A COMPASS to Model Comparison and Simulation-Based Inference in Galactic Chemical Evolution

Berkay Günes [1] **Sven Buder** [2 3] **Tobias Buck** [1 4]

Abstract

We present COMPASS, a novel simulation-based inference framework that combines score-based diffusion models with transformer architectures to jointly perform parameter estimation and Bayesian model comparison across competing Galactic Chemical Evolution (GCE) models. COMPASS handles high-dimensional, incomplete, and variable-size stellar abundance datasets. Applied to high-precision elemental abundance measurements, COMPASS evaluates 40 combinations of nucleosynthetic yield tables. The model strongly favours Asymptotic Giant Branch yields from NuGrid and core-collapse SN yields used in the IllustrisTNG simulation, achieving near-unity cumulative posterior probability. Using the preferred model, we infer a steep high-mass IMF slope and an elevated Supernova Ia normalization, consistent with prior solar neighbourhood studies but now derived from fully amortized Bayesian inference. Our results demonstrate that modern SBI methods can robustly constrain uncertain physics in astrophysical simulators and enable principled model selection when analysing complex, simulation-based data.

1. Motivation

Cosmological simulations of galaxy formation (e.g. Sawala et al., 2016; Hopkins et al., 2018; Pillepich et al., 2018; Buck, 2020; Buck et al., 2020; Font et al., 2020; Agertz et al., 2021) rely on a small set of galactic parameters to capture key processes in stellar evolution, such as star formation and supernova feedback. Two particularly uncertain inputs are the shape of the initial mass function (IMF), which determines the mass distribution of stars formed from the interstellar medium (ISM), and the rate and delay-time distribution of Type Ia supernovae (SN Ia).

These parameters strongly influence chemical enrichment histories (Romano et al., 2005; Vincenzo et al., 2015; Mollá et al., 2015), yet remain poorly constrained by observations. Moreover, GCE models vary significantly in their choice of nucleosynthetic yields and assumptions about enrichment sources, including asymptotic giant branch (AGB) stars, core-collapse supernovae (cc-SNe), and SN Ia. For instance, high-mass IMF slopes steeper than canonical values have been proposed by various studies (Côté et al., 2016; Weisz et al., 2015; Rybizki & Just, 2015a; Chabrier et al., 2014, Tab. 7). Likewise, SN Ia normalization and delay-time distribution remain active areas of debate (Maoz et al., 2010; 2012; Jiménez et al., 2015; Buck et al., 2021). Different stellar evolution models also yield a broad range of element production rates (e.g. Nomoto et al., 1997; Kobayashi et al., 2006; Portinari et al., 1998; Karakas, 2010; Doherty et al., 2014; Fishlock et al., 2014; Karakas & Lugaro, 2016a).

Here, we introduce COMPASS, a scalable simulation-based inference (SBI; Cranmer et al., 2020) framework to overcome these challenges. Building on the work of Philcox & Rybizki (2019), who used computationally expensive Hamiltonian Monte Carlo (HMC) methods, COMPASS is designed for robust Bayesian model comparison and parameter inference from large, complex stellar abundance datasets.

2. Related Work

Recent work has increasingly applied machine learning to model comparison in astrophysics. Karchev et al. (2023) used a deep learning method for Bayesian model comparison (Elsemüller et al., 2023) to evaluate simulation-based SN Ia light curve models. Zanisi et al. (2021) compared Illustris (Vogelsberger et al., 2014) and IllustrisTNG (Pillepich et al., 2018) simulations to r-band Sloan Digital Sky Survey (SDSS) images (Kollmeier et al., 2019) by using two PixelCNNs (Van Den Oord et al., 2016) to

[1]Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 205, D-69120 Heidelberg, Germany [2]Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia [3]ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), Australia [4]Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Albert-Ueberle-Straße 2, D-69120 Heidelberg, Germany. Correspondence to: Berkay Günes <b.guenes@stud.uni-heidelberg.de>, Tobias Buck <tobias.buck@iwr.uni-heidelberg.de>.

ML4Astro 2025, Vancouver, CA.

generate pixel-wise anomaly scores. Jin et al. (2024) applied GANomaly (Akcay et al., 2019), an anomaly detection model based on Generative Adversarial Networks (GANs; Goodfellow et al., 2020), to evaluate NIHAO galaxy simulations (Wang et al., 2015; Buck et al., 2019; Buck et al., 2020) via anomaly scores relative to SDSS images. Similarly, Zhou et al. (2024) combined out-of-distribution detection with amortized Bayesian model comparison to assess simulated galaxy images against SDSS observations.

3. Data

Observational Data and Elemental Abundances The chemical composition of stars is typical measured in logarithmic abundance ratios of the number fraction of elements [X/Fe] and [Fe/H], defined as: $[\mathrm{X/Y}] = \log_{10}(N_\mathrm{X}/N_\mathrm{Y})_\mathrm{star} - \log_{10}(N_\mathrm{X}/N_\mathrm{Y})_\odot$, where N_X is the number density of element X, and $\odot$ refers to solar abundances from Asplund et al. (2009).

Real data is taken from the high-precision stellar abundance measurements from Nissen et al. (2020), which provide detailed compositions of solar-type stars. This dataset offers a clean, well-characterized sample for evaluating `COMPASS` inference fidelity.

Galactic Chemical Evolution Model and Mock Data We use the `CHEMPYMulti` simulator (Philcox & Rybizki, 2019), based on the original `CHEMPY` GCE model (Rybizki et al., 2017), to generate stellar abundance predictions given parameterized stellar and ISM physics. The model evolves chemical abundances over cosmic time using published yield tables for SN Ia, SN II, and AGB feedback (Philcox et al., 2018). These nucleosynthetic yield tables are theoretical predictions that quantify the mass of each chemical element produced and ejected by a star of a given initial mass and metallicity over its lifetime. Our inference operates over six free parameters, grouped as: (1) **Global parameters** ($\vec{\Lambda}$): the IMF slope α_IMF (Chabrier, 2003) and the SN Ia normalization $\log_{10}(N_\mathrm{Ia})$, assumed constant across stars. (2) **Local parameters** ($\{\vec{\Theta}_i\}$): star-formation efficiency $\log_{10}(\mathrm{SFE})$, SFR peak time $\log_{10}(\mathrm{SFR_{peak}})$, and outflow fraction x_out, each specific to a star's birth environment (Rybizki et al., 2017; Philcox & Rybizki, 2019). (3) **Birth times** ($\{T_i\}$): formation times in Gyr, setting ISM conditions from which stellar abundances are drawn. Because H is used for normalization, only $n_\mathrm{el} = 8$ independent elemental abundances are required for modelling with `Chempy`. To simulate observational uncertainties, we perturb synthetic abundances with 5% Gaussian noise.

GCE Models for comparison We compare 40 different combinations of nucleosynthetic yield tables, spanning a range of theoretical prescriptions for AGB and core-collapse SN yields. These combinations span literature-derived SN Ia, SN II, and AGB yields (e.g. Nomoto et al., 1997; Kobayashi et al., 2006; Portinari et al., 1998; Karakas, 2010; Doherty et al., 2014; Fishlock et al., 2014; Karakas & Lugaro, 2016b). A full list of combinations is provided in Table 3 in the Appendix. The goal is to identify which yield models most accurately reproduce the observed chemical abundance patterns, based on the posterior model probability $P(\mathcal{M}_k \mid \mathbf{x})$.

4. Methods

Our goal is to compare GCE models based on how well they reproduce observed stellar abundance data. Each competing model $\mathcal{M}_i$ corresponds to a different nucleosynthesis yield set implemented in the CHEMPY simulator. We frame this as a Bayesian model comparison problem and use SBI to estimate posterior model probabilities. A key challenge in stellar data is the variable and potentially large number of observations. To address this, we design a machine learning pipeline (Figure 1) capable of handling sets of varying cardinality by combining score-based diffusion models with a transformer architecture (Peebles & Xie, 2023; Gloeckler et al., 2024).

Simulation-Based Inference SBI methods (e.g. Cranmer et al., 2020; Papamakarios et al., 2021; Gloeckler et al., 2024) estimate posteriors without requiring an explicit likelihood. Given a generative model $\mathcal{M}$ with parameters θ and synthetic observations $\vec{X}$ (e.g., simulated stellar abundances), we train a neural network to approximate $p(\theta \mid \vec{X}, \mathcal{M})$. Once trained, this posterior estimator can be applied to real observations $\vec{X}_\mathrm{R}$. Our implementation uses a conditional diffusion model trained on joint pairs $(\theta, \vec{x})$, along with a binary condition mask $\mathcal{M}_C$ indicating which values are observed or latent. This allows us to operate as both a Neural Likelihood Estimator (NLE) and Neural Posterior Estimator (NPE). From the posterior samples, we estimate MAP parameters $\hat{\theta}$ and then generate samples from the likelihood $p(\vec{x} \mid \hat{\theta})$ by inverting the mask $\mathcal{M}_C$. A Gaussian kernel density estimator (KDE) is used to approximate the likelihood at the observed $\vec{x}$, enabling model comparison via KDE-estimated log-likelihoods.

Score-Based Transformer Architecture To model the score function, we follow Gloeckler et al. (2024) and use a transformer-based diffusion model. Transformers overcome limitations of feed-forward networks in handling set-structured or sequential data. We use the adaLN-Zero DiT architecture as proposed in Peebles & Xie (2023), adapted for continuous data. An attention mask prevents latent tokens (drawn from noise) from attending to each other, focusing instead on observed inputs. Inputs are embedded in a high-dimensional space; diffusion timesteps are encoded via

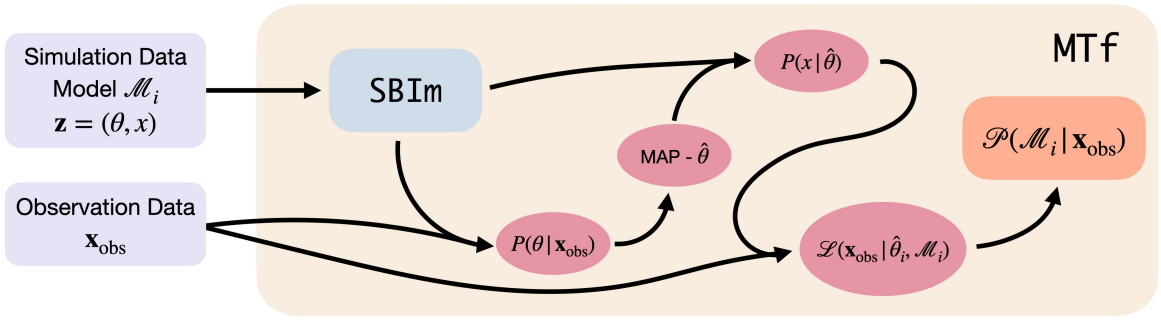

Figure 1. Flow chart of our model comparison workflow. Each of the 40 candidate models ($\mathcal{M}_i$) is used to train a separate Score-Based Inference Model (SBIm). For a given observation $\mathbf{x}_{\text{obs}}$, the corresponding trained `SBIm` infers the posterior $P(\theta|\mathbf{x}_{\text{obs}})$ to find the MAP parameters $\hat{\theta}$. The same model is then used to estimate the maximized likelihood $L(\mathbf{x}_{\text{obs}}|\hat{\theta}_i, \mathcal{M}_i)$. The set of likelihoods from all 40 models is then used to derive the final model posterior probabilities, allowing for a principled comparison of the underlying physical assumptions.

a sinusoidal timestep embedding (Vaswani et al., 2023). The condition mask $\mathcal{M}_C$ guides the network to infer unobserved values based on observed ones. More details are given in Appendix B.

