass and redshift distribution of the halo population (encoded in the halo mass function and comoving volume), $P(\lambda_{\text{obs}} | m, z)$ is the cluster scaling relation, $P(z_{\text{obs}} | z)$ is the observed redshift-true redshift relation, and $p(\lambda_{\text{obs}}, z_{\text{obs}})$ (respectively, $c(m, z)$) is the purity (respectively, the completeness) of the cluster finder algorithm. From Agüena & Lima (2018), completeness can be modeled as purity with a smoothed step function

$$c(m, z) = \frac{(m/m_c)^{n_{\text{comp}}(z)}}{1 + (m/m_c)^{n_{\text{comp}}(z)}}. \quad (\text{D.2})$$

From a simulation point of view, for each halo with mass and redshift m, z , we generate $u \sim U(0, 1)$, if $u > c(m, z)$, the halo is removed from the catalog. Along with the implementation of purity in Sect. 2.2, completeness reproduces the Eq. (D.1).

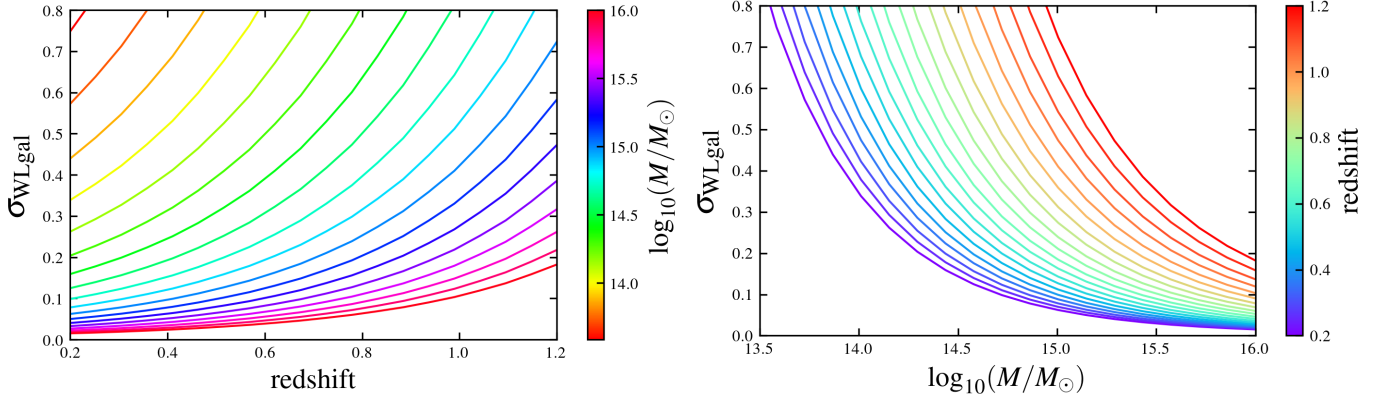


Fig. B.1. *Left:* Dispersion on the $\log_{10} m_{\text{WL}}$ as a function of redshift for different halo mass values. *Right:* Dispersion on the $\log_{10} m_{\text{WL}}$ as a function of halo mass for different redshift values.

