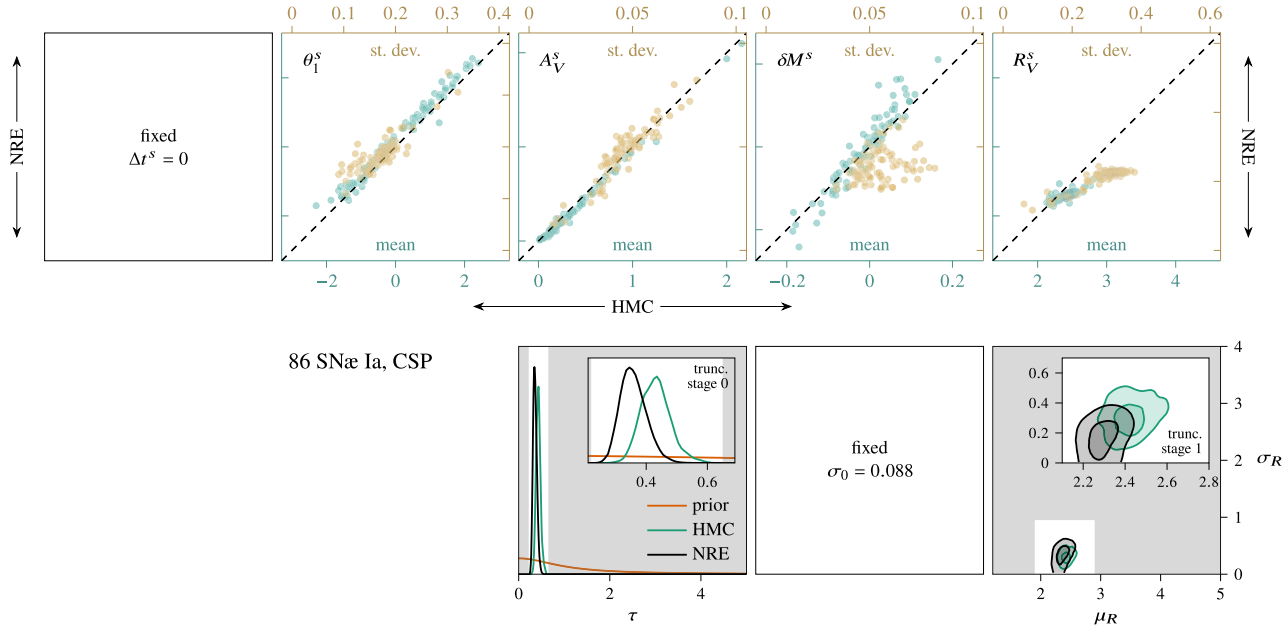

Figure 7. Inference results from the real data set. *Top*: comparison of marginal local-parameter posterior moments derived with NRE and HMC. *Bottom*: posteriors for the global parameters. See Fig. 5 for more details.

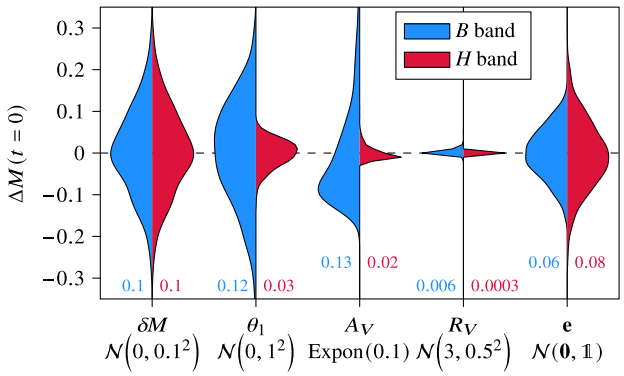

Figure 8. Variations in rest-frame B and H absolute magnitudes at phase 0 (around maximum), as simulated with the BayeSN trained by M20, induced by varying each of the free local parameters according to its fiducial hierarchical prior, with respect to a reference value with $\delta M = 0$, $\theta_1 = 0$, $A_V = 0.1$, $R_V = 3$, $\epsilon = 0$. Numbers along the bottom specify the standard deviation for the two bands.

into a common-dimensional space; a *shared* fully connected SN post-processing subnetwork applied in parallel to the embeddings of all SNæ; and a fully connected summarizer combining the results. It is as expressive and fast to evaluate and train as conventional fully connected networks (taking a few hours to converge with training data generated in ∼30 mins, in the same ballpark as highly optimized likelihood codes) but manages to fully extract the relevant information before overfitting.