Training Each of the 40 models is trained on 10^6 simulated pairs and validated on 10^5 examples. Parameters (Λ, Θ_i) are sampled from a uniform prior spanning a range of $\pm 5\sigma_{\text{Prior}}$ centred on the prior means μ_{Prior}, following the setup of Rybizki et al. (2017). This wide support mitigates bias toward the prior. To simulate realistic measurement errors, we perturb the synthetic abundances with 5% Gaussian noise. This value is chosen to be representative of the typical $1-\sigma$ statistical uncertainties reported in the Nissen et al. (2020, Table 3) dataset. Further details on hyperparameter tuning, diffusion scheduling, and calibration are provided in Appendix B.

Model Comparison In simulation-based settings where the likelihood is implicit and model evidence is unavailable, we perform model comparison using the Akaike Information Criterion (AIC) (Akaike, 1978; Burnham & Anderson, 2002). AIC allows us to estimate the relative quality of models based on their maximized log-likelihood $\ln \mathcal{L}(\mathbf{x} \mid \hat{\theta}_i, \mathcal{M}_i)$ and is suitable for non-nested models derived from simulators. To compare a set of candidate models $\{\mathcal{M}_i\}$, we first compute AIC differences $\Delta_i(\text{AIC}) = \text{AIC}_i - \text{AIC}_{\min}$, where $\text{AIC}_{\min}$ is the minimum AIC value among the models. The relative likelihood of model $\mathcal{M}_i$ given the data $\mathbf{x}$ is then proportional to $\exp(-\frac{1}{2}\Delta_i(\text{AIC}))$. Normalizing these relative likelihoods yields posterior model probabilities (Wagenmakers & Farrell, 2004; AKAIKE, 1979; Bozdogan, 1987). As detailed in Appendix A, when all compared models have the same number of parameters (as is the case for the yield set comparisons in `COMPASS`), these probabilities simplify to:

$$\mathcal{P}(\mathcal{M}_i \mid \mathbf{x}) = \frac{\mathcal{L}(\mathbf{x} \mid \hat{\theta}_i, \mathcal{M}_i)}{\sum_j \mathcal{L}(\mathbf{x} \mid \hat{\theta}_j, \mathcal{M}_j)}. \tag{1}$$

This softmax-based comparison ranks models by their plausibility given the data. To assess whether a preferred model $\mathcal{M}_j$ is statistically supported, we perform hypothesis testing using the Bayes factor relative to a null model H_0, following Morey et al. (2016). Details are provided in Appendix B.6.

5. Results

The `COMPASS` framework is applied to observational data to address two primary astrophysical objectives: first, performing Bayesian model comparison across competing Galactic Chemical Evolution (GCE) models, and second, inferring key galactic parameters using the model(s) identified as most plausible. The rigorous evaluation of `COMPASS`'s performance, including its consistent recovery of ground-truth models and parameters on mock datasets, is presented in Appendix C.

Model Comparison We apply `COMPASS` to individual stellar abundances from the Nissen et al. (2020) dataset, comparing 40 GCE models based on different combinations of AGB and CC-SN nucleosynthetic yield tables. These posterior model probabilities are then used to rank yield sets and guide downstream parameter inference.

Figure 2 summarizes the model comparison results: Top panel: Violin plots of single-star model posteriors $P(\mathcal{M}_k \mid x_i)$ for selected competitive yield combinations show strong preference for specific models on a per-star basis. Middle panel: All 40 yield sets, sorted by median model posterior across stars, reveal overall model performance distribution. Bottom panel: The cumulative posterior probability

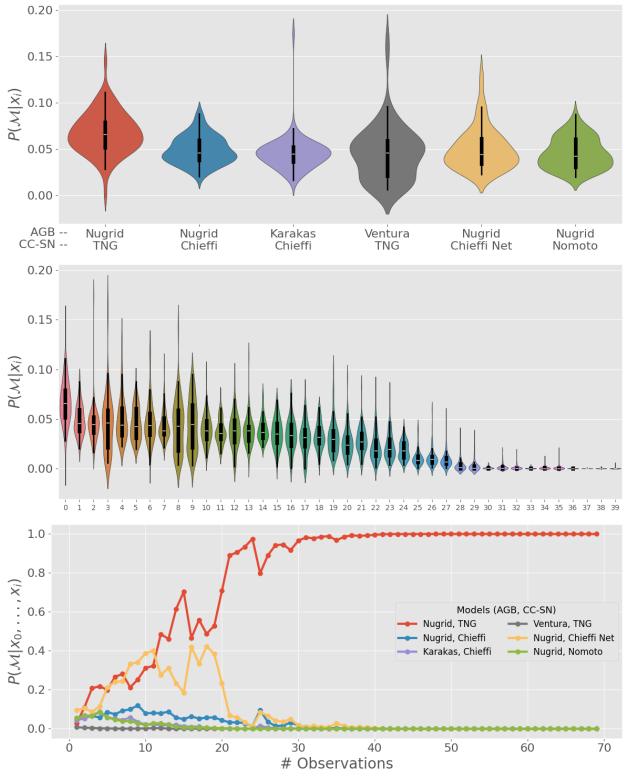

Figure 2. Observational Yield-Set Inference Bayesian comparison of 40 nucleosynthetic yield set combinations using observational data from Nissen et al. (2020). **Top:** Top six single-star model posteriors $P(\mathcal{M}_k|x_i)$. **Middle:** Single-star posterior probabilities for all 40 tested combinations. **Bottom:** Cumulative model posterior probability $P(\mathcal{M}_k|x_0, \ldots, x_i)$.

$P(\mathcal{M}_k \mid x_0, \ldots, x_i)$ increases as more stars are considered, highlighting the growing statistical significance.

A clear preference emerges for models combining Ritter et al. (2018) (NuGrid) AGB yields with the CC-SN yields used in the IllustrisTNG simulation (Pillepich et al., 2017; Kobayashi et al., 2006; Portinari et al., 1997). This combination achieves a relative cumulative posterior probability nearing 100% after incorporating all 69 stars. While some models perform well for individual stars, their posterior probability drops when aggregated over the full dataset. A full ranking is provided in Table 3 (Appendix). These results underscore the strength of COMPASS in distinguishing physically motivated GCE models using real observational data, enabling robust model selection for chemical evolution studies.

Inferring Galactic Parameters Before applying COMPASS to real data, we benchmarked it against both the SBI pipeline from Buck et al. (2025) and Hamiltonian Monte Carlo (HMC) inference used in Philcox & Rybizki (2019). Across tests involving matched and mismatched yield sets—including more complex data from the IllustrisTNG simulation—COMPASS reliably recovered ground-truth parameters (Appendix D).

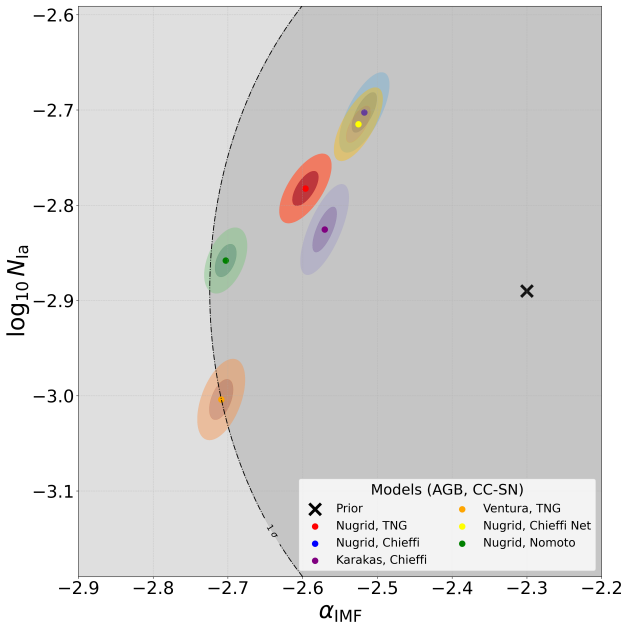

Figure 3. Inferred Galactic Parameters from Observational Data Joint posterior contours ($1\sigma, 2\sigma$) for the IMF high-mass slope $\alpha_{\rm IMF}$ and SN Ia normalization $\log_{10} N_{\rm Ia}$, from Nissen et al. (2020) using COMPASS. Colors denote the top six yield set combinations identified in Fig. 2 and discussed in Tab. 4). The black 'X' and dashed line indicate the CHEMPY prior mean and standard deviation, respectively.

We then used the best-performing yield sets to infer global parameters from the Nissen et al. (2020) sample. Figure 3 shows the joint posterior distributions of the IMF slope $\alpha_{\rm IMF}$ and the SN Ia normalization $\log_{10} N_{\rm Ia}$ for the top six yield combinations. Table 4 in the Appendix summarizes their means and uncertainties. Despite minor shifts depending on the yield set, the top-performing models yield consistent constraints with posterior distributions that are well-separated from prior means, indicating strong data-driven constraints. The best model (NuGrid AGB + TNG CC-SN yields) gives $\alpha_{\rm IMF} = -2.60 \pm 0.02$ and $\log_{10} N_{\rm Ia} = -2.78 \pm 0.02$. These values point to a systematically steeper IMF slope than the canonical Salpeter value of -2.35 or the Chabrier prior mean of -2.3, and a SN Ia normalization significantly elevated relative to the prior mean of -2.89 consistent with earlier studies of the solar neighborhood (e.g., Rybizki & Just, 2015b).

Scientific Impact Our results demonstrate that combining high-fidelity observational data with SBI and diffusion-based inference enables precise discrimination among theoretical models of stellar nucleosynthesis. COMPASS not only selects the most plausible yield prescriptions but

also delivers tight, interpretable posteriors on key galactic parameters—$\alpha_{\rm IMF}$ and $N_{\rm Ia}$—critical for understanding galaxy-scale feedback and enrichment. This framework provides a powerful tool for constraining uncertain physics in cosmological simulations and stellar population synthesis.

6. Conclusion and Summary

We introduced COMPASS, a simulation-based inference framework that performs joint model comparison and parameter estimation. Here, we have employed it in the context of Galactic Chemical Evolution (GCE). Built on a score-based diffusion model and transformer architecture, COMPASS handles high-dimensional, partially observed data while maintaining scalability and statistical rigor.