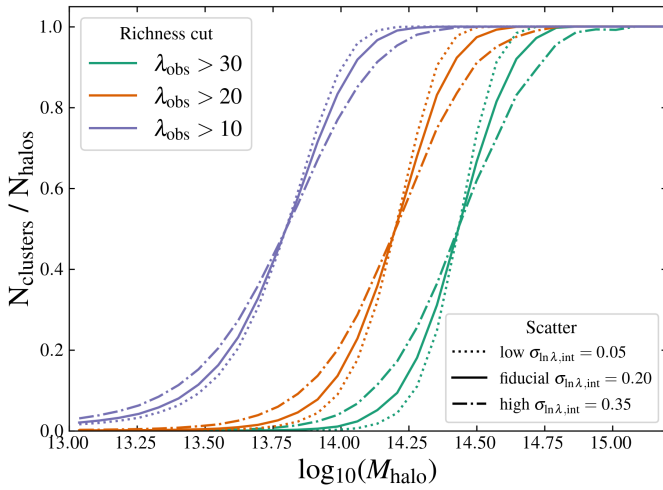


Fig. C.1. Completeness of the simulated cluster catalog as a function of halo mass for different observed richness thresholds, and intrinsic richness–mass scatter $\sigma_{\ln \lambda, \text{int}}$ in Eq. (14). Lower richness thresholds and smaller intrinsic richness–mass scatter yield higher completeness at fixed mass, with all curves asymptoting to unity at high mass.

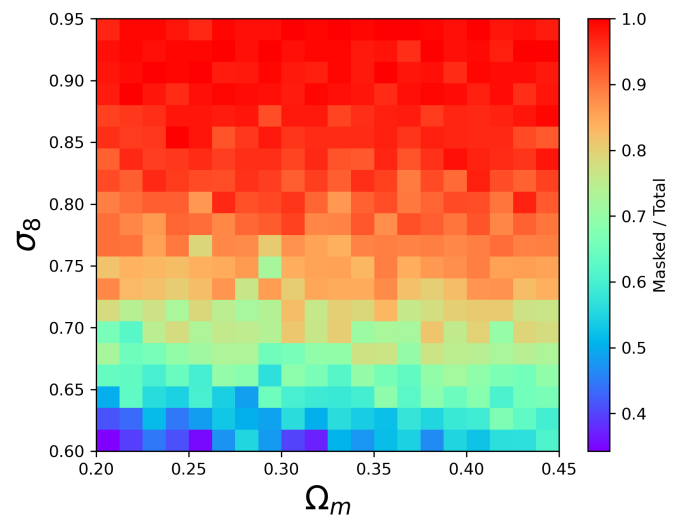


Fig. E.1. Ratio between the two-dimensional distribution of sampled parameters in the Ω_m – σ_8 plane and the same distribution after masking simulations that contain at least one empty richness–redshift bin.

Appendix E: Dealing with empty bins

In Fig. E.1 we show the ratio map between (i) the full set of 30 000 parameter samples θ_k drawn from the priors defined in Sect. 3.2 and (ii) the same parameter samples after masking those that produce binned cluster counts with at least one empty bin. We find that most of the excluded samples are concentrated in the lower-left region of the two-dimensional parameter space, corresponding to low values of Ω_m and σ_8 .

This selective removal affects the posterior training by effectively imposing a smooth prior in the low- Ω_m –low- σ_8 region. In Fig. E.2 we show the posterior distributions obtained for the idealized data vector using three configurations: `count_masked`, `log10m`, and `count_log10m`, where the counts are masked, and the mean logarithmic mass can therefore be computed. Compared to the unmasked case using the appropriate Nm statistics, the resulting posteriors exhibit a small but noticeable shift.

Appendix F: Probability coverage test

As discussed in Sect. 3.2, Bayesian robustness tests are crucial to assess the performance and reliability of NDE for cosmological

parameter inference. For each of the 60 000 simulations drawn from the prior defined in Sect. 3.2, we draw 5000 samples from the corresponding posterior estimates to obtain a γ -credible set, for $\gamma \in [0, 1]$. The fraction of sets that contain the true parameter converges to the expected coverage probability (ECP) p_γ , whose calibration curves are shown in Fig. F.1. We observe a tendency toward overconfidence (empirical coverage p_γ lower than nominal coverage γ). This is an effect of NDE training when the posterior estimation is performed close to the edge of the parameter prior. This problem is not seen when the posterior estimation is performed away from the prior edge. One solution is to train the NDE within large priors, and evaluate the ECP with the posterior estimates within smaller priors (as mentioned in 3.2), giving Fig. F.2 (left panel, shown for the joint `count_Nm` setup).

