Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **518,** 1106–1127 (2023)
Advance Access publication 2022 June 3

https://doi.org/10.1093/mnras/stac1404

The Dark Energy Survey supernova program: cosmological biases from supernova photometric classification

M. Vincenzi [1,2,3]★ M. Sullivan [1] A. Möller [4,5] P. Armstrong [6] B. A. Bassett,[7,8,9] D. Brout [10]† D. Carollo,[11] A. Carr,[12] T. M. Davis [12] C. Frohmaier [1,2] L. Galbany [13,14] K. Glazebrook,[4] O. Graur [2,15] L. Kelsey [1,2] R. Kessler,[16,17] E. Kovacs,[18] G. F. Lewis [19] C. Lidman [6,20] U. Malik,[6] R. C. Nichol,[2] B. Popovic,[3] M. Sako,[21] D. Scolnic [3] M. Smith [1] G. Taylor [6] B. E. Tucker,[6] P. Wiseman [1] M. Aguena,[22] S. Allam,[23] J. Annis,[23] J. Asorey,[24] D. Bacon,[2] E. Bertin,[25,26] D. Brooks,[27] D. L. Burke,[28,29] A. Carnero Rosell,[22] J. Carretero,[30] F. J. Castander,[13,14] M. Costanzi,[31,32,33] L. N. da Costa,[22,34] M. E. S. Pereira,[35,36] J. De Vicente,[24] S. Desai,[37] H. T. Diehl,[23] P. Doel,[27] S. Everett,[38] I. Ferrero,[39] B. Flaugher,[23] P. Fosalba,[13,14] J. Frieman,[17,23] J. García-Bellido,[40] D. W. Gerdes,[35,41] D. Gruen,[42] G. Gutierrez,[23] S. R. Hinton,[12] D. L. Hollowood,[38] K. Honscheid,[43,44] D. J. James,[10] K. Kuehn,[45,46] N. Kuropatkin,[23] O. Lahav,[27] T. S. Li,[47,48] M. Lima,[22,49] M. A. G. Maia,[22,34] J. L. Marshall,[50] R. Miquel,[30,51] R. Morgan,[52] R. L. C. Ogando,[34] A. Palmese,[53] F. Paz-Chinchón,[54,55] A. Pieres,[22,34] A. A. Plazas Malagón,[56] K. Reil,[29] A. Roodman,[28,29] E. Sanchez,[24] M. Schubnell,[35] S. Serrano,[13,14] I. Sevilla-Noarbe,[24] E. Suchyta,[57] G. Tarle,[35] C. To,[28,29,58] T. N. Varga,[59,60] J. Weller,[59,60] and R. D. Wilkinson[61] (DES Collaboration)

Affiliations are listed at the end of the paper

Accepted 2022 April 22. Received 2022 March 30; in original form 2021 December 1

ABSTRACT

Cosmological analyses of samples of photometrically identified type Ia supernovae (SNe Ia) depend on understanding the effects of 'contamination' from core-collapse and peculiar SN Ia events. We employ a rigorous analysis using the photometric classifier SuperNNova on state-of-the-art simulations of SN samples to determine cosmological biases due to such 'non-Ia' contamination in the Dark Energy Survey (DES) 5-yr SN sample. Depending on the non-Ia SN models used in the SuperNNova training and testing samples, contamination ranges from 0.8 to 3.5 per cent, with a classification efficiency of 97.7–99.5 per cent. Using the Bayesian Estimation Applied to Multiple Species (BEAMS) framework and its extension BBC ('BEAMS with Bias Correction'), we produce a redshift-binned Hubble diagram marginalized over contamination and corrected for selection effects, and use it to constrain the dark energy equation-of-state, w. Assuming a flat universe with Gaussian $\Omega_{\rm M}$ prior of 0.311 ± 0.010, we show that biases on w are <0.008 when using SuperNNova, with systematic uncertainties associated with contamination around 10 per cent of the statistical uncertainty on w for the DES-SN sample. An alternative approach of discarding contaminants using outlier rejection techniques (e.g. Chauvenet's criterion) in place of SuperNNova leads to biases on w that are larger but still modest (0.015–0.03). Finally, we measure biases due to contamination on w_0 and w_a (assuming a flat universe), and find these to be <0.009 in w_0 and <0.108 in w_a, 5 to 10 times smaller than the statistical uncertainties for the DES-SN sample.

Key words: surveys – supernovae: general – cosmology: observations.

1 INTRODUCTION

Type Ia supernovae (SNe Ia) are widely used in cosmology to directly measure the accelerating expansion rate of the universe, and to characterize the properties of the 'dark energy' thought to cause it. Following the original detection of the accelerating cosmic expansion using SNe Ia (Riess et al. 1998; Perlmutter et al. 1999), two decades of time-domain surveys have discovered and followed up thousands of cosmologically useful SNe Ia, from the local universe to redshifts beyond $z \sim 1$. As the statistical power of these samples has improved, there has been a commensurate reduction in systematic uncertainties that has broadly tracked the increase in SN Ia numbers (Astier et al. 2006; Kessler et al. 2009b; Sullivan et al. 2011; Betoule et al. 2014; Rest et al. 2014; Riess et al. 2018; Scolnic et al. 2018; Abbott et al. 2019b). However, unlocking the full constraining power of current and future samples of SNe Ia requires a new level of controlling systematic uncertainties introduced by the use of photometric SN

★ E-mail: maria.vincenzi@duke.edu
† NASA Einstein Fellow.

classification. Modelling and assessing systematic biases introduced by SN classification is the main focus of this paper.

Photometric SN classification methods are needed when candidate SNe detected by a survey lack a spectroscopic confirmation of their type. In these cases, most cosmological analyses to date have been restricted to SN events with spectroscopic redshift from the likely host galaxy, and SN classification is based on the characteristics of the observed light curve. Early approaches were frequently used for individual high-redshift SN events forming part of relatively small samples (e.g. Perlmutter et al. 1999; Riess et al. 2007), albeit often using other contextual information such as host galaxy type. More general approaches include selecting candidate SNe Ia based on their light-curve fit properties (Bazin et al. 2011) and classifying SNe based on both template fitting (e.g. pSNID; Sako et al. 2011, 2018 or González-Gaitán et al. 2014) and machine-learning approaches (Lochner et al. 2016; Möller et al. 2016; Möller & de Boissière 2020).

The outputs from SN photometric classifiers require a careful interpretation, as instead of the simple binary classification associated with spectroscopic classification (i.e. SN Ia or not a SN Ia), photometric classifiers return the probability of each event being a SN Ia, $P_{\rm Ia}$. A framework is needed to marginalize over the contamination from events that are not SNe Ia. The Bayesian Estimation Applied to Multiple Species (BEAMS) method (Kunz, Bassett & Hlozek 2007), and its extension 'BEAMS with Bias Corrections' (BBC; Kessler & Scolnic 2017), are frequently used in this context, the latter also incorporating corrections due to selection effects based on high-quality survey simulations.

The development of photometric classification has been motivated by the recent and future large SN surveys like the Sloan Digital Sky Survey (SDSS) SN Survey (Sako et al. 2018), the Pan-STARRS Medium Deep Survey (Jones et al. 2017, 2018), the Dark Energy Survey SN programme (Bernstein et al. 2012; Smith et al. 2020b), and the future Legacy Survey of Space and Time (LSST; Ivezić et al. 2019). These SN imaging surveys motivated large spectroscopic follow-up programmes to measure host-galaxy redshifts for the majority of discovered SNe, and use them for cosmological measurements. The first measurement of the equation-of-state of dark energy, w, with a photometric SN Ia sample was performed by Campbell et al. (2013) using data from the SDSS SN Survey. They used pSNID, together with a selection of events based on their SN Ia light-curve fit properties, which together reduced contamination in the SN Ia sample to an estimated 3.9 per cent. However, the systematic effects of this contamination on the final measurement of w was not estimated. Hlozek et al. (2012) first demonstrated the application of BEAMS on the SDSS SN sample (similar to the sample used by Campbell et al. 2013), but also lacked an assessment of systematic uncertainties in the analysis.

The cosmological analysis of the Pan-STARRS (PS1) photometric SN sample (Jones et al. 2017, 2018) was the first to include an evaluation of the cosmological biases and systematic uncertainties introduced by contamination in the photometrically classified SN Ia sample. Using several simple classification approaches that do not rely on machine learning, including pSNID, the biases on measurements of w due to contamination were estimated to be small, and the associated systematic uncertainty was estimated to be $\sigma_{w,\rm syst} = 0.012$. This uncertainty is significantly smaller than the total systematic uncertainty on w of 0.043, illustrating that, under the assumptions of this analysis, contamination resulted in a small contribution to the total uncertainty budget.

Recently developed photometric classifiers (Lochner et al. 2016; Möller & de Boissière 2020) have shown a good performance on simulated samples of SNe developed for various classification challenges (Kessler et al. 2010b, 2019a; Hložek et al. 2020). However, a critical issue remains: the training and validation of these classifiers are often performed on the same sample of simulated SN events. These simulated samples are generated either applying the same selection function of the test set, or assuming the training sample is biased towards brighter events due to spectroscopic selection effects. These simulations may not reflect the true diversity of the transient universe, and may require tuning in their input astrophysics to reproduce the observed characteristics of the selected SN sample (Jones et al. 2017, 2018). This procedure can potentially lead to an over-estimation of the classifier performance and thus underestimate systematic uncertainties in measured cosmological parameters. Ultimately, the development of accurate SN survey simulations for the training and validation of these photometric classifiers is at least as important as the development of the classifiers themselves.

This paper investigates biases in the measurement of cosmological parameters that are introduced in the use of photometric SN classification algorithms within the BBC framework. Our focus is on the Dark Energy Survey[1] (DES) SN programme (DES-SN; Smith et al. 2020b) data set. DES-SN is a state-of-the-art sample for SN Ia cosmology analysis, with approximately 2000 likely SNe Ia in the final '5-yr' sample: ∼20 per cent of the SNe have follow-up spectroscopy of the SN itself (e.g. Smith et al. 2020b), and most of the remaining events have a host galaxy spectroscopic redshift (see Lidman et al. 2020).

Vincenzi et al. (2021, hereafter V21) previously presented large simulations of DES-SN that generate realistic samples of transients that accurately describe DES-SN data. The simulation includes the 'normal' SNe Ia, improved core-collapse SN spectral templates (Vincenzi et al. 2019, hereafter V19) and peculiar SNe Ia (SN1991bg-like SNe and SN2002cx-like SNe; Kessler et al. 2019a), as well as the DES survey characteristics, to make accurate predictions for the expected populations of SNe in DES-SN. These simulations demonstrated an excellent agreement between data and simulated SN properties across many parameter distributions, including Hubble residuals and Hubble residual distribution tails. Analysing these simulated samples in detail, and fitting all the detected events with the SALT2 SN Ia light-curve model (Guy et al. 2007), V21 predicted 6–8 per cent of the sample to be comprised of events that are not SNe Ia, after an event selection based on the light-curve properties and fitted SALT2 parameters. No photometric classification algorithm was used.

Here we generate simulations as in V21 to assess the performance of the SuperNNova (SNN) photometric SN classifier (Möller & de Boissière 2020) when applied to DES-SN data. SNN is a deep learning classifier that identifies SNe Ia with high accuracy (see analyses presented by Möller & de Boissière 2020, and Möller et al. 2022). We exploit the BEAMS implementation in the BBC framework to assess the impact of contamination on the cosmological analysis of the DES-SN photometric sample. The strength of our analysis lies in the fact that we use realistic simulations of SNe Ia and non-Ia SN contamination that have been shown to reproduce the general photometric properties of the DES-SN data to high accuracy (V21). We also test the effect of a range of astrophysically plausible core-collapse SN model variations on the final cosmological measurements.

The paper is outlined as follows. In Section 2, we review the DES-SN data set and the simulation infrastructure used in our analysis. Section 3 details our cosmological analysis framework, including distance estimation, BEAMS, and bias corrections. Section 4

[1] https://www.darkenergysurvey.org/

introduces the SNN classifier and assesses its performance on our simulated data sets, and in Section 5 we present an analysis of the cosmological biases introduced by the photometric classification of the DES-SN sample. We conclude in Section 6.

2 DES-SN DATA AND SIMULATIONS

DES is an optical imaging survey designed to constrain the properties of dark energy and other cosmological parameters by combining four different astrophysical probes: weak gravitational lensing, large scale structure, galaxy clusters, and SNe Ia (Abbott et al. 2019a). DES ran for 6 yr and used the Dark Energy Camera (DECam; Flaugher et al. 2015), mounted on the *Blanco* 4-m telescope at the Cerro Tololo Inter-American Observatory. For time-domain science, DES monitored 10 3-deg^2 fields with an average cadence of 7 d in the *griz* filters. Eight of the 10 fields were surveyed to a depth of ∼23.5 mag per visit ('shallow fields'), and the remaining two to a deeper limit of $m \sim 24.5$ mag per visit ('deep fields'), thus extending to $z \sim 1.2$ the redshift limit to detect SNe Ia.

2.1 The DES photometric SN sample

The primary goal of the DES-SN programme is to measure the light curves of a sample of SNe Ia for use in cosmological analyses. In this paper, we use the same DES photometric SN sample as described in V21. This sample includes ∼3600 events that have an identified host galaxy and accurately measured host galaxy spectroscopic redshift, and that pass light-curve quality selection: observations in two filters with at least one epoch with a signal-to-noise ratio (SNR) >5, at least one observation before the estimated time of peak brightness, and one observation after 10 d (rest-frame) after peak brightness.

Following V21, SN host information is derived from the deep coadded images of (Wiseman et al. 2020), and SN light-curve photometry measured using the DES Difference Imaging pipeline (DIFFIMG; Kessler et al. 2015). The quality of the DIFFIMG light curves is adequate for the analysis presented in this paper, but we highlight that the final DES SN light-curves with a more accurate and precise scene modelling photometry (SMP) approach (Astier et al. 2013; Brout et al. 2019a) is in the process of being applied to all DES-SN data. We also note that approximately 200 new host galaxy spectroscopic redshifts have been processed and incorporated into the sample while this analysis was developed. However, in this work we use the V21 sample to maintain consistency with that analysis.

2.1.1 Low-z SN sample

As this paper considers the cosmological impact of our modelling choices and photometric classification methods, we include a 'low-z' (i.e. $z < 0.1$) external SN Ia sample to combine with our DES-SN sample. We include five publicly available low-z samples from the Harvard-Smithsonian Center for Astrophysics (CfA3S, CfA3K, and CfA4; Hicken et al. 2009, 2012), the Carnegie Supernova Project (CSP-1; Contreras et al. 2010), and the Foundation Supernova sample (DR1; Foley et al. 2017). These samples include spectroscopically confirmed SNe Ia only, therefore they are not affected by contamination.

2.2 Simulations

Our SN simulations use SN time-series spectrophotometric templates, rates, luminosity functions, and empirical relationships between SNe and their host galaxies, as well as the DES survey characteristics, to simulate the transient populations detected in the 5 yr of DES-SN. The simulations are presented in detail in V21 and are generated using the supernova analysis software package (SNANA; Kessler et al. 2009a) as described in V21. The simulation and analysis code were orchestrated by the PIPPIN (Hinton & Brout 2020)[2] pipeline.

V21 presented nine DES-SN simulations testing different modelling choices and assumptions. The analysis presented in this paper has been tested for the full set of simulations presented in V21. However, for simplicity we focus on a reduced sample of five simulations, that encapsulate a wide range of scenarios and provides the most informative results. These simulations are:

(i) 'Baseline' a simulation built using the core-collapse SN templates of V19, and luminosity functions presented by Li et al. (2011) and revised as described by V19;
(ii) 'LFs+Offset' same as Baseline, but with the core-collapse SN luminosity functions brightened by 0.5 mag;
(iii) 'Dust(H98)' uses the host-galaxy dust extinction-corrected core-collapse SN templates of V19, using the revised Li et al. (2011) luminosity functions and a dust distribution presented by Hatano et al. (1998);
(iv) 'J17' uses the core-collapse SN templates of Jones et al. (2017, hereafter J17) together with their adjusted luminosity;
(v) 'DES-CC' simulations: uses a new set of core-collapse templates of Hounsel et al. (in preparation; hereafter, DES-CC), built from a magnitude-limited sample ($i < 21.5$) of spectroscopically and photometrically identified non type Ia SNe from DES-SN.

The main characteristics of each simulation are summarized in Table 1. We also consider two simulation subsets, one that includes only SNe Ia and one that includes only SNe Ia and peculiar SNe Ia ('Only pec Ia'). These subsets exclude exclude core-collapse SNe, and are used to disentangle the effects of core-collapse SN contamination from other sources of systematic biases in the analysis.

In all DES-SN simulations, host galaxies are associated with SNe using published SN rates as a function of global galaxy properties (stellar mass and star formation rate). We use separate rates for SNe Ia, peculiar SNe Ia, stripped envelope SNe (type Ib, type Ic, and type IIb SNe) and hydrogen-rich SNe (type II and type IIn SNe; see section 4.5 in V21). We also include the dependence of the SN Ia light-curve shape on host galaxy properties.

We combine the DES-SN simulations with simulations of the low-z SN Ia samples introduced in Section 2.1.1. These samples are simulated following Kessler et al. (2019b, section 7.2) and Jones et al. (2019, section 3.1) and simulate mocks of the CfA (CfA3S, CfA3K, CfA4), CSP-1, and the Foundation Supernova samples.