Our method was first validated on synthetic data, where it reliably recovered ground-truth models and parameters. Applied to real stellar abundance data from Nissen et al. (2020), COMPASS identified a strongly preferred nucleosynthetic yield combination—NuGrid AGB yields and TNG CC-SN yields—and inferred global galactic parameters with high precision. The results favor an IMF slope significantly steeper than canonical models and an elevated SN Ia normalization, both of which have critical implications for galaxy formation and enrichment modeling.

COMPASS represents a novel integration of modern simulation-based inference techniques with physically motivated astrophysical models. Unlike traditional MCMC-based GCE studies, our framework enables efficient, amortized inference and principled Bayesian model comparison across a large space of competing simulators.

Looking ahead, we anticipate that COMPASS can revolutionize the modeling of chemical evolution in galaxies. By enabling rigorous, data-driven selection among yield prescriptions and accurate inference of physical parameters, it offers a scalable pathway to calibrate and test subgrid physics in cosmological simulations. Future extensions to hierarchical inference and time-resolved abundance datasets could further improve our understanding of galactic feedback and chemical enrichment over cosmic time.

Impact Statement

The authors are not aware of any immediate ethical or societal implications of this work. This work purely aims to aid scientific research and proposes a method for SBI in challenging astrophysical settings. While there will certainly be many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agertz, O., Renaud, F., Feltzing, S., Read, J. I., Ryde, N., Andersson, E. P., Rey, M. P., Bensby, T., and Feuillet, D. K. VINTERGATAN - I. The origins of chemically, kinematically, and structurally distinct discs in a simulated Milky Way-mass galaxy. *MNRAS*, 503(4):5826–5845, June 2021. doi: 10.1093/mnras/stab322.

Akaike, H. On newer statistical approaches to parameter estimation and structure determination. *IFAC Proceedings Volumes*, 11(1):1877–1884, 1978. ISSN 1474-6670. doi: https://doi.org/10.1016/S1474-6670(17)66162-7. URL https://www.sciencedirect.com/science/article/pii/S1474667017661627. 7th Triennial World Congress of the IFAC on A Link Between Science and Applications of Automatic Control, Helsinki, Finland, 12-16 June.

AKAIKE, H. A bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika*, 66(2):237–242, 08 1979. ISSN 0006-3444. doi: 10.1093/biomet/66.2.237. URL https://doi.org/10.1093/biomet/66.2.237.

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 622–637. Springer, 2019.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework, 2019. URL https://arxiv.org/abs/1907.10902.

Andrews, B. H., Weinberg, D. H., Schönrich, R., and Johnson, J. A. Inflow, Outflow, Yields, and Stellar Population Mixing in Chemical Evolution Models. *ApJ*, 835(2):224, Feb 2017. doi: 10.3847/1538-4357/835/2/224.

Asplund, M., Grevesse, N., Sauval, A. J., and Scott, P. The Chemical Composition of the Sun. *ARA&A*, 47(1):481–522, Sep 2009. doi: 10.1146/annurev.astro.46.060407.145222.

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., Ginsburg, A., Price-Whelan, A. M., Kerzendorf, W. E., Conley, A., Crighton, N., Barbary, K., Muna, D., Ferguson, H., Grollier, F., Parikh, M. M., Nair, P. H., Unther, H. M., Deil, C., Woillez, J., Conseil, S., Kramer, R., Turner, J. E. H., Singer, L., Fox, R., Weaver, B. A., Zabalza, V., Edwards, Z. I., Azalee Bostroem, K., Burke, D. J., Casey, A. R., Crawford, S. M., Dencheva, N.,

Ely, J., Jenness, T., Labrie, K., Lim, P. L., Pierfederici, F., Pontzen, A., Ptak, A., Refsdal, B., Servillat, M., and Streicher, O. Astropy: A community Python package for astronomy. *aap*, 558:A33, October 2013. doi: 10.1051/0004-6361/201322068.

Austin, S. M., West, C., and Heger, A. Reducing uncertainties in the production of the gamma-emitting nuclei 26al, 44ti, and 60fe in core-collapse supernovae by using effective helium burning rates. *The Astrophysical Journal Letters*, 839(1):L9, April 2017. ISSN 2041-8213. doi: 10.3847/2041-8213/aa68e7. URL http://dx.doi.org/10.3847/2041-8213/aa68e7.

Bozdogan, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):353–356, 1987. doi: 10.1007/BF02294361.

Buck, T. On the origin of the chemical bimodality of disc stars: a tale of merger and migration. *MNRAS*, 491(4): 5435–5446, February 2020. doi: 10.1093/mnras/stz3289.

Buck, T., Macciò, A. V., Dutton, A. A., Obreja, A., and Frings, J. Nihao xv: the environmental impact of the host galaxy on galactic satellite and field dwarf galaxies. *Monthly Notices of the Royal Astronomical Society*, 483 (1):1314–1341, 2019.

Buck, T., Obreja, A., Macciò, A. V., Minchev, I., Dutton, A. A., and Ostriker, J. P. NIHAO-UHD: the properties of MW-like stellar discs in high-resolution cosmological simulations. *MNRAS*, 491(3):3461–3478, January 2020. doi: 10.1093/mnras/stz3241.

Buck, T., Rybizki, J., Buder, S., Obreja, A., Macciò, A. V., Pfrommer, C., Steinmetz, M., and Ness, M. The challenge of simultaneously matching the observed diversity of chemical abundance patterns in cosmological hydrodynamical simulations. *MNRAS*, 508(3):3365–3387, December 2021. doi: 10.1093/mnras/stab2736.

Buck, T., Obreja, A., Ratcliffe, B., Lu, Y., Minchev, I., and Macciò, A. V. The impact of early massive mergers on the chemical evolution of Milky Way-like galaxies: insights from NIHAO-UHD simulations. *MNRAS*, 523 (1):1565–1576, July 2023. doi: 10.1093/mnras/stad1503.

Buck, T., Günes, B., Viterbo, G., Oliver, W. H., and Buder, S. Inferring galactic parameters from chemical abundances with simulation-based inference, 2025. URL https://arxiv.org/abs/2503.02456.

Burnham, K. and Anderson, D. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag, 2002.

Chabrier, G. Galactic Stellar and Substellar Initial Mass Function. *PASP*, 115:763–795, July 2003. doi: 10.1086/376392.

Chabrier, G., Hennebelle, P., and Charlot, S. Variations of the Stellar Initial Mass Function in the Progenitors of Massive Early-type Galaxies and in Extreme Starburst Environments. *ApJ*, 796(2):75, Dec 2014. doi: 10.1088/0004-637X/796/2/75.

Chieffi, A. and Limongi, M. Explosive Yields of Massive Stars from Z = 0 to Z = Z_{solar}. *ApJ*, 608(1):405–410, June 2004. doi: 10.1086/392523.

Clarke, A. J., Debattista, V. P., Nidever, D. L., Loebman, S. R., Simons, R. C., Kassin, S., Du, M., Ness, M., Fisher, D. B., Quinn, T. R., Wadsley, J., Freeman, K. C., and Popescu, C. C. The imprint of clump formation at high redshift - I. A disc α-abundance dichotomy. *MNRAS*, 484 (3):3476–3490, Apr 2019. doi: 10.1093/mnras/stz104.

Côté, B., Ritter, C., O'Shea, B. W., Herwig, F., Pignatari, M., Jones, S., and Fryer, C. L. Uncertainties in Galactic Chemical Evolution Models. *ApJ*, 824(2):82, Jun 2016. doi: 10.3847/0004-637X/824/2/82.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1912789117.

Doherty, C. L., Gil-Pons, P., Lau, H. H. B., Lattanzio, J. C., and Siess, L. Super and massive AGB stars - II. Nucleosynthesis and yields - Z = 0.02, 0.008 and 0.004. *MNRAS*, 437(1):195–214, Jan 2014. doi: 10.1093/mnras/stt1877.

Elsemüller, L., Schnuerch, M., Bürkner, P.-C., and Radev, S. T. A deep learning method for comparing bayesian hierarchical models, 2023.

Fishlock, C. K., Karakas, A. I., Lugaro, M., and Yong, D. Evolution and Nucleosynthesis of Asymptotic Giant Branch Stellar Models of Low Metallicity. *ApJ*, 797(1): 44, Dec 2014. doi: 10.1088/0004-637X/797/1/44.

Font, A. S., McCarthy, I. G., Poole-Mckenzie, R., Stafford, S. G., Brown, S. T., Schaye, J., Crain, R. A., Theuns, T., and Schaller, M. The ARTEMIS simulations: stellar haloes of Milky Way-mass galaxies. *MNRAS*, 498(2): 1765–1785, October 2020. doi: 10.1093/mnras/staa2463.

Gloeckler, M., Deistler, M., Weilbach, C., Wood, F., and Macke, J. H. All-in-one simulation-based inference, 2024. URL https://arxiv.org/abs/2404.09636.

Glover, S. and Dixon, P. Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, 11(5):791–806, 2004. doi: 10.3758/BF03196706. URL https://doi.org/10.3758/BF03196706.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Grand, R. J. J., Bustamante, S., Gómez, F. A., Kawata, D., Marinacci, F., Pakmor, R., Rix, H.-W., Simpson, C. M., Sparre, M., and Springel, V. Origin of chemically distinct discs in the Auriga cosmological simulations. *MNRAS*, 474:3629–3639, March 2018. doi: 10.1093/mnras/stx3025.

Hopkins, P. F., Wetzel, A., Kereš, D., Faucher-Giguère, C.-A., Quataert, E., Boylan-Kolchin, M., Murray, N., Hayward, C. C., Garrison-Kimmel, S., Hummels, C., Feldmann, R., Torrey, P., Ma, X., Anglés-Alcázar, D., Su, K.-Y., Orr, M., Schmitz, D., Escala, I., Sanderson, R., Grudić, M. Y., Hafen, Z., Kim, J.-H., Fitts, A., Bullock, J. S., Wheeler, C., Chan, T. K., Elbert, O. D., and Narayanan, D. FIRE-2 simulations: physics versus numerics in galaxy formation. *MNRAS*, 480(1):800–863, October 2018. doi: 10.1093/mnras/sty1690.

Jiménez, N., Tissera, P. B., and Matteucci, F. Type Ia Supernova Progenitors and Chemical Enrichment in Hydrodynamical Simulations. I. The Single-degenerate Scenario. *ApJ*, 810(2):137, Sep 2015. doi: 10.1088/0004-637X/810/2/137.

Jin, Z., Macciò, A. V., Faucher, N., Pasquato, M., Buck, T., Dixon, K. L., Arora, N., Blank, M., and Vulanovic, P. Quantitatively rating galaxy simulations against real observations with anomaly detection. *Monthly Notices of the Royal Astronomical Society*, 529(4):3536–3549, 2024.

Karakas, A. I. Updated stellar yields from asymptotic giant branch models. *MNRAS*, 403(3):1413–1425, Apr 2010. doi: 10.1111/j.1365-2966.2009.16198.x.