Appendix G: Recovered parameters versus inputs

Figure G.1 compares the recovered parameters computed as the mean of Capish posteriors (y -axis) versus their input values (x -axis).

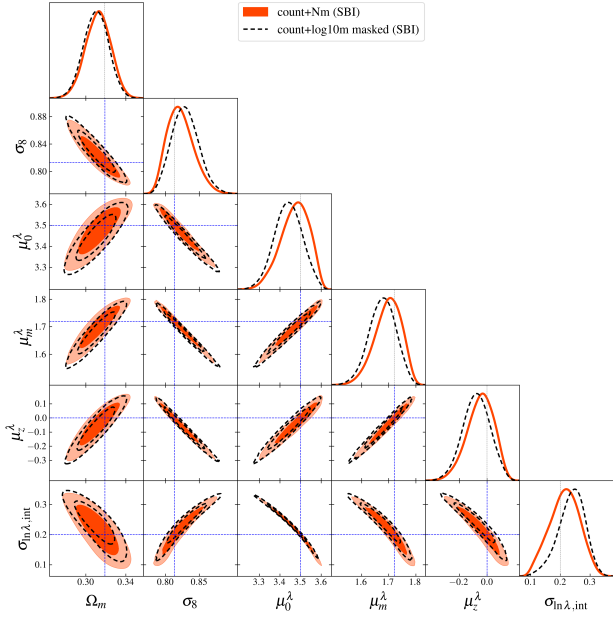


Fig. E.2. Posterior distributions obtained from the trained posterior generator for the three configurations `count`, `Nm`, and `count_Nm`, but masking sampled parameters which generated simulations with at least one empty bin.

Appendix H: Analytical likelihood approach

H.1. Count likelihood

The analytical prediction for the cluster number count is given by (Payerne et al. 2023)

$$N_{ij} = \Omega_S \int_{\lambda_i}^{\lambda_{i+1}} \int_{z_j}^{z_{j+1}} \int_{m_{\min}}^{m_{\max}} dz dm d\lambda_{\text{obs}} P(\lambda_{\text{obs}}|m, z) n_h(m, z). \quad (\text{H.1})$$

In the above equation, we consider the mass-richness relation $\ln \lambda_{\text{obs}} \sim \mathcal{N}(\langle \ln \lambda_{\text{obs}} | m, z \rangle, \sigma_{\ln \lambda}^2)$ where $\langle \ln \lambda_{\text{obs}} | m, z \rangle$ is given in Eq. (13) and $\sigma_{\ln \lambda}^2$ is given in Eq. (14), no selection function and no photometric redshift, i.e., $z_{\text{true}} = z_{\text{phot}}$. We use the Gaussian likelihood given by

$$\mathcal{L}_{\text{BLC}}^{\text{Gauss-SN+SSC}} \propto |\Sigma_N|^{-2} \exp \left[-\frac{1}{2} [N - \hat{N}]^T \Sigma_N^{-1} [N - \hat{N}] \right] \quad (\text{H.2})$$

where

$$(\Sigma_N)_{ijkl} = N_{ij} \delta_{ik}^K \delta_{jl}^K + N_{ij} N_{kl} \langle b \rangle_{ij} \langle b \rangle_{kl} S_{jl}, \quad (\text{H.3})$$

is the binned count covariance (Lacasa et al. 2023). The first term is the Poisson shot noise, and the second is the binned SSC contribution, and $\langle b \rangle_{ij}$ is the average halo bias in the given richness-redshift bin.