In the present work, we have made a number of simplifying assumptions that do not affect significantly the results in the low-redshift, fairly small-size, high-signal-to-noise case we consider. For example, we employ the simplified instrumental model that summarizes observational uncertainties in a `FLUXCALERR`; to properly investigate the impact of measurement-related systematics, one must introduce calibration parameters (e.g. $\mathrm{ZP}^{s,i}$) for each data point, which would significantly impact runtimes of likelihood-based methods.[15] The same applies to the hierarchical parameters we have kept fixed in this study: namely, the MW dust law (and its variation along different lines of sight) and extinction amount and the total and cosmological redshifts of the supernovæ. In contrast, with SBI all extra parameters can be marginalized implicitly by stochastically sampling them in the simulator, or marginally inferred in the same way (and at the same time) as the parameters we have already considered, provided suitable models: e.g. Zhang, Yuan & Chen (2023) for MW dust and Rahman et al. (2022) for redshifts induced by peculiar velocities consistent with models of the galactic bulk flow.

Along the same line of thought, one can also consider training the SN Ia template used to simulate light curves – represented in BayeSN through $\boldsymbol{W}_k$ and $\boldsymbol{\Sigma}_\mathbf{e}$ – together with the properties of individual SNæ Ia (notably, distances used for cosmology). Even though this is the main reason behind using a hierarchical model, for BayeSN this is yet to be attempted because of subtleties related to selection effects. Likelihood-based analyses are thus typically split in two stages, with the second one (cosmological inference) using a fixed mean/median SN Ia spectral energy distribution (SED) model. To SBI, on the other hand, $\boldsymbol{W}_k$ and $\boldsymbol{\Sigma}_\mathbf{e}$ are just another set of global parameters implicitly marginalized in the training data. Sampling them from the very general priors used by M22 (and successors) will introduce tremendous and difficult-to-handle variability in the training data, which, however, can be remedied through high-dimensional truncated SBI (see, e.g. Anau Montel, Alvey & Weniger 2023; List, Montel & Weniger 2023). Furthermore, SBI opens up the possibility of using *implicit* (data-driven) priors and/or SN templates implemented as neural generative models trained (or fine-tuned)

[15]Current state-of-the-art analyses (e.g. Brout et al. 2022a, b; Vincenzi et al. 2024) employ a simplistic yet laborious procedure of linear propagation of systematic 'uncertainties' (in fact, offsets) to the final results (e.g. estimates of cosmological parameters). This can be streamlined and made more principled with the SBI approach introduced here.

simultaneously with the inference, as recently done by Alsing et al. (2024) for galaxy photometry.

Beyond individual objects, numerous qualitative enhancements of the population modelling of SNæ Ia and their hosts have been explored in the literature. These follow two main threads: considering correlations of SN properties with those of their hosts, and allowing for their evolution with time (with redshift used as a proxy). Due to the high-dimensional nature of the required analysis, confronting the different models using likelihood-based pipelines has been cumbersome and reliant mainly on visual inspection of the results of parameter inference, e.g. comparing the dust-law distributions of high- and low-mass galaxies or examining trends of inferred properties with redshift. SBI, however, provides an avenue towards principled Bayesian model comparison by giving direct access to marginalized model probabilities (and Bayes factors) even for models with thousands of dimensions like the one in this study, as recently demonstrated by Karchev et al. (2023a). Furthermore, amortization allows for exploring the results as a function of the values of the underlying parameters on mock data – unthinkable with likelihood-based methods.

Selection effects and non-Ia contamination thus remain the two major hurdles left before a fully-fledged application of SBI to SN Ia cosmology becomes viable. Accounting for them requires simulating the way transients are identified and classified into a survey release, leading to different-sized surveys in the training set and transients observed in different time/band configurations. One might imagine two ways to circumvent these problems. One is to take a step back in the data processing pipeline, simulating and subsequently feeding in the neural ratio estimator raw telescope images, which have a set number, order, and characteristics after the survey has been performed (see, e.g. Sánchez et al. (2022) for a similar approach but using the conventional SN cosmology pipeline). This is equivalent to padding the light curves, which would be the standard approach to unequal-length sequences in machine learning, and including light curves for undetected objects. The downside is clear: the network must learn selection effects and contamination from a prohibitively large amount of mostly uninformative data. Alternatively, one might still try to condition the simulator on the observed light curves of detected objects by smartly modifying the hierarchical priors. However, this might cause issues when many objects are considered, as we discuss in Karchev et al. (in preparation).

