Università degli Studi di Padova

Physics and Astronomy Department "Galileo Galilei"

Master Degree in Physics of Data

CASBI: Chemical Abundance Simulation Based Inference

Supervisor:

Prof. Antonino Milone

Co-Supervisor:

Dr. Tobias Buck

Co-Supervisor:

Dr. Giuliano Iorio

Candidate:

Giuseppe Viterbo

Academic Year 2023/2024

"Home is behind, the world ahead

And there are many paths to tread

Through shadow, to the edge of night

Until the stars are all alight."

—J.R.R. Tolkien, *The Lord of the Rings*

Contents

Chapter 1

Abstract

Galaxies evolve hierarchically through merging with lower-mass systems, and the "remnants" of these events are a key indicator of their past assembly history. These "remnants" are star systems that, during their lifetime, were captured by the gravitational potential of a larger galaxy, and ended up in the star halo of their host. Accurately measuring the properties of the accreted galaxies and hence unraveling the Milky Way (MW) formation history is a challenging task. In this work I present CASBI (Chemical Abundance Simulation Based Inferece), a novel inference pipeline for Galactic Archaeology based on Simulation Based Inference methods. CASBI leverages on the fact that there is a well defined mass-metallicity relation for galaxies. CASBI performs inference of key galaxy properties based on multidimensional chemical abundances of stars in the stellar halo. Hence, I recast the task of unraveling the merger history of the MW into a SBI problem to recover the properties (e.g. total stellar mass and infall time) of the halo building blocks using the multidimensional chemical abundances of stars in the stellar halo as observable. I highlight CASBI's potential by inferring posteriors for the stellar masses of completely phase mixed dwarf galaxies solely from the 2d-distributions of stellar abundance in the iron vs. oxygen plane and find accurate and precise inference results.

Chapter 2

Introduction

2.1 Morphology of the Milky Way: disk and halo

This section in inspired by [BT08]. The majority of stars in the Milky Way lies in a flattened, roughly axisymmetric structure known as the Galactic disk. The midplane of this disk is called the Galactic plane. In the direction z perpendicular to the Galactic plane the density of stars follows an exponential,

$$\rho(R, z) = \rho(R, 0)e^{-|z|/z_d(R)}, \tag{2.1}$$

where $z_d(R)$ is the scale height at radius R. A good representation of the vertical structure of the disk is obtained by superimposing two populations described by different scale heights: the thin disk with $z_d(R) \simeq 300$ pc and the thick disk with $z_d(R) \simeq 1$ kpc. Age-wise the stars in the thin disk are younger. The thick disk is characterized by stars with lower metallicity and, at a given metallicity, they have higher abundances of the α elements relative to iron, suggesting that the thick disk formed earlier. The outermost stellar structure is the stellar halo, where 1% of the stars of the Galaxy reside. These stars are old and have low metallicity, suggesting that this structure was one of the first to form in the history of the Galaxy. It is also characterized by no mean rotation and an almost spherical mass distribution that follows a power-law function of the radius ($\rho \propto r^{-3}$). The stellar halo is where the majority of disrupted stellar systems reside, like globular clusters and small satellite galaxies,

which are the main protagonists of this thesis. As described in [Col+18], the classical kinematic cuts to select different components are for thin disk stars a total velocity $V_{\text{tot}} < 50\,\text{km}\,\text{s}^{-1}$, for thick disk $70\,\text{km}\,\text{s}^{-1} < V_{\text{tot}} < 180\,\text{km}\,\text{s}^{-1}$ and for the halo stars $V_{\text{tot}} > 200\,\text{km}\,\text{s}^{-1}$. It is also possible to make cuts based on tangetial velocity: $V_T < 40\,\text{km}\,\text{s}^{-1}$ for the thin disk and $60\,\text{km}\,\text{s}^{-1} < V_T < 150\,\text{km}\,\text{s}^{-1}$ for the thick disk, but still $V_T > 200\,\text{km}\,\text{s}^{-1}$ for the halo.

2.2 Galactic Archaeology

Milky Way like galaxies have an eating habit of destroying and digesting hundreds of lower mass systems. The heaviest of these "meals" possess a star content that adds up to form an extended galactic halo of stars. The long orbital time scale in the halo has allowed these stellar systems to retain information of the progenitor orbits in the phase space, especially in the Energy-Angular momentum plane. Thus, as described in [DB24], the galactic halo is not only a "fossil record" of the assembly history of the MW, but it is also a great tool to study high redshift (now destroyed) dwarf galaxies, opening a cosmological window in our MW neighborhood, with all of its observational advantages. For these reasons, the flourishing field of Galactic Archaeology has become increasingly popular, aiming, as the name "Archaeology" suggests, to probe information on the early Universe by studying old and near relicts of the past. The main limitation in the study of the galactic halo has been the limited and incomplete amount of velocity information available.

The scenery changed with the launch of the astrometric Gaia mission and its goal of creating an all-sky phase space map of the Galaxy. A data-heavy period has begun, especially with the Gaia second data release (DR2) which contains position, parallax and proper motion of 1.3 billion sources, with a limiting magnitude of G = 21. If we take nearby stars using the tangential velocity cut to select mostly halo objects, it is possible to obtain the Hertzsprung-Russell diagram (HRD)[1] in Fig. 2.1. This HRD presents two clearly distinct sequences, separated at the level of the turn-off point, which are referred to as blue and red sequences with respect to the Gaia color index $\mathbf{G_{BP} - G_{RP}}$. As proposed in [Col+18], this bimodality is interpreted as the presence of two stellar

---

[1]The HRD is a scatter plot of color index and absolute magnitude of stars, which are respectively linked to their stellar luminosity and their temperature. This kind of plots are common tools for astronomers to study stellar population and star evolution.
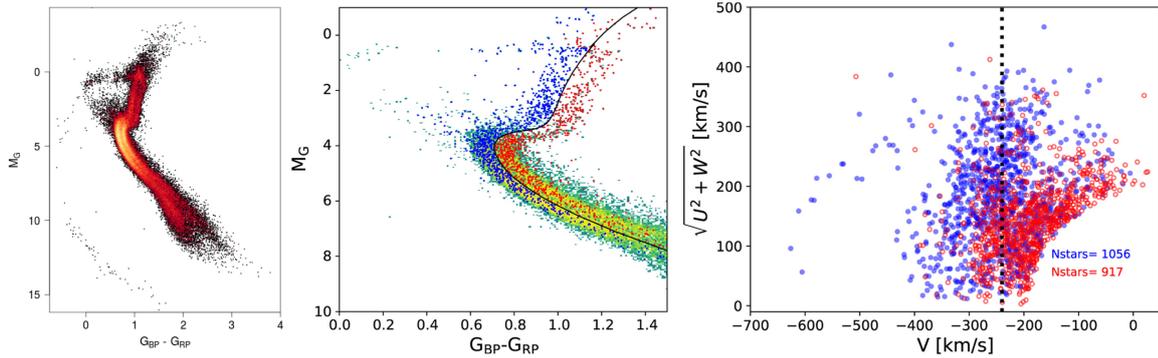
Figure 2.1: Left: HRD for stars in halo like orbit from the DR2. Center: Stars in halo like orbit divided into the blue and red sequence. Right: Toomroe diagram for the stars in the two sequences. This Figure is taken from [DB24].

populations with different mean metallicities, [Fe/H] $\sim$ -1.3 and [Fe/H] $\sim$ -0.5. Further studies show that chemically the red sequence has higher α-element abundances with respect to the blue sequence, while dinamically its orbits are mostly prograde and characterized by low eccentricity; on the other hand, stars belonging to the blue sequence are on high eccentricity orbits that reach further into the halo. This kind of chemodynamical analysis was further tested by [Hel+18], and finally the red sequence was associated with thick disc stars and the blue sequence with an ancient massive merger.

In Fig. 2.2 I report the most recent Gaia release (GD3) and a summary of its exceptional and challenging dataset which has given new fuel to the field of Galactic Archaeology.

Inferring the assembly history of the Milky Way is a challenging task, even in the era of the Gaia mission and its six-dimensional phase space data, and the complementary chemical information obtained from wide-field spectroscopic programs such as GALAH [De +15], H3 [Con+19], APOGEE [Maj+17], RAVE [Ste+06], SEGUE [Yan+09], or LAMOST [Cui+12]. Although the dynamical times of the accreted objects are on the order of the age of the host galaxy, phase mixing of accreted and in situ stars will occur, and phase space only retains part of the information on the original infall parameters. Hence, robustly identifying distinct structures is challenging, and disentangling the components in fully phased mixed situations is nearly impossible. However, stellar chemical abundances remain unchanged throughout the lifetime of a star, serving as unique labels to tag stars. In order to understand why chemical tagging is so effective, we need to keep in mind some important results: the chemical abundance space is dependent on the star formation history and
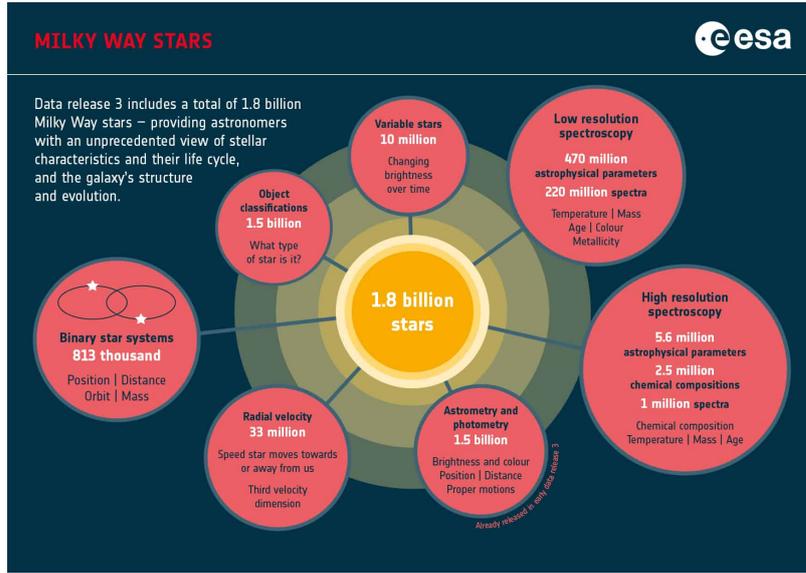
Figure 2.2: Gaia third data release in number.

the total stellar mass of galaxies, leading to distinct differences in the abundance distribution of different galaxies (e.g. [Buc+23]). In particular, the type II (core collapse) supernove (SNe) of massive stars produce α-elements and iron in an almost constant ratio over short timescales (Myr), while type Ia SNe produce iron more efficiently over long time scales period (Gyr). As previously stated, the chemical space is dominated by the total mass of the galaxy, since the more massive galaxies are more capable to resist the expulsion of metals due to supernova feedback, stellar wind, or ram pressure. In particular for dwarf satellite galaxies, as extensively shown in [Kir+13], there exist a mass metallicity relation that holds for several orders of magnitude. In figure 2.3 I report this relation for MW and M31 satellite galaxies.

A great tool to study the star formation history, having access to the metallicity of individual stars, is the Metallicity Distribution Function (MDF). The shape of this distribution can tell us whether a galaxy evolved isolated ("closed box" model) or with gas flow during its star formation lifetime. To obtain a faithful model for the MDF for satellite dwarf galaxies, the introduction of ram pressure stripping is required. During their orbit, satellite dwarf galaxies pass through the hot corona, which can exert ram pressure on the galaxies' gas, and this kind of stripping is able to efficiently remove all the gas after a couple of pericentric passages, effectively halting the star
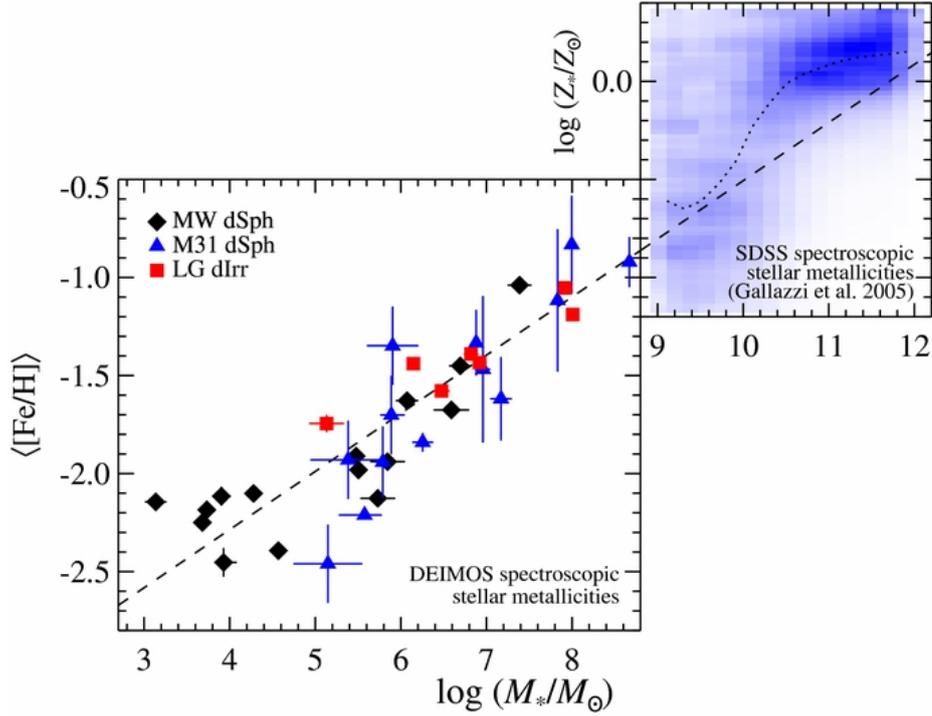
Figure 2.3: Left: the stellar mass stellar metallicity relation for MW and M31 satellites dwarf galaxies. The average metallicity were obtained using spettroscopy instead of photomoetry since they are not effected by the age-metallicity degeneracy. The functional form obtain by fitting the MW satellite is: $< [Fe/H] >= (-1.69 \pm 0.04) + (0.3 \pm 0.02) \log(\frac{M_\star}{10^6 M_\odot})$. It can be appreciated that this relation holds also for more massive galaxies (right), like the sample from the Sload Digital Sky Survey (SDSS). This Figure is taken from [Kir+13].
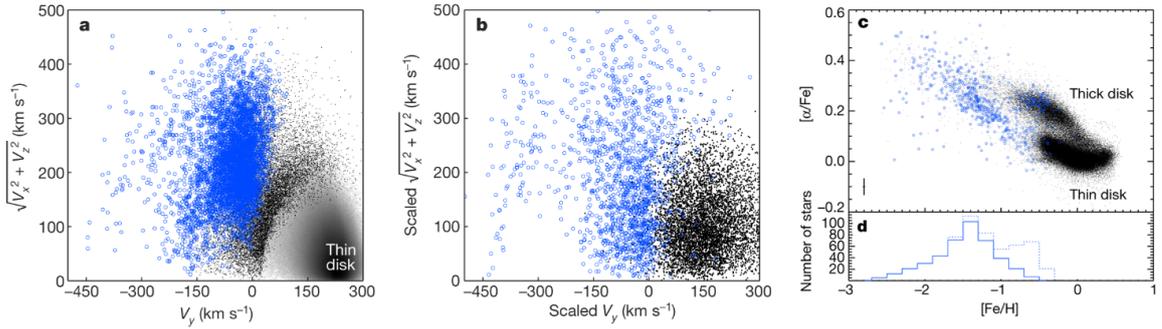
Figure 2.4: Left panel: Toomre diagram for Gaia DR2 data where in blue are stars selected to pick out the GSE structure. Middle panel: similar to the left panel but with simulated data of a minor merger, resulting in a less concentrated structure, mostly due to the fact that a more massive structure is more able to retain its original orbital properties. Right panel: the star in blue is the same as in the left panel cross-mated with the APOGEE chemical information. This Figure is taken from [Hel+18].

formation history.