For both the DES-SN and low-z simulations, we assume the SN Ia intrinsic brightness in rest-frame B-band to be $M_B = -19.365$ and we set the nuisance parameters applied for stretch and colour corrections, α and β, equal to $\alpha = 0.167$, $\beta = 3.1$. Moreover, we use a flat 'Λ cold dark matter' (ΛCDM) cosmological model as input, with a Hubble constant $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and $\Omega_\mathrm{M} = 0.311$ (e.g. Planck Collaboration 2020). We generate 50 realizations of the DES-SN survey and pair these with 50 realizations of the low-z sample. Throughout, the statistical properties of the simulated samples are presented as the mean of the 50 realizations, and uncertainties are measured as the standard deviation.

[2] https://github.com/Samreay/Pippin

Table 1. Summary of core-collapse SN assumptions in the DES-SN simulations.

Label	Template library	Luminosity functions	Dust model	Avg. number of SNe after light-curve selection †	Percentage of Ia, PecIa, II, Ibc
Baseline	V19	Revised Li et al. (2011), Gaussian parameterization	N/A*	1650	93.4, 1.4, 4.5, 0.8
LFs+Offset	V19	Revised Li et al. (2011) + 0.5 mag brightening offset	N/A	1722	90.5, 1.4, 6.8, 1.3
Dust(H98)	dereddened V19	Revised Li et al. (2011), Gaussian parameterization	H98‡	1687	93.2, 1.4, 4.2, 1.1
J17	J17	Adjusted LFs from Li et al. (2011)	N/A	1667	94.3, 1.4, 3.0, 1.4
DES-CC	DES-CC	DES-CC	N/A	1687	91.6, 1.4, 0.5, 6.5

*N/A: Not applicable: core-collapse SN templates are not corrected for host galaxy extinction, and the simulation does not include extinction.
†Selection criteria from Section 2.3, without classification. Numbers are calculated as the mean over 50 realizations of the DES-SN survey. Each simulation include SNe Ia, peculiar SNe Ia and core-collapse SNe. Normal SNe Ia alone account for 1522 events on average.
‡Hatano, Branch & Deaton (1998).

2.3 SN light-curve fitting and selection

We fit all simulated and observed SN light-curves with the SALT2 SN Ia light-curve model (Guy et al. 2007, 2010) using the trained model parameters from Betoule et al. (2014) and a χ^2-minimization program in SNANA. This fit determines several rest-frame parameters under the assumption that the event is a SN Ia: the time of SN peak brightness t_0, a stretch-like (Perlmutter et al. 1997) parameter x_1, a colour parameter c, and the light-curve normalization parameter x_0, as well as their uncertainties (i.e. σ_{t_0} etc.). We select SN events in both simulations and data that are well described by this SALT2 model. This selection is based on the fit parameters, their uncertainties, and the goodness of the light-curve fit ('FitProb')[3]. This is the same selection as used in V21 and in the Joint Light-Curve Analysis sample (JLA; Betoule et al. 2014). In detail, the selection requirements are:

(i) $|x_1| < 3$ and $|c| < 0.3$,
(ii) $\sigma_{x_1} < 1$ and $\sigma_{t_0} < 2$,
(iii) FitProb > 0.001.

The outcome of applying this selection to our data and simulations can be found in Table 2. The result is a data sample of 1676 SNe from DES-SN and 312 low-z SNe (155 SNe from the CfA and CSP samples and 157 from the Foundation sample). Averaging our 50 Baseline simulations, we have 1650 SNe from DES-SN and 400 at low-z (161 SNe Ia from the CfA and CSP samples, and 238 SNe Ia from Foundation).

We also explore a tighter selection on the SN colour c, removing redder SNe using a selection of $-0.3 < c < 0.15$. This further reduces contamination from core-collapse SNe, with a minimal and easy-to-model loss of SNe Ia (see Table 2). This asymmetric colour selection is also motivated by the fact that several analyses have shown that redder SNe Ia exhibit larger scatter on the Hubble diagram (Kelsey et al. 2020; Brout & Scolnic 2021).

3 COSMOLOGICAL ANALYSIS FRAMEWORK

Next, we briefly review the framework used to measure the SN Ia redshift–distance relation ('Hubble diagram') and estimate cosmological parameters from our SN data and simulations. We begin by describing the method used to estimate distances from the SN Ia light curve parameters (Section 3.1). We then present the Hubble diagram fitting method called 'BEAMS with Bias Corrections' (BBC; Kessler & Scolnic 2017). In the BBC method, we implement (i) the method presented by Marriner et al. (2011) to determine SN distances and nuisance parameters (Section 3.1), (ii) the BEAMS formalism (Kunz, Bassett & Hlozek 2007) to marginalize over the contamination from non-Ia SNe (Section 3.2), and (iii) simulated bias corrections to account for survey selection effects (Section 3.4). The main output of the BBC framework is a redshift-binned SN distance–redshift relation corrected for selection effects and core-collapse SN contamination, from which the cosmological parameters can be estimated (Section 3.6). BBC also produces fitted nuisance parameters (Section 2.3). The cosmological analyses framework discussed in this section is illustrated in Fig. 1.

3.1 Distance estimation

The SN Ia distance modulus, $\mu_{\rm obs}$, is (e.g. Tripp 1998; Astier et al. 2006)

$$\mu_{\rm obs} = m_B + \alpha x_1 - \beta c + \mathcal{M}_B + \Delta\mu_{\rm bias}, \quad (1)$$

where $m_B = -2.5\log_{10}(x_0)$ and $\mathcal{M}_B$ is the absolute magnitude of a SN Ia with $x_1 = 0$ and $c = 0$. The global nuisance parameters α and β are determined following the approach presented by Marriner et al. (2011), i.e. fixing the cosmological parameters to some reference values (e.g. $\Omega_{\rm M} = 0.3$, $w = -1$) and fitting for distance modulus offsets, $\Delta\mu^b$, evaluated at different (log-spaced) redshift bins. A correction, $\Delta\mu_{\rm bias}$, is applied to each SN to correct for selection effects from the survey and analysis (see Section 3.4).

We neglect the dependence between $\mu_{\rm obs}$ and host galaxy properties in our simulations and fitting (e.g. Sullivan et al. 2010). These correlations can shift the dark energy equation-of-state w by approximately 1 per cent (Smith et al. 2020a) but ignoring them has negligible impact on studies of systematics related to contamination.

3.2 The BEAMS likelihood

BEAMS is a Bayesian framework for using photometric classifications of SNe Ia, and their probabilities, in cosmology. The BEAMS likelihood requires for each SN an estimate of its probability of being a SN Ia, $P_{\rm Ia}$. This set of probabilities are generally determined using photometric classifiers.

The BEAMS formalism is implemented in BBC, and used to fit for a binned Hubble diagram. We define the binned Hubble diagram as a set of binned distance modulii, $\mu^b_{\rm Ia}$, evaluated for each of the $N_{\rm bins}$ redshift bins.[4] The binned distance modulii $\mu^b_{\rm Ia}$ are estimated

[3]FitProb $\in$ [0,1] and is the computed probability from χ^2 and number of degrees of freedom, and assuming Gaussian-distributed errors. It quantifies how well each light curve is described by the SALT2 model.

[4]We note that this binned Hubble diagram $\mu^b_{\rm Ia}$ is distinct from the distance modulus for individual events in equation (1).

Table 2. Number of observed and simulated SNe following the application of various selection criteria.

Selection criteria	Data			Simulations (avg over 50 realizations [c])		
	DES-SN	Low-z	Total	DES-SN	Low-z	Total
SALT2 selection	1676	312	1995	1650	400	2050
SALT2 selection + valid bias correction[a]	1603	288	1891	1588	380	1969
SALT2 selection + Chauvenet's criterion[b]	1561	309	1870	1572	400	1972
SALT2 + valid bias corr + Chauvenet	1533	286	1819	1545	380	1926
SALT2 + valid bias corr + Chauvenet + SALT2 c <0.15	1353	273	1626	1336	361	1697

[a]See Section 3.4 for the definition of valid bias corrections.
[b]See Section 3.5 for a discussion about Chauvenet's criterion and outlier rejection methods.
[c]Number of SNe averaged over 50 realizations ($N_{\rm SNe}$). The typical r.m.s. measured over the 50 realizations is $\sqrt{N_{\rm SNe}}$.

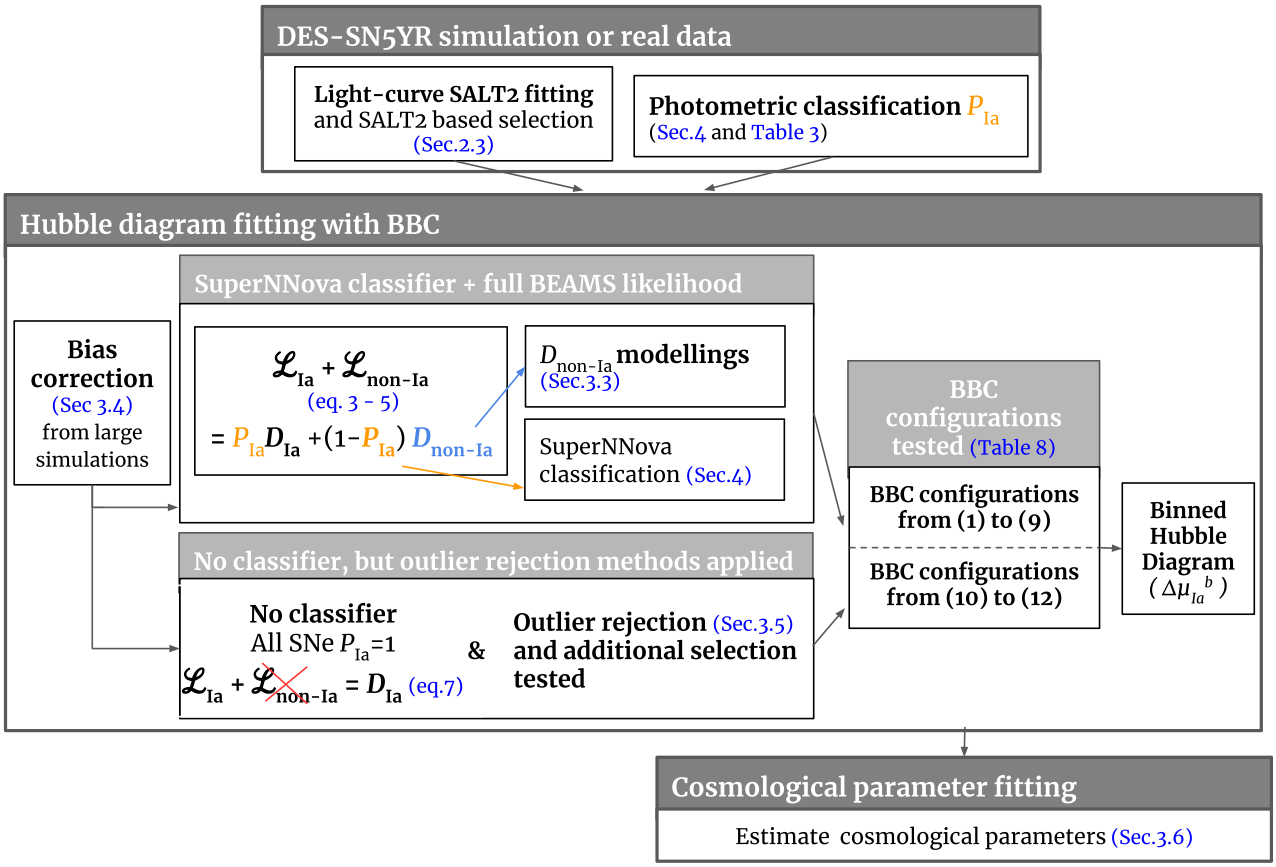

Figure 1. Flow chart of the cosmological analysis framework BBC (Kessler & Scolnic 2017), exploited in this work. BBC is specifically designed to estimate cosmological parameters from samples of photometrically identified SNe Ia. Photometric classifiers are introduced in Section 4, while the different BBC configurations tested in this work are listed in Table 9 and discussed in Section 5.

by maximizing the BEAMS likelihood. This is defined as the sum of two terms, one that models the SN Ia population, $\mathcal{L}_{\rm Ia}$, and the other that models a population of contaminants,

$$\sum_{i=1}^{N_{\rm SNe}} \left(\mathcal{L}^i_{\rm Ia} + \mathcal{L}^i_{\rm non\text{-}Ia}\right). \tag{2}$$

The two terms of the likelihood, $\mathcal{L}^i_{\rm Ia}$ and $\mathcal{L}^i_{\rm non\text{-}Ia}$, are defined as

$$\mathcal{L}^i_{\rm Ia} = \tilde{P}^i_{\rm Ia} \times \exp\left(-\frac{(\mu_{{\rm obs},i} + \Delta\mu^b - \mu_{\rm ref}(z_i))^2}{\sigma^2_{\mu,i}}\right)$$
$$\mathcal{L}^i_{\rm non\text{-}Ia} = (1 - \tilde{P}^i_{\rm Ia}) \times D_{\rm non\text{-}Ia}(z_i, \mu_{{\rm obs},i}, \mu_{{\rm ref},i}), \tag{3}$$

where $\mu_{{\rm ref}(z_i)}$ is the distance modulus of the i-th SN as predicted assuming a fixed reference cosmology ($\Omega_{\rm M} = 0.3, w = -1$), and $\Delta\mu^b$ are the offsets quantifying by how much observations deviate from the reference cosmology in each redshift bin. By construction, the binned Hubble diagram, $\mu^b_{\rm Ia}$ is equal to $\mu_{{\rm ref}(\langle z\rangle_b)} - \Delta\mu^b$. The distance modulus uncertainties $\sigma_{\mu,i}$ include the uncertainties propagated from the SALT2 light-curve fit (σ_{m_B}, σ_{x_1}, σ_c and relative covariances), the intrinsic SN Ia scatter ($\sigma_{\rm Ia,int}$), and peculiar velocity corrections uncertainties. The SN Ia intrinsic scatter term is determined as discussed by Kessler & Scolnic (2017, section 5.5).

In equation (3), the terms $\tilde{P}^i_{\rm Ia}$ and $(1\text{-}\tilde{P}^i_{\rm Ia})$ are weighting factors applied to the two likelihoods, and represent the 'scaled' probabilities of the i-th SN being a SN Ia and a core-collapse SN or peculiar SN Ia, respectively. The scaled probabilities are defined as

$$\tilde{P}^i_{\rm Ia} = \frac{P^i_{\rm Ia}}{P^i_{\rm tot}} \quad \text{and} \quad \tilde{P}^i_{\rm non\text{-}Ia} = \frac{S_{\rm non\text{-}Ia}\left(1 - P^i_{\rm Ia}\right)}{P^i_{\rm tot}}$$
$$P^i_{\rm tot} = \left(P^i_{\rm Ia} + S_{\rm non\text{-}Ia}\left(1 - P^i_{\rm Ia}\right)\right), \tag{4}$$

where $P_{\rm Ia}^i$ is the probability of the i-th SN being a SN Ia as predicted by a classifier, and $S_{\rm non\text{-}Ia}$ is a scaling factor and an additional free parameter in the minimization of the likelihood. This additional factor enables correcting for inaccurate probabilities[5] and it is equal to one for perfectly calibrated probabilities (see Kunz, Bassett & Hlozek 2007; Jones et al. 2018, for a discussion on the necessity of scaling probabilities). As a result, the free parameters in the BEAMS likelihood minimization are the $N_{\rm bins}$ offset terms Δ_μ^b, the nuisance parameters α and β, the SN Ia intrinsic scatter term $\sigma_{\rm Ia,int}$, and the scaling factor $S_{\rm non\text{-}Ia}$. In this analysis, we use 20 logarithmically equally spaced redshift bins.

Modelling the contamination likelihood term $D_{\rm non\text{-}Ia}$ (equation 3) is more difficult because core-collapse SNe are not standardized by the SALT2 framework. Qualitatively, we expect the distribution of non-Ia SN distance moduli to have a larger scatter and to be shifted from $\mu_{\rm ref}$ by a positive offset because non-Ia SNe are generally fainter than SNe Ia.

As BEAMS is designed to handle both SNe Ia and non SNe Ia, we do not apply a $P_{\rm Ia}$ cut prior to the BBC fit. However, in Appendix A, we discuss the effects (and disadvantages) of combining BEAMS with (for example) a $P_{\rm Ia} > 0.5$ selection and motivate the absence of this cut.

3.3 Modelling the contamination likelihood

We test two different approaches to describe $D_{\rm non\text{-}Ia}$ analytically. The first follows Hlozek et al. (2012), who tested an approximation in which core-collapse SN distance moduli and intrinsic scatter are parametrized similarly to SNe Ia

$$D_{\rm non\text{-}Ia} = \exp\left(-\frac{(\mu_{{\rm obs},i} - \mu_{\rm ref,\ non\text{-}Ia}(z_i))^2}{\sigma_{\mu,i}^2}\right) \tag{5}$$

where

$$\mu_{\rm ref,\ non\text{-}Ia} = \mu_{\rm ref} + \Psi(z) \text{ and } \sigma_{\rm Ia,\ int} \rightarrow \sigma_{\rm non\text{-}Ia,\ int}(z), \tag{6}$$

and $\Psi(z)$ describes the brightness offset of the population of contaminants, and $\sigma_{\rm non\text{-}Ia,int}$ is the redshift dependent intrinsic scatter of contaminants that is included in $\sigma_{\mu,i}$ in equation (5). Both terms are modelled as second order polynomials, the coefficients of which are fitted during the BBC fit. This parametrization introduces six additional free parameters in the likelihood in equation (2). Fig. 2(a) shows an example of the best fit $\Psi(z)$ (and relative $\sigma_{\rm non\text{-}Ia,int}(z)$) measured from the Baseline simulations (Section 2.2).

Kessler & Scolnic (2017) introduced an alternative approach, and determine the term $\mu_{\rm ref,non\text{-}Ia}$ in equation (6) from simulation of core-collapse SNe. The mean and dispersion of the core-collapse SN distance moduli are measured from the simulation at different redshift bins. In this approach, there are no extra free parameters in the BBC fit.