In light of this, our current approach – conditioning on the number and order of SNæ in a survey and on the number and order of observations of each SN – has limited prospects of solving selection effects. In upcoming work, we will demonstrate cosmological inference in the presence of selection effects and variable-length data by using tools that have already been exploited in the SN literature: e.g. Gaussian process regression for regularizing the light curves (e.g. Revsbech et al. 2018; Boone 2019), in combination with cutting-edge techniques for permutation-invariant SBI (Rodrigues et al. 2021; Campeau-Poirier et al. 2023; Heinrich et al. 2024; Makinen, Alsing & Wandelt 2023).

We make use of `Clipppy`,[16] a Python package based on `pyro` (Bingham et al. 2019) and `PyTorch` (Paszke et al. 2019), for the probabilistic part of our forward simulator and `PyTorch Lightning` (Falcon & The PyTorch Lightning team 2023) for training the inference network.

[16] https://github.com/kosiokarchev/clipppy

ACKNOWLEDGEMENTS

RT acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 - 'Project FAIR Future Artificial Intelligence Research'. This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22]. RT is partially supported by the Fondazione ICSC, Spoke 3 'Astrophysics and Cosmos Observations', Piano Nazionale di Ripresa e Resilienza Project ID CN00000013 'Italian Research Center on High-Performance Computing, Big Data and Quantum Computing' funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di 'campioni nazionali di R&S (M4C2-19)' - Next Generation EU (NGEU).

MG and KSM are supported by the European Union's Horizon 2020 research and innovation programme under ERC Grant Agreement No. 101002652 and Marie Sklodowska-Curie Grant Agreement No. 873089. BMB is supported by the Cambridge Centre for Doctoral Training in Data-Intensive Science funded by the UK Science and Technology Facilities Council (STFC).

CW has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 864035).

Part of this work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

DATA AVAILABILITY

This article makes use of data released with `SNANA`, available on Zenodo at https://dx.doi.org/10.5281/zenodo.4001177. The mock data generated in this research will be shared upon reasonable request to the corresponding author.

REFERENCES

Alsing J., Charnock T., Feeney S., Wandelt B., 2019, MNRAS, 488, 4440

Alsing J., Thorp S., Deger S., Peiris H., Leistedt B., Mortlock D., Leja J., 2024, pop-cosmos: A comprehensive picture of the galaxy population from COSMOS data. preprint (arxiv:2402.00935)

Alsing J., Wandelt B., 2019, MNRAS, 488, 5093

Alvey J., Bhardwaj U., Domcke V., Pieroni M., Weniger C., 2024, Phys. Rev. D, 109, 083008

Alvey J., Bhardwaj U., Nissanke S., Weniger C., 2023b, What to do when things get crowded? scalable joint analysis of overlapping gravitational wave signals. preprint (arxiv:2308.06318)

Alvey J., Gerdes M., Weniger C., 2023a, MNRAS, 525, 3662

Anau Montel N., Alvey J., Weniger C., 2023, Scalable Inference with Autoregressive Neural Ratio Estimation. preprint (arxiv:2308.08597)

Anau Montel N., Coogan A., Correa C., Karchev K., Weniger C., 2022, MNRAS, 518, 2746

Anau Montel N., Weniger C., 2022, Detection is truncation: studying source populations with truncated marginal neural ratio estimation. preprint (arxiv:2211.04291)

Autenrieth M., van Dyk D. A., Trotta R., Stenning D. C., 2023, Statistical Analysis and Data Mining, 19

Autenrieth M., Wright A. H., Trotta R., van Dyk D. A., Stenning D. C., Joachimi B., 2024, Improved Weak Lensing Photometric Redshift Calibration via Stratlearn and Hierarchical Modeling. preprint (arxiv:2401.04687)

Avelino A., Friedman A. S., Mandel K. S., Jones D. O., Challis P. J., Kirshner R. P., 2019, ApJ, 887, 106