The crossmatch between Gaia and spectroscopic data allowed for the discovery of the "Gaia-Sausage-Enceladus" (GSE) ([Bel+18], [Hel+18]), a massive accretion event whose remnant now dominates the observation of the inner stellar halo of our Galaxy. In the early works, the GSE is described as a major structure with mostly highly eccentric retrograde orbit with a chemical abundance distribution of stars that is highly distinct from the thin and thick disc stars of the Milky Way, as shown in Fig. 2.4. Today, we know that some of the GSE stars selected in [Hel+18] belong to the Sequoia event, as describe in [Mye+22], and that the the GSE has a zero net velocity. It is important to note that in both [Bel+18] and [Hel+18], the use of numerical simulations has been crucial to interpret the observational results, relying on N-body simulation of a massive merger (20% of the mass of the host) with a MW like galaxy.

As well described in [Kyl20], numerical simulations are not well suited for statistical inference, often due to the intractability of the likelihood, and in general scientists rely on powerful ad hoc or field-specific summary statistics to overcome this problem. With the increasing growth of Machine Learning capabilities and the interconnection between probabilistic models and likelihood-free methods, I decided to test the Simulation Based Inference (SBI) technique on the challenging task of inferring MW satellites parameters (e.g. stellar mass and infall time) from the chemical abundance plane, while also facing the intrinsic problem of not knowing at prior how many parameters we have

to infer in a given galactic halo. Hence CASBI (Chemical Abundance Simulation Based Inference) is a prof-of-concept work that aims to take advantage of cosmological simulations to guide the inference process and obtain accurate and calibrated posterior distribution via Neural Posterior Estimate (NPE). As reported in Chapter 5, CASBI is able to strongly constrain the stellar mass parameters, probably due to the previously cited mass-metallicity relation, but lacks the information to brake the degeneracy of the infall time.

This thesis is structured as follows. Chapter 3 reports two methods that try to bridge the gap between simulations and observations in the field of Galactic Archaeology, and also a brief introduction to the SBI technique. In chapter 4 the main CASBI pipeline is presented. In chapter 5 the results obtained on the test set and metrics to evaluate the accuracy and calibration of this method on both a subhalo and a halo basis are reported. In chapter 6, I summarize the major findings, describe limitations, and propose some future prospects.

Chapter 3

Previous work

In this chapter, I present two inspiring methods for studying galactic halo. CASBI tries to build upon the core ideas and results presented in these two works, leveraging the automatic and accurate tools available in the Machine Learning realm to perform inference. Lastly, a short summary of the general framework of Simulation Based Inference is presented, with a focus on the Neural Posterior Estimate technique.

3.1 The reconstruction of the Assembly history of the Milky Way

Robustly identify distinct structures in a galactic halo is challenging, and disentangling components in fully phase mix situations is nearly impossible. In order to characterize the assembly history [Cun+22] propose using the "CARDs", the Chemical Abundance Ratio Distribution of the stars, obtained from a subsample of accreted object candidate from the FIRE-2 zoom-in cosmological simulations of MW-mass galaxies [Wet+16]. This method does not recover posteriors for the parameters of the accreted objects, but rather considers the host halo as a linear combinations of templates CARDs

$$\text{CARD}_{\text{halo, model}}(x_d) = \sum_i \sum_j A_{ij} \text{CARD}_{\text{temp},ij}(x_d | M_{\text{sat},i}, t_{100,j}),\qquad(3.1)$$

treating each coefficient A_{ij} as the fraction of mass contribution to the chemical abundance ratio x_d from the accretion event of the template satellite with mass $M_{\mathrm{sat},i}$ and quenching time $t_{100,j}$ [1], and it tries to recover those coefficients by minimizing the difference between the observed CARDs and the linear combination of the templates. An example of template iron and magnesium joint distributions constructed from dwarf galaxies is presented in Fig. 3.1. The templates that were used belong to the catalog of star particles in the FIRE simulations belonging to dwarf galaxies, stellar streams, and phase-mixed debris constructed in [Pan+21]. This method relies on the assumption that the chemical space of accreted and isolated dwarf galaxies is similar up to the quenching time, due to ram pressure quenching the star formation history of the accreted object and hence 'freezing' these abundance ratios. A possible limitation of this assumption, which might also affect CASBI results, is that a distinction between destroyed and surviving accreted dwarf galaxies is not taken into account. This can lead, as described in [Nai+20], to a -0.3 dex offset of the mass metallicity relation [Kir+11] that might be necessary.

The second work is described in [Dea+23]. This method takes advantage of the mass-metallicity relation to decompose the MDF of the host galaxy as a mixture of accreted halo's MDF, assumed gaussian for each of these building blocks. This decomposition is based on [Kir+11] that demonstrated that at the dwarf mass scale, not only does the average metallicity varies with the mass, but also the width of the MDF, with the lowest mass dwarf having a wider spread of metallicities. The Likelihood that is used in [Dea+23] for the [Fe/H] distribution, indicated as **z** is then:

$$P(z|N, L_i, \mu_i, \sigma_i) = \frac{1}{\sum L_i} \sum_{i=1}^{N} L_i \mathcal{N}(z|\mu_i, \sigma_i), \tag{3.2}$$

where they have assumed that the number of stars in the sample scales linearly with galaxy luminosity L_i, that the μ_i follows a mass-metallicity relation and that the σ_i depend on the galaxy luminosity as described in [Kir+11].

Similarly to CASBI, this method faces the challenge of having a variable number of parameters, making them difficult to sample in practice. In order to tackle this problem they decided to bin the luminosity L_j and count the number of contributions from each bin N_j. In order to perform

---

[1]The $t_{100,j}$ is defined as the time for a dwarf or halo progenitor to form 100% of its stars. In CASBI, it is also referred to as infall time.
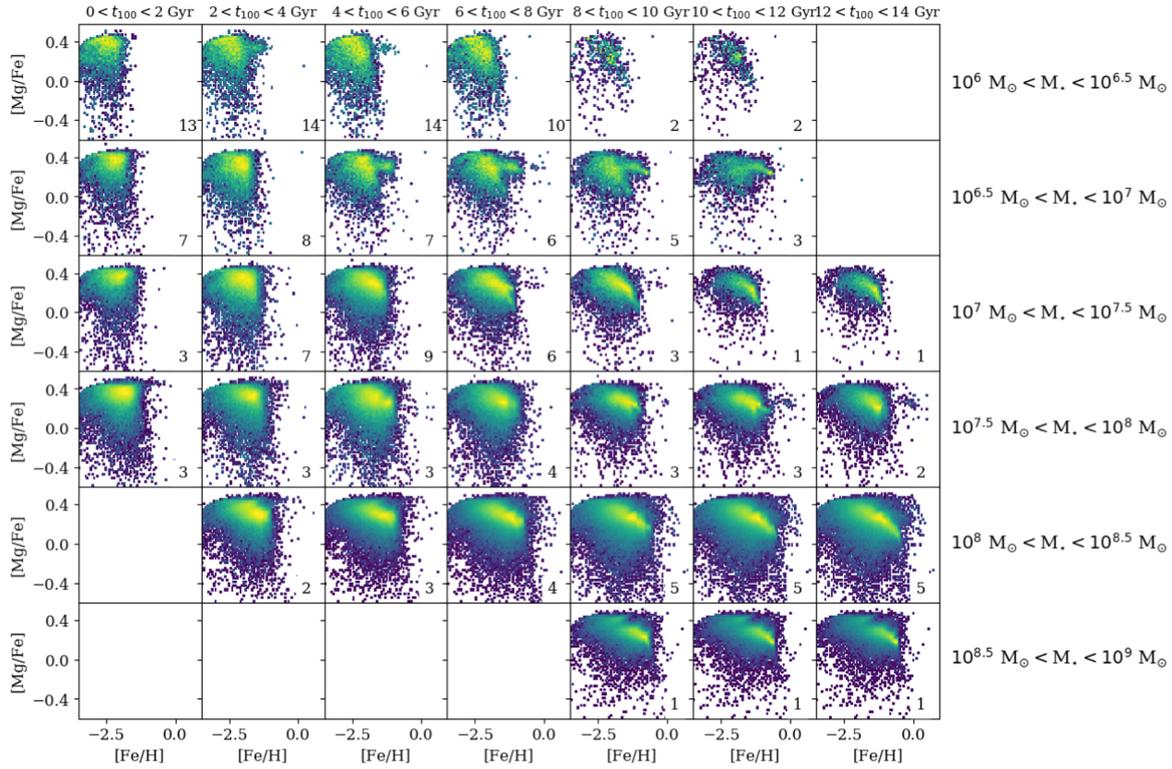
Figure 3.1: Template for accretion events constructed using dwarf galaxy in [Cun+22]. More massive dwarf galaxies have CARDs that extend to higher metallicities. At fixed stellar mass, galaxies that assemble more quickly (lower t_{100}) have more density at higher [Mg/Fe] than the component with a more extended star formation history.

the inference, they adopted a nested sampling scheme to obtain a posterior distribution for the number of galaxies in each luminosity bin. Using the inferred posterior probability, they were able to draw the Milky Way halo number of satellite as a function of their luminosity, reported in Fig. 3.2. The samples used in this posterior were obtained from different spectroscopic surveys after applying various cuts to avoid contamination from thick disc stars. The cuts were made based on parallax distance, radial distance, height with respect to the plane of the galaxy, and only stars with retrograde orbits were selected.

CASBI adopts the same superimposition of the components contribution, without neither a prefix nor an analytical form for the joint distribution of the chemical abundances, relaxing these assumptions and relying only on the available samples from the N-body simulations.
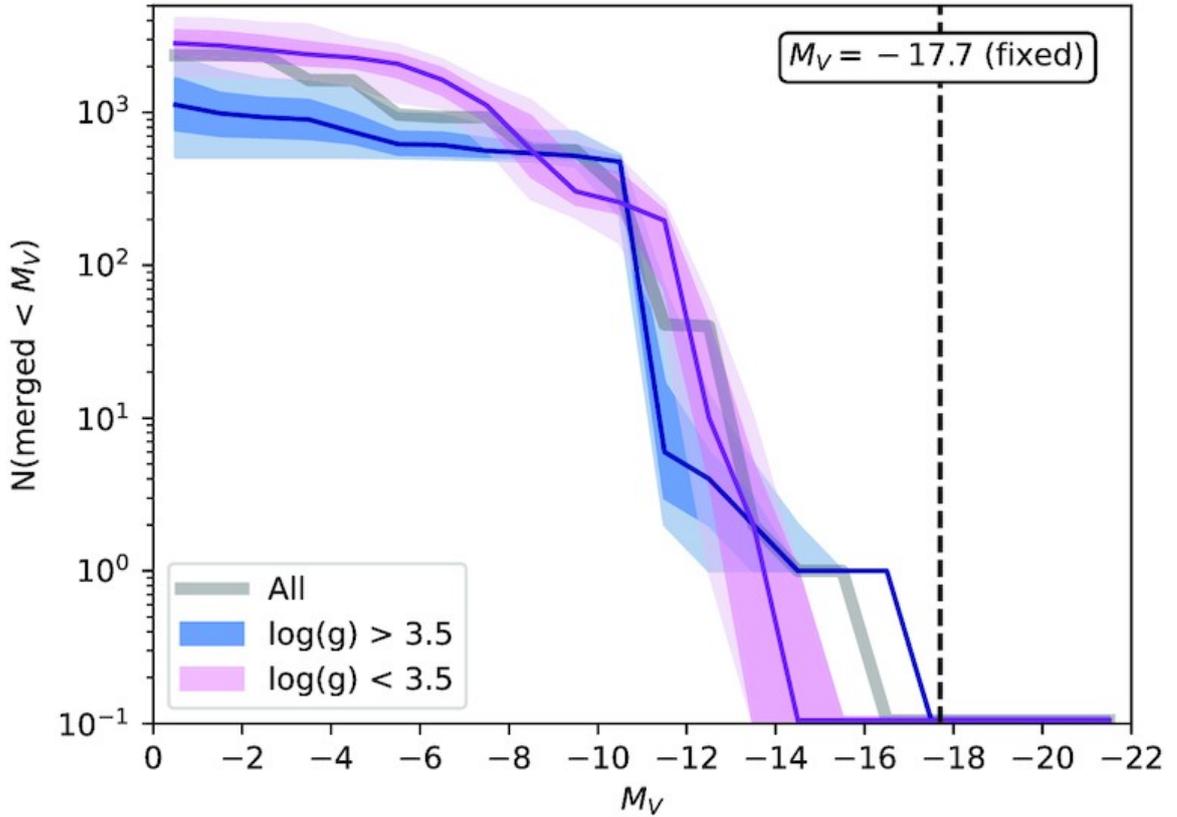
Figure 3.2: The estimated number of destroyed dwarf galaxy in the MW halo as a function of their V band absolute magnitude. The separation in two bins of log(g) is because dwarf stars and giants can have different metallicity biases. This Fig. is taken from [Dea+23].