H.2. Mass likelihood

For the mean weak-lensing mass, we consider the Gaussian likelihood

$$\mathcal{L}_{\text{MWL}} \propto |\Sigma_M|^{-2} \exp \left[-\frac{1}{2} \sum_{ij} \left(\frac{\log_{10}(M_{ij}) - \log_{10}(\hat{M}_{ij})}{\sigma_{\log_{10} M_{ij}}} \right)^2 \right], \quad (\text{H.4})$$

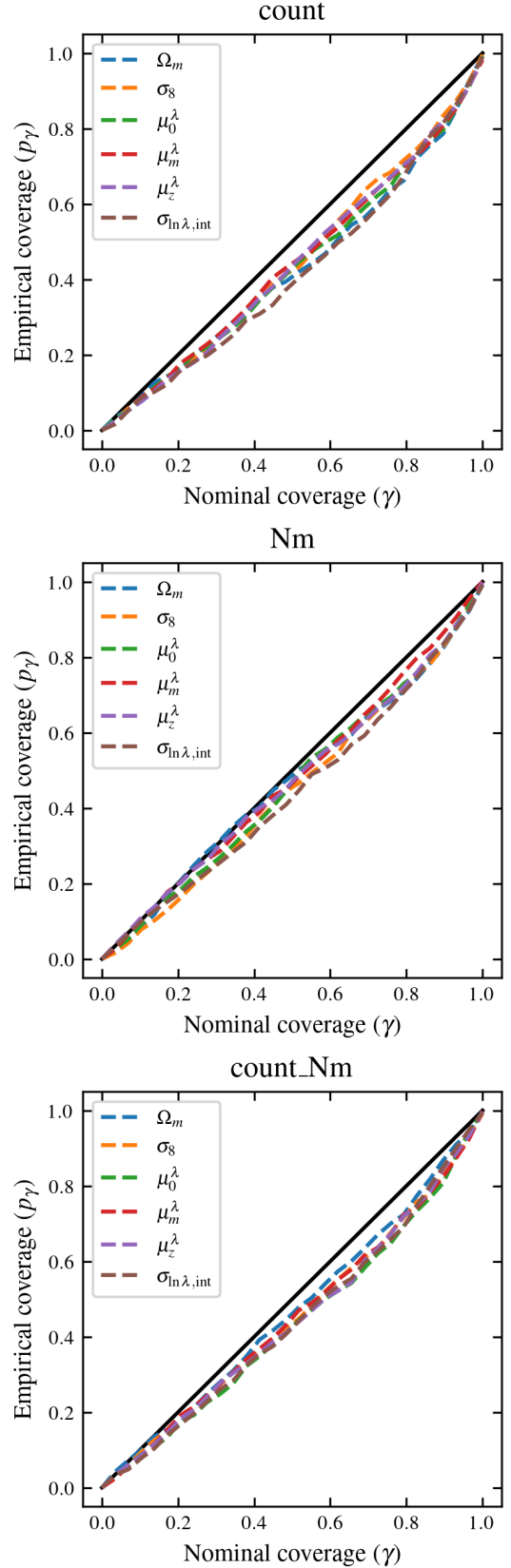


Fig. F.1. ECP computed for the three posterior estimates. *Top*: Trained on counts only (`count`); *middle*: trained on the product of counts and mean masses (`Nm`). *Bottom*: Trained on the combination of counts and mean masses (`count_Nm`). Each panel shows the fraction of true parameter values falling within the corresponding credible intervals, illustrating the calibration of the SBI posteriors.