Barbary K. et al., 2016, Sncosmo, Zenodo

Bernardo R. C., Grandón D., Levi Said J., Cárdenas V. H., 2023, Phys. Dark Univ., 40, 101213
Betoule M. et al., 2014, A&A, 568, A22
Bhardwaj U., Alvey J., Miller B. K., Nissanke S., Weniger C., 2023, Phys. Rev. D, 108, 042004
Bingham E. et al., 2019, J. Mach. Learn. Res., 20, 973
Boone K., 2019, AJ, 158, 257
Brehmer J., Louppe G., Pavez J., Cranmer Kyle, 2020, Proc. Natl. Acad. Sci., 117, 5242
Brout D. et al., 2022a, ApJ, 938, 110
Brout D. et al., 2022b, ApJ, 938, 111
Brout D., Riess A., 2023, in Hubble Constant Tension. preprint (arxiv:2311.08253)
Brout D., Scolnic D., 2021, ApJ, 909, 26
Campeau-Poirier È., Perreault-Levasseur L., Coogan A., Hezaveh Y., 2023, Time Delay Cosmography with a Neural Ratio Estimator. preprint (arxiv:2309.16063)
Carrick J., Turnbull S. J., Lavaux G., Hudson M. J., 2015, MNRAS, 450, 317
Charnock T., Lavaux G., Wandelt B. D., 2018, Phys. Rev. D, 97, 083004
Chen J.-F., Wang Y.-C., Zhang T., Zhang T.-J., 2023, Phys. Rev. D, 107, 063517
Childress M. et al., 2013, ApJ, 770, 108
Chung C., Yoon S.-J., Park S., An S., Son J., Cho H., Lee Y.-W., 2023, ApJ, 959, 94
Coogan A., Anau Montel N., Karchev K., Grootes M. W., Nattino F., Weniger C., 2024, MNRAS, 527, 66
Cranmer K., Brehmer J., Louppe G., 2020, PNAS, 117, 30055
Davis T. M. et al., 2011, ApJ, 741, 67
Delaunoy A., Hermans J., Rozet F., Wehenkel A., Louppe G., 2022, Towards Reliable Simulation-based Inference with Balanced Neural Ratio Estimation. preprint (arxiv:2208.13624)
Di Valentino E. et al., 2021, Class. Quant. Grav., 38, 153001
Draine B. T., 2003, ARA&A, 41, 241
Falcon W., The PyTorch Lightning Team, 2023, Pytorch Lightning, Zenodo
Fitzpatrick E. L., 1999, PASP, 111, 63
Grayling M., Thorp S., Mandel K. S., Dhawan S., Uzsoy A. S., Boyd B. M., Hayesn E. E., Ward S. M., 2024, Scalable hierarchical BayeSN inference: Investigating dependence of SN Ia host galaxy dust properties on stellar mass and redshift. preprint (arxiv:2401.08755)
Guy J. et al., 2007, A&A, 466, 11
Guy J., Astier P., Nobili S., Regnault N., Pain R., 2005, A&A, 443, 781
Handley W. J., Hobson M. P., Lasenby A. N., 2015, MNRAS, 450, L61
Heinrich L., Mishra-Sharma S., Pollard C., Windischhofer P., 2024, Transactions on Machine Learning Research
Hermans J., Begy V., Louppe G., 2020, in Proceedings of the 37th International Conference on Machine Learning, PMLR. p. 4239
Hermans J., Delaunoy A., Rozet F., Wehenkel A., Begy V., Louppe G., 2022, Transactions on Machine Learning Research.
Hill R. et al., 2018, MNRAS, 481, 2766
Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R., 2012, Improving neural networks by preventing co-adaptation of feature detectors. preprint (arxiv:1207.0580)
Hinton S. R. et al., 2019, ApJ, 876, 15
Hoffman M. D., Gelman A., 2014, J. Mach. Learn. Res., 15, 1593
Hogg D. W., 2000, preprint(astro-ph/9905116)
Hounsell R. et al., 2018, ApJ, 867, 23
Hsiao E. Y., Conley A., Howell D. A., Sullivan M., Pritchet C. J., Carlberg R. G., Nugent P. E., Phillips M. M., 2007, ApJ, 663, 1187
Huterer D., Shafer D. L., 2018, Rep. Prog. Phys., 81, 016901
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jeffrey N., Wandelt B. D., 2024, Mach. Learn.: Sci. Technol., 5, 015008
Jennings E., Wolf R., Sako M., 2016, A new approach for obtaining cosmological constraints from Type Ia Supernovae using Approximate Bayesian Computation. preprint (arxiv:1611.03087)
Jones D. O. et al., 2018, ApJ, 867, 108
Jones D. O., Riess A. G., Scolnic D. M., 2015, ApJ, 812, 31
Karchev K., 2023, J. Cosmol. Astropart. Phys., 07, 065