Karakas, A. I. and Lugaro, M. Stellar Yields from Metal-rich Asymptotic Giant Branch Models. *ApJ*, 825(1):26, Jul 2016a. doi: 10.3847/0004-637X/825/1/26.

Karakas, A. I. and Lugaro, M. Stellar Yields from Metal-rich Asymptotic Giant Branch Models. *ApJ*, 825(1):26, July 2016b. doi: 10.3847/0004-637X/825/1/26.

Karchev, K., Trotta, R., and Weniger, C. Simsims: Simulation-based supernova ia model selection with thousands of latent variables. *arXiv preprint arXiv:2311.15650*, 2023.

Kobayashi, C., Umeda, H., Nomoto, K., Tominaga, N., and Ohkubo, T. Galactic chemical evolution: Carbon through zinc. *The Astrophysical Journal*, 653(2):1145–1171, December 2006. ISSN 1538-4357. doi: 10.1086/508914. URL http://dx.doi.org/10.1086/508914.

Kollmeier, J., Anderson, S., Blanc, G., Blanton, M., Covey, K., Crane, J., Drory, N., Frinchaboy, P., Froning, C., Johnson, J., et al. Sdss-v pioneering panoptic spectroscopy. *Bulletin of the American Astronomical Society*, 2019.

Limongi, M. and Chieffi, A. Presupernova evolution and explosive nucleosynthesis of rotating massive stars in the metallicity range $-3 \leq [fe/h] \leq 0$. *The Astrophysical Journal Supplement Series*, 237(1):13, July 2018. ISSN 1538-4365. doi: 10.3847/1538-4365/aacb24. URL http://dx.doi.org/10.3847/1538-4365/aacb24.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, 2022. URL https://arxiv.org/abs/2206.00927.

Mackereth, J. T., Crain, R. A., Schiavon, R. P., Schaye, J., Theuns, T., and Schaller, M. The origin of diverse α-element abundances in galaxy discs. *MNRAS*, 477:5072–5089, July 2018. doi: 10.1093/mnras/sty972.

Maoz, D., Sharon, K., and Gal-Yam, A. The Supernova Delay Time Distribution in Galaxy Clusters and Implications for Type-Ia Progenitors and Metal Enrichment. *ApJ*, 722(2):1879–1894, Oct 2010. doi: 10.1088/0004-637X/722/2/1879.

Maoz, D., Mannucci, F., and Brandt, T. D. The delay-time distribution of Type Ia supernovae from Sloan II. *MNRAS*, 426(4):3282–3294, Nov 2012. doi: 10.1111/j.1365-2966.2012.21871.x.

Minchev, I., Chiappini, C., and Martig, M. Chemodynamical evolution of the Milky Way disk. I. The solar vicinity. *A&A*, 558:A9, October 2013. doi: 10.1051/0004-6361/201220189.

Mollá, M., Cavichia, O., Gavilán, M., and Gibson, B. K. Galactic chemical evolution: stellar yields and the initial mass function. *MNRAS*, 451(4):3693–3708, Aug 2015. doi: 10.1093/mnras/stv1102.

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18, 2016. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2015.11.001. URL https://www.sciencedirect.com/science/article/pii/S0022249615000723.

Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments.

Nissen, P. E., Christensen-Dalsgaard, J., Mosumgaard, J. R., Silva Aguirre, V., Spitoni, E., and Verma, K. High-precision abundances of elements in solar-type stars: Evidence of two distinct sequences in abundance-age relations. *Astronomy & Astrophysics*, 640:A81, August 2020. ISSN 1432-0746. doi: 10.1051/0004-6361/202038300. URL http://dx.doi.org/10.1051/0004-6361/202038300.

Nomoto, K., Iwamoto, K., Nakasato, N., Thielemann, F. K., Brachwitz, F., Tsujimoto, T., Kubo, Y., and Kishimoto, N. Nucleosynthesis in type Ia supernovae. *Nucl. Phys. A*, 621:467–476, Feb 1997. doi: 10.1016/S0375-9474(97)00291-1.

Nomoto, K., Kobayashi, C., and Tominaga, N. Nucleosynthesis in Stars and the Chemical Enrichment of Galaxies. *ARA&A*, 51(1):457–509, Aug 2013. doi: 10.1146/annurev-astro-082812-140956.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference, 2021.

Peebles, W. and Xie, S. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Philcox, O., Rybizki, J., and Gutcke, T. A. On the optimal choice of nucleosynthetic yields, initial mass function, and number of sne ia for chemical evolution modeling. *The Astrophysical Journal*, 861(1):40, June 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aac6e4. URL http://dx.doi.org/10.3847/1538-4357/aac6e4.

Philcox, O. H. E. and Rybizki, J. Inferring galactic parameters from chemical abundances: A multi-star approach. *The Astrophysical Journal*, 887(1):9, December 2019. ISSN 1538-4357. doi: 10.3847/1538-4357/ab5186. URL http://dx.doi.org/10.3847/1538-4357/ab5186.

Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., and Marinacci, F. Simulating galaxy formation with the illustristng model. *Monthly Notices of the Royal Astronomical Society*, 473(3):4077–4106, 10 2017. ISSN 0035-8711. doi: 10.1093/mnras/stx2656. URL https://doi.org/10.1093/mnras/stx2656.

Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., and Marinacci, F. Simulating galaxy formation with the IllustrisTNG model. *MNRAS*, 473:4077–4106, January 2018. doi: 10.1093/mnras/stx2656.

Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., et al. Simulating galaxy formation with the illustristng model. *Monthly Notices of the Royal Astronomical Society*, 473(3):4077–4106, 2018.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., Bartolo, N., Battaner, E., Battye, R., Benabed, K., Benoît, A., Benoit-Lévy, A., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bonaldi, A., Bonavera, L., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Catalano, A., Challinor, A., Chamballu, A., Chary, R. R., Chiang, H. C., Chluba, J., Christensen, P. R., Church, S., Clements, D. L., Colombi, S., Colombo, L. P. L., Combet, C., Coulais, A., Crill, B. P., Curto, A., Cuttaia, F., Danese, L., Davies, R. D., Davis, R. J., de Bernardis, P., de Rosa, A., de Zotti, G., Delabrouille, J., Désert, F. X., Di Valentino, E., Dickinson, C., Diego, J. M., Dolag, K., Dole, H., Donzelli, S., Doré, O., Douspis, M., Ducout, A., Dunkley, J., Dupac, X., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Farhang, M., Fergusson, J., Finelli, F., Forni, O., Frailis, M., Fraisse, A. A., Franceschi, E., Frejsel, A., Galeotta, S., Galli, S., Ganga, K., Gauthier, C., Gerbino, M., Ghosh, T., Giard, M., Giraud-Héraud, Y., Giusarma, E., Gjerløw, E., González-Nuevo, J., Górski, K. M., Gratton, S., Gregorio, A., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Hansen, F. K., Hanson, D., Harrison, D. L., Helou, G., Henrot- Versillé, S., Hernández-Monteagudo, C., Herranz, D., Hildebrandt, S. R., Hivon, E., Hobson, M., Holmes, W. A., Hornstrup, A., Hovest, W., Huang, Z., Huffenberger, K. M., Hurier, G., Jaffe, A. H., Jaffe, T. R., Jones, W. C., Juvela, M., Keihänen, E., Keskitalo, R., Kisner, T. S., Kneissl, R., Knoche, J., Knox, L., Kunz, M., Kurki-Suonio, H., Lagache, G., Lähteenmäki, A., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Leahy, J. P., Leonardi, R., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Linden-Vørnle, M., López-Caniego, M., Lubin, P. M., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marchini, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Masi, S., Matarrese, S., McGehee, P., Meinhold, P. R., Melchiorri, A., Melin, J. B., Mendes, L., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Moneti, A., Montier, L., Morgante, G., Mortlock, D., Moss, A., Munshi, D., Murphy, J. A., Naselsky, P., Nati, F., Natoli, P., Netterfield, C. B., Nørgaard-Nielsen, H. U., Noviello, F., Novikov, D., Novikov, I., Oxborrow, C. A., Paci, F., Pagano, L., Pajot, F., Paladini, R., Paoletti, D., Partridge,

B., Pasian, F., Patanchon, G., Pearson, T. J., Perdereau, O., Perotto, L., Perrotta, F., Pettorino, V., Piacentini, F., Piat, M., Pierpaoli, E., Pietrobon, D., Plaszczynski, S., Pointecouteau, E., Polenta, G., Popa, L., Pratt, G. W., Prézeau, G., Prunet, S., Puget, J. L., Rachen, J. P., Reach, W. T., Rebolo, R., Reinecke, M., Remazeilles, M., Renault, C., Renzi, A., Ristorcelli, I., Rocha, G., Rosset, C., Rossetti, M., Roudier, G., Rouillé d'Orfeuil, B., Rowan-Robinson, M., Rubiño-Martín, J. A., Rusholme, B., Said, N., Salvatelli, V., Salvati, L., Sandri, M., Santos, D., Savelainen, M., Savini, G., Scott, D., Seiffert, M. D., Serra, P., Shellard, E. P. S., Spencer, L. D., Spinelli, M., Stolyarov, V., Stompor, R., Sudiwala, R., Sunyaev, R., Sutton, D., Suur-Uski, A. S., Sygnet, J. F., Tauber, J. A., Terenzi, L., Toffolatti, L., Tomasi, M., Tristram, M., Trombetti, T., Tucci, M., Tuovinen, J., Türler, M., Umana, G., Valenziano, L., Valiviita, J., Van Tent, F., Vielva, P., Villa, F., Wade, L. A., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Wilkinson, A., Yvon, D., Zacchei, A., and Zonca, A. Planck 2015 results. XIII. Cosmological parameters. *A&A*, 594:A13, September 2016. doi: 10.1051/0004-6361/201525830.

Portinari, L., Chiosi, C., and Bressan, A. Galactic chemical enrichment with new metallicity dependent yields, 1997. URL https://arxiv.org/abs/astro-ph/9711337.

Portinari, L., Chiosi, C., and Bressan, A. Galactic chemical enrichment with new metallicity dependent stellar yields. *A&A*, 334:505–539, Jun 1998.