3.2   Simulation Based Inference

The SBI framework has existed along with the more traditional likelihood-based inference methods for quite some years already, having its roots in Approximate Bayes Computation (ABC) [Rub84], and has been used in a variety of fields, from cosmology to particle physics. The main difference between SBI and likelihood-based methods, such as Markov Chain Monte Carlo (MCMC), is that the former does not require the likelihood function to be known, but rather relies on a simulator to generate synthetic data $\mathbf{x}$ once the input parameters $\boldsymbol{\theta}$ are passed to it, and the inference pipeline is trained based on data-parameters pairs $(\mathbf{x}, \boldsymbol{\theta})$.

The recent advance of this technique was made possible by the use of ML models to emulate con-

ditional probability distributions, a technique known as Neural Density Estimation (NDE) [Pap19]. NDE is achieved by training a Normalizing Flow architecture[2], a generative model that allows to obtain samples from a complex distribution $p(x)$ by constructing a series of **bijective** transformations $f^i_{\phi_i}$ that map x to a latent space z that is distributed as a simple distribution, like a Gaussian. Accordingly to [KD18], by approximating the transformation f^i with the i-th layer of a Neural Network with parameters ϕ_i, the model learns how to sample from $p(x)$ using the schema presented in Eq.3.3.

$$p(x) \sim x \equiv h_0 \xleftrightarrow{f^1_{\phi_1}} h_1 \xleftrightarrow{f^2_{\phi_2}} h_2 \dots \xleftrightarrow{f^K_{\phi_K}} h_K \equiv z \sim \mathcal{N}(z; 0, \mathcal{I}) \tag{3.3}$$

Using the change of variable formula, the probability density function (pdf) can be written as:

$$\begin{aligned}
\log p(x) &= \log p(z) + \log \left| \det \left(\frac{\partial z}{\partial x} \right) \right| \\
&= \log p(z) + \sum_{i=1}^{K} \log \left| \det \left(\frac{\partial h_i}{\partial h_{i-1}} \right) \right| \\
&= \log p(z) + \sum_{i=1}^{K} \log \left| \det \left(\frac{\partial f^i_{\phi_i}(h_{i-1})}{\partial h_{i-1}} \right) \right|,
\end{aligned} \tag{3.4}$$

where the last term is the sum of the log determinant of the Jacobian of the transformations $f^i_{\phi_i}$. In order to train the model, the negative log pdf is used as loss function. Once the model is trained, it is easy to sample from the distribution $p(x)$ by sampling from the latent space z and applying the inverse transformations $(f^1_{\phi_1})^{-1} \circ \cdots \circ (f^K_{\phi_K})^{-1}$. In order to keep the computation tractable, the use of `coupling layers` allows to obtain a triangular Jacobian. A simple yet effective coupling layer that acts separately along dimensions d is the Affine coupling layer, presented in Eq. 3.5. The functions s and t are usually parameterized by a neural network, and the fixed and transformed dimensions are swapped after each layer to achieve an overall permutation-invariant transformation.

$$\begin{cases} h_i^{1:d} &= h_{i-1}^{1:d} \\ h_i^{d:D} &= h_{i-1}^{d:D} \times \exp(s(h_{i-1}^{1:d})) + t(h_{i-1}^{1:d}) \end{cases} \tag{3.5}$$

---

[2]This has been by far the more popular option for NDE, but new and more performing models are beeing tested in the active field of SBI methods. An example is described in [Dax+23].

The choice of the invertible function affects the expressivity of the model, defined as the ability to approximate a more complex multivariate distribution, at the cost of more parameters, computational time, and inference time.

Following the discussion presented in [Ho+24], when using NDE for Bayesian analysis, one can choose to approximate the Posterior, the Likelihood or the Likelihood ratio[3], and this choice depends mostly on the problem that one wants to solve. As a rule of thumb, one can consider the dimensionality of the observations and the parameters as the complexity of the Likelihood and the Posterior respectively.

In this case, due to the complexity of the Likelihood distribution of the chemical abundance space[4], I choose to approximate the Posterior distributions, and so I adopted the Neural Posterior Estimate (method **F** in Figure 3.3) that can be trained using the negative log-Posterior as loss function:

$$\begin{aligned}
\mathcal{L}_{NPE}(\boldsymbol{\theta}) &= -\mathbb{E}_{\mathcal{D}_{train}} \log \hat{\mathcal{P}}(\theta_i | x_i) \\
&= -\mathbb{E}_{\mathcal{D}_{train}} \log \left(\frac{p(\theta)}{\tilde{p}(\theta)} q_\omega(\theta_i, x_i) \right),
\end{aligned} \tag{3.6}$$

where the Posterior distribution $\hat{\mathcal{P}}(\theta_i | x_i)$ is approximated by the product of the ratio of the prior $p(\theta)$ and proposal distribution $\tilde{p}(\theta)$ and the neural conditional distribution $q_\omega(\theta_i, x_i)$, parameterized by the parameters ω. Usually, the conditioning of the posterior approximation is obtained by making the transformation f dependent on the observation x, but due to its high dimensionality, especially in cosmological simulations, an embedding network is used to compress the observations in a latent space that is past to the NPE to guide the inference. Many excellent frameworks are already available to handle SBI analysis, and CASBI is built on top of the `ltu-ili`[5] python package [Ho+24]. In particular, CASBI analysis was performed with the `lampe` backend[6] to train a Neural Posterior Estimate of the posteriors of the parameters.

---

[3]They are referred to as Neural Posterior Estimate (NDE), Neural Likelihood Estimate (NLE) and Neural Ration Estimate (NRE) respectively. All of these techniques aim at the same goal: sampling the Posterior distribution.

[4]We would need to evaluate the Likelihood on the 64×64 pixel space, while the vector parameter space has only two dimensions.

[5]https://github.com/maho3/ltu-ili

[6]https://github.com/probabilists/lampe?tab=readme-ov-file
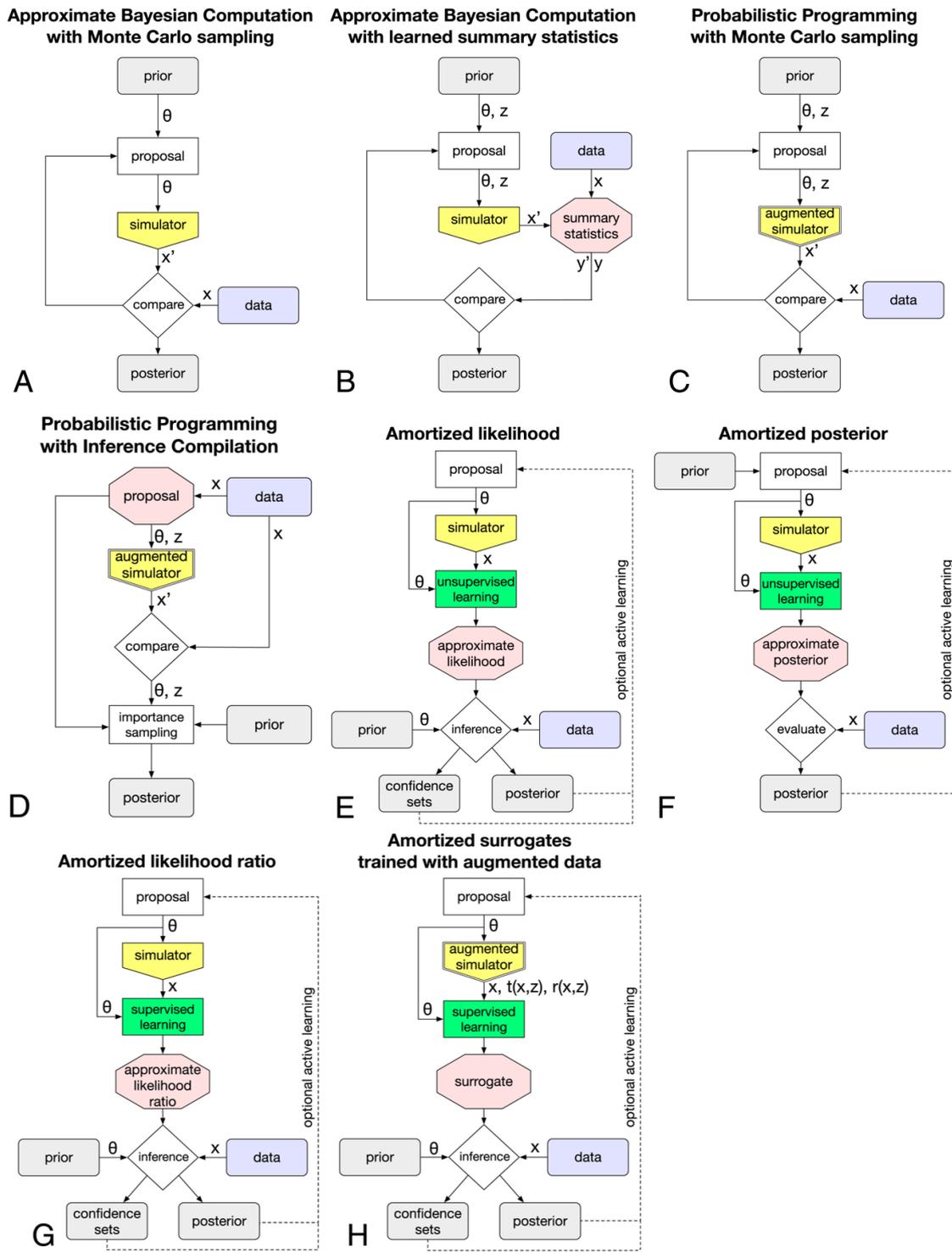
Figure 3.3: Different approaches to Simulation Based Inference, from [Kyl20].

Chapter 4

CASBI: Chemical Abundance Simulation Based Inference

CASBI is a Simulation Based Inference (SBI) Python package to recover the properties of building blocks of the Milky Way like galaxy's halo from spectroscopic observations of chemical abundances. In this chapter, I present the components and the pipeline of the inference process in Sections 4.1 and 4.2, while also reporting some of the early tests in Appendix A.1 and A.2 that are not present in the final version of this project.

4.1 Simulator and Preprocessing

The data-parameters pairs $(\mathbf{x}, \boldsymbol{\theta})$ needed to train the NPE are obtained from the Numerical Investigation of a Hundred Astrophysical Objects (**NIHAO**) project [Wan+15]. The **NIHAO** is a set of $\sim$100 cosmological zoom-in hydrodynamical simulations evolved using the N-body SPH solver `Gasoline` [WSQ04], with halos that range from dwarf ($M_{star} \sim 5 \times 10^9 M_\odot$) to Milky Way like ($M_{star} \sim 2 \times 10^{12} M_\odot$). The halos were identified using the Amiga Halo Finder code (AHF), presented in [KK09], by looking for overdensity in an adaptively smoothed density field. In order to handle these simulations, in CASBI the pre-processing is done using the tools available in `pynbody` [Pon+13]. Fig. 4.1 shows face on samples of galaxies in the **NIHAO** simulation.
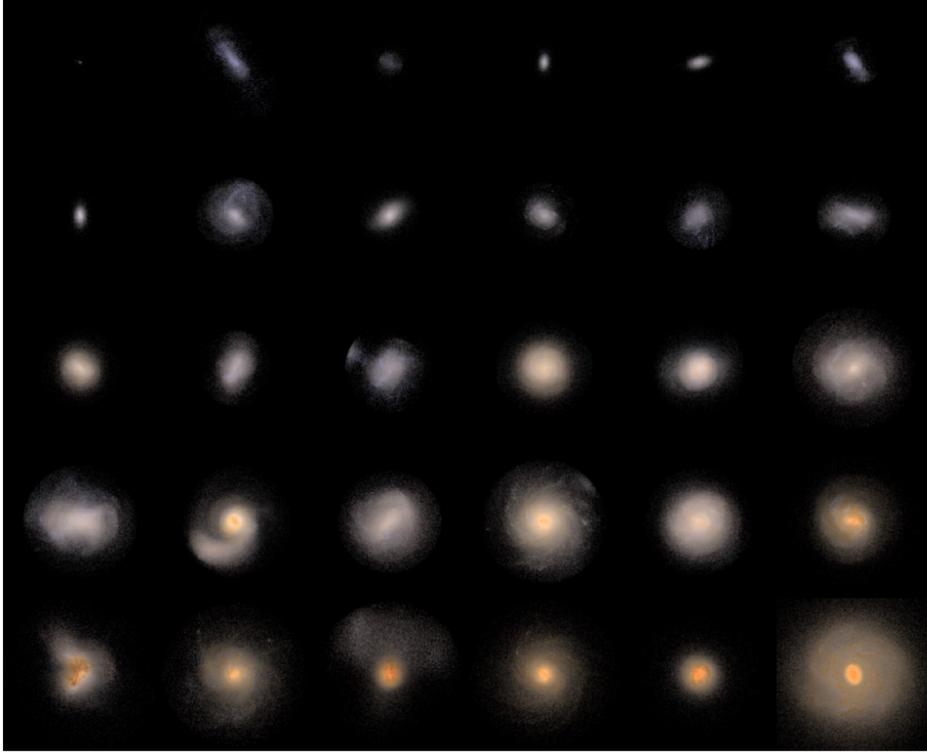
Figure 4.1: Face on **NIHAO** galaxies from [Wan+15]. Images are 50 kpc on a side.

Similarly to [Cun+22] and [Dea+23], CASBI relies on the assumption that once the accreted object falls into the gravitational potential of the Milky Way like galaxy its star formation rate is halted, so each of the snapshots in these simulations is treated as a possible building block of galactic halo.