Karchev K., Trotta R., Weniger C., 2023a, SimSIMS: Simulation-based Supernova Ia Model Selection with thousands of latent variables. preprint (arxiv:2311.15650)
Karchev K., Trotta R., Weniger C., 2023b, MNRAS, 520, 2209.06733
Kelly P. L., Hicken M., Burke D. L., Mandel K. S., Kirshner R. P., 2010, ApJ, 715, 743
Kelsey L. et al., 2021, MNRAS, 501, 4861
Kenworthy W. D. et al., 2021, ApJ, 923, 265
Kessler R. et al., 2009, PASP, 121, 1028
Kessler R., Scolnic D., 2017, ApJ, 836, 56
Kim Y.-L., Kang Y., Lee Y.-W., 2019, J. Kor. Astron. Soc., 52, 181
Kim Y.-L., Smith M., Sullivan M., Lee Y.-W., 2018, ApJ, 854, 24
Kingma D. P., Ba J., 2017, Adam: A Method for Stochastic Optimization. preprint (arxiv:1412.6980)
Krisciunas K. et al., 2017, AJ, 154, 211
Kunz M., Bassett B. A., Hlozek R. A., 2007, Phys. Rev. D, 75, 103508
Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258
Lemos P., Coogan A., Hezaveh Y., Perreault-Levasseur L., 2023, in Proceedings of the 40th International Conference on Machine Learning. PMLR, p. 19256
List F., Montel N. A., Weniger C., 2023, Bayesian Simulation-based inference for Cosmological Initial Conditions. preprint (arxiv:2310.19910)
LSST Science Collaboration, 2009, LSST Science Book, version 2.0. preprint (arxiv:0912.0201)
Lueckmann J.-M., Boelts J., Greenberg D., Goncalves P., Macke J., 2021, in Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. PMLR, p. 343 https://proceedings.mlr.press/v130/lueckmann21a.html
Ma C., Corasaniti P.-S., Bassett B. A., 2016, MNRAS, 463, 1651
Makinen T. L., Alsing J., Wandelt B. D., 2023, Fishnets: Information-Optimal, Scalable Aggregation for Sets and Graphs. preprint (arxiv:2310.03812)
Malmquist K. G., 1922, Meddelanden fran Lunds Astronomiska Observatorium Serie I, 100, 1
Malmquist K. G., 1925, Meddelanden fran Lunds Astronomiska Observatorium Serie I, 106, 1
Mandel K. S., Narayan G., Kirshner R. P., 2011, ApJ, 731, 120
Mandel K. S., Scolnic D. M., Shariff H., Foley R. J., Kirshner R. P., 2017, ApJ, 842, 93
Mandel K. S., Thorp S., Narayan G., Friedman A. S., Avelino A., 2022, MNRAS, 510, 3939
Mandel K. S., Wood-Vasey W. M., Friedman A. S., Kirshner R. P., 2009, ApJ, 704, 629
March M. C., Trotta R., Berkes P., Starkman G. D., Vaudrevange P. M., 2011, MNRAS, 418, 2308
March M. C., Wolf R. C., Sako m., D'Andrea C., Brout D., 2018, A Bayesian approach to truncated data sets: An application to Malmquist bias in Supernova Cosmology. preprint (arxiv:1804.02474)
Miller B. K., Cole A., Forré P., Louppe G., Weniger C., 2021, in Advances in Neural Information Processing Systems. Curran Associates, Inc., Morehouse Lane Red Hook, NY 12571 USA, p. 129
Moreno-Raya M. E., López-Sánchez Á. R., Mollá M., Galbany L., Vílchez J. M., Carnero A., 2016a, MNRAS, 462, 1281
Moreno-Raya M. E., Mollá M., López-Sánchez Á. R., Galbany L., Vílchez J. M., Carnero Rosell A., Domínguez I., 2016b, ApJ, 818, L19
Moreno-Torres J. G., Raeder T., Alaiz-Rodríguez R., Chawla N. V., Herrera F., 2012, Pattern Recogn., 45, 521
Paszke A. et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems 32. Curran Associates, Inc., Morehouse Lane Red Hook, NY 12571 USA, p. 8024
Perlmutter S. et al., 1999, ApJ, 517, 565
Phan D., Pradhan N., Jankowiak M., 2019, Composable Effects for Flexible and Accelerated Probabilistic Programming in Numpyro. preprint (arxiv:1912.11554)
Phillips M. M., 1993, ApJ, 413, L105
Popovic B. et al., 2024, MNRAS, 529, 2100
Popovic B., Brout D., Kessler R., Scolnic D., 2023, ApJ, 945, 84
Popovic B., Brout D., Kessler R., Scolnic D., Lu L., 2021, ApJ, 913, 49