Following this approach, we use our Baseline simulation to derive the core-collapse distribution on the Hubble diagram and we show the simulated $\mu_{\rm ref,non\text{-}Ia}$ versus redshift in Fig. 2b.

3.4 Bias corrections

All SN surveys are affected by selection effects introduced by their flux-limited nature. These effects introduce systematic biases in cosmological analyses of SN Ia samples, and thus SN Ia distances are corrected for such biases (equation 1). The corrections are generally

[5]Photometric classifiers often do not provide calibrated probabilities.

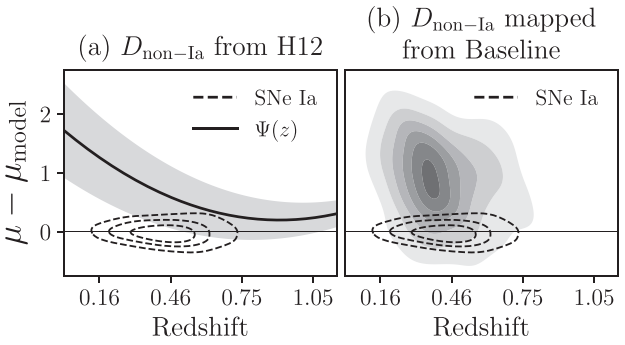

Figure 2. Modelling of core-collapse SN Hubble residuals versus redshift using the different approaches discussed in Section 3.3. Panel (a): the modelling of Hubble residuals for the H12 approach. The black curve and grey shaded region show the best-fitting polynomial $\Psi(z)$ and intrinsic scatter, $\sigma_{\rm non\text{-}Ia}$. For comparison, we also show the Hubble residual distribution for a sample of simulated SNe Ia (dashed contours). Panels (b): as Panel (a), but when applying the approach by Kessler & Scolnic (2017) and using simulations.

estimated using large SN Ia Monte Carlo simulations that accurately model the survey detection efficiency and other potential selection effects (Hamuy & Pinto 1999; Kessler et al. 2009b; Perrett et al. 2010; Betoule et al. 2014). Early use of simulations modelled distance bias corrections as a function of redshift only (Kessler et al. 2009b; Jones et al. 2018; Betoule et al. 2014), but Scolnic & Kessler (2016) showed that this approach is not adequate because distance biases also depend on colour and stretch.

We estimate bias corrections, $\Delta\mu_{\rm bias}$, using the BBC framework and the simulations following Section 2.2, but including only normal SNe Ia. BBC determines an average $\Delta\mu_{\rm bias}$ in a five dimensional grid $\{z, x_1, c, \alpha, \beta\}$. For each event, the bias is interpolated between neighboring bins in the subspace of $\{z, x_1, c\}$, and also interpolated in a 2×2 grid of α and β (α in [0.12, 0.20] and β in [2.3, 3.6]). The simulations are used to bias correct both the real DES-SN sample and the simulated DES-SN samples. We note that bias corrections are applied prior to the BEAMS likelihood minimization presented in Section 3.2 and they have been shown to have a weak dependence over α and β.

The simulations used to model bias corrections include 770 000 DES-SN events and 145 000 low-z SN events (this corresponds to 500 realizations of the DES-SN sample and 500 realizations of the low-z sample). The underlying assumption of BBC is that the bias correction simulation accurately describes the intrinsic properties of the SNe Ia and survey selection effects. Incomplete modelling of one of these aspects may result in inaccurate bias corrections (see Smith et al. 2020a; Popovic et al. 2021, for example). The degree to which core-collapse SN contamination can affect the modelling of the SN Ia intrinsic population (and therefore bias corrections and cosmology) will be explored in future analyses.

In the BBC approach, some cells in the five-dimensional parameter space have too few events (or no events) to reliably estimate bias corrections. SNe in these cells cannot be bias corrected and are rejected from the sample and the cosmological fit. This implicit cut further reduces the sample size, and affects SNe Ia and core-collapse SNe differently. The requirement of a valid bias correction is therefore an implicit photometric classifier for our sample. In Table 2, we report the numbers of SNe for which a valid bias correction cannot be estimated. In the low-z sample, 24 observed SNe Ia do not have valid bias corrections (approximately 8 per cent of the low redshift sample), and the simulated prediction is 18 SNe Ia on average, in good agreement with the data. In the DES-SN samples, there are 73 SNe without valid bias corrections in the

observed sample (<4 per cent) and the simulated prediction is 61 SNe on average. In our simulations, we find that almost 65 per cent of the SNe without valid bias corrections are core-collapse SNe or peculiar SNe Ia, illustrating the implicit classifier in BBC. We discuss this further in Section 4.4.

3.5 Outlier rejection: Chauvenet's criterion

Following Conley et al. (2011) many cosmological analyses use Chauvenet's criterion (Taylor 1997) to reject outliers on the Hubble diagram (Foley et al. 2017; Scolnic et al. 2018; Brout et al. 2019b), i.e. outliers in $\Delta\mu$. Given the number of SNe in the Hubble diagram and assuming their Hubble residuals are normally distributed around zero, Chauvenet's criterion can be used to identify the probability threshold (or σ cut) above which the expected number of data points is below unity (i.e. less the one event is expected to have such a large deviation from zero).

This approach has been used for samples of spectroscopically confirmed SNe Ia. In analyses of pure SNe Ia samples, Chauvenet's criterion selects normal SNe Ia and rejects atypical events or those that have poorly modelled peculiar velocities (for low redshift SNe especially).

In a photometric SN sample like the DES-SN sample, applying Chauvenet's criterion primarily rejects core-collapse SN contaminants that, in this case, are the main source of outliers in the Hubble diagram. Since we mainly focus on exploiting photometric classifiers to describe contamination (see Section 4), rather than outlier rejection or other sigma-clipping methods, we do not apply Chauvenet's criterion by default. However, we examine the difference in cosmological parameters between using photometric classifiers and applying Chauvenet's criteria with $P_{\rm Ia} = 1$ for all events (see Fig. 1). This second approach is effectively the same approach applied to analyses of spectroscopic SN sample, and it enables us to quantify cosmological biases from naively analysing a contaminated SN sample as a pure sample of spectroscopically confirmed SNe Ia.

For simplicity, we apply Chauvenet's criterion before the BBC fit, using approximate Hubble residuals computed from initial values of the nuisance parameters ($\alpha = 0.14$, $\beta = 3.1$) and our reference cosmology.

For our sample of 1995 SNe following SALT2 selection (Section 2.3), Chauvenet's criterion corresponds to a 4σ cut. This cut may affect the low-z and DES-SN samples in different ways. For the low-z sample, Chauvenet's criterion selects normal SNe Ia and rejects atypical events or those that have poorly modelled peculiar velocities. In the DES-SN sample, the criterion primarily affects core-collapse SN contaminants. To avoid conflating the different effects of Chauvenet's criterion, we *always* apply Chauvenet's criterion to the low-z sample, effectively freezing these samples across our tests.

Applying Chauvenet's criterion to our observed samples removes no SNe Ia from the Foundation sample,[6] three SNe Ia from the CfA+CSP samples, and 122 SNe from the DES-SN sample (approximately 7 per cent of the sample). From our simulated low-z samples, we predict no loss of low-z SNe after applying the criterion because our low-z simulation consists of normal SNe Ia without contamination. For the DES-SN sample we predict a reduction from an average of 1650 SNe to 1572 SNe (a loss of 78 SNe, approximately 5 per cent of the sample) using the 'Baseline' simulation, in slight tension with the data. Table 2 summarizes these numbers.

[6]Chauvenet's criterion has already been applied to the Foundation DR1 sample and removes nine SNe Ia (5 per cent of the sample). See table 7 by Foley et al. (2017).

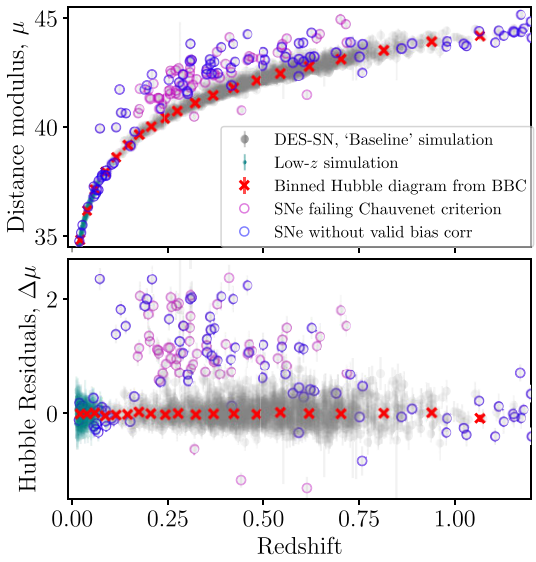

Figure 3. Simulated Hubble diagram (upper panel) and Hubble residuals (lower panel) for a single realization of the DES-SN sample (Baseline simulation, grey symbols) and the low-z sample (teal symbols). We apply the SALT2 selection criteria described in Section 2.3, but no other selection. SNe without a valid bias correction (Section 3.4) and/or failing Chauvenet's criterion (Section 3.5) are indicated with different colours.

In Fig. 3, we show an example of how Chauvenet's criterion and BBC bias corrections affect a simulated sample of SNe. The figure presents one realization of the DES and low-z simulations and highlights SNe that do not pass Chauvenet's criterion and that do not have a valid BBC bias correction.

3.6 Cosmological parameter estimation

The output of the BBC fit is a redshift-binned Hubble diagram corrected for selection effects and contamination, and the associated diagonal covariance matrix, $C_{\rm stat}$, that includes statistical uncertainties only. As a result of the binning, the dimension of the covariance matrix is reduced from $N_{\rm SNe}$ to $N_{\rm bins}$.

We note that binning the Hubble diagram may inflate systematic uncertainties that are not primarily redshift dependent (Brout, Hinton & Scolnic 2020). We will illustrate this uncertainty inflation for some systematics associated with SN photometric classification (Section 4.3), which may be self-calibrated in an unbinned approach.

Finally, we estimate cosmological parameters. We test two cosmological models: a flat wCDM model and a flat w_0w_aCDM model. In both models, the dark energy equation-of-state is parametrized as $\rho \propto a^{-3(1+w)}$, where ρ is the dark energy density and a is the scale factor and it is $a = (1+z)^{-1}$; however, while a wCDM model assumes constant w, a w_0w_aCDM model assumes $w = w_0 + w_a(1-a)$. Unless otherwise stated, we measure cosmological parameters assuming a prior on $\Omega_{\rm M}$ of 0.311 ± 0.010, following the cosmic microwave background measurements published by Planck Collaboration (2020). In future cosmological analyses of the DES photometric SN sample, SN constraints will be combined with the full cosmic microwave background (CMB) likelihood from Planck Collaboration (2020). In Section 5.1.3, we will show that CMB constraints constitute a more stringent prior compared to a

Gaussian Ω_M prior, and thus contribute to reduce both w-biases due to contamination and statistical uncertainty on w.

When testing a flat wCDM model, we measure cosmological parameters using a simple χ^2-minimization program that has evolved from the analysis of Conley et al. (2011). This program evaluates the χ^2 between $\mu^b_{\rm Ia}$ produced by BBC and $\mu_{\rm ref}$ over a grid of Ω_M, w and $\mathcal{M}_B$ values (assuming a flat universe) and estimate Ω_M and w marginalized over $\mathcal{M}_B$ (see Goliath et al. 2001 for a description of the χ^2 definition and marginalization). This program does not provide the full posterior distribution of the cosmological parameters we are interested to constrain. However, it is faster than most cosmological fitting programs and it is adequate for measuring biases on w.

To measure cosmological contours and to test a flat w_0w_aCDM model, we use the Cosmological Monte Carlo software COSMOMC (Lewis & Bridle 2002). For the DES-SN data, absolute estimates of the cosmological parameters are blinded and only relative *differences* between cosmological fits are examined.

4 PHOTOMETRIC CLASSIFICATION

We use the SuperNNova (SNN; Möller & de Boissière 2020) framework to perform photometric classification of our observed and simulated SN data sets, and measure for each SN event its probability of being a SN Ia, $P_{\rm Ia}$. We choose SNN as the code is publicly available, and SNN has demonstrated good classification performance in the literature. For comparison with SNN, we also use two simple algorithms to assign $P_{\rm Ia}$:

(i) `Perfect`: an ideal classifier, that assigns $P_{\rm Ia} = 1$ to SNe Ia and $P_{\rm Ia} = 0$ to peculiar SNe Ia or core-collapse SNe. This approach can only be used in simulations, where the true types are known;
(ii) `AllSNIa`: a classifier that assigns $P_{\rm Ia} = 1$ to every SN.

4.1 SuperNNova

SNN is an open-source[7] machine learning algorithm that implements Recurrent Neural Networks for photometric classification of SNe. It is trained to classify different types of transients using photometric data only (i.e. fluxes and flux uncertainties in different filters) and, optionally, redshift information. It does not rely on feature extraction or light-curve fitting.

Several metrics can be used to assess the performance of SNN. In the binary classification method, these are based on the number of true positives (TPs; SNe Ia correctly classified as such), true negatives (TNs; core-collapse SNe correctly classified as such), false positives (FPs; core-collapse SNe incorrectly identified as SNe Ia), and false negatives (FNs; SNe Ia identified as core-collapse). Following Möller & de Boissière (2020), the contamination (by core-collapse SNe, or peculiar SNe Ia) of the classified photometric SN Ia sample and the classification efficiency are defined as

$$\text{Contamination} = \frac{\text{FP}}{\text{FP} + \text{TP}} \tag{7}$$

and

$$\text{Efficiency} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

We implement SNN using the same hyper-parameters as Möller & de Boissière (2020), and include spectroscopic redshift information.

For our analysis, we normalize the input fluxes using the 'cosmo' method (Möller et al. 2022). In this method, each SN multiband light curve is normalized independently and the normalization factor is the SN maximum flux (in any filter). This method makes SNN agnostic to the relative differences in apparent brightness between SNe, while preserving colour and signal-to-noise information (flux uncertainties are normalized using the same factor as for fluxes). With this normalization, rescaled fluxes close to zero correspond to early/late data points and rescaled fluxes close to one correspond to data points around peak brightness.

We also test an alternative normalization method labelled as 'global'. In this method, the normalization factors are estimated from the full sample of light curves and the same normalization is applied to all light curves. This method preserves the relative brightnesses between different SNe and the full range of magnitudes. As a result, the brightest (lower redshift) SNe have rescaled fluxes closer to one, while faintest SNe have rescaled fluxes closer to zero.

4.2 Training of SNN

SNN requires training on very large samples of SNe (>100 000 events). Combining all SN surveys from the last 15 yr, the sample of spectroscopically confirmed SNe available is around 10 000 events;[8] it is an inhomogeneous sample with an uncertain selection function and biased towards bright, lower-redshift events. To obtain a training sample with sufficient statistics, SNN relies on large simulations where the SN Ia and SN non-Ia rest-frame SED models are derived from spectroscopically confirmed events.

To generate the training samples we combine 100 realizations of our Baseline simulation, apply a simple selection to the simulated events (at least two detections, applying the detection efficiency presented by Kessler et al. 2015), and apply the host galaxy spectroscopic efficiency of V21. We do not apply any additional spectroscopic classification efficiency like the one applied to the training samples generated for the SN classification challenges presented by Kessler et al. (2010a) and The PLAsTiCC team et al. (2018). Moreover, we do not perform SALT2 fits for SNN. We also generate three additional training samples, using the J17 simulation (`SNN(J17)`), the DES-CC simulation (`SNN(DES-CC)`), and the Baseline simulation with host galaxies assigned randomly (`SNN(randomHost)`).

To compare the two different normalizations in SNN, we also train a model using the Baseline simulation and the global normalization method instead of the cosmo normalization (`SNN(global)`). This tests the effects of a classifier that has knowledge of the relative brightnesses between SNe Ia and core-collapse SNe. A summary of the five SNN models and the assumptions in their training simulations is in Table 3.

4.3 Contamination and efficiency

We test SNN on the simulations summarized in Section 2.2, measuring the average contamination and efficiency after our standard selection (Section 2.3) and after requiring $P_{\rm Ia} > 0.5$ cut. As already mentioned in Section 3.2, BBC is designed to handle both SNe Ia and non SNe Ia therefore we do not require a $P_{\rm Ia} > 0.5$ cut in the cosmological sample (see Appendix A).

We first examine the case of no classifier (i.e. `AllSNIa` in Table 4) and SALT2-based selection. Applying only SALT2-based selection

[7] https://github.com/supernnova/SuperNNova

[8] Source: Transient Name Server, https://wis-tns.weizmann.ac.il/

Table 3. Details of the different SNN training samples.

SNN model name	Simulation used for SNN training	Core-collapse SN template library	Normalization	Number of SNe in training sample	Percentage of Ia, pec Ia and core-collapse in the training sample
`SNN(Base)`	Baseline	V19	cosmo	287 000	50, 6, 44
`SNN(J17)`	J17	J17	cosmo	287 000	50, 3, 47
`SNN(DES-CC)`	DES-CC	DES-CC	cosmo	240 000	50, 5, 45
`SNN(global)`	Baseline	V19	global	287 000	50, 6, 44
`SNN(randomHost)`	Baseline, random host association	V19	cosmo	155 700	50, 5, 45

Table 4. Contamination and efficiency measured for the `AllSNIa` classifier (rows) on different simulations (columns) after applying a $P_{\rm Ia} > 0.5$ cut.