The construction of the observables is done by aggregating multiple subhalo into a single stellar halo. In order to create subhalo I construct 2D histogram, referred to as $\mathbf{x_i}$, by binning the chemical abundance plane ([O/Fe], [Fe/H])[1] for each of the snapshot available in **NIHAO**. In Fig. 4.2 some of these galaxies abundance plane are reported. Hence, the observations are multidimensional distributions of chemical abundances. I have also filtered the galaxies that I use as building blocks to retain only those with a total stellar mass lower than the stellar mass of the Large Magellanic Cloud ($M_{star} < 6 \times 10^9 M_\odot$), the largest accreted object by the Milky Way. The 2D histograms have 64×64 pixels, and minimum and maximum values set after filtering all stars that were outside

---

[1]They are respectively proxy for α elements abundance and metallcity
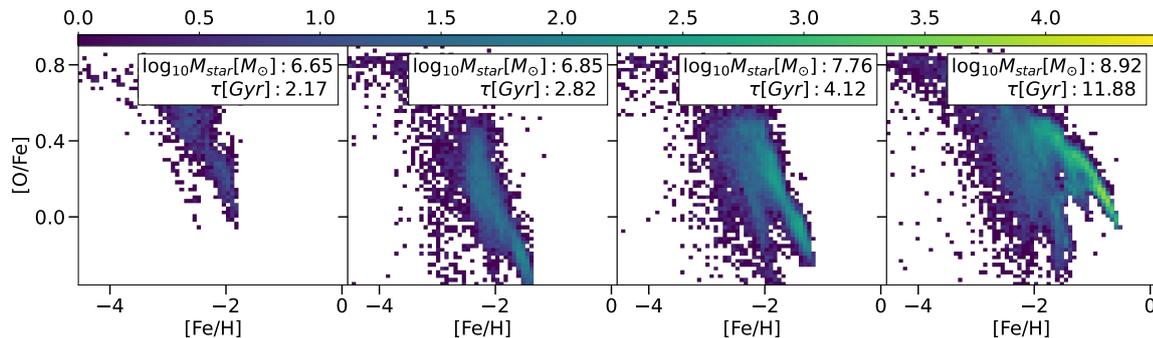
Figure 4.2: Example of chemical abundance plane used in CASBI with different masses and infall time. The logarithm number density is reported in the colorbar.

the 0.01 percentile in either metallicity or α element abundance[2]. All the star particles are collected into a `pandas` dataframe, and each of the x_i is uniquely identifiable through the `Galaxy name` attribute[3]. The set of all possible subhalos is defined as "Template Library". The actual stellar halo observable $\mathbf{x^j} = \sum_i^{N^j_{sub}} x_i^j$ used in CASBI is then a super imposition of N_{sub} of these 2D histograms, where N^j_{sub} is the number of accreted objects that are present in the j-th galaxy halo. The actual choice of how to sample the template library created from the **NIHAO** simulations can be adapted. Firstly, I tested to randomly sample in A.2, but eventually CASBI uses a more physically informed approach by using a luminosity function and a total stellar mass budget in Section 4.2. In CASBI the `template library` is a Python class that is implemented to handle the training and test sets, returning different data formats depending on whether they need to be passed to the inference pipeline or to be inspected by the user. The user can decide what kind of luminosity function to adopt, the stellar mass budget, and the noise level to apply to the observations.

The goal of CASBI is to recover the parameters $\theta_{\mathbf{i}}^{\mathbf{j}}$ for each of the subhalos in the galactic halo from the observable $\mathbf{x^j} = \sum_i x_i^j$, and to gain insight into the number of subhalos. Among all the possible parameters available from the simulations, I have decided to limit the parameters to stellar mass M_{star} and age of the accreted galaxy τ, also called infall time due to their equivalence in the assumption of quenched star formation after accretion.

---

[2]This filtering was used to remove strong outlier star particles present in some of the NIHAO snapshots.
[3]The x_i that are used in this work are conceptually equivalent to the CARDs in [Cun+22]

4.2   Realistic halo and one step Inference
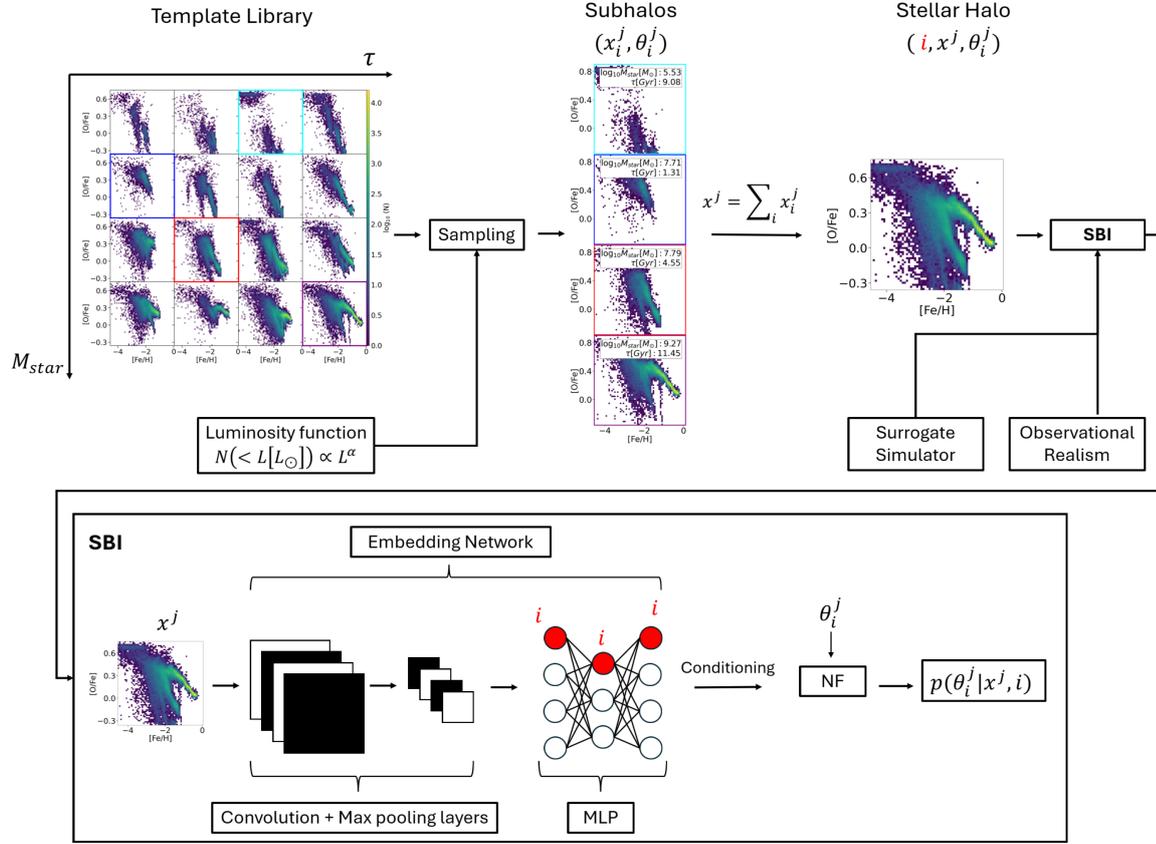
Figure 4.3: CASBI pipeline. In this thesis analysis I have fixed the **Template Library** to be a subsample of the **NIHAO** simulations, as described in Section 4.1, but it can be swapped with user-choice simulations. The choice of the Template Library incorporate all the assumption that one make on the chemical enrichment history of Galaxies, the dynamical effects that accreted objects undergo, and the cosmology, making this part the principal cause of possible mispecification. During **sampling** non-repeated subhalos are sampled, aiming to reproduce a **Luminosity function** that can be set by the user. For the analysis, the Luminosity function $N(< L)$ was taken it from [Kop+08]. Moreover a **stellar mass budget** can be set to generate realistic stellar halo without the need of fixing the number of subhalos (even though a maximum number of subhalos is set in the analysis for computational reasons). In the analysis, the stellar mass budget of the halo was fixed accordingly to [DBS19], but it can be set by the user accordingly to the estimated total stellar mass available in the observation. The **SBI** pipeline can incorporate a **Surrogate Simulator** to perform the sequential version of NPE. In the analysis I do not adopt this component as described in Section A.1. The **Observational Realism** encapsulate all that concerns bridging the gap between simulation and observation, i.e. uncertainties of spectroscopic surveys, selection functions, etc. In my analysis I have only injected the possibility to add instrumental noise, leaving background and foreground contamination and selection function as future work. During inference, the user can pass a galaxy abundance plane x^j and obtain posterior θ_i^j for the the i-th most massive subhalo. The model used to infer needs to be trained to the correct amount of noise that one suppose to be present in the observation of x^j.

The objective of the inference is not trivial, since in order to recover the parameters of the building blocks of the Milky Way like galaxy we need to fix the dimensionality of the priors. This is equivalent to having complete knowledge on the number of substructures that are present in the galactic halo. In order to avoid the need for a two-step inference[4] and still retain the possibility to access the information on how many subhalos populate a given abundance plane, I have decided to condition the SBI model to retrieve the i-most massive subhalo of the j-th stellar halo. In this way, the NPE is trained with pairs of $(x, \theta) = ((i, x^j), \theta_i^j)$, where $x^j = \sum_i x_i^j$ and x_i^j are ordered accordingly to their stellar mass M_{star}. Hence, the j-th stellar halo abundance plane is shown as many times as the number of subhalos present in it, and to guide the model in inferring the right parameters θ_i^j, the embedding is conditioned on the integer i. The subhalo index i is concatenated to each input of the fully connected layers of the CNN used to embed the observations, and it is also concatenated before passing the embedded information to the Normalizing Flow. This conditioning is summarized in the lower part of Fig. 4.3, in which the CASBI SBI pipeline shows the integer i as a red node concatenated to each layer of the embedding network.

Before proceeding, it is necessary to define the concept of magnitude. In astronomy, the magnitude is a measure of the brightness of an object. The apparent magnitude of an object, is defined as

$$m_1 - m_{\text{ref}} = -2.5 \log_{10} \left(\frac{F_1}{F_{\text{ref}}} \right), \tag{4.1}$$

where F is the flux (i.e. the brightness), and m is the apparent magnitude with respect to another object that is used as reference. If the distance of the object d (in parsec) is known, then it is also possible to define the absolute magnitude M as

$$m - M = 2.5 \log_{10} \left(\frac{d}{10} \right), \tag{4.2}$$

which can be interpreted as the apparent magnitude of the object as seen from a distance of 10 parsec. If we have observation in a particular band, for example the V band[5], we refer to the absolute magnitude in that band using M_V.

---

[4]This approach is presented in Appendix A.2.
[5]The V ("Visual") band, has an effective wavelength of $\lambda = 551$nm and a bandwidth of $\Delta\lambda = 88$ nm.

As described in section 4.1, in order to create a more realistic mock galaxy halo, I have decided to adopt a sampling scheme for the subhalos that is based on the luminosity function described in [Kop+08]. The luminosity function describes the subhalos distribution in a range of luminosities that spans from $M_V = -2$ all the way to the luminosity of the Large Magellanic cloud

$$\frac{dN}{dM_V} = 10 \times 10^{0.1(M_V + 5)}. \tag{4.3}$$

We can manipulate Eq. 4.3 to express it as a function of the Luminosity L

$$\frac{dN}{dL} = \frac{dN}{dM_V} \times \frac{dM_V}{dL} = 10^{0.1(M_{V,\odot} - 2.5\log_{10}(L) + 5) + 1} \times (-2.5L^{-1}) \sim L^{-1.25}, \tag{4.4}$$

which can be integrated to obtain the number of subhalos with luminosity lower then L

$$N(< L) = K \times L^{1+\alpha}, \tag{4.5}$$

where K is an integration constant and $\alpha = -1.25$ is the single power law exponent obtained by [Kop+08]. Other work based not only on SDSS observations like [Kop+08] but also on ΛCDM N-body simulation set $\alpha = -1.9 \pm 0.2$ ([Tol+08]). In this work, the value of α is fixed at -1.25 and leave the analysis of the impact of this choice to future work. Assuming a stellar-to-light mass ratio $L/M_{star} \sim 2$, I normalize the equation 4.5 after setting the support to be the interval of masses that are available in the catalog of NIHAO simulations ($10^{4.7} M_\odot < M_{star} < M_{star}^{halo}$) and sample from this distribution using an inverse scheme, with M_{star}^{halo} being the mass budget for the mock halo ($M_{star}^{halo} = 1.4 \pm 0.2 \times 10^9 M_\odot$ based on [DBS19]). After obtaining the analytic samples, the first and second Nearest Neighbors (NN) in the catalog are taken, making sure that the distance between the analytic and catalog sample is lower than 10% of the mass of the analytic one, and one of those NN is randomly selected as a subhalo for the mock halos. Lastly, the total mass budget is reduced by the mass of the NN that was used. The choice of the mass budget can be adapted and comes from observations that do not take into account the Large and Small Magellanic Cloud, but it can be customized or even set into a range of stellar-budget mass at each generation of a mock stellar halo. During this iterative procedure I make sure to sample non-repeated subhalos within the same mock

galaxy, while also avoiding repetitions of the same combinations between subhalos both within the training and test sets and across these two sets.

In Fig. 4.3, I report the CASBI pipeline. The modularity of the SBI technique is fully integrated, allowing to change all the components of this pipeline. The `template library` can be set to use a different suite of simulated galaxies (e.g. [Pil+23]), the sampling scheme can incorporate different luminosity function and stellar halo budget, the NPE, the embedding network architecture and hyperparameters can be modified to allow for higher accuracy and posterior coverage thanks to the `optuna` grid search implementation, and surrogate models (Free Form Flow FFF [Dra+24]) or semi-analytic model (GRUMPY [KM22]) can be implemented to allow for the Sequential version of the NPE. For the analysis presented in Chapter 5, the results are obtained using 1000 training set and 100 test set data pairs.

Chapter 5

Results

In this chapter the results of the inference pipeline on the test set are reported. I also tried to bridge the gap between simulations and observations by adding various levels of noise to the observable x and checked the response of both the chosen accuracy and calibration metrics. In the opening section 5.1, the calibration metrics that were used to seize the constraint capability and the possible biases of the posterior distributions are described.

5.1   Calibration

This section is strongly inspired by [Ho+24]. In posterior estimation, our aim is to maximize the constraining of θ given the observable x and to determine whether the uncertainties are calibrated to the training data. These criteria are naturally adversarial, and this problem can be interpreted as another instance of the bias-variance trade-off. It is possible to confront various NPE precisions by comparing the cumulative posterior value of the test set $\hat{\mathcal{P}}(D_{test}) = \prod_i^{N_{test}} \hat{\mathcal{P}}(\theta_i|x_i)$, since a larger posterior value concentrates more probability mass around the true value, which directly results in a higher constraining power. The calibration of the model uncertainties can be obtained using the Probability Integral Transformation (PIT), defined as the cumulative density function of the posterior given x_0:

$$\mathrm{PIT}(\theta|x_0) = \int_{-\infty}^{\theta} \hat{\mathcal{P}}(\theta|x_0)d\theta. \tag{5.1}$$

Due to the poor scaling of the PIT to higher dimensions, it is better to construct an estimate of the PIT value as follows:

$$\mathrm{PIT}(\theta|x_0) = \mathbb{E}_{\hat{\theta}\sim\hat{\mathcal{P}}(\theta|x)}[\Theta(\hat{\theta}-\theta)], \tag{5.2}$$

where Θ is the Heaviside step function. The PIT counts the number of times that the posterior samples $\hat{\theta}$ fall below the true parameter value θ. If we match the true posterior everywhere, we expect the PIT value to be uniformly distributed in the range $[0, 1]$ for each of the test set samples. Usually, the PIT distribution is studied using percentile-percentile (P-P) plots[1], comparing the cumulative distribution function (CDF) of the PIT value to the CDF of a uniform distribution. This tool can be used to constrain the over- and under-confidence regions of the posteriors. In Fig. 5.1 I report an example of uncalibrated posteriors approximation and the corresponding P-P plot [2]. But, the percentile-percentile plot can also be an indicator of biases that are not compensated by the uncertanties, as it is clear in Fig. 5.2 (a), leading to posteriors that are not scientifically accurate. In addition, the number of modes mismatches between the inferred and true samples can be captured using this kind of plot, as shown in the toy model of Fig. 5.3.