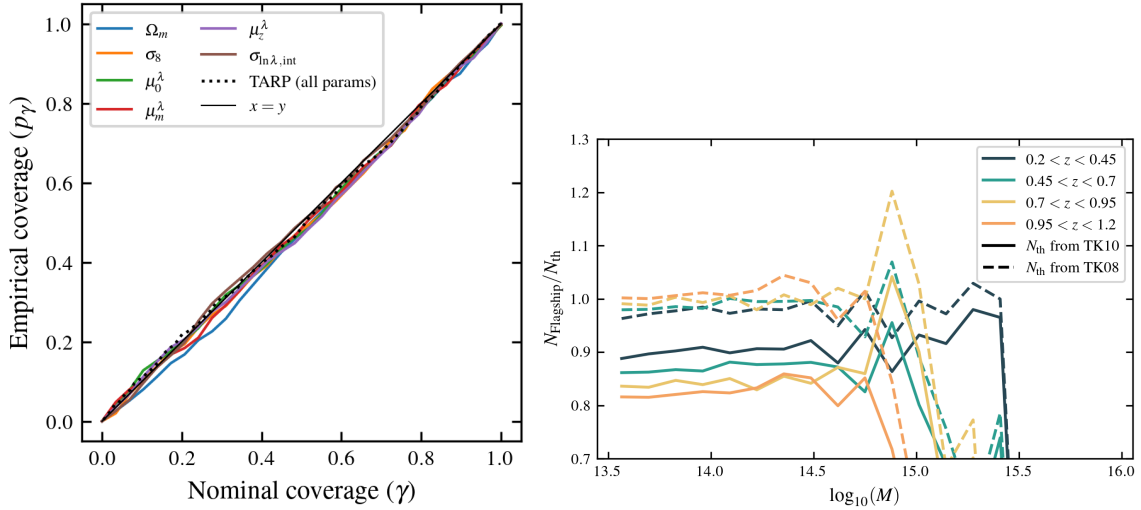


Fig. F.2. *Left:* Probability coverage tests using the estimated posterior over 6 Capish parameters, given the observation of counts and masses (count_Nm). To prevent inaccuracies in the density estimation near prior boundaries, the neural posterior is trained on a broad prior, then restricted to a tighter one. The ECP calibration is computed for each parameter, and the TARP calibration is computed for all parameters at once. *Right:* Halo abundance measured in the *Euclid* Flagship simulation, compared to halo abundance prediction based on Tinker et al. (2010, 2008).