Rahman W., Trotta R., Boruah S. S., Hudson M. J., van Dyk D. A., 2022, MNRAS, 514, 139
Revsbech E. A., Trotta R., van Dyk D. A., 2018, MNRAS, 473, 3969
Riess A. G. et al., 1998, AJ, 116, 1009
Riess A. G. et al., 2022, ApJ, 934, L7
Rigault M. et al., 2013, A&A, 560, A66
Rigault M. et al., 2015, ApJ, 802, 20
Rigault M. et al., 2020, A&A, 644, A176
Rodrigues P., Moreau T., Louppe G., Gramfort A., 2021, in Advances in Neural Information Processing Systems. Curran Associates, Inc., Morehouse Lane Red Hook, NY 12571 USA, p. 13432
Roman M. et al., 2018, A&A, 615, A68
Rose B. M., Garnavich P. M., Berg M. A., 2019, ApJ, 874, 32
Rubin D. et al., 2015, ApJ, 813, 137
Rubin D. et al., 2023, Union Through UNITY: Cosmology with 2,000 SNe Using a Unified Bayesian Framework. preprint (arxiv:2311.12098)
Sánchez B. O. et al., 2022, ApJ, 934, 96
Saunders C. et al., 2018, ApJ, 869, 167
Saxena A., Cole A., Gazagnes S., Meerburg P. D., Weniger C., Witte S. J., 2023, MNRAS, 525, 6097
Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103
Shariff H., Jiao X., Trotta R., van Dyk D. A., 2016, ApJ, 827, 1
Sisson S., Fan Y., Beaumont M., 2018, Handbook of Approximate Bayesian Computation, 1st edn., Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, New York
Stein G. et al., 2022, ApJ, 935, 5
Sullivan M. et al., 2010, MNRAS, 406, 782
Taylor G., Lidman C., Tucker B. E., Brout D., Hinton S. R., Kessler R., 2021, MNRAS, 504, 4111
Thorp S., Mandel K. S., 2022, MNRAS, 517, 2360
Thorp S., Mandel K. S., Jones D. O., Kirshner R. P., Challis P. M., 2024, Using Rest-Frame Optical and NIR Data from the RAISIN survey to Explore the Redshift Evolution of Dust Laws in SN ia Host Galaxies. preprint (arxiv:2402.18624)
Thorp S., Mandel K. S., Jones D. O., Ward S. M., Narayan G., 2021, MNRAS, 508, 4310
Tripp R., 1997, A&A, 325, 871
Tripp R., 1998, A&A, 331, 815
Villar V. A., 2022, Amortized Bayesian Inference for Supernovae in the Era of the Vera Rubin Observatory using Normalizing Flows. preprint (arxiv:2211.04480)
Vincenzi M. et al., 2024, The Dark Energy Survey Supernova Program: Cosmological Analysis and Systematic Uncertainties. preprint (arxiv:2401.02945)
Wang G.-J., Cheng C., Ma Y.-Z., Xia J.-Q., 2022, ApJS, 262, 24
Wang G.-J., Cheng C., Ma Y.-Z., Xia J.-Q., Abebe A., Beesham A., 2023, ApJS, 268, 7
Ward S. M. et al., 2023a, ApJ, 956, 111
Ward S. M., Dhawan S., Mandel K. S., Grayling M., Thorp S., 2023b, MNRAS, 526, 5715
Weyant A., Schafer C., Wood-Vasey W. M., 2013, ApJ, 764, 116
Wojtak R., Davis T. M., Wiis J., 2015, J. Cosmol. Astropart. Phys., 07, 025
Zeghal J., Lanusse F., Boucaud A., Remy B., Aubourg E., 2022, Neural Posterior Estimation with Differentiable Simulators. preprint (arxiv:2207.05636)
Zhang R., Yuan H., Chen B., 2023, ApJS, 269, 6