Selection criteria	Contamination						Efficiency
	Only pec Ia	Baseline	LFs+Offset	Dust(H98)	J17	DES-CC	(Baseline)
`AllSNIa`, no SALT2 selection [†]	2.6	22.5	31.7	22.0	28.5	25.8	-
`AllSNIa`	2.1	8.2	11.6	8.5	8.7	9.8	100.0
`AllSNIa+Chauvenet`	1.0	3.1	5.3	3.4	3.7	3.2	98.7
`AllSNIa+Chauvenet`, c <0.15	0.7	2.2	4.0	2.3	1.6	2.5	89.4

[†]Fraction of contaminants after SALT2 fit loose cuts of $x_1 \in [-4.9, 4.9]$ and $c \in [-0.49, 0.49]$ (i.e. without applying the SALT2-based selection discussed in Section 2.3; see V21).

Table 5. Contamination and efficiency measured for different SNN models (rows) tested on different simulations (columns).

SNN model[a]	Contamination after *testing* SNN on different simulations						Efficiency
	Only pec Ia	Baseline	LFs+Offset	Dust(H98)	J17	DES-CC	(Baseline)
`SNN(Base)`	0.4	**0.8** [b]	1.1	0.9	1.0	1.4	99.5
`SNN(J17)`	0.7	1.7	2.8	1.9	**1.0** [b]	2.1	99.2
`SNN(DES-CC)`	0.9	2.0	3.2	2.3	1.9	**1.6** [b]	99.0
`SNN(global)`	0.8	2.1	3.5	2.1	1.4	2.3	97.7
`SNN(randomHost)`	0.7	1.3	1.9	1.5	1.3	1.6	98.1

[a]See Table 3 for a description of the training approach utilized for each SNN model.
[b]We highlight in bold the contamination measured using the same simulation both for training and testing.

reduces contamination to less than 12 per cent, a factor of two smaller compared to SN samples before SALT2-based selection. When combined with outlier rejection (`AllSNIa+Chauvenet`, see Section 3.5), the contamination reduces to 4.0–6.6 per cent. A tighter SALT2 colour selection (Section 2.3) combined with Chauvenet's criterion (`AllSNIa+Chauvenet,`c `< 0.15`), reduces the contamination further to 1.6–4.0 per cent. These results set a level of comparison for assessing the performance of SNN.

The performance of the SNN models is shown in Table 5. For the SNN models `SNN(Base)`, `SNN(J17)`, and `SNN(DES-CC)`, the performance is improved compared to outlier rejection methods only, with contamination of 0.8–3.2 per cent and an efficiency equal or above 99 per cent. `SNN(Base)`, trained on our Baseline simulation, performs well not only when tested on Baseline simulations (0.8 per cent contamination), but also when tested on the simulations J17 and DES-CC, with contamination of 1.0 and 1.4 per cent, respectively. In these two cases, the `SNN(Base)` classifier is trained on core-collapse SN templates that are independent from the ones used to generate the simulations, suggesting that the `SNN(Base)` model generalizes well.

By contrast, the `SNN(J17)` and `SNN(DES-CC)` classifiers perform well when tested on simulations generated using the same core-collapse SN models (in bold in Table 5), but when tested on Baseline simulations they predict levels of contamination that are two and three times larger compared to using the `SNN(Base)` model. This difference reflects the increased diversity of contaminants in the Baseline simulation compared to the J17 and DES-CC simulations.

We make two further observations. The first is that, following the application of SNN, peculiar SNe Ia account for around a third to a half of the contamination (Table 5), suggesting that this class of transients plays an important role in our analysis, and that they are as difficult to identify as core-collapse SNe with the current training set and configuration.[9] The second is that, comparing the Baseline and Dust(H98) simulations, we do not observe large differences in the contamination even though none of the SNN models have been trained using the full range of dust extinction included in the Dust(H98) simulation. This result suggests that including dust extinction in the simulations that is unmodelled in the training samples does not significantly affect classification performance.

4.3.1 *Performance as a function of SN Ia properties*

Fig. 4 shows the contamination and efficiency for the Baseline simulation as a function of redshift, fitted x_1 and c, and $\Delta\mu$. These plots identify regions of parameter space where non-Ia SN contamination is higher (or efficiency is lower). The poorest performance in terms of contamination *per-bin* is observed at the extremes of the SALT2 parameter distributions.

Focusing on SALT2 c, contamination increases significantly for very blue events (>20 per cent for $c < -0.2$), mainly due to fast-declining type II and type IIn SNe that are generally bluer than SNe Ia at peak. Similarly, classification is more difficult for redder SNe (>10 per cent contamination and <95 per cent efficiency for $c > 0.2$), where intrinsically redder and lower signal-to-noise stripped envelope SNe are more easily misclassified as red (and therefore also faint) SNe Ia, and vice versa (see Table 6). Contamination is less than 2 per cent for $-0.1 < c < 0.1$, even when only applying the `AllSNIa` classifier and Chauvenet's criterion. For stretch, contamination at

[9]To improve classification of peculiar SNe Ia, the fraction of this sub-type of SNe could be augmented in the training set.

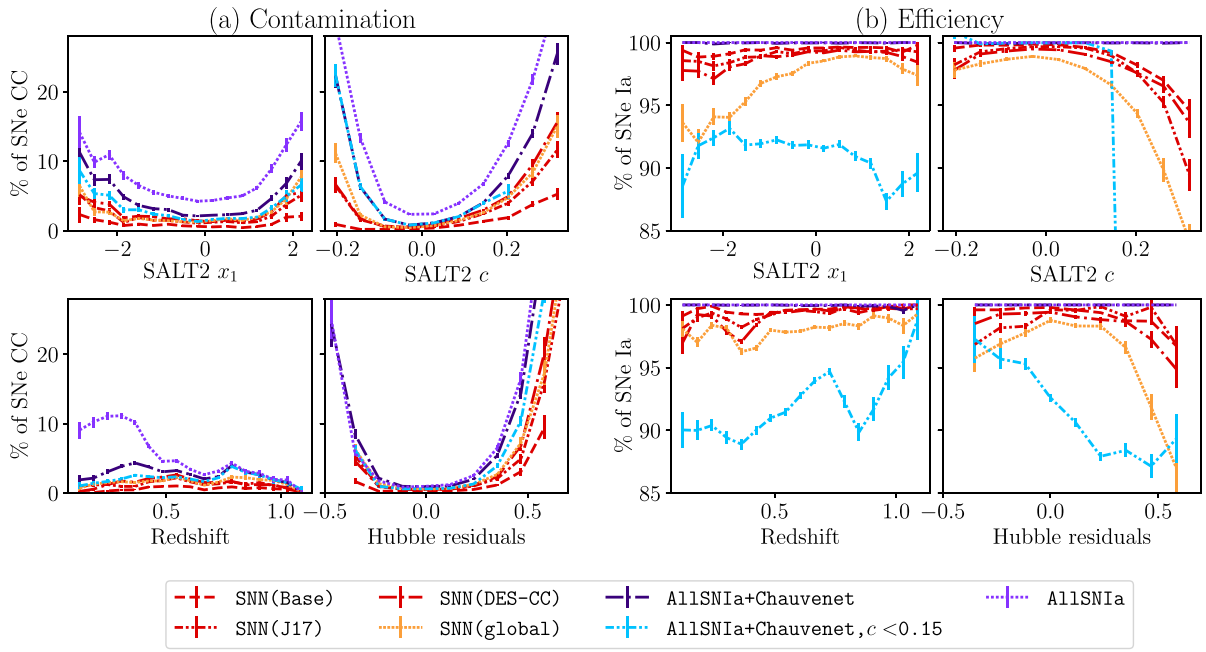

Figure 4. Contamination (panel (a)) and efficiency (panel (b)) using three SNN models `SNN(Base)`, `SNN(J17)`, and `SNN(DES-CC)` measured on our Baseline simulation. All contamination and efficiency percentages are measured relative to the bin, *not* relative to the total sample. In panel (a), we present contamination as a function of SALT2 x_1 (upper left), c (upper right), redshift (lower left), and Hubble residual (lower right). Panel (b) is the same as panel (a), but showing efficiency. Contamination and efficiency are defined in Section 4.1.

Table 6. Fraction of different sub-types of contaminants for different selection cuts. Contamination is measured on the Baseline simulation, after applying the SALT2-based selection described in Section 2.3 and Chauvenet's criterion. All contamination percentages are measured relative to the bin, *not* relative to the total sample.

Selection	Per cent non-Ia SNe	Per cent Pec Ia	Per cent II	Per cent Ibc
$c > 0.2$	16.6	5.3	0.6	10.7
$c < -0.2$	24.1	0.1	22.2	1.8
$x_1 > 2$	12.0	2.3	2.8	6.9
$x_1 < -2$	6.9	0.7	1.6	4.6
$\log_{10}M_\star/M_\odot < 10$	2.8	0.8	0.8	1.1
$\log_{10}M_\star/M_\odot > 10$	3.6	1.5	1.0	1.1

higher x_1 values is mainly due to slower declining stripped-envelope SNe, while contamination at the low x_1 is dominated by faster declining SNe Ic (see Table 6).

4.3.2 *Performance of global versus cosmo normalizations*

Contamination after using SNN models trained with the global SNN normalization (`SNN(global)`) is similar to the other SNN models trained using the cosmo normalization. However, `SNN(global)` has a significantly lower efficiency – less than 98.5 per cent – and it decreases significantly for positive Hubble residuals.

In the `SNN(global)` model, the relative brightness between SNe Ia and core-collapse SNe is preserved both in the training and testing phase. Our results show that encoding SN relative brightnesses in the classification does not result in a significant decrease in contamination, and mainly affects the classification of faint SNe Ia. Approximately 10–15 per cent of SNe Ia in the faint tail of the Hubble residual distribution ($\Delta\mu > 0.25$ mag) are misclassified as non-SNe Ia.

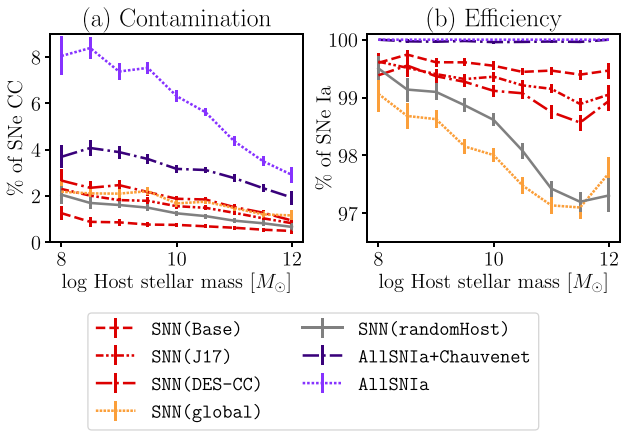

Figure 5. Contamination (left-hand panel) and efficiency (right-hand panel) as a function of SN host-galaxy stellar mass. We measure contamination and efficiency for different SNN models (see Section 4.2) and before/after applying the Chauvenet's criterion (see Section 3.5).

4.3.3 *Performance as a function of host galaxy properties*

Our simulations are designed to account for the differing properties and rates of SNe in different host galaxies. This allows us to predict contamination in our photometric SN Ia samples as a function of host galaxy properties. As a reference, the `SNN(randomHost)` does not use these intrinsic rates and assigns host galaxies randomly.

In Fig. 5, we present contamination and efficiency as a function of host galaxy stellar mass before applying any classification algorithm (i.e. applying only the `AllSNIa` classifier and Chauvenet's criterion) and after applying SNN. Contamination is not equally distributed across host galaxies of different mass, but is always larger in lower

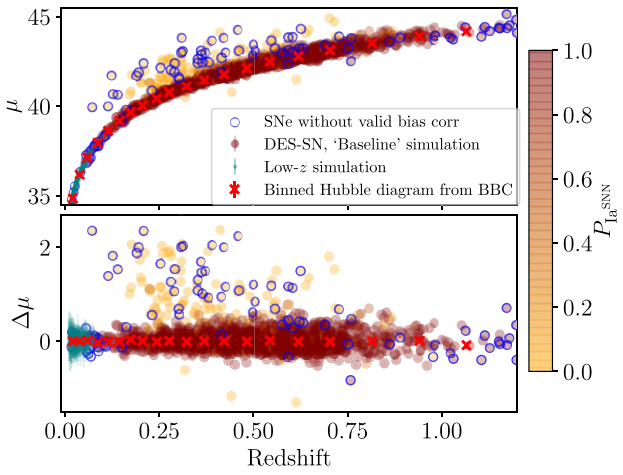

Figure 6. Same as Fig. 3, but here we highlight for each simulated event, its probability of being a type SN as estimated with `SNN(Base)`.

mass galaxies. This variation is expected as most of the hosts in the highest mass bin consists of more passive galaxies, with a preference towards SNe Ia and only small numbers of core-collapse SNe. Therefore, the fraction of contamination in these environments is low (less than 2 per cent) even with no photometric classification.

The efficiency of classification is mostly insensitive to host galaxy stellar mass, with two exceptions: efficiencies of the models `SNN(global)` and `SNN(randomHost)` drop significantly in higher mass galaxies. For the `SNN(global)` model using the 'global' normalization (Section 4.2), the training retains information about the relative brightnesses between SNe Ia and SN contaminants. This model is likely to heavily 'fit' on the information that core-collapse SNe are generally fainter than SNe Ia. This means that faint SNe Ia (i.e. SNe Ia with positive Hubble residuals, see Fig. 4b) in massive hosts (with lower signal-to-noise due to a brighter host galaxy background) are more easily misclassified as core-collapse SNe.

The `SNN(randomHost)` model is trained on a set of SN Ia light curves that have been assigned randomly to host galaxies. V21 demonstrated that the random association of host galaxies to simulated SNe produces a distribution of host brightnesses and masses in disagreement with the data (fig. 9 in V21). Therefore, host galaxies in the training sample of `SNN(randomHost)` are on average fainter than those in the DES-SN sample or simulations. When the `SNN(randomHost)` model is tested on realistic SN samples, a significant fraction of SNe Ia in bright and high mass galaxies is misclassified as core-collapse SNe. This test demonstrates the importance of training machine learning algorithms like SNN on simulations that include a realistic SN-host association. Sub-populations of SNe Ia (e.g. SNe in bright galaxies) can be reduced or removed by classification simply because they are not modelled in the training sample, with a potential impact on studies of SN Ia populations and on SN Ia cosmology in general.

Similarly to Fig. 3, we show the Hubble diagram for a simulated sample of SNe in Fig. 6 and we highlight SN probabilities, P_{Ia}, estimated applying `SNN(Base)`.

4.4 Effects of BBC bias corrections on contamination

In the BBC framework, there are cells of the three-dimensional subspace $\{z, x_1, c\}$ that have no SN Ia (or too few events). Real events in those cells are rejected prior to the BBC fit and this systematically disfavours SNe that lie in regions that are atypical for SNe Ia. As a result, the BBC bias corrections naturally reduce contamination from peculiar SNe Ia and core-collapse SNe. Tables 7 and 8 presents contamination and efficiency after BBC bias corrections are applied (cf. Tables 4 and 5, the contamination and efficiency before BBC). As expected, the number of SNe Ia is reduced by less than 1 per cent, while the number of core-collapse SNe is reduced by 20–30 per cent.

When analysing contamination after a $P_{Ia} > 0.5$ cut from SNN, the effect of bias corrections on the contamination is almost negligible because SNN is very efficient at removing contamination (see Table 8). However, when using no classifier (i.e. `AllSNIa`; Table 7) the bias corrections have a larger impact on reducing contamination. In Appendix B, we consider the sub-sample of events that are rejected from the sample only due to the lack of a valid bias correction, and investigate the impact of including these events in the analysis by fixing their bias correction to zero.

4.5 Comparison with the data

We apply bias corrections, Chauvenet's criterion and the SNN classifier to the DES photometric SN sample. In Fig. 7, we compare the results obtained from data and from simulations for different sets of selection cuts.[10]

First, we consider Hubble residuals measured after applying SALT2-based selection (Section 2.3), the Chauvenet's criterion (Section 3.5), and requiring a valid bias correction (Section 3.4). Simulations and data are in very good agreement (Fig. 7a); the asymmetry in the Hubble residual distribution due to the small fraction of core-collapse contamination (<3.8 per cent, see Table 7) is well reproduced by simulations and the reduced χ^2 between data and the Baseline simulation is approximately 1.1.

Secondly, we repeat the test above and additionally require $P_{Ia} > 0.5$, where P_{Ia} is estimated from the SNNclassifier trained on the Baseline simulation (`SNN(Base)`). The agreement between data and simulations is also good (reduced χ^2 of 0.7) and the tail of SNe with faint Hubble residuals is significantly reduced both in the data and in the simulations (Fig. 7b).

We note the presence of a few outliers (Hubble residuals larger than 1 mag) in the observed Hubble residuals distribution, that are not reproduced in the simulations. This could be due to a small fraction of SNe in the DES-SN sample (less than 1.1 per cent according to Wiseman et al. 2020) that is mismatched to a closer and brighter galaxy, and thus appear as faint outliers on the Hubble diagram. We remind the reader that host mismatch is not included in our simulations.

5 BIASES ON COSMOLOGICAL PARAMETERS

The BBC framework requires several modelling choices, each causing a potential bias on the binned SN Ia distance moduli, μ_{Ia}^b, and on the resulting fitted cosmological parameters. We explore these choices in this section. The BBC configurations we test are listed in Table 9 and illustrated in Fig. 1. Each is a different combination of classifier and $\mathcal{L}_{non\text{-}Ia}$. Specifically, we test:

(i) P_{Ia} measured from the five different SNN classifiers (Table 3), as well as the `Perfect` and `AllSNIa` approaches;

[10] A version of the same comparison *before* classification-based cuts is available in V21, fig. 13

Table 7. As Table 4, but following the application of BBC bias corrections.