Price-Whelan, A. M., Sip'ocz, B. M., G"unther, H. M., Lim, P. L., Crawford, S. M., Conseil, S., Shupe, D. L., Craig, M. W., Dencheva, N., Ginsburg, A., VanderPlas, J. T., Bradley, L. D., P'erez-Su'arez, D., de Val-Borro, M., Paper Contributors, P., Aldcroft, T. L., Cruz, K. L., Robitaille, T. P., Tollerud, E. J., Coordination Committee, A., Ardelean, C., Babej, T., Bach, Y. P., Bachetti, M., Bakanov, A. V., Bamford, S. P., Barentsen, G., Barmby, P., Baumbach, A., Berry, K. L., Biscani, F., Boquien, M., Bostroem, K. A., Bouma, L. G., Brammer, G. B., Bray, E. M., Breytenbach, H., Buddelmeijer, H., Burke, D. J., Calderone, G., Cano Rodr'iguez, J. L., Cara, M., Cardoso, J. V. M., Cheedella, S., Copin, Y., Corrales, L., Crichton, D., D extquoterightAvella, D., Deil, C., Depagne, E., Dietrich, J. P., Donath, A., Droettboom, M., Earl, N., Erben, T., Fabbro, S., Ferreira, L. A., Finethy, T., Fox, R. T., Garrison, L. H., Gibbons, S. L. J., Goldstein, D. A., Gommers, R., Greco, J. P., Greenfield, P., Groener, A. M., Grollier, F., Hagen, A., Hirst, P., Homeier, D., Horton, A. J., Hosseinzadeh, G., Hu, L., Hunkeler, J. S., Ivezi'c, Z., Jain, A., Jenness, T., Kanarek, G., Kendrew, S., Kern, N. S., Kerzendorf, W. E., Khvalko, A., King, J., Kirkby, D., Kulkarni, A. M., Kumar, A., Lee, A., Lenz, D., Littlefair, S. P., Ma, Z., Macleod, D. M., Mastropietro, M., McCully, C., Montagnac, S., Morris, B. M., Mueller, M., Mumford, S. J., Muna, D., Murphy, N. A., Nelson, S., Nguyen, G. H., Ninan, J. P., N"othe, M., Ogaz, S., Oh, S., Parejko, J. K., Parley, N., Pascual, S., Patil, R., Patil, A. A., Plunkett, A. L., Prochaska, J. X., Rastogi, T., Reddy Janga, V., Sabater, J., Sakurikar, P., Seifert, M., Sherbert, L. E., Sherwood-Taylor, H., Shih, A. Y., Sick, J., Silbiger, M. T., Singanamalla, S., Singer, L. P., Sladen, P. H., Sooley, K. A., Sornarajah, S., Streicher, O., Teuben, P., Thomas, S. W., Tremblay, G. R., Turner, J. E. H., Terr'on, V., van Kerkwijk, M. H., de la Vega, A., Watkins, L. L., Weaver, B. A., Whitmore, J. B., Woillez, J., Zabalza, V., and Contributors, A. The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *aj*, 156:123, September 2018. doi: 10.3847/1538-3881/aabc4f.

Ritter, C., Herwig, F., Jones, S., Pignatari, M., Fryer, C., and Hirschi, R. Nugrid stellar data set – ii. stellar yields from h to bi for stellar models with mzams = 1–25 $m_{\odot}$ and z = 0.0001–0.02. *Monthly Notices of the Royal Astronomical Society*, 480(1):538–571, June 2018. ISSN 1365-2966. doi: 10.1093/mnras/sty1729. URL http://dx.doi.org/10.1093/mnras/sty1729.

Romano, D., Chiappini, C., Matteucci, F., and Tosi, M. Quantifying the uncertainties of chemical evolution studies. I. Stellar lifetimes and initial mass function. *A&A*, 430:491–505, Feb 2005. doi: 10.1051/0004-6361:20048222.

Rybizki, J. and Just, A. Towards a fully consistent Milky Way disc model - III. Constraining the initial mass function. *MNRAS*, 447(4):3880–3891, Mar 2015a. doi: 10.1093/mnras/stu2734.

Rybizki, J. and Just, A. Towards a fully consistent Milky Way disc model - III. Constraining the initial mass function. *MNRAS*, 447(4):3880–3891, March 2015b. doi: 10.1093/mnras/stu2734.

Rybizki, J., Just, A., and Rix, H.-W. Chempy: A flexible chemical evolution model for abundance fitting: Do the sun's abundances alone constrain chemical evolution models? *Astronomy & Astrophysics*, 605:A59, September 2017. ISSN 1432-0746. doi: 10.1051/0004-6361/201730522. URL http://dx.doi.org/10.1051/0004-6361/201730522.

Sawala, T., Frenk, C. S., Fattahi, A., Navarro, J. F., Bower, R. G., Crain, R. A., Dalla Vecchia, C., Furlong, M., Helly, J. C., Jenkins, A., Oman, K. A., Schaller, M., Schaye, J., Theuns, T., Trayford, J., and White, S. D. M. The APOSTLE simulations: solutions to the Local Group's cosmic puzzles. *MNRAS*, 457:1931–1943, April 2016. doi: 10.1093/mnras/stw145.

Schönrich, R. and Binney, J. Chemical evolution with radial mixing. *MNRAS*, 396:203–222, June 2009. doi: 10.1111/j.1365-2966.2009.14750.x.

Seitenzahl, I. R., Ciaraldi-Schoolmann, F., Röpke, F. K., Fink, M., Hillebrandt, W., Kromer, M., Pakmor, R., Ruiter, A. J., Sim, S. A., and Taubenberger, S. Three-dimensional delayed-detonation models with nucleosynthesis for type ia supernovae. *Monthly Notices of the Royal Astronomical Society*, 429(2):1156–1172, 12 2012. ISSN 0035-8711. doi: 10.1093/mnras/sts402. URL https://doi.org/10.1093/mnras/sts402.

Thielemann, F. K., Argast, D., Brachwitz, F., Hix, W. R., Höflich, P., Liebendörfer, M., Martinez-Pinedo, G., Mezzacappa, A., Panov, I., and Rauscher, T. Nuclear cross sections, nuclear structure and stellar nucleosynthesis. *Nucl. Phys. A*, 718:139–146, May 2003. doi: 10.1016/S0375-9474(03)00704-8.

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Vincenzo, F., Matteucci, F., Recchi, S., Calura, F., McWilliam, A., and Lanfranchi, G. A. The IGIMF and other IMFs in dSphs: the case of Sagittarius. *MNRAS*, 449 (2):1327–1339, May 2015. doi: 10.1093/mnras/stv357.

Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Nelson, D., and Hernquist, L. Introducing the illustris project: simulating the coevolution of dark and visible matter in the universe. *Monthly Notices of the Royal Astronomical Society*, 444(2):1518–1547, 2014.

Wagenmakers, E.-J. and Farrell, S. Aic model selection using akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, 2004. doi: 10.3758/BF03206482. URL https://doi.org/10.3758/BF03206482.

Wang, L., Dutton, A. A., Stinson, G. S., Macciò, A. V., Penzo, C., Kang, X., Keller, B. W., and Wadsley, J. Nihao project–i. reproducing the inefficiency of galaxy formation across cosmic time with a large sample of cosmological hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 454(1):83–94, 2015.

Weisz, D. R., Johnson, L. C., Foreman-Mackey, D., Dolphin, A. E., Beerman, L. C., Williams, B. F., Dalcanton, J. J., Rix, H.-W., Hogg, D. W., Fouesneau, M., Johnson, B. D., Bell, E. F., Boyer, M. L., Gouliermis, D., Guhathakurta, P., Kalirai, J. S., Lewis, A. R., Seth, A. C., and Skillman, E. D. The High-mass Stellar Initial Mass Function in M31 Clusters. *ApJ*, 806(2):198, Jun 2015. doi: 10.1088/0004-637X/806/2/198.

Zanisi, L., Huertas-Company, M., Lanusse, F., Bottrell, C., Pillepich, A., Nelson, D., Rodriguez-Gomez, V., Shankar, F., Hernquist, L., Dekel, A., et al. A deep learning approach to test the small-scale galaxy morphology and its relationship with star formation activity in hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 501(3):4359–4382, 2021.

Zhou, L., Radev, S. T., Oliver, W. H., Obreja, A., Jin, Z., and Buck, T. Evaluating Sparse Galaxy Simulations via Out-of-Distribution Detection and Amortized Bayesian Model Comparison. *arXiv e-prints*, art. arXiv:2410.10606, October 2024. doi: 10.48550/arXiv.2410.10606.

A. Derivation of Model Posterior

Model comparison typically relies on model evidence. However, in simulator-based settings where the likelihood function is often intractable, calculating model evidence directly is not feasible. An alternative approach for model comparison, particularly suitable for non-nested models, is to use the Akaike Information Criterion (AIC) (Akaike, 1978). AIC estimates the prediction error and thereby the relative quality of statistical models for a given set of data.

Given the maximized log-likelihood for model $\mathcal{M}_i$, denoted as $\ln \mathcal{L}(\mathbf{x}|\hat{\theta}_i, \mathcal{M}_i)$ (where $\hat{\theta}_i$ are the parameter values that maximize the likelihood for model $\mathcal{M}_i$), the AIC is defined as:

$$\text{AIC}_i = -2 \ln \mathcal{L}(\mathbf{x}|\hat{\theta}_i, \mathcal{M}_i) + 2k_i \tag{2}$$

where k_i is the number of estimable parameters in model $\mathcal{M}_i$. The model with the lowest AIC is generally preferred.

To compare a set of N models, we first calculate the AIC difference for each model $\mathcal{M}_i$ relative to the model with the minimum AIC in the set ($\text{AIC}_{\min}$): (Wagenmakers & Farrell, 2004; Akaike, 1978; AKAIKE, 1979; Bozdogan, 1987):

$$\Delta_i(\text{AIC}) = \text{AIC}_i - \text{AIC}_{\min} \tag{3}$$

The likelihood of model $\mathcal{M}_i$ being the best model (in the Kullback-Leibler information sense), given the data $\mathbf{x}$, can be estimated relative to the other models using these AIC differences. The relative likelihood of model $\mathcal{M}_i$, sometimes called an "Akaike weight", is given by (Wagenmakers & Farrell, 2004; Burnham & Anderson, 2002):

$$\mathcal{L}_{\text{rel}}(\mathcal{M}_i|\mathbf{x}) \propto \exp\left(-\frac{1}{2}\Delta_i(\text{AIC})\right) \tag{4}$$

To obtain posterior probabilities for each model, $\mathcal{P}(\mathcal{M}_i|\mathbf{x})$, these relative likelihoods are normalized by summing over all models in the candidate set:

$$\mathcal{P}(\mathcal{M}_i|\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}\Delta_i(\text{AIC})\right)}{\sum_{j=1}^{N} \exp\left(-\frac{1}{2}\Delta_j(\text{AIC})\right)} \tag{5}$$

Substituting Eq. (3) into Eq. (5):

$$\mathcal{P}(\mathcal{M}_i|\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\text{AIC}_i - \text{AIC}_{\min})\right)}{\sum_{j=1}^{N} \exp\left(-\frac{1}{2}(\text{AIC}_j - \text{AIC}_{\min})\right)} \tag{6}$$

$$= \frac{\exp\left(-\frac{1}{2}\text{AIC}_i\right)}{\sum_{j=1}^{N} \exp\left(-\frac{1}{2}\text{AIC}_j\right)} \tag{7}$$

Now, substituting the definition of AIC_i from Eq. (2) and under consideration that all models in `COMPASS` under comparison have the same number of parameters, i.e., $k_i = k_j = k$ for all i, j, then the additional parameter term is common to the numerator and all terms in the sum in the denominator, and thus cancels out. In this specific scenario, the posterior model probability simplifies to a direct ratio of the maximized likelihoods of the data given each model:

$$\mathcal{P}(\mathcal{M}_i|\mathbf{x}) = \frac{\mathcal{L}(\mathbf{x}|\hat{\theta}_i, \mathcal{M}_i)}{\sum_{j=1}^{N} \mathcal{L}(\mathbf{x}|\hat{\theta}_j, \mathcal{M}_j)} \quad (\text{if } k_i = k_j \text{ for all } i, j) \tag{8}$$

The model with the highest posterior probability $\mathcal{P}(\mathcal{M}_i|\mathbf{x})$ is then considered the best-supported model by the data, according to this criterion.