APPENDIX A: BAYESIAN VALIDATION AND FREQUENTIST CALIBRATION

Amortized inference allows validating the coverage properties of the approximate posteriors (in a Bayesian sense) and producing confidence regions with exact frequentist coverage, as we detailed in SICRET Subsection 3.4. Here, we briefly explain the two procedures and related concepts and implement them, respectively, for Figs A1 and A2.

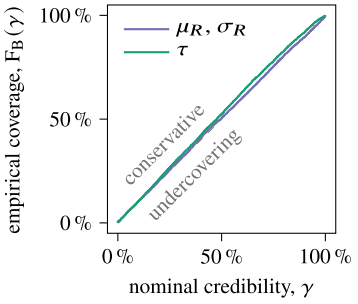

Figure A1. Bayesian P–P plot for global-parameter approximate posteriors trained for inference from the real data set.

We consider approximate posteriors $q(\boldsymbol{\Theta}|\boldsymbol{d}) \equiv \hat{\mathrm{r}}(\boldsymbol{\Theta}, \boldsymbol{d})\, \mathrm{p}(\boldsymbol{\Theta})$ evaluated for data $\boldsymbol{d}$ simulated from true parameter values $\boldsymbol{\Theta}_0$ and the corresponding credibilities

$$\gamma(\boldsymbol{\Theta}_0, \boldsymbol{d}) \equiv \int_{\Gamma_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_0, \boldsymbol{d})} q(\boldsymbol{\Theta}|\boldsymbol{d})\, \mathrm{d}\boldsymbol{\Theta} \tag{A1}$$

of highest-likelihood regions (HLRs)[17] $\Gamma_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_0, \boldsymbol{d}) \equiv \{\boldsymbol{\Theta} : \hat{\mathrm{r}}(\boldsymbol{\Theta}\boldsymbol{d}) > \hat{\mathrm{r}}(\boldsymbol{\Theta}_0, \boldsymbol{d})\}$.

Bayesian validation uses random parameters sampled from the prior: $\boldsymbol{\Theta}_0 \sim \mathrm{p}(\boldsymbol{\Theta}_0)$, and the self-consistency of the prior and data-averaged posterior – the fact that, when averaged over data sets sampled from the marginal likelihood $\mathrm{p}(\boldsymbol{d}) \equiv \int \mathrm{p}(\boldsymbol{d}|\boldsymbol{\Theta})\, \mathrm{p}|\boldsymbol{\Theta})\, \mathrm{d}\boldsymbol{\Theta}$, the exact posterior $\mathrm{p}(\boldsymbol{\Theta}|\boldsymbol{d})$ reverts to the prior $\mathrm{p}(\boldsymbol{\Theta})$:

$$\mathbb{E}_{\mathrm{p}(\boldsymbol{d})}[\mathrm{p}(\boldsymbol{\Theta}|\boldsymbol{d})] \equiv \int \mathrm{p}(\boldsymbol{\Theta}|\boldsymbol{d})\mathrm{p}(\boldsymbol{d})\, \mathrm{d}\boldsymbol{d} = \int \mathrm{p}(\boldsymbol{\Theta}, \boldsymbol{d})\, \mathrm{d}\boldsymbol{d} = \mathrm{p}(\boldsymbol{\Theta}), \tag{A2}$$

and so credibilities computed with $q(\boldsymbol{\Theta}|\boldsymbol{d}) \to \mathrm{p}(\boldsymbol{\Theta}|\boldsymbol{d})$ are uniformly distributed over [0; 1], or equivalently, have a cumulative distribution $F_{\mathrm{B}}(\gamma) = \gamma$. On a probability–probability (P–P) plot, depicting F versus γ, this manifests as a diagonal line. If, empirically, $F_{\mathrm{B}}(\gamma) > \gamma$, the posteriors $q(\boldsymbol{\Theta}|\boldsymbol{d})$ are, on average, conservative: i.e. they cover true values more often than expected. And conversely, $F_{\mathrm{B}}(\gamma) < \gamma$ implies they are overconfident, i.e. exhibit a greater scatter around the true value (or, possibly, even a bias) than expected from their sizes. However, $\mathrm{p}(\boldsymbol{\Theta}|\boldsymbol{d})$ is not the only distribution which has perfect Bayesian coverage: in fact, using even the prior for $q(\boldsymbol{\Theta}|\boldsymbol{d})$ would lead to $F_{\mathrm{B}}(\gamma) = \gamma$.

Instead of averaging over the prior, one can examine the distribution of credibilities, conditioning on a fixed parameter value $\boldsymbol{\Theta}_0$, which leads to $\boldsymbol{d} \sim \mathrm{p}(\boldsymbol{d}|\boldsymbol{\Theta}_0)$ (instead of $\boldsymbol{d} \sim \mathrm{p}(\boldsymbol{d})$) and so can be used for frequentist inference. In this scenario, in general, $\mathrm{F_f}(\gamma|\boldsymbol{\Theta}_0) \neq \gamma$ due to the approximate nature of $\mathrm{q}(\boldsymbol{\Theta}|\boldsymbol{d})$ on one hand, as before, but also because of the influence of a non-uniform prior. Calculating $\mathrm{F_f}(\gamma|\boldsymbol{\Theta}_0)$ as a function of $\boldsymbol{\Theta}_0$, e.g. on a grid or by using nearest neighbours among prior samples (we use the latter), allows one to derive the *required credibility* $\hat{\gamma}(\boldsymbol{\Theta}_0\tilde{\gamma})$, for any desired confidence $\tilde{\gamma}$, as the $\tilde{\gamma}^{\mathrm{th}}$ quantile of $\mathrm{F_f}(\gamma|\boldsymbol{\Theta}_0)$: see Karchev et al. (2023a, fig. 3) for an illustration.