Selection criteria	Contamination after BBC						Efficiency
	Only pec Ia	Baseline	LFs+Offset	Dust(H98)	J17	DES-CC	(Baseline)
`AllSNIa`	1.9	5.7	8.0	6.4	6.4	7.2	100.0
`AllSNIa+Chauvenet`	1.1	3.1	5.0	3.6	3.5	3.2	100.0
`AllSNIa+Chauvenet`, $c < 0.15$	0.8	2.2	3.8	2.6	1.8	2.6	91.5

Table 8. As Table 5, but following the application of BBC bias corrections.

SNN model[a]	Contamination after testing SNN on different simulations						Efficiency
	Only pec Ia	Baseline	LFs+Offset	Dust(H98)	J17	DES-CC	(Baseline)
`SNN(Base)`	0.4	**0.7** [b]	1.0	0.9	0.9	1.3	99.5
`SNN(J17)`	0.6	1.5	2.4	1.7	**1.0** [b]	2.0	99.2
`SNN(DES-CC)`	0.9	1.8	2.9	2.1	1.7	**1.5** [b]	99.0
`SNN(global)`	0.8	1.7	2.8	1.9	1.3	2.0	97.8
`SNN(randomHost)`	0.7	1.2	1.8	1.4	1.2	1.6	98.1

[a]See Table 3 for a description of the training approach utilized for each SNN model.
[b]We highlight in bold the contamination measured using the same simulation both for training and testing.

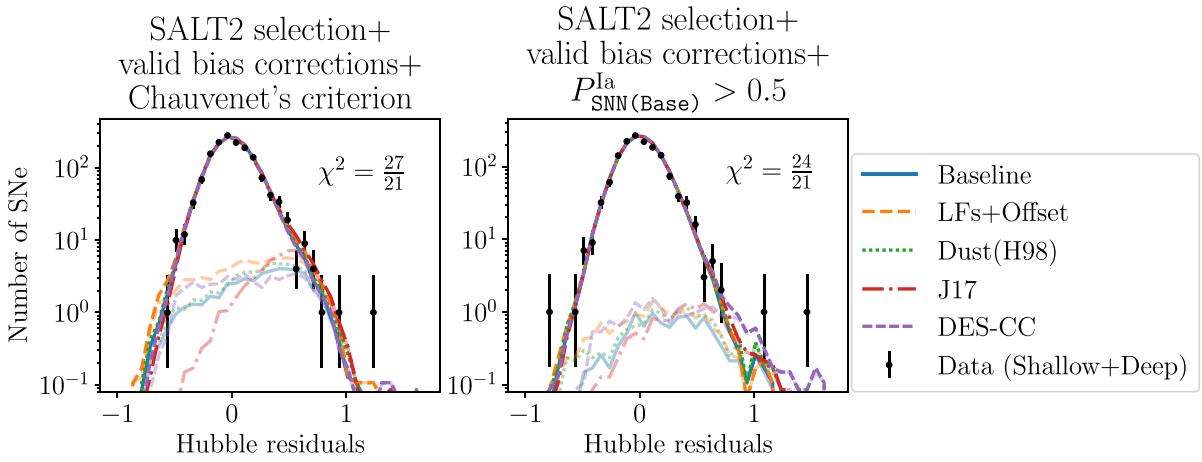

Figure 7. Distributions of observed and simulated Hubble residuals for different selection criteria. Distributions are presented for the data (including Poisson uncertainties, black symbols) and for the five simulations summarized in Table 1: both SNe Ia and core-collapse SNe are combined in the darker lines, and simulated core-collapse SNe only are shown in the partially transparent lines. *Left:* Sample selected applying SALT2-based selection discussed in Section 2.3, Chauvenet's criterion discussed in Section 3.5, and requiring a valid bias correction (Section 3.4). *Right:* Sample selected applying SALT2-based selection discussed in Section 2.3, a probability cut $P_{\rm Ia} > 0.5$ (where $P_{\rm Ia}$ are determined using the SNN classifier trained on the Baseline simulation, `SNN(Base)`) and requiring a valid bias correction. In each panel, we report the reduced χ^2 between data and simulations.

(ii) Two approaches for the modelling of $\mathcal{L}_{\rm non\text{-}Ia}$ (Section 3.3): the polynomial fitting method of H12 ($\mathtt{D_{non\text{-}Ia}(H12)}$), and the Kessler & Scolnic (2017) method implemented using the Baseline simulation ($\mathtt{D_{non\text{-}Ia}(Base)}$).

We test combining Chauvenet's criterion (Section 3.5) with the `AllSNIa` approach.

We consider as our reference the configuration that uses the classifier `SNN(Base)`, and for which the core collapse SN likelihood is modelled from the Baseline simulation. This has the label '$\mathtt{SNN(Base)\ D_{non\text{-}Ia}(Base)}$', and is used as the benchmark to evaluate other BBC configurations.

All our tests are run on the simulations presented in Section 2.2, reproducing the realistic scenario of testing classifiers on samples of light curves that are not in the samples used to train the classifier. This allows a verification that our modelling of $\mathcal{L}_{\rm CC}$ is sufficiently generalized to be applied to any population of core-collapse SN contaminants. Both are critical to robustly validate our results.

For each simulation, we estimate different cosmology-related parameters averaged over 50 realizations: $\mu^b_{\rm Ia}$, nuisance parameters (α, β, $\sigma_{\rm Ia,int}$, $S_{\rm non\text{-}Ia}$), w, and the time-varying dark energy equation-of-state parameters w_0 and w_a. We then calculate biases due to contamination as

$$\Delta X = \langle X_{\rm Ia+CC} - X_{\rm Ia\ only,\ perfect\ classification}\rangle_{\rm (50\ realizations)}, \quad (9)$$

where X represents either $\mu^b_{\rm Ia}$ or the nuisance parameters or cosmological parameters w, w_0, w_a depending on the context. Essentially, we define a bias ΔX on a cosmological parameter X due to contamination as the average difference between the value of the parameter fitted including contamination, and the value of the parameter fitted with no contamination and assuming a perfect classification. Uncertainties on ΔX are estimated as standard errors on the mean.

5.1 Biases for a flat *w*CDM model

We first consider fits in a wCDM model. Our key results are in Fig. 8, showing Δw estimated using different BBC options and simulations. The cosmological results presented from the data are preliminary and are blinded (i.e. the best-fitting cosmology is not known) and are therefore also shown as shifts Δw with respect to the (arbitrary) BBC

Table 9. Summary of BBC configurations (see also Fig. 1). The second line (highlighted) lists the reference configuration.

	BBC configuration[(a)]	Classifier	Modelling of $D_{\text{non-Ia}}$	Δw using Baseline simulation[(b)]	Δw using DES-SN Data[(c)]
1)	`Perfect` $D_{\text{non-Ia}}$`(Base)`	Perfect	Baseline	0.0001 ± 0.0002	–
2)*	`SNN(Base)` $D_{\text{non-Ia}}$`(Base)`	SNN(Base)	Baseline	0.0045 ± 0.0008	0.0000 (0.0338)
3)	`SNN(J17)` $D_{\text{non-Ia}}$`(Base)`	SNN(J17)	Baseline	0.0109 ± 0.0009	0.0059 (0.0342)
4)	`SNN(DES-CC)` $D_{\text{non-Ia}}$`(Base)`	SNN(DES-CC)	Baseline	0.0045 ± 0.0008	0.0101 (0.0324)
5)	`SNN(Base)` $D_{\text{non-Ia}}$`(H12)`	SNN(Base)	Fit (H12)	0.0048 ± 0.0008	-0.0015 (0.0338)
6)	`SNN(J17)` $D_{\text{non-Ia}}$`(H12)`	SNN(J17)	Fit (H12)	0.0135 ± 0.0012	0.0025 (0.0331)
7)	`SNN(DES-CC)` $D_{\text{non-Ia}}$`(H12)`	SNN(DES-CC)	Fit (H12)	0.0048 ± 0.0008	0.0070 (0.0329)
8)	`SNN(global)` $D_{\text{non-Ia}}$`(Base)`	SNN(global)	Baseline	0.0128 ± 0.0010	0.0253(0.0319)
9)	`SNN(randHost)` $D_{\text{non-Ia}}$`(Base)`	SNN(randHost)	Baseline	0.0043 ± 0.0007	0.0095 (0.0328)
10)	`AllSNIa`	$P_{\text{Ia}} = 1\ \forall$ SN	‡	-0.0252 ± 0.0046	0.0407 (0.0517)
11)	`AllSNIa+Chauvenet`	$P_{\text{Ia}} = 1\ \forall$ SN	‡	-0.0152 ± 0.0014	-0.0018 (0.0346)
12)	`AllSNIa+Chauvenet, c<0.15`	$P_{\text{Ia}} = 1\ \forall$ SN	‡	-0.0139 ± 0.0020	-0.0005 (0.0345)

(a) The numbers of selected SNe are in Table 2. The SALT2 selection and the requirement of a valid bias correction is always applied. Any additional selection criteria are indicated in the name of the BBC configuration.
(b) Calculated using equation (9).
(c) Biases measured from the DES-SN sample. Shifts are with respect to the value estimated using our BBC reference `SNN(Base)` $D_{\text{non-Ia}}$`(Base)`. Errors reported in parenthesis are the *statistical* uncertainties on w only.
‡Assuming all SNe have $P_{\text{Ia}} = 1$ means that the core collapse SN term in the BEAMS likelihood is always zero (equation 3).
*Reference BBC configuration. For this BBC configuration, we obtain Δw of 0.0045 ± 0.0008 for Baseline simulation, 0.0082 ± 0.0008 for LFs+Offset simulation, 0.0046 ± 0.0009 for Dust(H98) simulation, 0.0019 ± 0.0007 for J17, and 0.0076 ± 0.0009 for DES-CC.

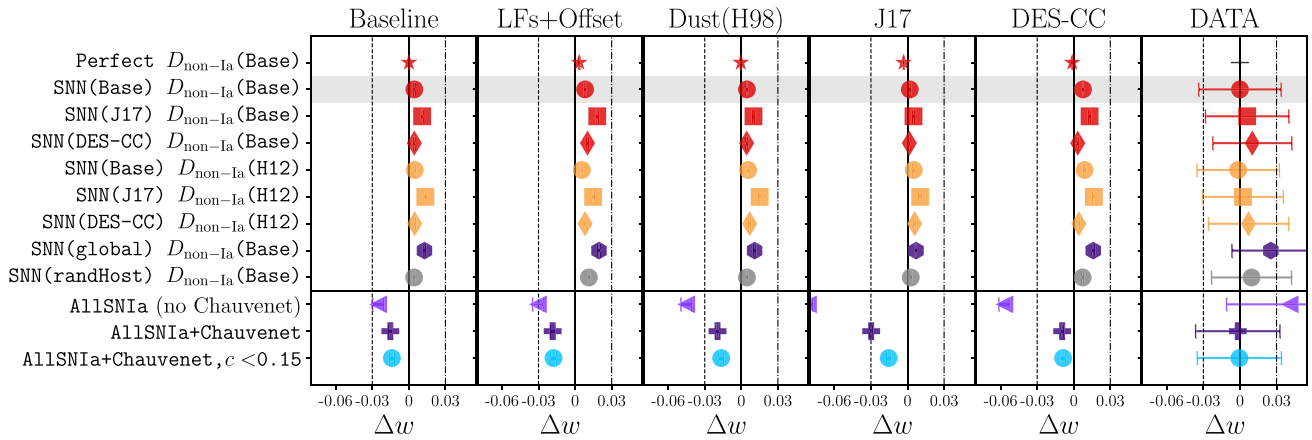

Figure 8. Biases on the recovered dark energy equation-of-state parameter, Δw, measured for each simulation (Table 1) and for different BBC configurations (Table 9). Different SNN models correspond to different symbols (circles for `SNN(Base)`, squares for `SNN(J17)`, diamonds for `SNN(DES-CC)`), and different $D_{\text{non-Ia}}$ modelling approaches correspond to different colours (red for $D_{\text{non-Ia}}$`(Base)` and orange for $D_{\text{non-Ia}}$`(H12)`). For simulations, we estimate Δw and relative uncertainties as described in equation (9). For the data (last column), we present Δw with respect to our reference BBC configuration (`SNN(Base)` D_{CC}`(Base)`, second from the top, highlighted). Data error bars are 1σ statistical uncertainties only, and are not independent for each BBC configuration.

reference configuration (`SNN(Base)` D_{CC}`(Base)`). Uncertainties on the data are the 1σ statistical uncertainties, while for simulations we average the results of 50 realizations.

5.1.1 Cosmological biases using the SNN classifier

Testing the different simulations presented in Section 2.2 with SNN, we find that the biases on w are <1 per cent (from a minimum of $\Delta w = 0.002$ for Baseline simulation to a maximum $\Delta w = 0.008$ for J17 simulation) for our BBC reference configuration, and <2 per cent for the other configurations in Table 9 (a maximum $\Delta w = 0.015$ is estimated for LFs+Offset simulation analyzed with `SNN(J17)` model). Across all the BBC configurations and simulations tested, the biases on the fitted nuisance parameters α and β are <1.5 and <1.8 per cent, respectively (see Fig. 12). Biases on SN Ia intrinsic scatter $\sigma_{\text{Ia,int}}$ are also consistent with zero and the recovered scaling parameter $S_{\text{non-Ia}}$ is consistent with one.

In Fig. 9, we present the full $\Omega_{\text{M}}-w$ cosmological contours[11] from a single realization of the DES-like sample (i.e. the same statistical constraining power as expected from the DES-SN photometric sample). We compare cosmological contours for the ideal scenario of a perfectly classified sample of SNe Ia and for the realistic scenario of a contaminated sample of SNe Ia analysed using the SNN classifier. The biases on cosmological constraints due to contamination are significantly smaller than the statistical uncertainties.

[11]As described in Section 3.6, we estimate contours using the cosmological fitter COSMOMC.

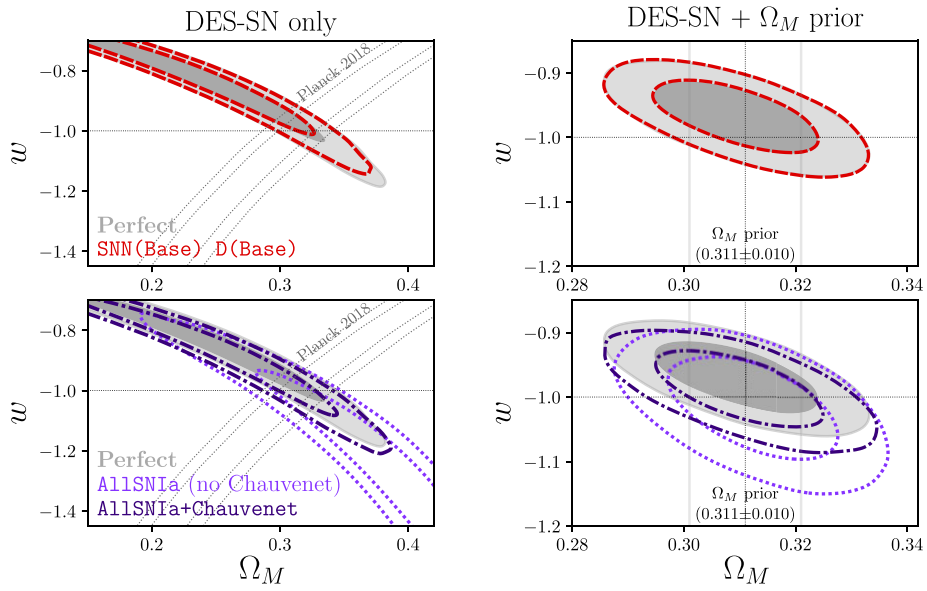

Figure 9. Cosmological contours estimated from a DES-SN simulated sample. The results show one realization of the Baseline simulation. We show the cosmological constraints (68 and 95 per cent confidence intervals) on $\Omega_M - w$ for a flat wCDM model, with and without an Ω_M prior of 0.311 ± 0.010 (right-hand and left-hand panels, respectively). We also present results obtained using a perfectly classified sample of SNe Ia (grey filled contours), a contaminated sample of SNe analyzed assuming all SNe passing SALT2 selection are SNe Ia (dotted purple contours, lower panel), assuming all SNe passing SALT2 selection and Chauvenet's criterion are SNe Ia (dot–dashed purple contours, lower panel) and using our reference BBC configuration `SNN(Base)` $\mathtt{D}_{\mathtt{non\text{-}Ia}}$`(Base)` (dashed red contours, top panel). The Ω_M prior of 0.311 ± 0.010 is in grey.

Fig. 10 shows the biases on the binned Hubble diagram ($\Delta\mu$) using different SNN models. Generally, the $|\Delta\mu|$ are less than 10 mmag across all tests and simulations (consistent with the small biases measured on w). We observe consistently across all simulations that SN Ia distances estimated from BBC are mostly unbiased ($\Delta\mu$ <4 mmag) at lower redshifts ($z < 0.5$), and the largest biases are observed at $z \sim 0.7$, towards negative values (i.e. brighter values). At these redshifts, the number of true SNe Ia decreases and thus the modelling of the core collapse SN population is both more critical and more uncertain. This makes the marginalization of core collapse SN contamination from BBC less accurate. The choice of the modelling approach adopted for the contamination likelihood can have a significant impact on $\mu^b_{\rm Ia}$. For the same SNN model, $\mu^b_{\rm Ia}$ can differ by >5 mmag when varying the modelling of the contamination likelihood. This is particularly evident in the simulation where contaminants are artificially brightened (LFs+Offset). This suggests that the choice of training sample for SNN is not the only driver of systematics.