B. Network Architecture & Calibration

B.1. Conditional Transformer Architecture

`COMPASS` uses a time-dependent transformer, `ConditionTransformer`, to approximate the conditional score function $s_\phi(\mathbf{z}_t, \mathcal{M}_C, t) \approx \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)$, following advances in diffusion-based generative modeling (Peebles & Xie, 2023; Gloeckler et al., 2024).

Built on DiT (Peebles & Xie, 2023), it applies adaLN-Zero and timestep embeddings, and takes as input $\mathbf{z} = (\theta, \mathbf{x}) \in \mathbb{R}^{D_\theta + D_x}$. A binary mask $\mathcal{M}_C \in \{0, 1\}^D$ specifies which dimensions are observed (1) or latent (0). Custom attention masks restrict each latent token to attend only to observed tokens, preventing leakage between latents. This supports two inference modes:

- **NPE**: Infer θ given $\mathbf{x}$ by masking $\mathcal{M}_C = (0, 1)$
- **NLE**: Infer $\mathbf{x}$ given θ by masking $\mathcal{M}_C = (1, 0)$

B.2. Hyperparameter Tuning

Key hyperparameters – transformer depth, hidden size, MLP ratio, and attention heads –were optimized using `OPTUNA` (Akiba et al., 2019) across 1000 trials to balance predictive accuracy $-\log P(\theta|x)$ and posterior coverage $\Delta_{\max}$TARP.

Our final architecture parameters are a batch size of 125, a sigma of 2.5, a depth of 5, 1 head and a hidden size of 65 with an MLP ratio of 3.

B.3. Diffusion Time

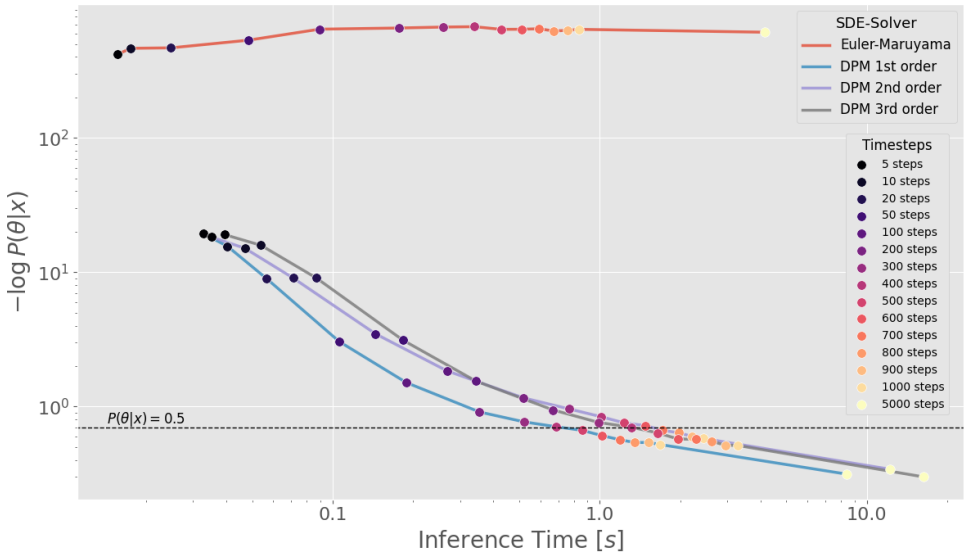

Figure 4. Accuracy of SDE-Solvers Comparison of SDE solver performance in terms of predictive accuracy ($-\log P(\theta|x)$) versus inference time per sample. Euler-Maruyama and DPM-Solvers (1st, 2nd, and 3rd order) were tested with varying numbers of diffusion steps (indicated by colored points, ranging from 5 to 5000 steps). Accuracy is averaged over 1000 mock observations. The dashed horizontal line at $-\log P(\theta|x) \approx 0.693$ represents a posterior probability of $P(\theta|x) = 0.5$ for the true parameter. Lower $-\log P(\theta|x)$ values indicate higher accuracy. The DPM-Solver (1st order) with 500 steps offers a good balance of accuracy and computational efficiency.

Inference quality depends on reverse SDE discretization. We evaluated Euler-Maruyama and 1st–3rd order DPM-Solvers (Lu et al., 2022) over 15 timestep schedules, measuring $-\log P(\theta|x)$ vs. runtime on 1000 test samples (see Fig. 4). The 1st-order DPM-Solver offers the best accuracy-efficiency trade-off, achieving reliable posteriors ($-\log P(\theta|x) < 0.693$) with 500 steps and 1s/sample inference time on 8× RTX 2080 Ti GPUs. This configuration is adopted as default.

B.4. Sampling with COMPASS

Once trained, s_ϕ is used to generate conditional samples from $p_0(\mathbf{z}_{\text{latent}}|\mathbf{z}_{\text{observed}})$ by solving the reverse SDE. COMPASS employs DPM-Solver (Lu et al., 2022) to integrate the deterministic term of the VESDE:

$$\frac{d\mathbf{z}}{dt} = -\sigma^{2t} s_\phi(\mathbf{z}, \mathcal{M}_C, t) \tag{9}$$

The first-order update for step $t \to t'$ is:

$$\mathbf{z}_{t'} = \mathbf{z}_t - (1 - \mathcal{M}_C)\, \sigma_t\, s_\phi(\mathbf{z}_t, \mathcal{M}_C, t)\, dt \tag{10}$$

Only latent dimensions are updated. Higher-order solvers improve this with intermediate evaluations. To enhance fidelity, COMPASS adds optional Langevin corrector steps:

$$\mathbf{z} \leftarrow \mathbf{z} + \delta_L\, \sigma_{t'}^2 s_\phi(\mathbf{z}, \mathcal{M}_C, t') + \sqrt{2\delta_L \sigma_{t'}^2} \cdot \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(0, \mathbf{I}) \tag{11}$$

With δ_L as the Langevin SNR, this stochastic refinement improves sample quality and avoids mode collapse. Correctors are applied every 5 steps (5 iterations per trigger, by default). This predictor-corrector scheme combines the efficiency of deterministic solvers with stochastic refinement, enabling accurate, robust inference.

B.5. Posterior Calibration

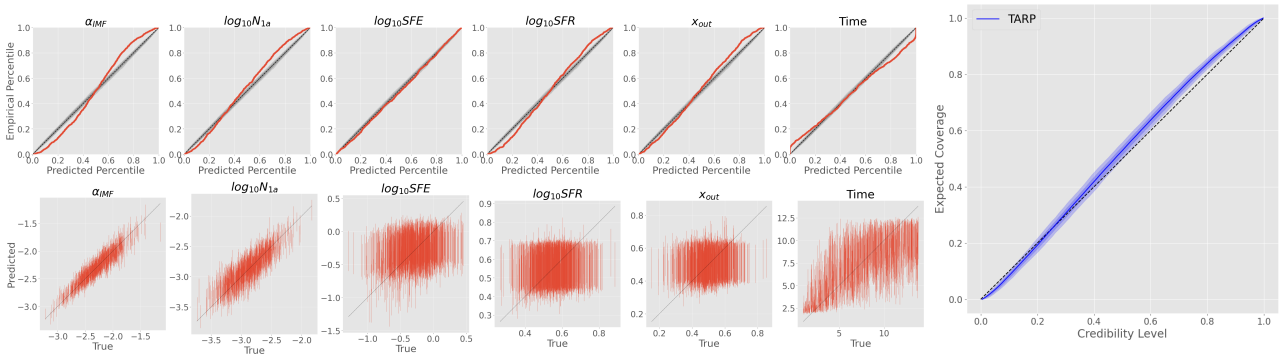

Figure 5. Posterior calibration diagnostics showing the TARP plots at the top and the true vs. predicted parameter plots for each of the six parameters at the bottom and the TARP over all parameters on the right

To ensure reliable uncertainty quantification, we evaluate posterior calibration using the TARP diagnostic and true-vs-predicted plots. Figure 5 presents calibration results for the six `CHEMPY` parameters. TARP plots (top row) show credibility intervals match empirical coverage well—curves align with the diagonal, confirming well-calibrated uncertainty estimates. Posterior mean vs. true parameter plots (bottom row) show strong agreement for global parameters (α_{IMF}, $\log_{10}(N_{Ia})$), which benefit from strong data constraints. Local parameters ($\log_{10}(\text{SFE})$, $\log_{10}(\text{SFR}_{\text{peak}})$, x_{out}, and T) show greater spread, reflecting increased uncertainty and weaker constraints—expected given their spatial and temporal variability.

On the right of figure 5 is the aggregated TARP across all parameters and test samples. The curve closely tracks $y = x$, confirming that the overall posterior coverage is statistically sound. For example, 90% intervals contain the true values approximately 90% of the time.

In summary, the selected architecture and solver yield accurate, calibrated posteriors—especially for global parameters—supporting robust Bayesian inference and model comparison. The broader posteriors for local parameters reflect inherent data limitations.

B.6. Significance test

To assess whether the data supports a candidate model $\mathcal{M}_j$ over a simpler baseline H_0, we use a likelihood-ratio test comparing their marginal likelihoods:

$$K = \frac{\mathcal{L}_{\mathcal{M}_j}(\mathbf{x})}{\mathcal{L}_{H_0}(\mathbf{x})} \quad \text{or} \quad K = \log \mathcal{L}_{\mathcal{M}_j}(\mathbf{x}) - \log \mathcal{L}_{H_0}(\mathbf{x}) \tag{12}$$

Here, K is the *Bayes Factor* (Morey et al., 2016), quantifying how much more likely the observed data $\mathbf{x}$ is under model $\mathcal{M}_j$ than under H_0. $K > 1$ supports $\mathcal{M}_j$ while $K < 1$ supports H_0. This Bayesian model comparison allows principled hypothesis testing beyond point estimates, reflecting both model fit and complexity (Glover & Dixon, 2004).

C. Testing on CHEMPY Mock Observational Data

To evaluate the model comparison capabilities of the COMPASS framework, an initial test was conducted using controlled mock observational data. Three distinct combinations of nucleosynthetic yield sets were selected for this experiment, with the SN Ia and AGB yields fixed to Seitenzahl et al. (2012) and Karakas & Lugaro (2016b), respectively. Only the core-collapse supernova (CC-SN) yields were varied among three widely used sets: Chieffi & Limongi (2004) (Chieffi), Limongi & Chieffi (2018) (Limongi), and Austin et al. (2017) (West), following the rationale in Buck et al. (2021), who highlighted the significant variation in predicted yields across these models.