---

[1]Also called quantile-quantile plot.
[2]The plot is not accurate, it is created to give a rough idea of the general relation between the posterior behavior and how it is reflected in the P-P plot.
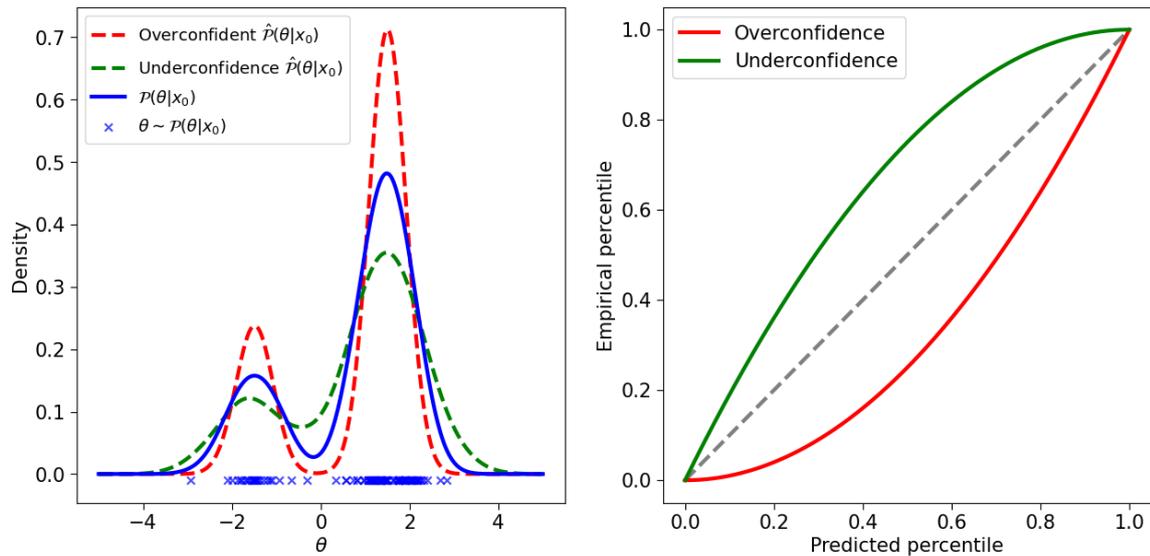
Figure 5.1: Examples of SBI uncalibrated posterior. Left: overconfident (red) and underconfident (green) posteriors obtained from the same data θ, for which the true posterior (blue) is shown. Right: P-P plot for the two uncalibrated posteriors. The goal of the calibration is to obtain a distribution that is underconfident and as close to the diagonal as possible. This figure is inspired by [Fal+].

Since as the dimensionality of θ increases, the proper coverage requires exponentially more samples, one can rely on the PIT of each component θ_i of the marginal posterior, and we expect that the value

$$\mathrm{PIT}(\theta_i|x_0) = \mathbb{E}_{\hat{\theta} \sim \hat{\mathcal{P}}(\theta_i|x)}[\Theta(\hat{\theta}_i - \theta_i)], \tag{5.3}$$

has the same properties as the PIT obtain in equation 5.2 if the model is globally consistent on the test set.

Lastly, as an approximation to verify multivariate posterior coverage, I use the Test of Accuracy with Random Points (TARP) [Lem+23]. TARP constructs, in the limit of sufficient samples, an estimate of posterior coverage which is guaranteed to converge to the true posterior coverage. All of the previously described posterior coverage methods are already integrated in the `ltu-ili` package.
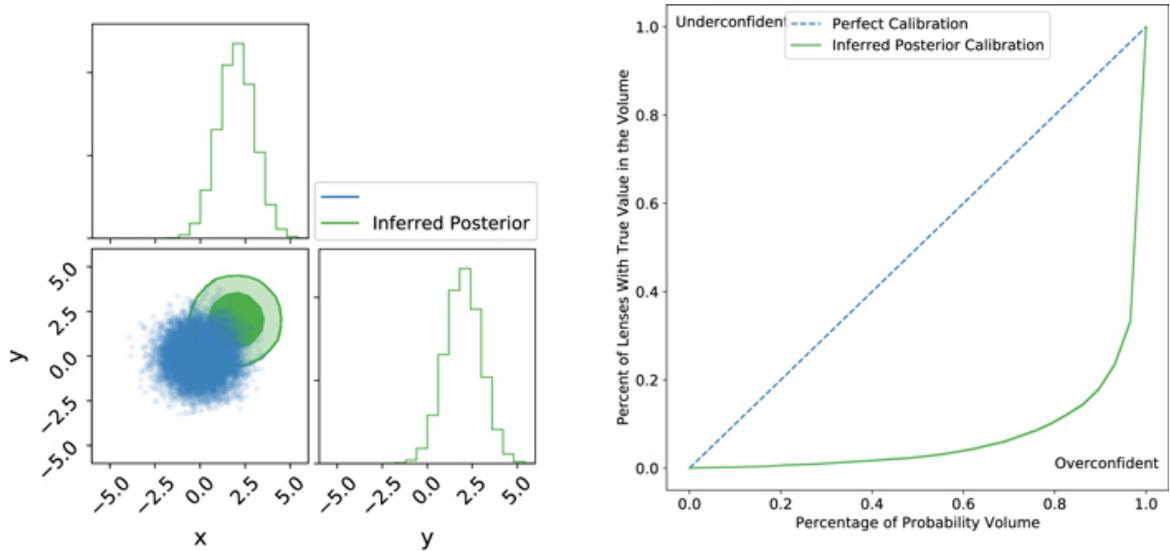
(a) Calibration of an offset inferred posterior
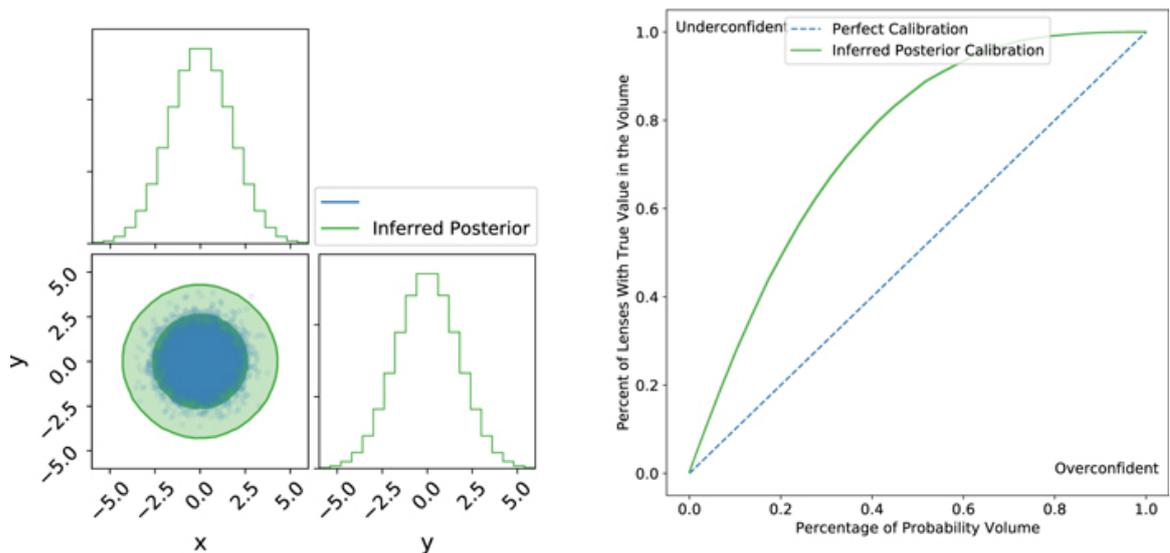
(b) Calibration of an inferred posterior with large uncertanties

Figure 5.2: A visualization of the quantile–quantile plot for two toy scenarios. In (a), the inferred posterior has the correct spread, but is offset from the correct mean, leading to significant overconfidence. In (b), the inferred posterior is correctly centered, but has a spread that is too large, leading to significant underconfidence. This Figure and its caption are taken from Appendix A of [Wag+21].
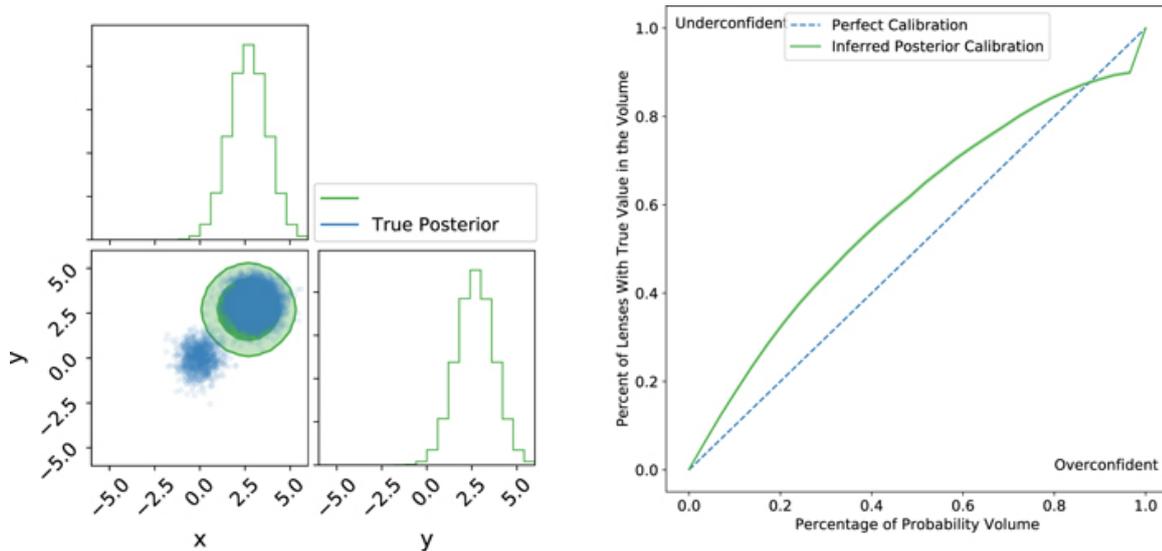
Figure 5.3: A visualization of the quantile–quantile plot for a toy model where the inferred posterior is univariate but the true posterior is bivariate. Note that the signal falls neither cleanly in the over or underconfident regions, but rather crosses from one to the other. This Figure and its caption are taken from Appendix A of [Wag+21].