Finally, we note that for all our tests with SNN we find that the binned Hubble diagram $\mu^b_{\rm Ia}$ is mainly biased towards negative values, and this in turn corresponds to positive biases on w. This suggests that combining SNN with the BEAMS formalism tends to slightly 'over-correct' for contamination and, therefore, preferentially biases the Hubble diagram towards brighter values. In the next section, we discuss cosmological biases when applying Chauvenet's criterion and no classification and we observe the opposite trend.

5.1.2 Cosmological biases using Chauvenet's criterion without a classifier

We next test the case of not using a classifier and assuming all SNe in the samples that pass the SALT2 selection are SNe Ia (`AllSNIa`), setting $P_{\rm Ia} = 1$ for every SN and the contamination term in the BEAMS likelihood to zero. We also test outlier rejection in combination with the `AllSNIa` approach, with the results in Fig. 11.

With no outlier rejection, the binned $\mu^b_{\rm Ia}$ are biased towards fainter values pulled by faint core collapse SN contaminants, especially at $z < 0.5$. At higher-z the biases are smaller (<10 mmag) as contamination is naturally reduced by Malmquist bias, and can either be brighter (e.g. for LFs+Offset) or fainter (e.g. J17) depending on the properties of the simulated core collapse SNe. As expected, this approach results in significant biases with $\Delta w = -0.025 \pm 0.009$ for Baseline up to $\Delta w = -0.082 \pm 0.008$ for J17 (see also Fig. 9). The biases from this no-classifier approach have the opposite sign compared to the biases found when combining SNN and the BEAMS approach. In the no-classifier approach, the fainter population of contamination is 'under-corrected' (or effectively not corrected at all as core collapse SNe are assume to have $P_{\rm Ia} = 1$) therefore the biases on $\mu^b_{\rm Ia}$ are mainly positive and w-bias is negative.

When we combine Chauvenet's criterion with `AllSNIa`, the biases in $\mu^b_{\rm Ia}$ are reduced, generally to <10 mmag, and are broadly consistent with the SNN results (Fig. 11). The w-biases range from -0.010 ± 0.002 for Baseline to -0.019 ± 0.001 for Dust(H98) (Fig. 8). However, in the J17 simulations, while the fraction of contaminants (mostly red type Ib SNe) is similar to the other simulations (Table 7), their distribution on the Hubble diagram is such that, even after applying Chauvenet's criterion, a significant trend in $\mu^b_{\rm Ia}$ is introduced biasing w by -0.030 ± 0.004. This is reduced by 50 per cent with a stricter SALT2 c selection (to -0.015 ± 0.02), suggesting that the bulk population of red and bright contaminants is the main driver of this cosmological bias. For the other simulations, applying stricter SALT2 c cuts does not reduce biases on w significantly, while it reduces the number of SNe Ia by 8 per cent.

Fig. 12 shows that the fitted nuisance parameters are also biased when using Chauvenet's criterion only. When applying Chauvenet's criterion, the residual population of red and faint core-collapse contaminants lead to an overestimate of the fitted values of β by

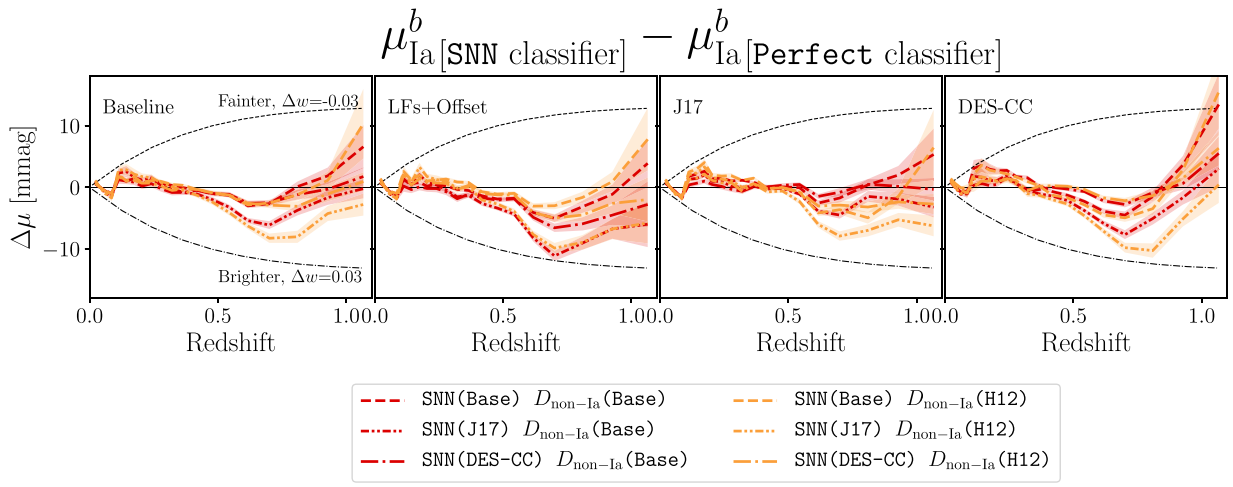

Figure 10. Differences in the binned distance modulus $\mu^b_{\rm Ia}$ when using SNN compared to using the `Perfect` classifier. Each panel presents the results when applied to a different simulation: Baseline (left), LFs+Offset (centre left), J17 (centre right), and DES-CC (right). We compare different SNN classifiers and different BBC configurations: each SNN model corresponds to a different line-style, and each $D_{\rm non\text{-}Ia}$ modelling approach corresponds to a different colour (see legend). Differences in distance modulus between $\Delta w = -0.03$ and $\Delta w = 0.03$ are presented as dashed and dot–dashed lines, respectively.

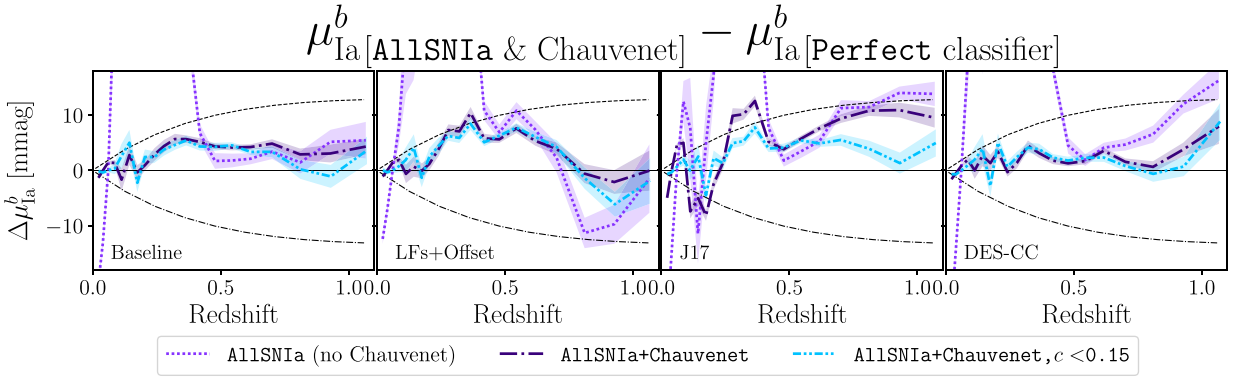

Figure 11. As Fig. 10, but comparing the `AllSNIa` approach with perfect classification. We combine the `AllSNIa` approach with different SN selection criteria: SALT2 selection only (`AllSNIa`), SALT2 selection, and Chauvenet's criterion (`AllSNIa+Chauvenet`), and finally including stricter SALT2 c cuts (`AllSNIa+Chauvenet,c < 0.15`).

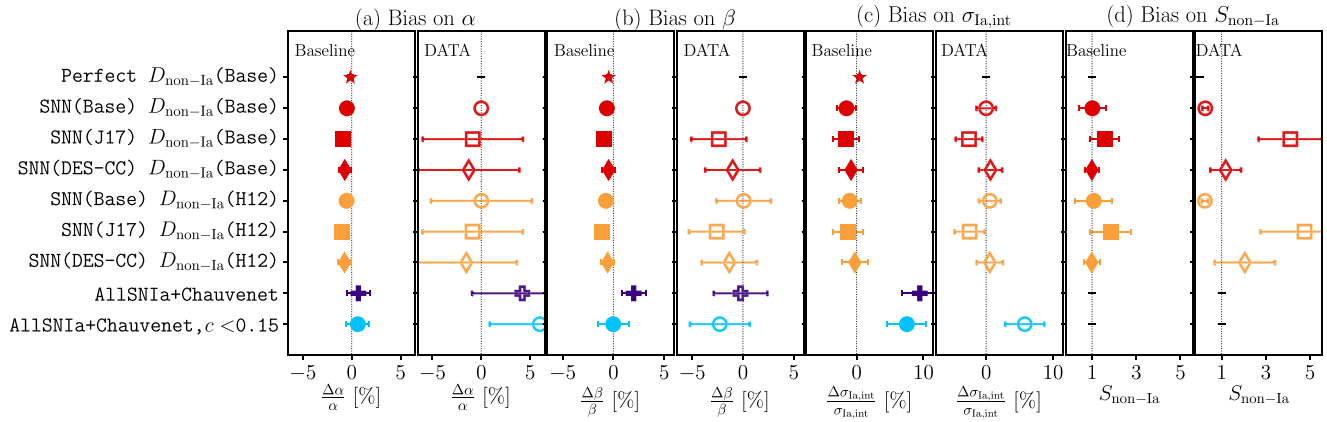

Figure 12. Relative differences in the fitted nuisance parameters α (panel a), β (panel b), and $\sigma_{\rm Ia,int}$ (panel c), and the deviation from one of the scaling parameter $S_{\rm non\text{-}Ia}$ (panel d). Each pair of panels presents the results for Baseline (left) and DES-SN5YR data (right). We compare different BBC configurations (see Table 9). The BBC fitting procedure does not return uncertainty on $\sigma_{\rm Ia,int}$. Therefore, for both data and simulations the uncertainties on $\sigma_{\rm Ia,int}$ are estimated as the r.m.s. spread in $\sigma_{\rm Ia,int}$ measured from the 50 realizations of the Baseline simulation.

approximately 3 per cent. These biases are reduced to <1 per cent when applying stricter SALT2 c cuts. Biases on α are <1 per cent. The SN Ia intrinsic scatter is also overestimated by 7–10 per cent.

The cosmological constraints presented in Fig. 9 highlight the power of outlier rejection methods like Chauvenet's criterion. For a DES-like simulated sample, when we assume all SNe passing SALT2 selection and Chauvenet's criterion are SNe Ia (`AllSNIa+Chauvenet`), the biases on the cosmological contours are small. These findings and the results presented in Figs 8 and 11 suggest that cosmological biases due to contamination can be small even without applying photometric classification algorithms and using only outlier rejection methods.

5.1.3 *The role of priors*

Besides SNN and Chauvenet's criterion, the Ω_M prior discussed in Section 3.6 is another element that indirectly contributes to reduce biases on w due to contamination. In SN cosmology, SNe Ia measurements and CMB measurements are typically combined in order to break the respective degeneracies on the Ω_M and w parameter space, and thus reduce the overall statistical uncertainty on w. As shown in Fig. 9 (left-hand panel), core collapse contamination shifts the SN-only cosmological contours *along* the 'banana-shaped' SN contours and *perpendicularly* to the CMB constraints and to a Gaussian Ω_M prior. Therefore, combining SNe with CMB measurements (left-hand panels in Fig. 9) or applying an Ω_M prior (right-hand panels in Fig. 9) not only reduces statistical uncertainties on w, but also significantly mitigates systematic biases on w due to contamination.

We highlight that, for w estimates, CMB constraints are more stringent (i.e. almost perfectly orthogonal to SN-only constraints) than a Gaussian Ω_M prior. For this reason, we anticipate that updating our prior with the latest CMB measurements from Planck Collaboration (2020) will further reduce statistical uncertainties on w and systematic biases on w due to contamination.

5.1.4 *Biases when applied to data*

We perform the same tests on the DES-SN data as applied to the simulations. Clearly, the true classification of each SN and the unbiased μ^b_{Ia} is not known, so we estimate relative biases between different BBC configurations.

Table 9 (last column on the right) and Fig. 8 (last column on the right) present Δw shifts measured from the data and estimated with respect to the value of w fitted from our reference BBC configuration. Using Chauvenet's criterion and assuming all events are SNe Ia, we obtain $\Delta w = -0.0018$ (r.m.s. on Δw estimated from 50 realizations of the Baseline simulation is 0.0076). This result suggests that our reference BBC configuration and the Chuavenet's criterion approach are consistent within the uncertainties. When comparing our reference BBC configurations with the BBC configurations that use SNN models `SNN(J17)` and `SNN(DES-CC)` (i.e. BBC configurations 3 and 4 in Table 9), we observe shifts on w of 0.0059 (r.m.s. from simulations is 0.0036) and 0.0101 (r.m.s. from simulations is 0.0036). The BBC configuration that implements the `SNN(global)` classifier results in the largest Δw, but given the caveats discussed in Section 4.3.2) we do not consider `SNN(global)` a robust classification method. The statistical uncertainty on w for our reference BBC configuration is 0.034, which is approximately three times the maximum $\Delta\mu$ observed in the data. These results confirm that for the cosmological analysis of the DES photometric SN sample, contamination is a subdominant systematic when compared to the statistical uncertainty.

In Fig. 12, we compare fitted nuisance parameters when using the reference BBC configuration and other BBC configurations. The parameters α and β fitted from the data are consistent between the different configurations tested. Large discrepancies are seen in the fitted values of the scaling factor $S_{\text{non-Ia}}$. $S_{\text{non-Ia}}$ for the data is 0.26 ± 0.13, 4.11 ± 1.44, and 1.18 ± 0.70 when using `SNN(Base)`, `SNN(J17)`, and `SNN(DES-CC)`, respectively, and the non-Ia likelihood approach $D_{\text{non-Ia}}$`(Base)`. Predicting $D_{\text{non-Ia}}$ and constraining the factor $S_{\text{non-Ia}}$ is difficult when the percentage of contaminants in the sample is already very low and this explains these large differences in the fitted values.

For comparison and a sanity check, we also test the performances of the SNN classifier `SNN(Base)` and Chauvenet's criterion on the DES-SN sample of spectroscopically confirmed SNe. After applying all the selection criteria discussed in Section 2.3, we have 401 spectroscopically classified SNe observed by DES. We find that 354 events are certain SNe Ia, 44 likely SNe Ia, and three are classified as non-Ia (two stripped envelope SNe and one hydrogen-rich SN). Only one out of the three non-Ia SNe satisfy Chauvenet's criterion. All three events have $P_{Ia} < 0.2$. The spectroscopic sample is significantly biased towards bright, high signal to noise ratio events, therefore it is not surprising that the contamination is extremely low (less than 1 per cent after SALT2-based cuts only and zero after probability cuts). However, it shows how efficiently a SALT2-based selection and Chauvenet's criterion can reduce contamination, as generally applied in the cosmological analysis of spectroscopic samples of SNe Ia (Foley et al. 2017; Scolnic et al. 2018; Brout et al. 2019b).

5.2 Systematic uncertainties associated with contamination

In this section, we estimate the contribution of contamination to the w systematic error budget from a DES-like cosmological analysis. In order to do this, we follow the approach presented by Conley et al. (2011) and Brout et al. (2019b, section 3.8.2) and define a systematic covariance matrix, C_{syst}, that can be included in the fit for cosmological parameters. The χ^2-minimization cosmological fitter introduced in Section 3.6 does not currently handle a systematic covariance matrix; for this reason, we use COSMOMC when estimating systematic uncertainties on w.

Given $\partial\mu^b_{Ia,s_k}$ the differences in the binned Hubble diagram after changing the systematic parameter s_k, the systematic covariance matrix, C^{ij}_{syst}, is defined as

$$C^{ij}_{syst} = \sum_{k=1}^{N_{syst}} \left(\frac{\partial\mu^i_{Ia,s_k}}{\partial s_k}\right)\left(\frac{\partial\mu^j_{Ia,s_k}}{\partial s_k}\right)\sigma^2_{s_k}, \quad (10)$$

where σ_{s_k} is the uncertainty of the systematic s_k and the indexes i and j are iterated over the N_{bins} redshift bins ($i, j = 1,..., N_{bins}$).

We build two different covariance matrices: one that includes variations over the three SNN models (`SNN(Base)`, `SNN(J17)`, and `SNN(DES-CC)`) but fixes the contamination likelihood to $D_{\text{non-Ia}}$`(Base)` (configurations 2, 3, and 4 in Table 9), and one that includes variations over the three SNN models (`SNN(Base)`, `SNN(J17)`, and `SNN(DES-CC)`) but fixes the contamination likelihood to $D_{\text{non-Ia}}$`(H12)` (configurations 5, 6, and 7 in Table 9). For each systematic, we estimate the contribution to the total error budget on w by applying the definition presented by Brout et al. (2019b, equation 22)

$$\sigma'_w = \sqrt{(\sigma^2_{stat+syst} - \sigma^2_{stat})}, \quad (11)$$

where $\sigma_{stat+syst}$ is the uncertainty estimated when considering only one (or a sub-group of) systematics and σ_{stat} is the statistical

Table 10. Uncertainty contributions to w for a wCDM model (SNe are combined with a Ω_M prior of 0.311 ± 0.010). See Table 9 for a detailed description of the BBC configurations listed in the first column.