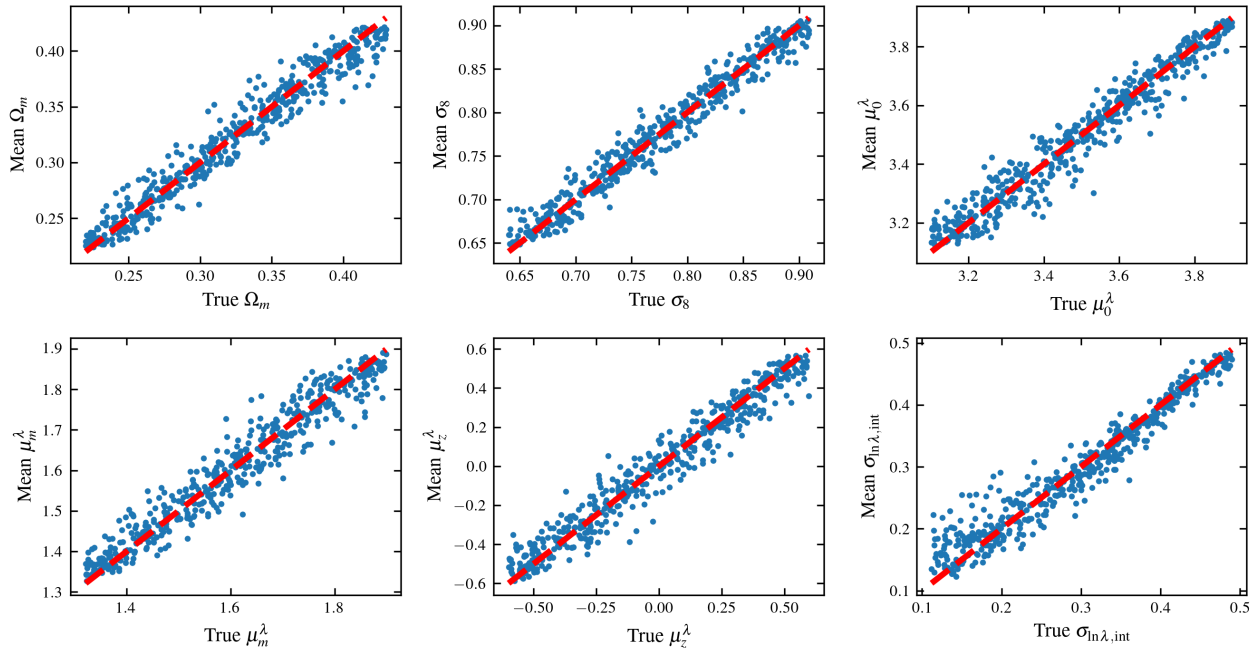


Fig. G.1. Recovered mean parameters versus true input values for the count_Nm configuration, computed from 500 simulated pairs $\{\widehat{\theta}_{\text{true},k}, \widehat{D}_k\}$. The red lines indicate the ideal one-to-one relation $x = y$.

where $\widehat{M}_{ij}$ is the observed mean mass in richness–redshift bin (i, j) and $\Sigma_{\log_{10} M}$ is the diagonal covariance of $\log_{10} M$ uncertainties. The theoretical mean mass in each bin is

$$M_{ij}^{\Gamma} = \frac{\Omega_S}{\bar{N}_{ij}} \int_{\lambda_i}^{\lambda_{i+1}} \int_{z_j}^{z_{j+1}} \int_{m_{\min}}^{m_{\max}} m^{\Gamma} W(z) \times P(\lambda_{\text{obs}}|m, z) n_h(m, z) dm d\lambda_{\text{obs}} dz, \quad (\text{H.5})$$

where $\bar{N}_{ij}$ is identical to N_{ij} in Eq. (H.1) with the addition of the window function $W(z)$. The uncertainty on the mean logarithmic mass is propagated from the Fisher information of the

weak-lensing profile for a single cluster:

$$\sigma_{\log_{10} M_{ij}}(M_{ij}) = \frac{\sigma_M(M_{ij})}{\ln(10) M_{ij}} \frac{1}{\sqrt{N_{ij}}}. \quad (\text{H.6})$$

The uncertainty on the cluster mass σ_M is obtained from the inverse Fisher matrix constructed for the parameters $\theta = (M, c, A)$, where c is the NFW concentration and A is a global multiplicative amplitude factor. The Fisher matrix is defined as

$$F_{\alpha\beta} = \int_{R_{\min}}^{R_{\max}} \frac{1}{\sigma_{\Delta\Sigma}^2(R)} \frac{\partial \Delta\Sigma(R)}{\partial \theta_{\alpha}} \frac{\partial \Delta\Sigma(R)}{\partial \theta_{\beta}} dR, \quad (\text{H.7})$$

with $\Delta\Sigma(R) = A \times \Delta\Sigma_{\text{NFW}}(R|m, c)$ the predicted excess surface density for an NFW halo of mass M , concentration c , and ampli-

tude A . The amplitude A denotes the possible mis-calibration of source photometric redshifts and shape measurement error. The per-radius dispersion of the excess surface density profile $\sigma_{\Delta\Sigma}^2(R)$ can be modeled as in Eq. (B.2). Gaussian priors with width σ_α^2 can be used to constrain better the systematics like $F_{\alpha\alpha} \rightarrow F_{\alpha\alpha} + 1/\sigma_\alpha^2$. The covariance matrix is then $C_{\alpha\beta} = (F^{-1})_{\alpha\beta}$ and $\sigma_M = \sqrt{C_{MM}}$. In this work, we put priors of zero width for the amplitude A and for the concentration c , to mimic the error model for individual lensing masses in Appendix B (which uses a fixed concentration-mass relation, and $A = 1$).

Appendix I: Flagship underlying halo mass function

We compute the halo abundance (as described in Sect. 2.1) and compare it with the abundance of dark matter halos identified in the *Euclid* Flagship simulation. The measured abundance is compared to predictions from the halo mass function implementations of Tinker et al. (2010) and Tinker et al. (2008). As shown in Fig. F.2 (right panel), the bias between the simulated and predicted abundances highlights—consistent with Euclid Collaboration: Castander et al. (2025)—that the Tinker et al. (2008) formulation provides a better match to the simulated dataset.