5.2   Test set

The `ltu-ili` accuracy and coverage plots from the `ltu-ili` package of the 100 galaxies halos of the test set are reported, respectively, in Figs. 5.4, 5.5 and 5.6. In Fig. 5.4, the error bars represent the 68% confidence interval, and I find good predictive performance on the stellar mass M_s throughout the mass range probed, while the infall time τ is poorly constrained. The lack of predictability of the infall times can be associated with the degeneracy of the chemical abundances between a massive system that was accreted early on and a less massive system that was accreted more recently. The integration of the orbital information or using more α elements (e.g. Mg, Si, Mn) could alleviate this limitation of the model. Both of these improvements are left for future work. The results presented in the remaining sections are obtained only for the stellar mass parameters. Figs. 5.5 and 5.6 percentile plots show that the model is well calibrated in both marginal and posterior probabilities. The `optuna` multi-objective optimization was essential to achieve these results, since I looked for the hyperparameters that were able to obtain the highest log-probability and the minimum distance from the diagonal at the 0.5 level for the TARP value.
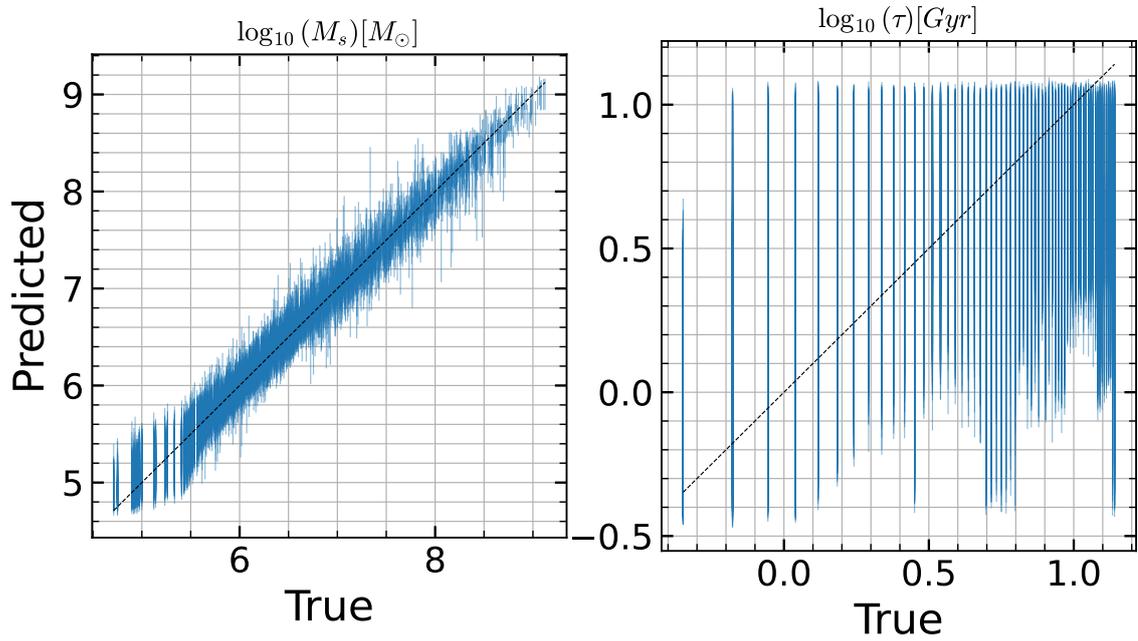
Figure 5.4: Predicted vs True values with error bars representing the 68% confidence interval
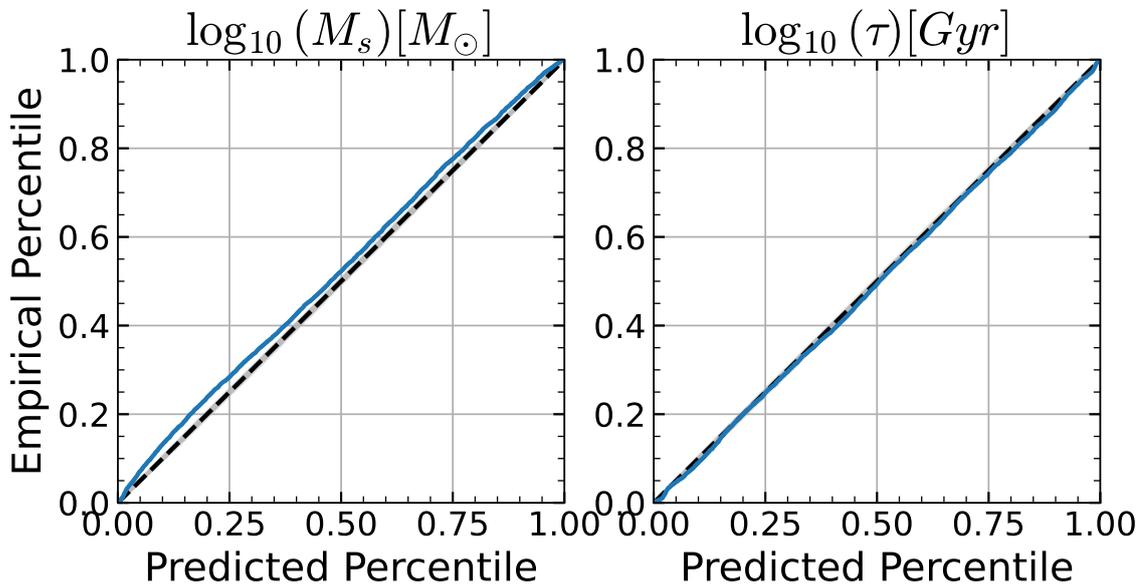
Figure 5.5: Coverage for the marginal probabilities.
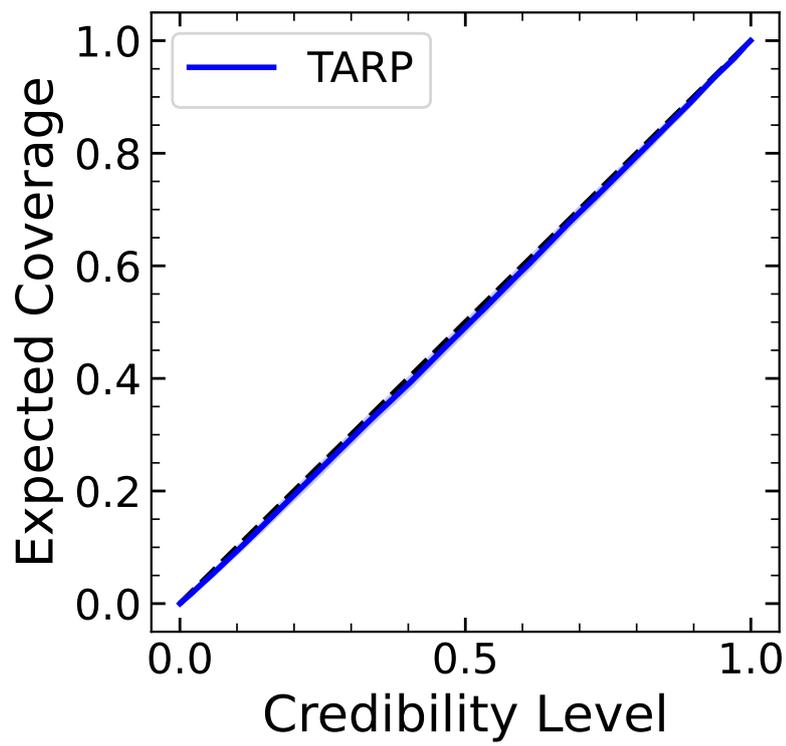
Figure 5.6: TARP percentile-percentile plot. Approximation of coverage plot for the joint distribution.

5.3   A first step toward inference on observations

Simulations, in general, have no uncertainties associated with their values. If we want to bridge the gap between the idealized realm of numerical simulations and the real world, we need to inject all the possible sources of noise and error that affect observations. Inspired by the work presented in [BMB24], I tested the robustness of the model by training it on observations contaminated by Gaussian noise with zero mean and increasing standard deviation σ_{Noise}[3]. In Fig. 5.7, we can see that as σ_{Noise} increases, the characteristic features of the abundance plane are blurred out, making it harder to recognize distinct contributions from different accretion events.
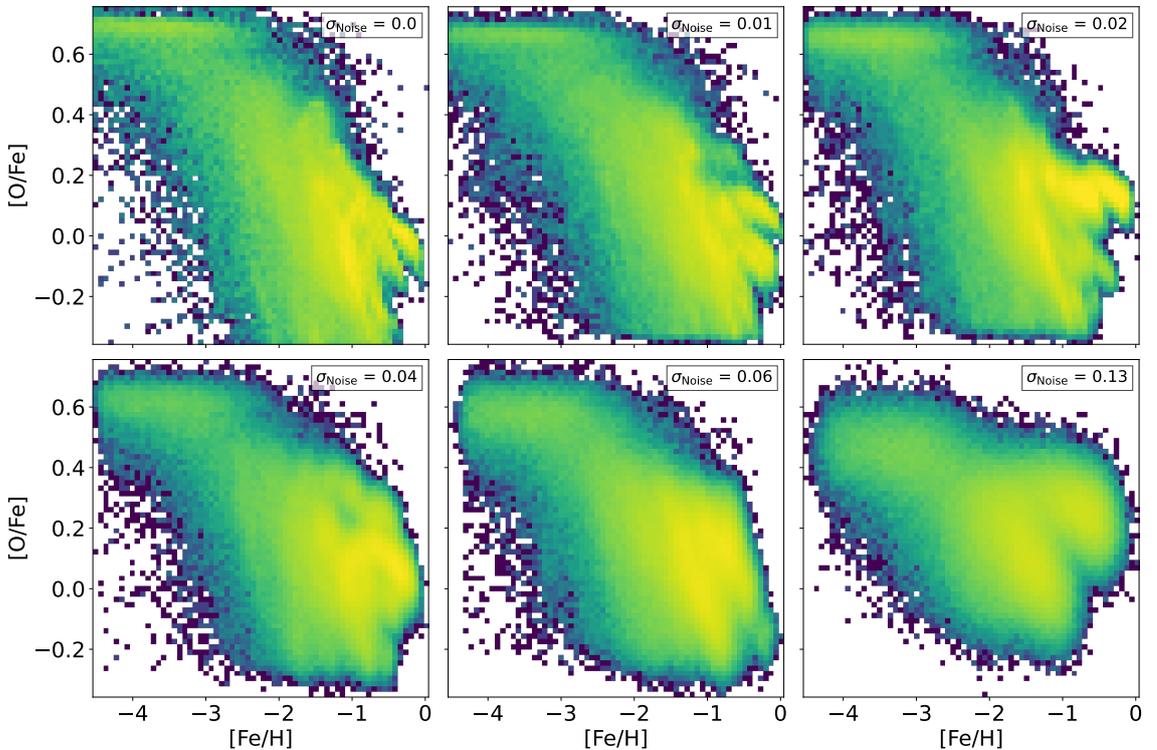
Figure 5.7: Logaritmic density distribution of particles in the abundance plane for the first test set galaxy halo (halo index $j = 0$) of `template library` generated with different levels of observational noise σ_{Noise}.

For each of the galaxy in Fig. 5.7, I report in Fig. 5.8 the posterior samples of the star mass of a few subhalo that populate those galactic halo. In this example, we can appreciate that the true

---

[3]I generated a new `template library` for each σ_{Noise}
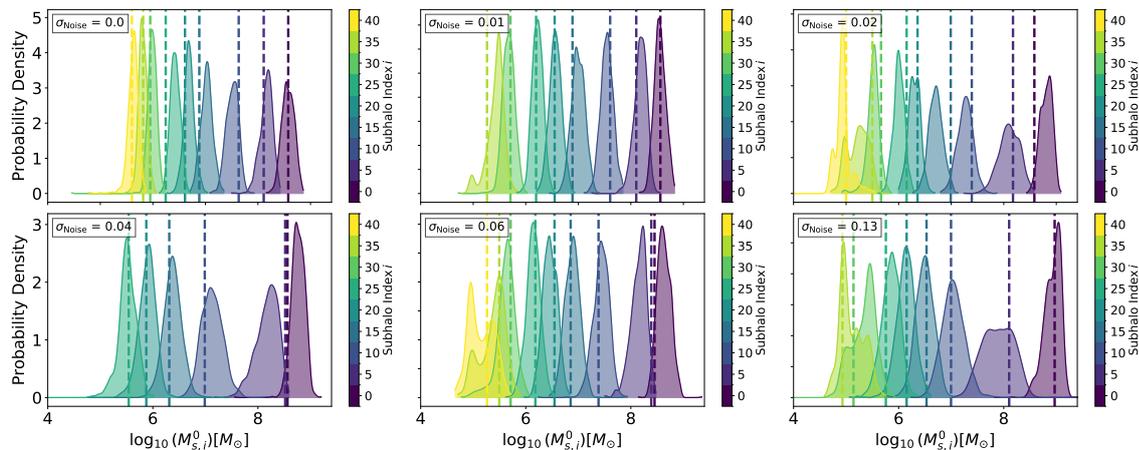
Figure 5.8: Posterior of the stellar mass for subhalos with index i belonging to the first stellar halo of the test set ($j = 0$), for each of the `template library` generated with different levels of observational noise σ_{Noise}. The vertical lines show the true values for each subhalo.

value for each of the subhalos falls close to the posterior estimate (i.e., the posterior mean) and that posterior predictions are relatively peaked around the true values. Also, it is clear that the model is able to recognize that a higher value of i corresponds to a distribution of samples that is shifted towards lower values of the star mass.

In the appendix, I report the same `ltu-ili` plots of Fig. 5.4, 5.5 and 5.6 for each of the noise levels that I have tested. All results are obtained on the test set of each `template library`. A general trend seems to emerge as the noise increases: the model becomes overconfident. From the accuracy plots, we can clearly see that the model starts to be biased toward inferring smaller stellar masses. This is actually expected since the peculiar and unique feature of major merger starts to become more and more blurred.

5.4 Subhalo level accuracy and uncertainty

Although all the performances on a subhalo level can be appreciated just by examining the appendix plots, I decided to give some average accuracy and uncertainty metrics over the test set to highlight the general trends that are visible when varying the noise level. The accuracy is calculated as the test set average of the mean absolute error (MAE) and the root mean square error (RMSE) between
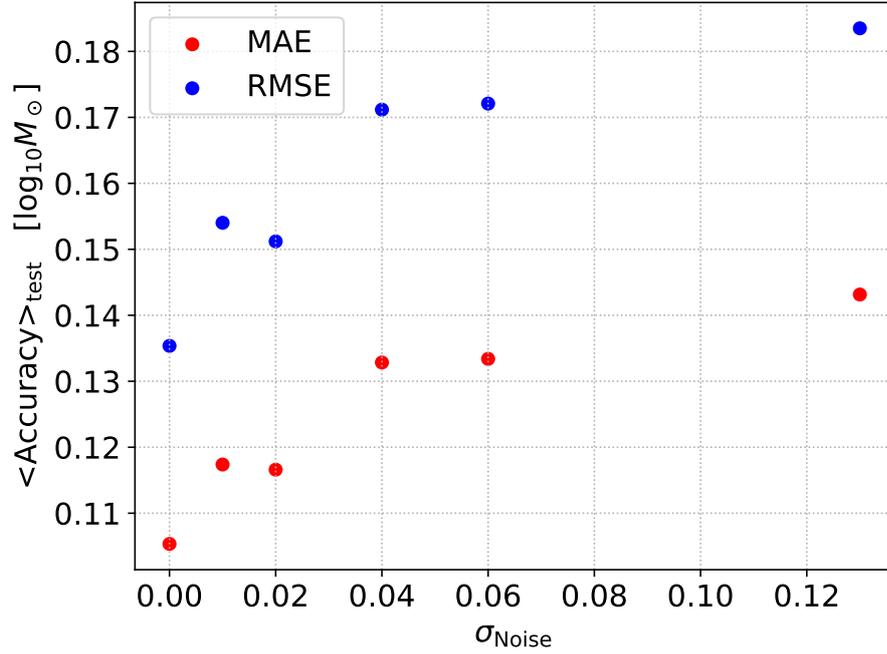
Figure 5.9: Average accuracy metric over the test set for different noise level.

the sample average and the true value of the stellar mass. From Fig. 5.9 we can see the expected degradation of accuracy with the noise level.

The uncertainty is calculated using the the test set average of the samples' standard deviation[4], the 68% and 99% percentiles. Also in this case the performances are degraded by the noise, especially when we consider the 99% percentile.

---

[4]I approximate the posterior with a Gaussian, reasonable assumption sustained by the spot check plot presented in Fig. 5.8.
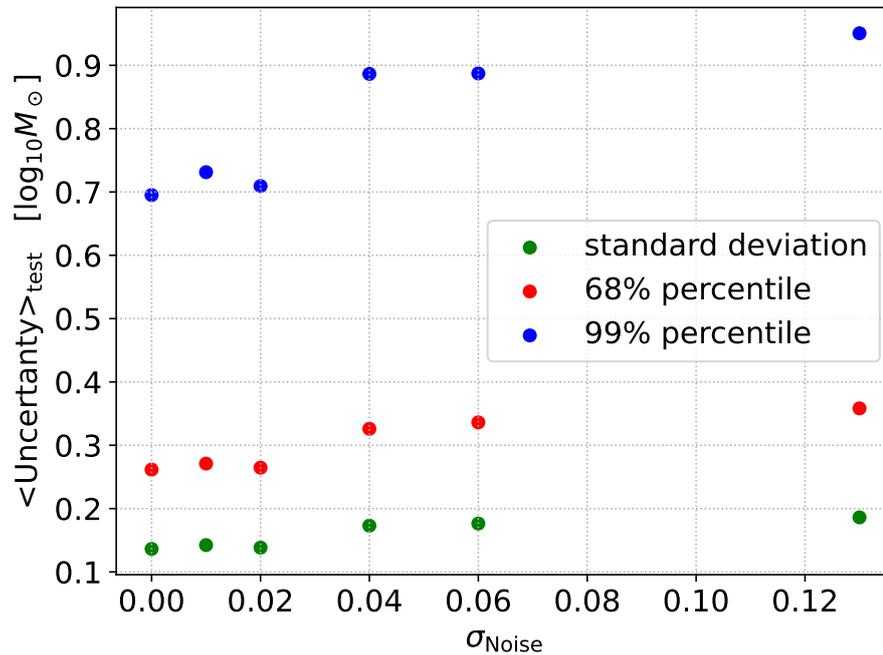
Figure 5.10: Average uncertanty metric over the test set for different noise level.