	σ'_w	$\sigma'_w/\sigma_{\rm stat}$	$\sigma_{\rm stat+syst}$
Total $\sigma_{\rm stat}$	–	–	0.039
2) SNN(Base) $D_{\rm non\text{-}Ia}$(Base)			
3) SNN(J17) $D_{\rm non\text{-}Ia}$(Base)	0.004	0.106	0.040
4) SNN(DES-CC) $D_{\rm non\text{-}Ia}$(Base)			
5) SNN(Base) $D_{\rm non\text{-}Ia}$(H12)			
6) SNN(J17) $D_{\rm non\text{-}Ia}$(H12)	0.007	0.171	0.040
7) SNN(DES-CC) $D_{\rm non\text{-}Ia}$(H12)			

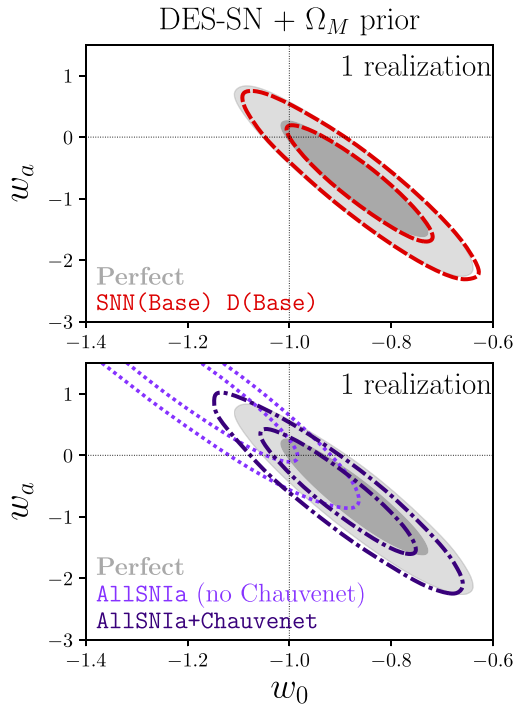

Figure 13. Same as Fig. 9 but in the w_0–w_a plane and assuming a flat w_0w_aCDM model with an $\Omega_{\rm M}$ prior of 0.311 ± 0.010.

uncertainty. The results are estimated for our Baseline simulation and presented in Table 10 (and obtain similar results when performing the same test on the other simulations). Systematic uncertainties associated with contamination are 0.004 for the $D_{\rm non\text{-}Ia}$(Base) method and 0.007 for the polynomial fitting method by H12. In general, systematics associated with contamination are at most a third of the statistical error, which corresponds to an increase of the overall w error budget by less than 5 per cent.

In Appendix A, we highlight some potential limitations related to the $D_{\rm non\text{-}Ia}$(H12) approach and to the choice of modelling the core-collapse likelihood term as a second order polynomial. Therefore, we consider the $D_{\rm non\text{-}Ia}$(Base) method as the most reliable one in our analysis and quote $\sigma'_w = 0.004$ to be our best estimate of systematic uncertainties associated with contamination.

5.3 Biases for a time-varying w_0/w_a model

We analyse the effects of contamination when fitting our simulated SN samples assuming a flat w_0w_a CDM model. In Fig 13, we present the $w_0 - w_a$ cosmological contours obtained from one realization of the Baseline simulation and assuming a Gaussian $\Omega_{\rm M}$ prior of 0.311 ± 0.010.

In Fig. 14, we present the average biases on w_0 and w_a measured for the Baseline simulation. For different BBC configurations (Table 9) and SNN, we find a −0.011–0.001 bias on w_0 and 0.008–0.166 bias on w_a. Using Chauvenet's criterion and AllSNIa, we find biases of −0.031 and 0.097 on w_0 and w_a, respectively. If we assume our reference BBC configuration is the most robust one, we measure biases across the different core collapse SN simulations of $-0.009 < \Delta w_0 < 0.000$ and $0.047 < \Delta w_a < 0.108$. This is shown in Fig. 14.

By comparison, the average statistical uncertainties on w_0 and w_a expected for a DES-like sample are 0.097 and 0.620, i.e. 5–10 times larger than the biases Δw_0 and Δw_a due to contamination.

Looking further to the future, these results can inform the planning of future time-domain experiments such as the optical Legacy Survey of Space and Time (LSST; Ivezić et al. 2019) that will be conducted using the Vera Rubin Observatory. Although the exact observational strategy is being developed, LSST is expected to discover more than 1000 new SNe Ia per night. Spectroscopic follow-up programmes such as the Time-Domain Extragalactic Survey (TiDES; Swann et al. 2019) and others, will provide host galaxy spectroscopic redshifts as well as spectroscopic classifications for a subset of these events. The photometric SN Ia sample is expected to include at least 25 times more cosmologically useful SNe Ia than the DES-SN photometric SN Ia sample, with similar redshift distributions (Frohmaier et al., in preparation). In parallel, low redshift SN samples are also expected to increase (approximately × 10 more SNe Ia than available in current low-z samples; see DESC Science Requirements Document; The LSST Dark Energy Science Collaboration 2018).

Following these forecasts, we estimate the statistical uncertainties on w_0 and w_a expected when combining 25× the DES-SN5YR photometric SN sample, 10× the current low-z samples, and an $\Omega_{\rm M}$ prior of 0.311 ± 0.010. These are found to be 0.03 and 0.19 for w_0 and w_a, respectively, i.e. approximately 3 and 2 times larger than the biases Δw_0 and Δw_a found when applying our reference BBC configuration on the full range of simulations. The contours are presented in Fig. 14. We conclude that contamination is not expected to degrade the figure of merit of the LSST SN Ia sample significantly, especially when implementing classification techniques like SNN.

6 CONCLUSIONS

In this paper, we have exploited state-of-the-art simulations of SN candidates detected by the Dark Energy Survey (DES) to quantify systematic effects in cosmological analyses introduced by the use of photometric SN classification methods. We focused on the testing of SuperNNova (SNN), a SN photometric classification tool based on machine learning techniques. In order to provide a robust assessment of the algorithm's performance and avoiding potential over-fitting, we have trained and tested SNN not only on our 'Baseline' simulation of DES (Table 1), but on a wider suite of DES simulations designed to explore different astrophysical assumptions in the core collapse SN population and different compilations of core collapse SN templates. We then perform a state-of-the-art analysis using SALT2 light curve fitting, BEAMS and its extension BBC to estimate bias corrections and correct for contamination, and cosmology fitting. In this way, we can propagate the effects of contamination to cosmological parameter estimation. Our main findings are:

(i) Across our DES simulations, contamination ranges from 0.8–3.2 per cent when using SNN, with the efficiency of the classification ranging from 99.0–99.5 per cent (Table 4). Therefore, on a sample of

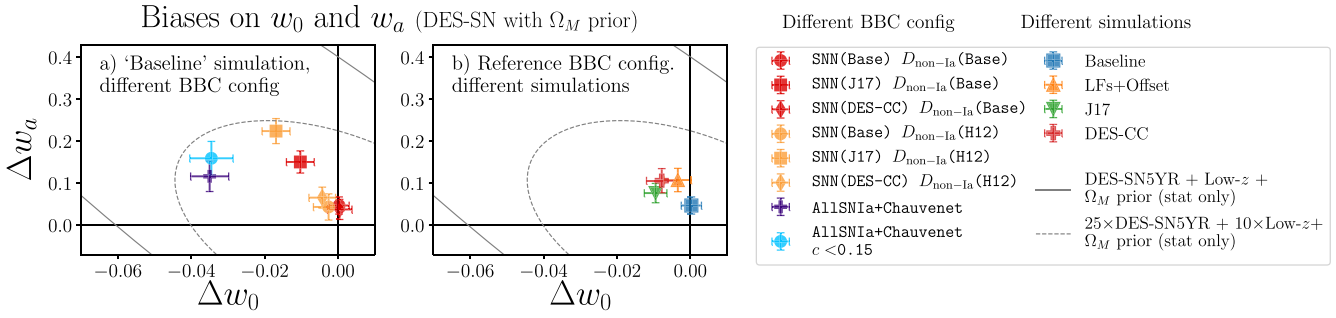

Figure 14. Biases on the dark energy equation-of-state parameters w_0 and w_a measured using (a) the Baseline simulation and varying the BBC configuration, and (b) the reference BBC configuration (Table 9) but varying the core collapse SN simulation (panel b). In panel (a), different SNN models correspond to different symbols (circles for `SNN(Base)`, squares for `SNN(J17)`, diamonds for `SNN(DES-CC)`), and different $D_{\rm non\text{-}Ia}$ modelling approaches correspond to different colours (red for `Dnon-Ia(Base)` and orange for `Dnon-Ia(H12)`). In panel (b), we present results for four different core collapse simulations (see Table 1). The biases Δw_0 and Δw_a and the relative uncertainties are measured as described in equation (9). As a comparison, we show the zero-centred 68 per cent w_0-w_a contours from DES-SN sample combined with an $\Omega_{\rm M}$ prior of 0.311 ± 0.010 and from a sample of 25 times the size of DES-SN sample.

approximately 1680 SNe (Table 2), we expect SNN to misclassify as SN Ia approximately 14–55 core collapse SNe and to exclude from the cosmological fit 9–17 true SNe Ia.

(ii) SNN trained on our Baseline simulation performs well across all simulated data samples, including those based on independent libraries of core collapse SN templates, with a contamination of $\leq$1.4 per cent. SNN classifiers trained on simulations using templates from J17 or DES-CC perform well when tested on simulations built using the same set of templates (<1 per cent contamination), but when tested on simulations built using independent core-collapse SN templates, contamination increases to 1.7–3.2 per cent.

(iii) Outlier rejection methods like Chauvenet's criterion can also significantly reduce contamination (to <3.1 per cent in the Baseline simulation, and <5.3 per cent for the other simulations, see Table 5). This can be further reduced with a tighter selection based on the SN Ia colour (<4.0 per cent).

(iv) We combine the BBC formalism with SNN trained on the Baseline simulation, and set this as our reference approach. Assuming a flat wCDM model, we find that biases on w are below 1 per cent ($|\Delta w| < 0.0082$), and the recovered nuisance parameters (α, β, $\sigma_{\rm Ia,\,int}$) are unbiased. When exploring additional BBC configurations and SNN training methods, we find that biases on w are at most 0.018. These biases are respectively 4 and 2 times smaller than the expected statistical uncertainty on w from DES-SN. The predicted systematic uncertainties related to contamination are <0.007 and this suggests that contamination increases by less than 5 per cent the total uncertainty on w and it is not a limiting systematic for the cosmological analysis of the DES-SN sample.

(v) When we implement Chauvenet's criterion and assume that all SNe that are not identified as outliers in the Hubble diagram are type Ia, this simplistic approach provides relatively small biases on w ($|\Delta w| < 0.018$ and $|\Delta w| < 0.033$ with and without stricter SALT2c-based selection). These results show that cosmological biases from contamination are small even without applying photometric classification algorithms. This suggests an alternative and promising path to carry out cosmological analysis of photometric samples, that avoid over-reliance on machine learning techniques. We recommend for future analyses to perform close comparisons between cosmological results obtained using outlier rejection techniques and machine learning classifiers.

(vi) Core-collapse contamination shifts the SN-only cosmological contours perpendicularly to CMB constraints (see Fig. 9). Therefore, combining SNe with CMB measurements (and not only with a Gaussian $\Omega_{\rm M}$ prior) will not only reduce the statistical uncertainty on w, but also further mitigate systematic biases on w due to contamination. In future cosmological analyses of the DES photometric SN sample, SN constraints will be combined with CMB constraints from Planck Collaboration (2020), therefore we anticipate our estimates of w-biases due to contamination and $\sigma_{\rm stat}$ on w to decrease compared to using the Gaussian $\Omega_{\rm M}$ prior in this paper. From a preliminary analysis, we forecast the contribution of contamination to the statistical error budget on w (i.e. $\sigma'_w/\sigma_{\rm stat}$, see Table 10) to change by less than 20 per cent.

(vii) We estimate biases due to contamination on w_0 and w_a. Combing the DES-SN sample with a Gaussian $\Omega_{\rm M}$ prior of 0.311 ± 0.010, we show the biases on w_0 to be less than 0.009, and the bias on w_a to be less than 0.108. These are 5–10 times smaller than the statistical uncertainties on w_0 and w_a expected from the DES-SN sample. When using outlier rejection techniques (e.g. Chauvenet's criterion), we find biases on w_0 of approximately 0.03 (approximately three times larger than biases found when implementing SNN) and biases on w_a of approximately 0.1 (comparable to biases found when implementing SNN).

In general, the results in this paper are encouraging for the ongoing DES-SN cosmological analysis, and demonstrate the tools to fully exploit the photometric DES-SN sample to constrain the dark energy equation-of-state. Our work lays the foundation for the cosmological analysis of the DES photometric SN sample and our results will be essential to assess the systematic error budget on cosmological parameters estimated from the DES-SN sample.

ACKNOWLEDGEMENTS

This work was supported by the Science and Technology Facilities Council (grant number ST/P006760/1) through the DISCnet Centre for Doctoral Training. MS acknowledges support from EU/FP7-ERC grant 615929, and PW acknowledges support from STFC grant ST/R000506/1. TMD acknowledges support from ARC grant FL180100168. LG acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities (MICIU) under the 2019 Ramón y Cajal program RYC2019-027683 and from the Spanish MICIU project PID2020-115253GA-I00. RH and MS were supported by DOE grant DE-FOA-0001781 and NASA grant NNH15ZDA001N-WFIRST. The material is based upon work supported by NASA under award number 80GSFC17M0002. LK thanks the UKRI Future Leaders Fellowship for support through the grant MR/T01881X/1.

This paper has gone through internal review by the DES collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft, and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, NFS's NOIRLab, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

Based in part on observations at Cerro Tololo Inter-American Observatory at NSF's NOIRLab (NOIRLab Prop. ID 2012B-0001; PI: J. Frieman), which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The DES data management system is supported by the National Science Foundation under Grant Numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MICINN under grants ESP2017-89838, PGC2018-094773, PGC2018-102021, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) do e-Universo (CNPq grant 465376/2014-2).

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

This work was completed in part with resources provided by the University of Chicago's Research Computing Center.

Finally, this work was based in part on data acquired at the Anglo-Australian Telescope, under program A/2013B/012. We acknowledge the traditional owners of the land on which the AAT stands, the Gamilaraay people, and pay our respects to elders past and present.

Software: NUMPY (Oliphant 2006), MATPLOTLIB (Hunter 2007), PANDAS (Wes McKinney 2010), SCIPY (Virtanen et al. 2020), SNANA (Kessler et al. 2009a), PIPPIN (Hinton & Brout 2020).

DATA AVAILABILITY STATEMENT

Input and configuration files needed to train and test the photometric classifier SNN and to run BBC are available at https://github.com/maria-vincenzi/DES_CC_simulations. Data relative to the DES photometric sample used in Fig. 7 are also available.

REFERENCES

Abbott T. M. C. et al., 2019a, Phys. Rev. Lett., 122, 171301
Abbott T. M. C. et al., 2019b, ApJ, 872, L30
Astier P. et al., 2006, A&A, 447, 31
Astier P. et al., 2013, A&A, 557, A55
Bazin G. et al., 2011, A&A, 534, A43
Bernstein J. P. et al., 2012, ApJ, 753, 152
Betoule M. et al., 2014, A&A, 568, A22
Brout D., Scolnic D., 2021, ApJ, 909, 26
Brout D. et al., 2019a, ApJ, 874, 106
Brout D. et al., 2019b, ApJ, 874, 150
Brout D., Hinton S., Scolnic D., 2021, ApJ, 912, L26
Campbell H. et al., 2013, ApJ, 763, 88
Conley A. et al., 2011, ApJS, 192, 1
Contreras C. et al., 2010, AJ, 139, 519
Flaugher B. et al., 2015, AJ, 150, 150
Foley R. J. et al., 2017, MNRAS, 475, 193
Goliath M., Amanullah R., Astier P., Goobar A., Pain R., 2001, A&A, 380, 6
González-Gaitán S. et al., 2014, ApJ, 795, 142
Guy J. et al., 2007, A&A, 466, 11
Guy J. et al., 2010, A&A, 523, A7
Hamuy M., Pinto P. A., 1999, AJ, 117, 1185
Hatano K., Branch D., Deaton J., 1998, ApJ, 502, 177
Hicken M. et al., 2009, ApJ, 700, 331
Hicken M. et al., 2012, ApJS, 200, 12
Hinton S., Brout D., 2020, J. Open Source Softw., 5, 2122
Hlozek R. et al., 2012, ApJ, 752, 79
Hložek R. et al., 2020, preprint (arXiv:2012.12392)
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jones D. O. et al., 2017, ApJ, 843, 6
Jones D. O. et al., 2018, ApJ, 857, 51
Jones D. O. et al., 2019, ApJ, 881, 19
Kelsey L. et al., 2020, MNRAS, 501, 4861
Kessler R., Scolnic D., 2017, ApJ, 836, 56
Kessler R. et al., 2009a, Publ. Astron. Soc. Pac., 121, 1028
Kessler R. et al., 2009b, ApJS, 185, 32
Kessler R., Conley A., Jha S., Kuhlmann S., 2010a, preprint (arXiv:1001.5210)
Kessler R. et al., 2010b, PASP, 122, 1415
Kessler R. et al., 2015, AJ, 150, 172
Kessler R. et al., 2019a, PASP, 131, 094501
Kessler R. et al., 2019b, MNRAS, 485, 1171
Kunz M., Bassett B. A., Hlozek R. A., 2007, Phys. Rev. D, 75, 103508
Lewis A., Bridle S., 2002, Phys. Rev. D, 66
Li W. et al., 2011, MNRAS, 412, 1441
Lidman C. et al., 2020, MNRAS, 496, 19
Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31
Marriner J. et al., 2011, ApJ, 740, 72
McKinney W., 2010, in van der Walt S., Millman J., eds, Proceedings of the 9th Python in Science Conference. p. 56
Möller A., de Boissière T., 2020, MNRAS, 491, 4277
Möller A. et al., 2016, J. Cosmology Astropart. Phys., 2016, 008
Möller A. et al., 2022, preprint (arXiv:2201.11142)

Oliphant T., 2006, Guide to NumPy. Available at https://web.mit.edu/dvp/Public/numpybook.pdf
Perlmutter S. et al., 1997, ApJ, 483, 565
Perlmutter S. et al., 1999, ApJ, 517, 565
Perrett K. et al., 2010, AJ, 140, 518
Planck Collaboration, 2020, A&A, 641, A6
Popovic B., Brout D., Kessler R., Scolnic D., Lu L., 2021, ApJ, 913, 49
Rest A. et al., 2014, ApJ, 795, 44
Riess A. G. et al., 1998, AJ, 116, 1009
Riess A. G. et al., 2007, ApJ, 659, 98
Riess A. G. et al., 2018, ApJ, 853, 126
Sako M. et al., 2011, ApJ, 738, 162
Sako M. et al., 2018, PASP, 130, 064002
Scolnic D., Kessler R., 2016, ApJ, 822, L35
Scolnic D. M. et al., 2018, ApJ, 859, 101
Smith M. et al., 2020a, MNRAS, 494, 4426
Smith M. et al., 2020b, AJ, 160, 267
Sullivan M. et al., 2010, MNRAS, 406, 782
Sullivan M. et al., 2011, ApJ, 737, 102
Swann E. et al., 2019, Messenger, 175, 58
Taylor J., 1997, Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements, 2nd edn. University Science Books, Mill Valley, California
The LSST Dark Energy Science Collaboration, 2018, preprint (arXiv:1809.01669)
The PLAsTiCC team et al., 2018, preprint (arXiv:1810.00001)
Tripp R., 1998, A&A, 331, 815
Vincenzi M., Sullivan M., Firth R. E., Gutiérrez C. P., Frohmaier C., Smith M., Angus C., Nichol R. C., 2019, MNRAS, 489, 5802
Vincenzi M. et al., 2021, MNRAS, 505, 2819
Virtanen P. et al., 2020, Nature Methods, 17, 261
Wiseman P. et al., 2020, MNRAS, 495, 4040

APPENDIX A: EFFECTS OF PROBABILITY CUTS

In Section 4.3, we showed that a probability cut of $P_{\rm Ia} > 0.5$ can reduce contamination in the DES-SN sample by a factor of 4–5, depending on the SNN classifier considered. However, the BEAMS/BBC framework is specifically designed to handle samples that include both SNe Ia and contaminants, with the BEAMS likelihood calculated using $P_{\rm Ia}$. Here we test the impact of combining BEAMS/BBC with probability-based cuts on cosmology.