Bayesian model comparison involves estimating the model posterior probability $P(\mathcal{M}|x)$, which typically requires full posterior inference of the parameters θ under each candidate model $\mathcal{M}$, followed by evaluation of the likelihood $P(x|\theta, \mathcal{M})$. This process can be computationally expensive if performed at high precision for every model. However, when the primary goal is model selection rather than precise parameter estimation, a high-fidelity posterior $P(\theta|x, \mathcal{M})$ may not be necessary. To reduce inference time while preserving sufficient accuracy for model discrimination, the number of diffusion timesteps was reduced. Based on the results from Figure 4, using 50 diffusion steps lowers the average inference time per sample to approximately 0.1 seconds, while still maintaining an acceptable level of accuracy.

Figure 6 presents the resulting distributions of relative model posterior probabilities $P(\mathcal{M}_k|x_i)$ for individual mock observations x_i, where $\mathcal{M}_k$ represents one of the three CC-SN yield models. Each of the three yield sets (Chieffi, Limongi, West) was used to generate a separate mock dataset, and the COMPASS framework was tasked with identifying the correct underlying model from the candidate set, using only 50 diffusion timesteps. The top panel of Figure 6 shows inference results for data generated using the Chieffi model; the middle panel corresponds to Limongi; and the bottom panel to West. In each case, COMPASS correctly identifies the true generative model with high posterior probability for the majority of individual observations, even at reduced computational cost.

While individual posteriors provide strong evidence for model discrimination, combining results across multiple observations further increases the statistical confidence. Figure 7 illustrates this cumulative effect, showing the aggregated posterior probability as a function of the number of observations, using the same test data from the middle panel (Limongi) in Figure 6. The cumulative curve demonstrates that, after just 10 independent observations, the posterior probability assigned to the true model (Limongi) reaches 100%, indicating decisive support.

These results confirm that COMPASS can perform efficient and accurate Bayesian model selection, even with significantly reduced diffusion timesteps. This makes the framework highly suitable for large-scale testing and comparative inference across astrophysical models without incurring prohibitive computational costs.

D. Testing simulation-based Inference with COMPASS

To rigorously assess the parameter inference capabilities of the COMPASS framework using Simulation-Based Inference (SBI), a series of tests were conducted with mock observational data generated by the one-zone Galactic Chemical Evolution (GCE) simulator CHEMPY. The primary objective is to benchmark COMPASS against the SBI methodology from Buck et al. (2025), which employs a Neural Posterior Estimator (NPE), and also against traditional Hamiltonian Monte Carlo (HMC) inference pipeline developed by (Philcox & Rybizki, 2019), which serves as the direct methodological predecessor to this work. This allows for evaluating whether COMPASS can reliably recover known ground-truth parameters under various inference scenarios.

All inference models discussed in this section were trained on the IllustrisTNG nucleosynthetic yield tables implemented in CHEMPY (see Table 1).

TNG Yield Sets For initial validation, mock data were generated using the same TNG yield tables as used in the training phase (see Table 1). The global parameters were fixed at $\alpha_{\rm IMF} = -2.3$ and $\log_{10} N_{\rm Ia} = -2.89$. Figure 8 compares inference results from COMPASS, the SBI pipeline from Buck et al. (2025), and the HMC method of Philcox & Rybizki (2019).

As seen in Figure 8, both COMPASS and SBI accurately recover the ground-truth parameters, with precision improving as more stars are included in the sample. Thanks to their amortized inference nature, both SBI and COMPASS can scale to larger sample sizes with minimal computational overhead – an important feature given the high sample variance when using small stellar datasets. COMPASS demonstrates inference accuracy comparable to the previous SBI implementation, with

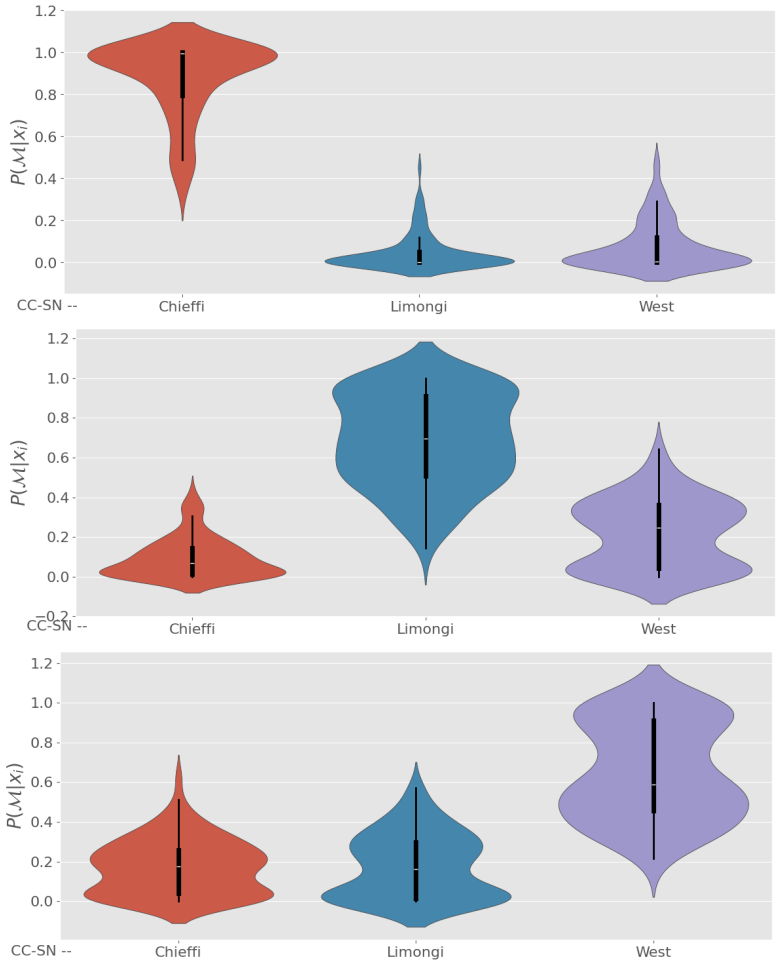

Figure 6. Violin plots showing the distribution of single-observation posterior probabilities $P(\mathcal{M}_k|x_i)$ for three competing CC-SN yield models (Chieffi, Limongi and West), when tested on mock data. Each panel corresponds to a different ground-truth model used to generate the mock observations: **(Top):** Data generated with Chieffi & Limongi (2004) yields. **(Middle):** Data generated with Limongi & Chieffi (2018) yields. **(Bottom):** Data generated with Austin et al. (2017) yields. The framework correctly identifies the true generating model with high probability in most individual observations, even with only 50 diffusion timesteps.

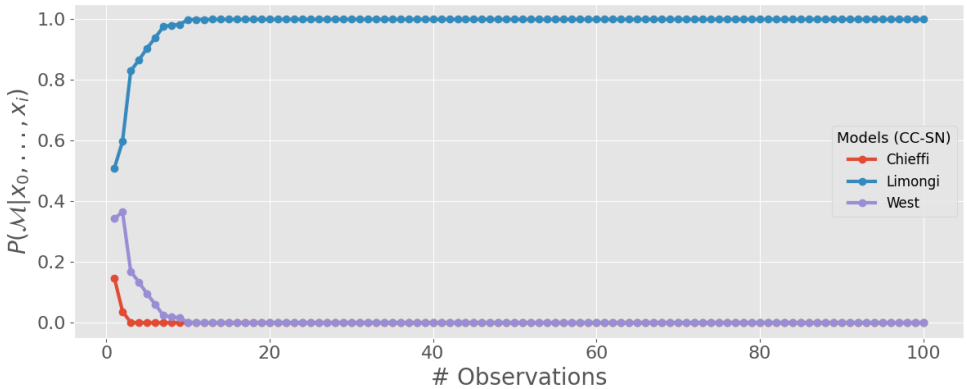

Figure 7. Cumulative relative model posterior probability $P(\mathcal{M}_k|x_0,\ldots,x_i)$ as a function of the number of combined mock observations i. This example illustrates the case where the mock data was generated using the Limongi & Chieffi (2018) CC-SN yields (corresponding to the middle panel of Fig. 6). The probability rapidly converges to 100% for the correct model (Limongi) after only 10 observations, demonstrating the increased confidence gained by combining evidence from multiple data points.

deviations of less than 0.5% for both $\alpha_{\rm IMF}$ and $\log_{10} N_{\rm Ia}$, and the ground truth consistently lying within the 1σ credible interval. This highlights COMPASS's utility not only for model comparison, but also for precise parameter inference.

Type	Yield Table
SN Ia	Nomoto et al. (1997)
CC-SN	Kobayashi et al. (2006); Portinari et al. (1997)
AGB	Karakas (2010); Doherty et al. (2014) Fishlock et al. (2014)

Table 1. TNG Yield Tables

Type	Yield Table
SN Ia	Thielemann et al. (2003)
CC-SN	Nomoto et al. (2013)
AGB	Karakas & Lugaro (2016b)

Table 2. Alternative Yield Tables

Alternative Yield-Sets Since no tabulated nucleosynthetic yield set perfectly reflects nature, and it is often unclear which yield combination is most realistic, this section investigates the sensitivity of inference results to yield set misspecification. To that end, mock data were generated using an alternative yield set (Table 2) while still fixing the global parameters at $\alpha_{\rm IMF} = -2.3$ and $\log_{10} N_{\rm Ia} = -2.89$. The local ISM parameters (Θ_i, T) were drawn from their priors. These alternative yields were chosen such that each nucleosynthetic channel differs by approximately $\mathcal{O}(10\%)$ in predicted abundance yields, providing a meaningful test of model misspecification.

Figure 9 presents inference results under these alternative yields. All three inference methods show degradation in accuracy due to the incorrect model assumptions, but COMPASS maintains closer alignment with the ground-truth values. These results reinforce the importance of correct yield-set selection for reliable parameter inference.

IllustrisTNG Simulation While CHEMPY provides a computationally efficient one-zone GCE framework, it simplifies several key physical processes in the interstellar medium (ISM), such as gas mixing, starbursts, and environmental coupling. To evaluate whether these simplifications bias parameter inference, a more physically realistic dataset was constructed from a full hydrodynamical simulation. Specifically, a Milky Way-like galaxy was selected from the $z = 0$ snapshot of the high-resolution TNG100-1 simulation. Subhalo index 5223071—with a halo mass close to $10^{12}\, M_\odot$—was chosen to represent a Milky Way analog. From this system, 1,000 stellar particles were randomly selected from a total of $\sim$ 40,000 available. Each stellar particle represents a population of stars and carries mass-weighted elemental mass fractions $\{d_i^j\}$ and a formation time given by the cosmological scale factor a_i.