5.5 Galactic halo reconstruction

To further understand the possible biases introduced by training on noisy data, I decided to study inference at the halo level. Given a certain galactic halo x^j I try to understand how well we can reconstruct its subhalo stellar mass distribution. The test set average of the Kullback-Lieber (KL) divergence and Jensen-Shannon (JS) entropy between the distribution of the sample estimate and the true values are presented in Figs. 5.11 and 5.12 respectively. Once again, as the chemical feature gets blurred, the performance to be able to disentangle the contribution from different subhalos in a given galaxy becomes more challenging.
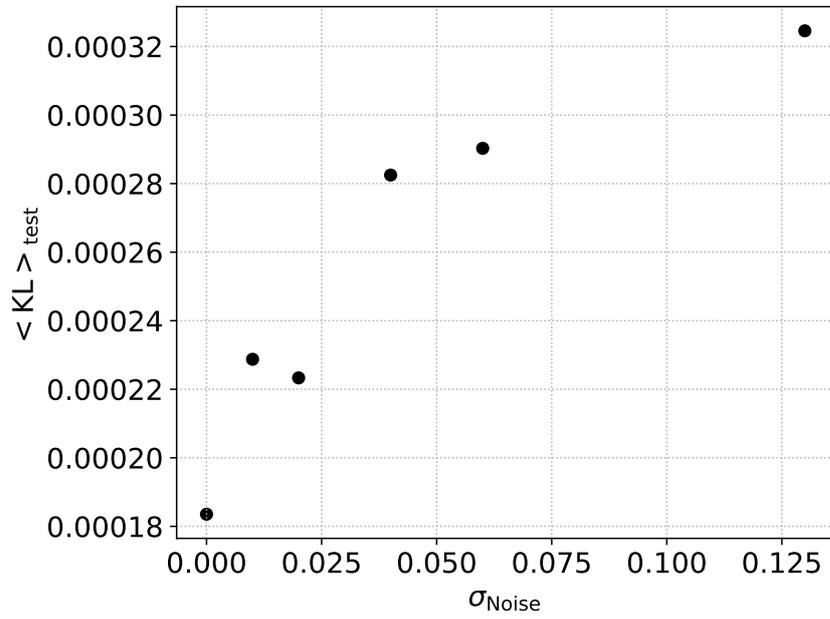
Figure 5.11: Test set KL divergence average between the distribution of sample estimate and the true values.
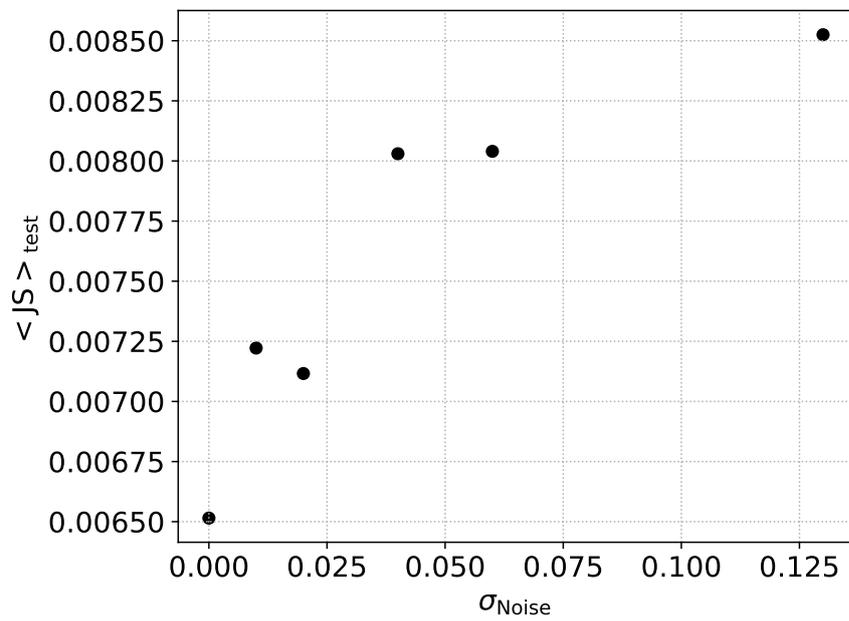
Figure 5.12: Test set JS entropy average between the distribution of sample estimate and the true values.

Chapter 6

Conclusion

I find that SBI can be applied to the complex task of parameter estimation for accreted dwarf galaxies in the MW, leveraging only the chemical information. This allows for the inference of infall dwarf galaxies properties even in the case of fully phase-mixed objects in the Galactic halo and opens up a completely new avenue for Galactic Archaeology. The information of multidimensional chemical abundance distributions is able to guide the model towards a well-calibrated and accurate reconstruction of the stellar mass of accreted dwarf galaxies when tested on noise-free data. Moreover, it is clear that after adding an observational noise σ_{Noise} higher than 0.01 dex, the pipeline starts to be affected by the blurring of the information and becomes biased to inferring smaller stellar masses. Furthermore, CASBI avoids the need to bin the subhalo information as in [Dea+23]. In this way with CASBI it is possible to directly perform Bayesian inference on the parameters of each individual subhalo. One limitation of this method is the simplified scenario in which I have tested CASBI. I have assumed to be able to perfectly remove background and foreground stars and perform inference on a clean halo sample. Additionally, no observational selection function has yet been assumed. These challenges will be implemented in the 'Observational Realism' component of the CASBI pipeline as future work (see Fig. 4.3). Moreover, the poor performance on the infall time might need to add either more α element abundances, or the dynamical information of the action angle plane. Lastly, the final goal of this prof-of-concept work would be to be applied to real spectroscopic observations.

The code for CASBI, including extensive documentation, is available on GitHub [1], as well as the

data [2] to reproduce the results of this work.

---

[1] https://github.com/vepe99/CASBI
[2] https://zenodo.org/records/13730400

Appendix A

Early Test

In this Appendix are reported two CASBI integration that are not present in the final structure of the package. In A.1 I tested whether the use of surrogate simulator for the sequential version of the NPE could allow the pipeline to achieve better results. In A.2 are reported the initial solution to fix the dimensionality of the prior.

A.1   Free Form Flow as a surrogate simulator

If new data pairs are generated at inference time, by sampling the prior, passing those samples to the simulator, and sequentially repeating the inference, it is possible to achieve better accuracy, as shown empirically in Fig. A.1, at the cost of losing the **amortize** property[1] .

---

[1]The amortize property is defined as the possibility to infer the posterior of a given observation without the need to simulate new data pairs.
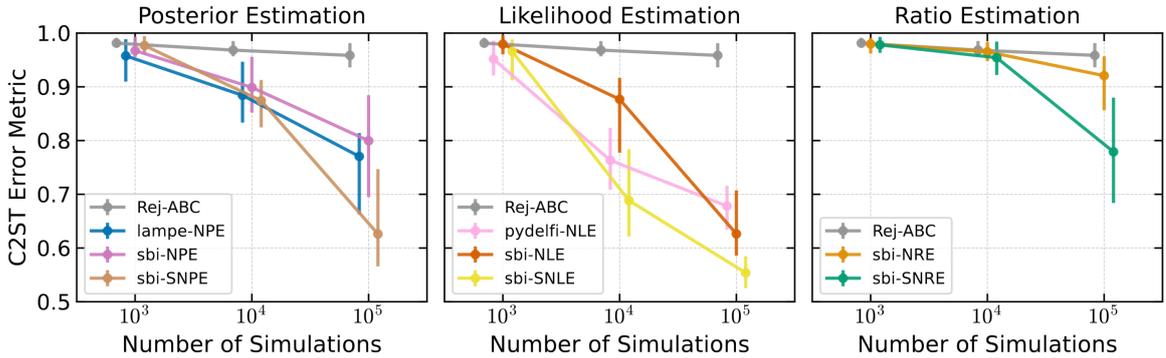
Figure A.1: Different SBI method accuracy over 10 independent test with similar simulation budget. The **S** stends for the Sequential version implementation of corresponding NDE used. The error is defined in terms of the Classifier Two Sample Test (C2ST) [LO18], which is defined as the accuracy of a classifier to distinguish between true and inferred posterior samples, with a C2ST value of 0.5 implying that the two sampled distributions are the same. This Figure is taken from [Ho+24].

Inspired by this result, I have decided to explore the possibility of implement Sequential Neural Posterior Estimate (SNPE) in CASBI. The use of the simulator at inference time is not usually possible for cosmological application; in fact, running those simulations is both computationally and time consuming. I have decided to implement a surrogate simulator that aims to learn how to sample new observations x from the Likelihood probability $p(x|\theta)$. The surrogate simulator that I adopted is the Free Form Flow (FFF) [Dra+24] model, a Normalizing Flow architecture that relaxes the need for invertible transformations and only requires dimensionality preservation at each step. After trying to use the more common Neural Spline Flow (NSF), as in [WB23], I decided to adopt FFF due to higher fidelity in the generation of new observations x, using the D-statistic [LRH] of the two-dimensional Kolmogorov-Smirnov test as fidelity metric. The architecture of the FFF is composed of an encoder and a decoder that take the role respectively of the forward transformation f_θ to the normally distributed latent space, and an approximation of the inverse transformation g_ϕ that maps the latent space to the observation space. The major innovations in Free Form Flow that are needed to understand the choice and flexibility of this architecture are:

- Gradient Trick: since the most computationally expensive part of the loss function of normal-izing flow is calculating the Jacobian of the transformation, the authors propose to estimate its gradient using a pair of vector-Jacobian and Jacobian-vector products easily available in

standard automatic differentiation libraries. The gradient trick is implement in the pipeline by rewriting the maximum likelihood loss derived from equation 3.4 as:

$$\mathcal{L}_{ML}^{f^{-1}} = \mathbb{E}_{x,v}[-\log p(f_\theta(x)) - v^T J_\theta \text{SG}(J^{-1}v)], \tag{A.1}$$

where v is a random vector with unit variance with the same dimensions as x, and SG is the Stop Gradient operation. This loss enable to train normalizing flow architecture with a tractable inverse function whose Jacobian determinant is not easily accessible.

- Inverse Approximation: classical normalizing flow architecture requires the access to the analytic inverse of the transformations f_θ^{-1}, either by constructing Invertible Neural Network or defining the flow with a differential equation with a known reverse time process. The authors propose to approximate the inverse with a learned inverse $g_\phi \approx f_\theta^{-1}$. The loss function is then modified to learn this approximation with the following contribution, called reconstruction loss:

$$\mathcal{L}_R = \frac{1}{2}\mathbb{E}_x[||x - g_\phi(f_\theta^{-1}(x))||^2]. \tag{A.2}$$

This part allows to remove the architectural constrains from f_θ and g_ϕ except for preserving the dimensions.

Combining both contributions from equations A.1 and A.2 leads to the following loss function:

$$\mathcal{L}_{FFF}^g = \mathcal{L}_{ML}^g + \beta\mathcal{L}_R, \tag{A.3}$$

where the $\mathcal{L}_{ML}^g$ is used in place of the $\mathcal{L}_{ML}^{f^{-1}}$ with the justification that they have the same critical points and the β is a trade off hyperparameter. For a more in depth explanation and a mathematical foundation of FFF architecture I refer to the original paper [Dra+24].

In CASBI, I have chosen to use a Skip Connection Multi Layer Perceptron (SC-MLP) as both encoder and decoder, and I have followed the suggestions in Appendix B.1 of [Dra+24] to make this architecture conditional by concatenating to each layer the parameters θ sampled from the prior distributions. Although the FFF architecture has good interpolation capability, returning

average D-static values lower than 0.3 when reconstructing the test set, the problem of sampling independent parameter values from the prior distributions makes the net extrapolate in regions of the conditional space where no data were shown, degrading the generate abundance subhalo. For example, it could happen that a very large M_{star} is sampled together with a very low infall time τ, which is a combination of parameters that are not physical and hence was not shown during training, so the generated halo is not realistic, that is, it has a very high D-statistic value. Due to the general poorer performance of SNPE using the FFF as a surrogate simulator, I have decided not to include this architecture in the analysis. Future work could incorporate some level of correlation in the prior distribution of the parameters, since currently the `ltu-ili` package allows only for independent priors.

A.2   Two step Inference

I decided to tackle the problem of fixing the dimensionality of the prior in the case of fully mixed remnants separating the inference in two steps, in the first it infers the number of subhalos and in the second the parameters of each of the subhalos:

1. **Inference of the number of substructure**: In this step I train a NPE to recover the posterior distribution of the number of substructure N_{sub}, by using the observable $\mathbf{x^j}$. The prior for the parameter is assumed to be uniform between 2 and 100. This boundaries were selected in accordance to the order of magnitude of substructures found in [Dea+23]. For each of the possible N_{sub} I extract 1000 $x^j = \sum_{i=1}^{N_{sub}} x_i^j$ from the **NIHAO** simulations, in order to have a total of almost 10^5 SBI training couples (N_{sub}^j, x^j), with 20% used as validation. I used the same process to generate almost 10^4 test set samples, making sure that the same combinations of `Galaxy_name` attribute weren't shown in training and test. The training of the NPE is done using the `sbi` backend, using 4 `nsf` (neural spline flow) with 10 layers and 100 neurons each. In order to take full advantage of the image-like structure of the data, I adopt as embedding network a Convolutional Neural Network (CNN) to reduce the dimensionality of the input of the NPE from 64×64 to 128. The CNN have 3 convolutional layers with 8, 16 and 32 filter, 3 Max Pooling layers and 3 fully connected layers with 512, 256, 128 neurons.

In this step I have not impose that the N_{sub} must be a discrete variable, and I have decide to just truncate the inferred value to the closest integer. To the knowledge of the author no SBI framework has implemented a way of dealing with the inference of discrete random variables, so I leave a more precise implementation as a future work. I propose instead another method to obtain the number of substructure, by casting this inference as a classification problem. I use a SkipConnection CNN [2], considering the number of substructure as the label to assign to each x^j.