A1 Core collapse SN likelihood and BBC configurations

We combine the $P_{\rm Ia} > 0.5$ selection and several different configurations of BBC, summarized in Table A1. Applying a probability cut removes all SNe with $P_{\rm Ia} < 0.5$ from the main sample and from the simulations used to estimate bias corrections (which only include SNe Ia) and in the core-collapse SN simulation used to map the core-collapse SN likelihood ($\mathcal{L}_{\rm CC}$). Probability cuts therefore have a complex impact on the analysis.

Table A1. BBC options tested with a SALT2 and $P_{\rm Ia}$>0.5 selection.

BBC configuration	Classifier	Modelling of $D_{\rm CC}$
SNN(Base) $D_{\rm CC}$(Base) PIa>.5	SNN(Base)	Baseline
SNN(J17) $D_{\rm CC}$(Base) PIa>.5	SNN(J17)	Baseline
SNN(H20) $D_{\rm CC}$(Base) PIa>.5	SNN(H20)	Baseline
SNN(Base) $D_{\rm CC}$(H12) PIa>.5	SNN(Base)	Fit (H12)
SNN(J17) $D_{\rm CC}$(H12) PIa>.5	SNN(J17)	Fit (H12)
SNN(H20) $D_{\rm CC}$(H12) PIa>.5	SNN(H20)	Fit (H12)

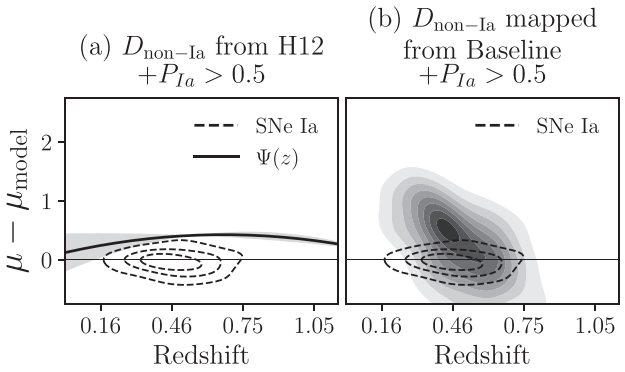

Figure A1. Same as Fig. 2 but when applying $P_{\rm Ia}$-based cuts.

Fig. A1 shows the effect of a probability selection on the two core-collapse SN likelihood models tested. Comparing Figs A1 a and 2(a), the best fit $\Psi(z)$ (and relative $\sigma_{\rm CC,\,int}(z)$) measured when a $P_{\rm Ia} > 0.5$ selection is applied is significantly different from the best-fitting $\Psi(z)$ estimate without such a selection. This is expected because the $P_{\rm Ia}$ cut only selects the brightest contaminants and significantly reshapes the distribution of contamination on the Hubble diagram. This is particularly evident at high redshift, where contamination sharply drops and $\Psi(z)$ is an extrapolation.

Similar differences are seen when comparing the core-collapse SN maps derived from the Baseline simulation *before* applying the $P_{\rm Ia}$ cut (Fig. 2b) and *after* (Fig. A1b). After $P_{\rm Ia}$ cuts, the distribution of core-collapse SN contamination is shifted to slightly higher redshifts ($0.3 > z > 0.7$), skewed towards the SN Ia likelihood (centred on $\mu_{\rm model}$) and sharply reduced at high redshift.

A2 Conclusions

Introducing a probability-based selection makes the modelling of $D_{\rm CC}$ more complex. This can lead to significant biases when using the H12 approach, where the core-collapse SN likelihood is fitted from the data and it is assumed to be fully described by a second-order polynomial. After probability cuts, this assumption is not adequate because the contamination likelihood at high redshift is essentially an extrapolation of the fitted polynomial and it does not reflect the drop in contamination seen in simulations.

In Fig. A2, we show that biases on fitted $\mu^b_{\rm Ia}$ when implementing SNN $D_{\rm non-Ia}$(Base) and $P_{\rm Ia>0.5}$ cut are still <10 mag. However, when applying the H12 approach, the contamination likelihood is not robustly modelled and many high-redshift, faint SNe Ia are assigned a higher likelihood of being contaminants and excluded from the cosmological fit. This biases $\mu^b_{\rm Ia}$ towards negative values and propagates to the estimate of cosmological parameters. In Fig. A3, we show the w-biases for the different BBC configurations tested and we find biases larger than 0.04 for the majority of the configurations where the H12 approach is used. We note that the main driver of the bias in this case is not the presence of contaminants in the sample but the loss of SNe Ia in the cosmological fit.

Finally, when the contaminants likelihood is modelled from the Baseline simulation, the recovered biases are equal to or lower than 2–3 per cent and generally consistent with the biases found when a $P_{\rm Ia}$ cut is not applied.

In summary, our tests show that applying a probability-based selection perhaps counter-intuitively provides equal or higher biases on cosmological parameters. The more accurate the classifier, the lower the residual contamination in the sample and the more uncertain the modelling of contamination in BEAMS. For these

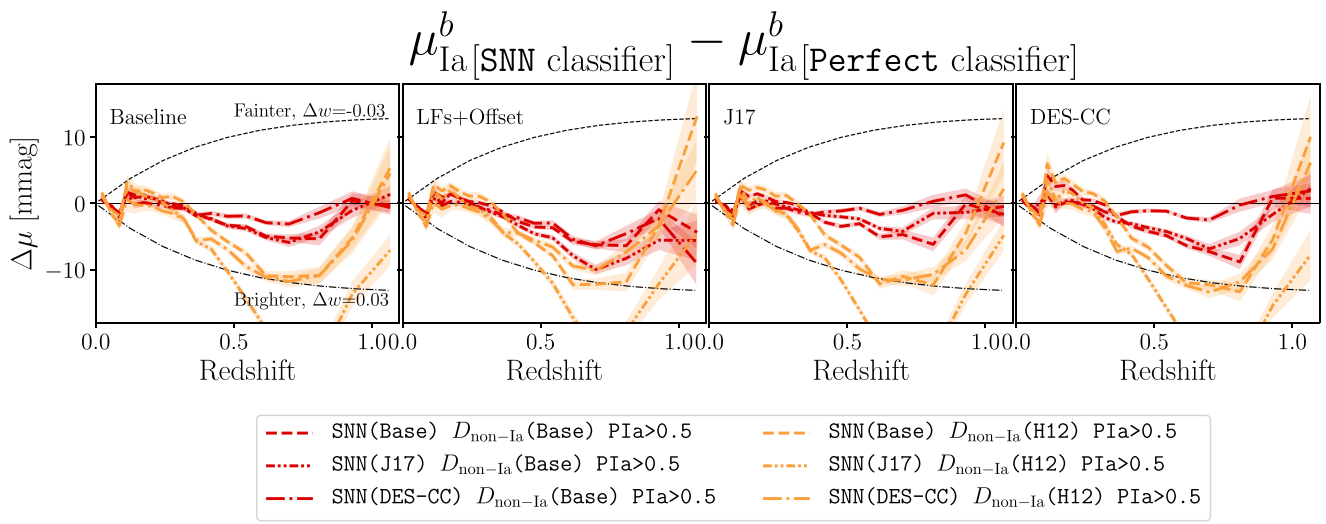

Figure A2. Same as Fig. 10, but applying a $P_{\rm Ia} > 0.5$ selection cut.

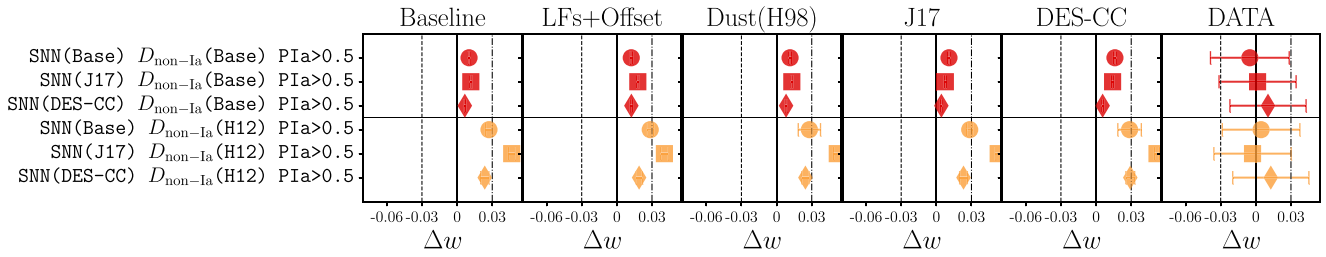

Figure A3. Same as Fig. 8 but for BBC configurations in Table A1.

reasons, a probability-based selection is not recommended and we do not implement it in our main analysis.

APPENDIX B: REJECTING SNE WITHOUT A VALID BIAS CORRECTION

With BBC it is not always possible to estimate valid bias corrections for every SN, particularly those in regions of parameter space where few SNe are simulated (see Section 3.4). These SNe are excluded from a cosmological analysis and this reduces contamination in the sample. Here we test how our results change if the requirement of a valid bias correction is relaxed, and if SNe without a valid bias correction are retained in the sample but with $\mu_{\rm bias}$ set to zero. Since setting $\Delta\mu_{\rm bias} = 0$ is clearly an incorrect approach in a cosmology analysis, we do not present updated results; rather, we compare the impact for data versus simulation to ensure that this effect is properly modelled.

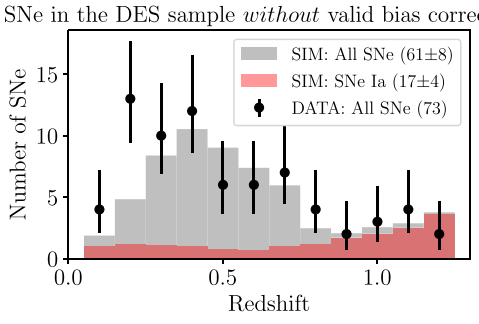

Figure B1. Redshift distribution of SN events for which BBC does not provide valid bias corrections. We compare the sample of such events in the DES data (open histogram) with the sample of such events in our DES-like simulations (filled grey histogram). In the simulations, only a third of the SNe without valid bias corrections are SNe Ia (red filled histogram).

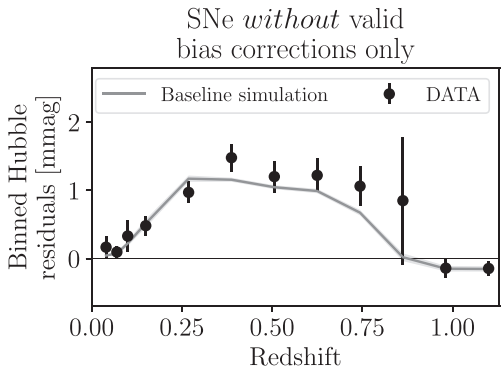

Figure B2. Difference in the observed and simulated binned Hubble diagram estimated when using our reference BBC configuration but including SNe without valid bias corrections. Simulations are generated using the Baseline approach. Uncertainties are estimated as the r.m.s. spread measured over the 50 realizations of the Baseline simulation.

The requirement of a valid bias correction significantly affects both the low-z and DES-SN samples (Table 2), but here we focus on the DES-SN sample and effects at higher redshifts. In Fig. B1, we present

the redshift distribution of observed and simulated DES SNe that pass the SALT2 selection, but do not have a valid bias correction. These distributions are generally consistent, but potential discrepancies are observed in the two lowest redshift bins. This suggests that the data include more atypical SNe than are modelled in the simulations.

Fig. B2 shows the observed and simulated binned Hubble residuals estimated when considering only SNe without valid bias corrections. The average Hubble residuals of this sub-population of uncorrected events is low (less the 2 mmag) and consistent between observations and simulations. This test confirms that we can model the selection effects introduced by BBC, and it further validates the results obtained using our simulations.

[1]*School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
[2]*Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK*
[3]*Department of Physics, Duke University, Durham, NC 27708, USA*
[4]*Centre for Astrophysics & Supercomputing, Swinburne University of Technology, Victoria 3122, Australia*
[5]*Université Clermont Auvergne, CNRS/IN2P3, LPC, F-63000 Clermont-Ferrand, France*
[6]*The Research School of Astronomy and Astrophysics, Australian National University, ACT 2601, Australia*
[7]*African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg 7945, South Africa*
[8]*Department of Maths and Applied Maths, University of Cape Town, Cape Town, 7700, South Africa*
[9]*South African Astronomical Observatory, Observatory, Cape Town, 7925, South Africa*
[10]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[11]*INAF, Astrophysical Observatory of Turin, I-10025 Pino Torinese, Italy*
[12]*School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia*
[13]*Institut d'Estudis Espacials de Catalunya (IEEC), E-08034 Barcelona, Spain*
[14]*Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, E-08193 Barcelona, Spain*
[15]*Department of Astrophysics, American Museum of Natural History, New York, NY 10024, USA*
[16]*Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA*
[17]*Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[18]*Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA*
[19]*Sydney Institute for Astronomy, School of Physics, A28, The University of Sydney, NSW 2006, Australia*
[20]*Centre for Gravitational Astrophysics, College of Science, The Australian National University, ACT 2601, Australia*
[21]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[22]*Laboratório Interinstitucional de e-Astronomia - LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil*
[23]*Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA*
[24]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid 28040, Spain*
[25]*CNRS, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France*
[26]*Sorbonne Universités, UPMC Univ Paris 06, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France*
[27]*Department of Physics & Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[28]*Kavli Institute for Particle Astrophysics & Cosmology, P.O. Box 2450, Stanford University, Stanford, CA 94305, USA*
[29]*SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[30]*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra (Barcelona), Spain*
[31]*Astronomy Unit, Department of Physics, University of Trieste, via Tiepolo 11, I-34131 Trieste, Italy*
[32]*INAF-Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[33]*Institute for Fundamental Physics of the Universe, Via Beirut 2, 34014 Trieste, Italy*
[34]*Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil*
[35]*Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[36]*Hamburger Sternwarte, Universität Hamburg, Gojenbergsweg 112, D-21029 Hamburg, Germany*
[37]*Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India*
[38]*Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA*
[39]*Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, NO-0315 Oslo, Norway*
[40]*Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, E-28049 Madrid, Spain*
[41]*Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA*
[42]*Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, D-81679 Munich, Germany*
[43]*Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA*
[44]*Department of Physics, The Ohio State University, Columbus, OH 43210, USA*
[45]*Australian Astronomical Optics, Macquarie University, North Ryde, NSW 2113, Australia*
[46]*Lowell Observatory, 1400 Mars Hill Rd, Flagstaff, AZ 86001, USA*
[47]*Observatories of the Carnegie Institution for Science, 813 Santa Barbara St., Pasadena, CA 91101, USA*
[48]*Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*
[49]*Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, São Paulo, SP, 05314-970, Brazil*
[50]*George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, and Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA*
[51]*Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain*
[52]*Physics Department, 2320 Chamberlin Hall, University of Wisconsin-Madison, 1150 University Avenue Madison, WI 53706-1390, USA*
[53]*Department of Astronomy, University of California, Berkeley, 501 Campbell Hall, Berkeley, CA 94720, USA*
[54]*Center for Astrophysical Surveys, National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA*
[55]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[56]*Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA*
[57]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[58]*Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[59]*Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse, D-85748 Garching, Germany*
[60]*Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians Universität München, Scheinerstr. 1, D-81679 München, Germany*
[61]*Department of Physics and Astronomy, Pevensey Building, University of Sussex, Brighton BN1 9QH, UK*

This paper has been typeset from a TEX/LATEX file prepared by the author.