The elemental abundances were converted to [X/Fe] using the solar reference values from Asplund et al. (2009), consistent with CHEMPY. Formation times T_i were derived from a_i using the astropy cosmology package (Astropy Collaboration et al., 2013; Price-Whelan et al., 2018), assuming a flat ΛCDM model with parameters from Planck Collaboration et al. (2016), as adopted in the TNG simulations. To ensure compatibility with the neural network training regime, particles with $T_i \notin [2, 12.8]$ Gyr were excluded, removing roughly 5% of the dataset. The final dataset mirrors the structure of the CHEMPY mock data, with observational uncertainties included in the same way.

This particular TNG galaxy was chosen by Philcox & Rybizki (2019) to contain a clear bimodal distribution in α-abundances—featuring both high- and low-α sequences—analogous to the Milky Way. The origin of this chemical bimodality is still debated and has been linked to various formation scenarios, including gas-rich mergers, starbursts (e.g. Grand et al., 2018; Mackereth et al., 2018; Clarke et al., 2019; Buck, 2020; Buck et al., 2023), and radial migration with selection effects (e.g. Schönrich & Binney, 2009; Minchev et al., 2013; Andrews et al., 2017).

Inference results using this TNG galaxy data are shown in Figure 10. The results confirm that increasing the number of stellar particles improves inference accuracy. Despite the mismatch between the complex TNG chemical enrichment model and the simpler training model (CHEMPY), both SBI and COMPASS successfully recover the underlying parameters. The posterior for $\log_{10}(N_{\rm Ia})$ is nearly unbiased, while $\alpha_{\rm IMF}$ shows a slight overestimation across all methods. Nonetheless, COMPASS and SBI deliver performance on par with HMC, and do so with significantly reduced computational cost, underscoring their practical advantage.

In summary, Figures 8, 9, and 10 collectively demonstrate the robustness of simulation-based inference for galactic parameter estimation from stellar chemical abundances. When interpreting these results, it is crucial to consider the methodological assumptions underlying the uncertainty quantification. Both COMPASS and the benchmark SBI pipeline rely on a Gaussian approximation of individual stellar posteriors to facilitate their analytical combination. This posterior factorization, while

computationally efficient, can lead to underestimated uncertainties and over-confident constraints in the regime of large stellar samples by not fully capturing the tails of the true posterior distributions.

We find that COMPASS consistently produces broader and thus more conservative posterior contours than the benchmark SBI pipeline. This is a direct consequence of our framework's explicit marginalization over observational uncertainties, a step that more realistically propagates measurement error into the final parameter constraints. While the Gaussian approximation remains a shared limitation, the overall accuracy of parameter recovery by COMPASS, combined with its more robust uncertainty estimation, provides strong validation for its use in both model selection and parameter inference.

E. Additional information in inference results

AGB Yields	**Core-Collapse Supernovae (CC-SN) Yields**									
	Chieffi	Nomoto	Portinari	Chieffi Net	Nomoto Net	NuGrid	West	TNG	CL18	Frischknecht
NuGrid	**1**	6	29	4	5	22	19	**0**	30	36
Karakas	**2**	18	31	14	21	26	23	9	35	39
Ventura	10	16	34	13	17	27	20	3	32	38
TNG	7	15	28	11	12	25	24	8	33	37

Table 3. Ranking of Nucleosynthetic Yield Set Combinations Based on Bayesian Model Comparison with Nissen et al. (2020) Data. Lower ranks indicate higher posterior probability and correspond to the numbers and colours in Figure 2.

AGB **CC-SN**	Nugrid TNG	Nugrid Chieffi	Karakas Chieffi	Ventura TNG	Nugrid Chieffi Net	Nugrid Nomoto
$\alpha_{\rm IMF}$	-2.60 ± 0.02	-2.52 ± 0.02	-2.57 ± 0.01	-2.71 ± 0.01	-2.53 ± 0.02	-2.70 ± 0.01
$\log_{10} N_{\rm Ia}$	-2.78 ± 0.02	-2.70 ± 0.02	-2.83 ± 0.02	-3.00 ± 0.02	-2.71 ± 0.02	-2.86 ± 0.02

Table 4. Inferred Galactic Parameters from Observational Data. Mean values and $\pm 1\sigma$ uncertainties for $\alpha_{\rm IMF}$ and $\log_{10}(N_{\rm Ia})$ for the top six yield set combinations.

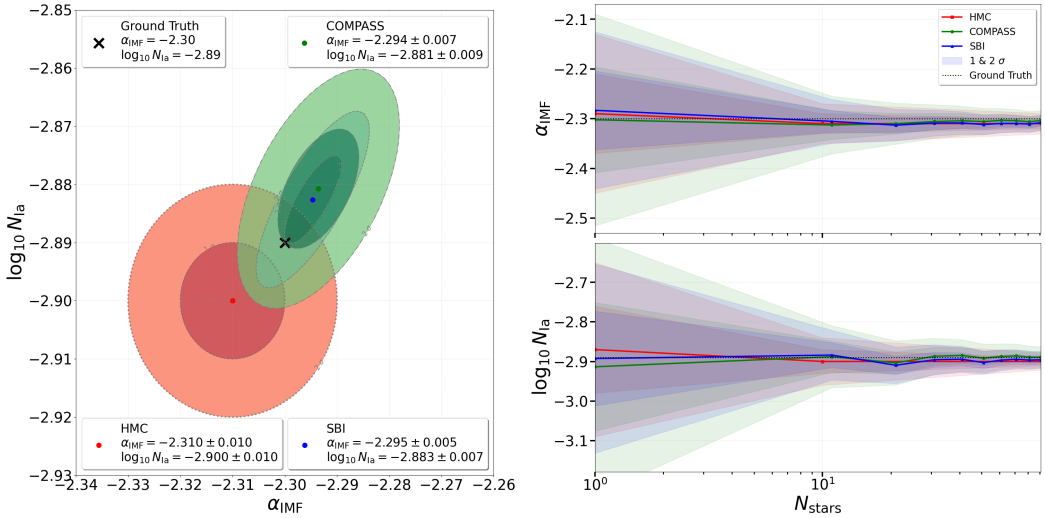

Figure 8. Parameter Inference with Matched TNG Yields. Comparison of inference results for the global galactic parameters Λ using COMPASS (green), SBI (blue) and HMC (red) on mock data generated with TNG yields. **(Left)**: Joint posterior distribution $P(\Lambda|\mathbf{x})$ inferred from 200 stars. **(Right):** Convergence of the inferred parameters as a function of $N_{\rm stars}$.

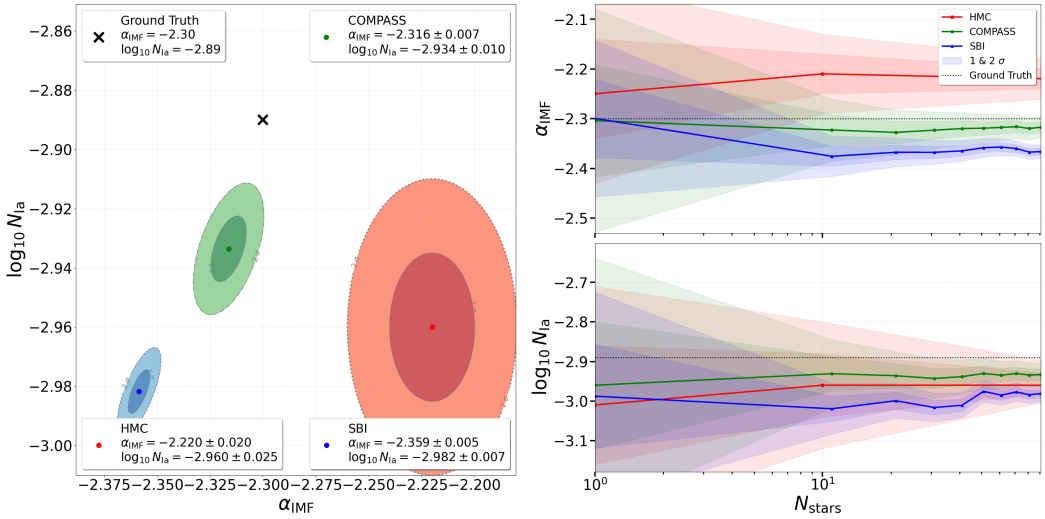

Figure 9. Parameter Inference Alternative Yields Comparison of inference results for the global galactic parameters Λ using COMPASS (green), SBI (blue) and HMC (red) on mock data generated with the alternative yield sets (Tab. 2) **(Left):** Joint posterior distribution $P(\Lambda|\mathbf{x})$ inferred from 200 stars. **(Right):** Convergence of the inferred parameters as a function of $N_{\rm stars}$.

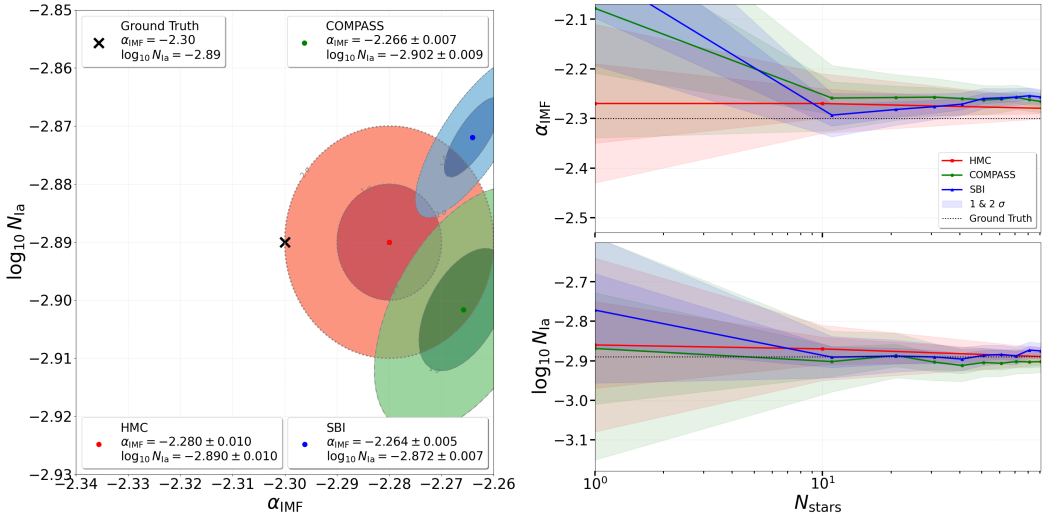

Figure 10. Parameter Inference from IllustrisTNG Simulation Comparison of inference results for the global galactic parameters Λ using COMPASS (green), SBI (blue) and HMC (red) on stellar abundances from the IllustrisTNG simulated galaxy. **(Left):** Joint posterior distribution $P(\Lambda|\mathbf{x})$ inferred from 200 stars. **(Right):** Convergence of the inferred parameters as a function of $N_{\rm stars}$.