[17] In SICRET, we instead used highest posterior density (HPD) regions. The difference is usually imperceptible, especially in well-constrained scenarios, and in fact any region definition can be used: here we use the approximate likelihood since it is directly accessible through $\hat{\mathrm{r}}$ and independent of parametrization, while posterior densities require evaluating the prior density as well (or density estimation from weighted samples). Of note is also an alternative method of defining credible regions using distances to random points (DRP), as described by Lemos et al. (2023), which can, in certain circumstances, detect a systematic bias in $q(\boldsymbol{\Theta}|\mathbf{d})$.

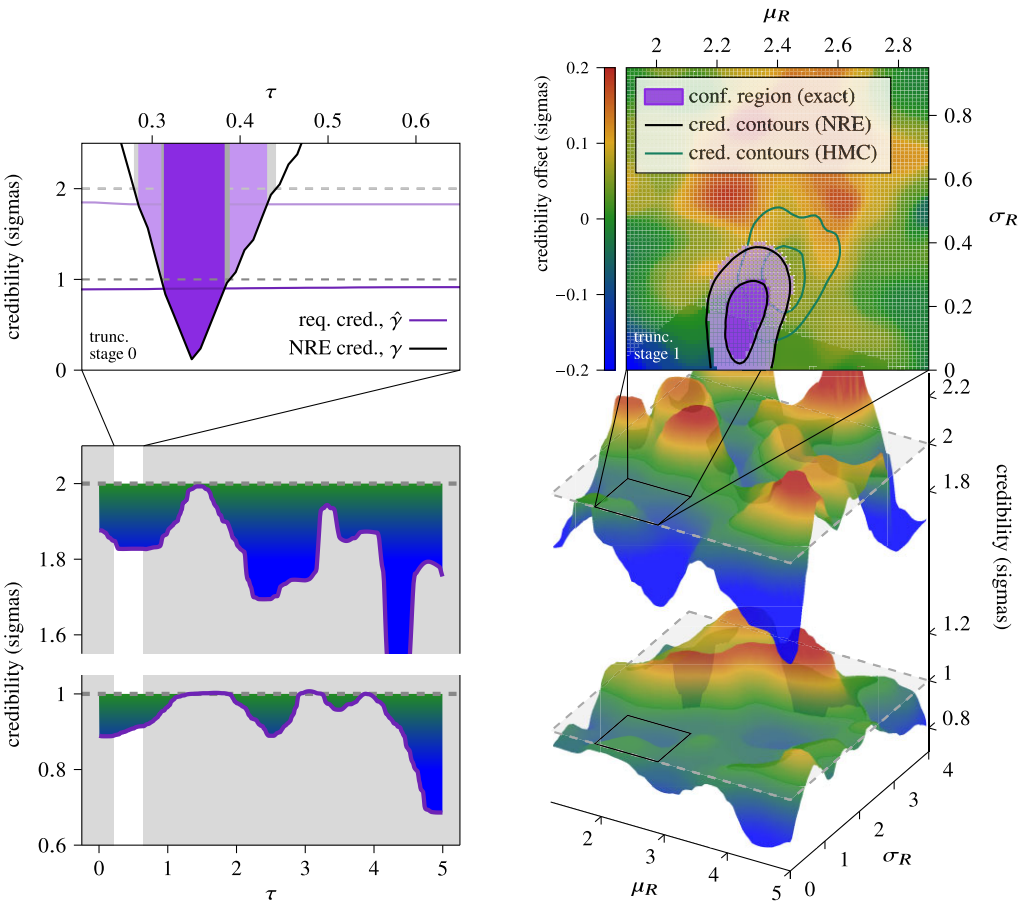

Figure A2. Calibrated frequentist global-parameter inference from the real CSP data set. As purple lines and coloured surfaces, we show the required credibility $\hat{\gamma}$ that achieves 1- or 2σ coverage (corresponding to 68.3 and 95.4 per cent in one dimension and ≈39 and ≈86 per cent in two dimensions), as a function of the parameter value. Everywhere, the colour axis represents the difference between the required credibility and the confidence level (empirical coverage). In the top row, using black lines, we depict nominal credibility from the NRE posterior evaluated on the observed data, which is used to derive calibrated 1- and 2σ confidence regions, filled in purple. The difference with the approximate posterior is insignificant, but the confidence region thus constructed has guaranteed coverage.

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