Appendix J: Cosmological and scaling relation parameter fits

We recap in Table J.1 the different constraints on the cosmological parameters Ω_m and σ_8 , as well as on the scaling-relation parameters, obtained in this work. The first line corresponds to the fiducial values, the second block to the SBI results, and the final block to the analytic likelihood results.

Table J.1. Means and 1σ dispersions of parameter posteriors for Ω_m , σ_8 , μ_0^λ , μ_m^λ , μ_c^λ , $\sigma_{\ln \lambda, \text{int}}$ obtained in this work.

Parameters	Ω_m	σ_8	μ_0^λ	μ_m^λ	μ_c^λ	$\sigma_{\ln \lambda, \text{int}}$
Fiducial	0.319	0.813	3.5	1.72	0	0.2
count (SBI)	0.343 ± 0.058	0.845 ± 0.055	3.396 ± 0.174	1.704 ± 0.078	-0.029 ± 0.176	0.214 ± 0.053
Nm (SBI)	0.314 ± 0.029	0.804 ± 0.068	3.533 ± 0.255	1.649 ± 0.136	-0.094 ± 0.221	0.243 ± 0.077
count+Nm (SBI)	0.312 ± 0.012	0.823 ± 0.019	3.472 ± 0.066	1.697 ± 0.048	-0.049 ± 0.093	0.215 ± 0.049
count-masked (SBI)	0.330 ± 0.061	0.842 ± 0.062	3.417 ± 0.172	1.694 ± 0.099	-0.049 ± 0.219	0.217 ± 0.054
log10m (SBI)	0.330 ± 0.059	0.871 ± 0.041	3.530 ± 0.061	1.732 ± 0.047	0.055 ± 0.083	0.197 ± 0.048
count+log10m masked (SBI)	0.310 ± 0.012	0.829 ± 0.019	3.441 ± 0.065	1.677 ± 0.047	-0.077 ± 0.094	0.237 ± 0.044
count+Nm with selection bias fixed (SBI)	0.317 ± 0.011	0.816 ± 0.015	3.491 ± 0.055	1.713 ± 0.040	-0.013 ± 0.075	0.202 ± 0.043
count+Nm with free selection bias (SBI)	0.317 ± 0.011	0.817 ± 0.019	3.489 ± 0.062	1.708 ± 0.043	-0.017 ± 0.084	0.203 ± 0.045
count+Nm (SBI, one sim.)	0.320 ± 0.013	0.822 ± 0.020	3.451 ± 0.071	1.701 ± 0.051	-0.058 ± 0.098	0.231 ± 0.049
count+Nm (SBI:TK10, Flagship)	0.307 ± 0.013	0.807 ± 0.019	3.482 ± 0.069	1.737 ± 0.050	-0.078 ± 0.099	0.206 ± 0.052
count+Nm (SBI:TK08, Flagship)	0.308 ± 0.015	0.831 ± 0.025	3.434 ± 0.083	1.670 ± 0.062	-0.084 ± 0.122	0.240 ± 0.055
count (analytic likelihood)	0.327 ± 0.048	0.824 ± 0.048	3.459 ± 0.141	1.710 ± 0.083	-0.012 ± 0.180	0.212 ± 0.058
log10m (analytic likelihood)	0.320 ± 0.072	0.823 ± 0.055	3.496 ± 0.057	1.721 ± 0.044	-0.002 ± 0.082	0.204 ± 0.047
count+log10m (analytic likelihood)	0.316 ± 0.008	0.816 ± 0.012	3.490 ± 0.044	1.713 ± 0.032	-0.012 ± 0.062	0.206 ± 0.033

Notes. The first block lists the fiducial values used in this work, the second block lists the constraints on the same parameters obtained with SBI, the third block lists analytic-likelihood constraints.