2. **Inference of $\theta^{\mathbf{j}}$:** Once I have the estimate $\tilde{N}_{sub}$, whether using dynamical information, the inference pipeline, or the classification method, I can proceed to the inference of parameters $\theta_{\mathbf{i}}^{\mathbf{j}}$. The priors for the parameters are assumed to be uniform between the minimum and maximum values available for the galaxies that I have filtered from the **NIHAO** simulations. I extract 10^5 random samples of $\tilde{N}_{sub}$ snapshots from the **NIHAO** simulations, and I construct the observable couples $(x^j, (\theta_1^j, \ldots, \theta_{\tilde{N}_{sub}}^j))$, with 20 % used as validation. I repeat the same process to generate 10^3 test set samples, making sure that the same combinations of `Galaxy_name` attribute weren't shown in training and test to perform calibration of the inference model. The NPE training is performed using the `sbi` backend, using 4 `nsf` (neural spline flow) with 10 layers and 100 neurons each. Once again I use the same CNN architecture of the previous step as embedding for the observation x^j.

Even though highly modular, this two-step inference has some limitations: the accuracy and calibration of the second step are heavily dependent on the ability of the first step to recover the number of sunhalos and hence to constrain the dimensionality of the prior for the second step. I expected the pipeline to be able to recover most of the information from the most massive subhalos, thanks to their more distinct feature in the abundance plane. The second problem is related to the linear scaling of the parameter dimension as a function of the number of subhalos. In a realistic case, I expect to have an order of ≈ 100 subhalos, resulting in a parameter space of dimension 2*100 = 200. Although I expect SBI to be able to handle high-dimensionality inference, it is clear that for each $\tilde{N}_{sub}$ a new training is required with its own hyperparameter tuning phase, resulting in the end in

---

[2]The architecture is the same as the embedding network described before with the addition of the Skip Connection layer in the fully connected layer, where the output of the previous layer gets added to the output before being passed trough the activation function, alleviating the vanishing gradient problem and allowing for better accuracy.

a collection of pre-trained models to be used depending on the number of subhalos[3].

In order to avoid this intense and time-consuming phase, I rethought the inference pipeline, which I present in section 4.2.

---

[3]Another possibility would be to limit the inference to the first M most massive subhalos, were I fix M and I assume that the remaining subhalos do not contribute significantly to the galactic halo.

Appendix B

Accuracy

In the following appendix I report the accuracy plots from the `ltu-ili` package for the test set with different noise levels σ_{Noise}.
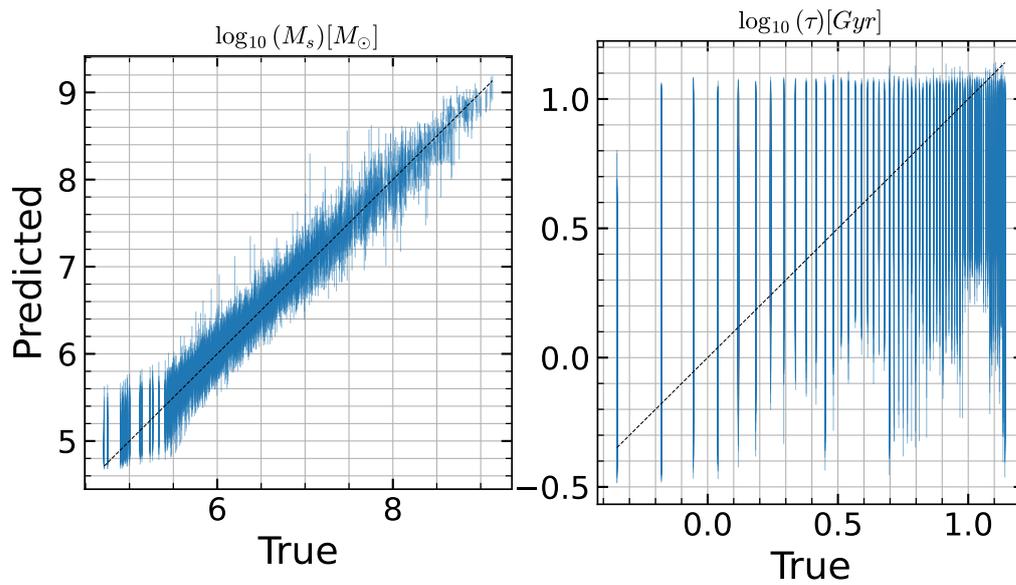
Figure B.1: Predicted vs True values with error bars representing the 68% confidence interval. The noise level is $\sigma_{Noise} = 0.01$
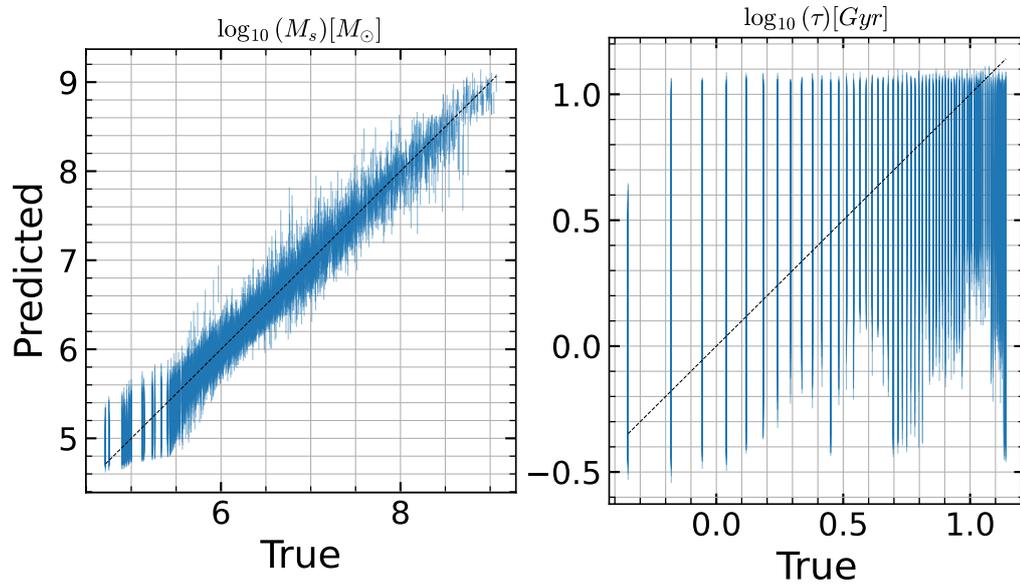
Figure B.2: Predicted vs True values with error bars representing the 68% confidence interval. The noise level is $\sigma_{Noise} = 0.02$
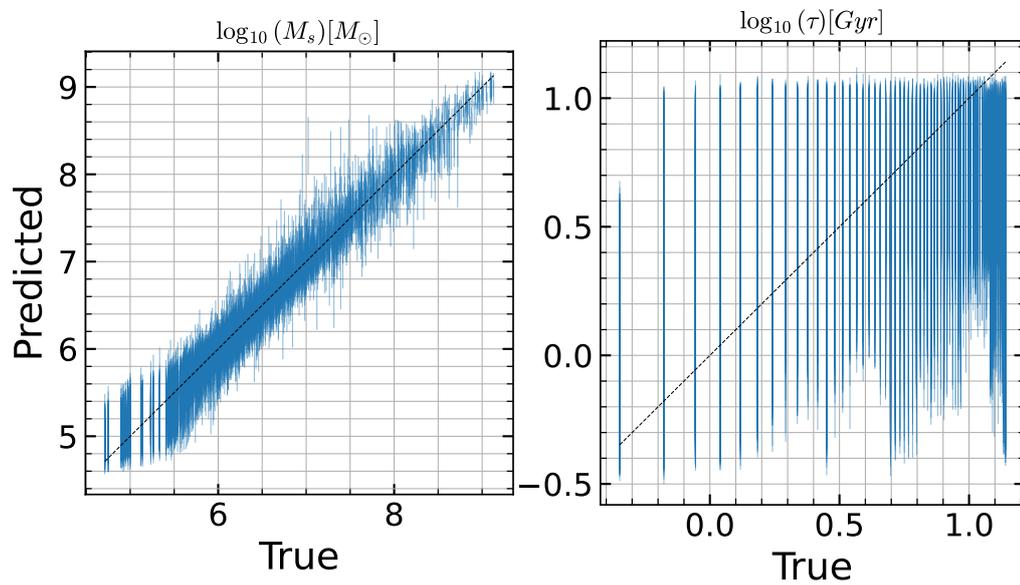
Figure B.3: Predicted vs True values with error bars representing the 68% confidence interval. The noise level is $\sigma_{Noise} = 0.04$
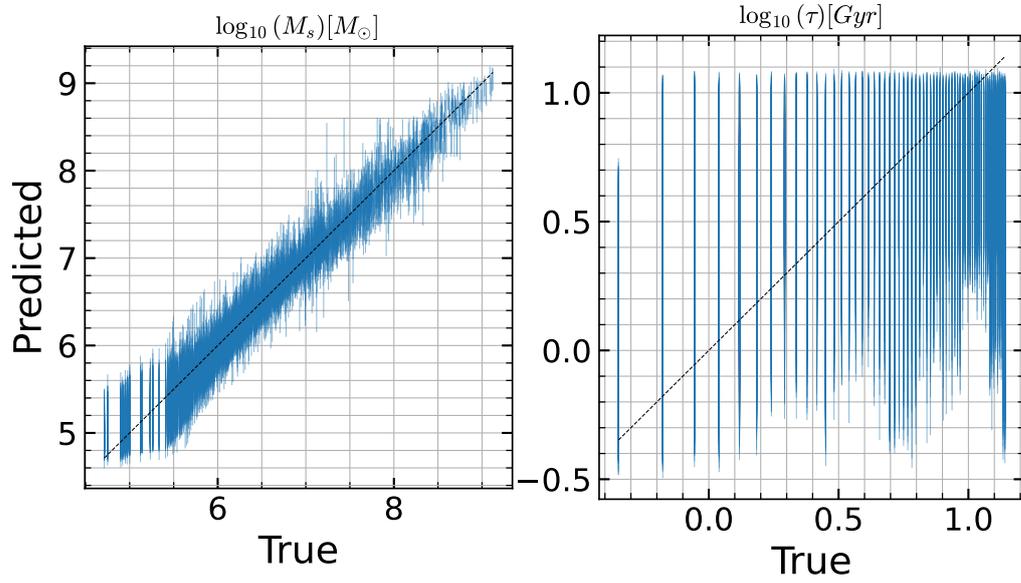
Figure B.4: Predicted vs True values with error bars representing the 68% confidence interval. The noise level is $\sigma_{Noise} = 0.06$
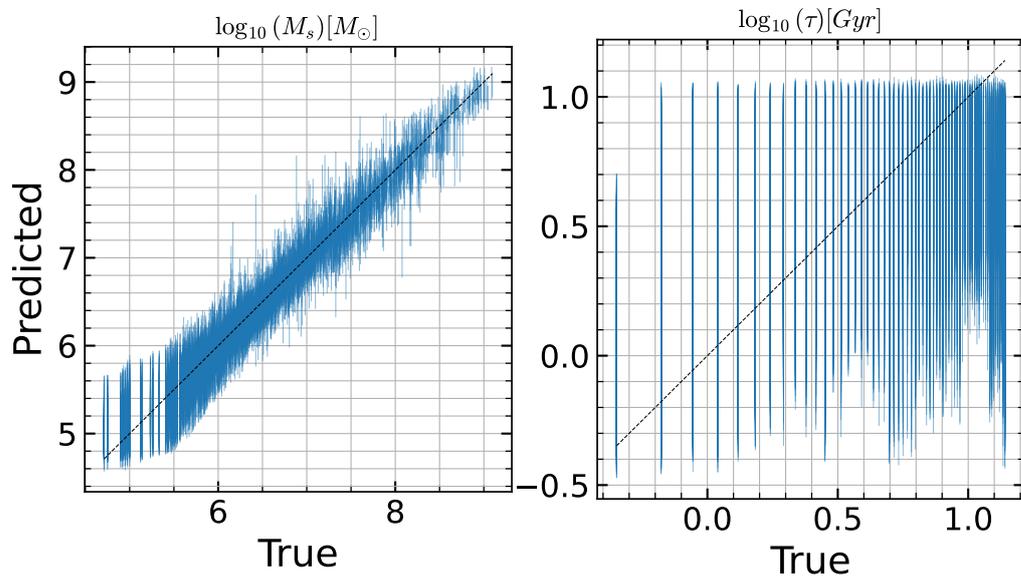
Figure B.5: Predicted vs True values with error bars representing the 68% confidence interval. The noise level is $\sigma_{Noise} = 0.13$

Appendix C

Coverage

In the following appendix I report the marginal coverage plots from the `ltu-ili` package for the test set with different noise levels σ_{Noise}.
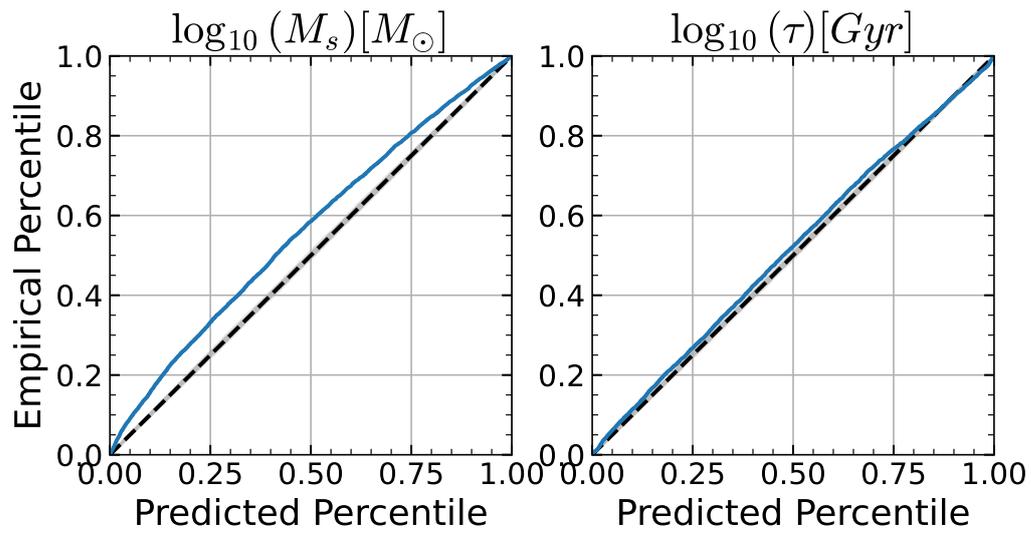
Figure C.1: Coverage for the marginal probabilities. The noise level is $\sigma_{Noise} = 0.01$
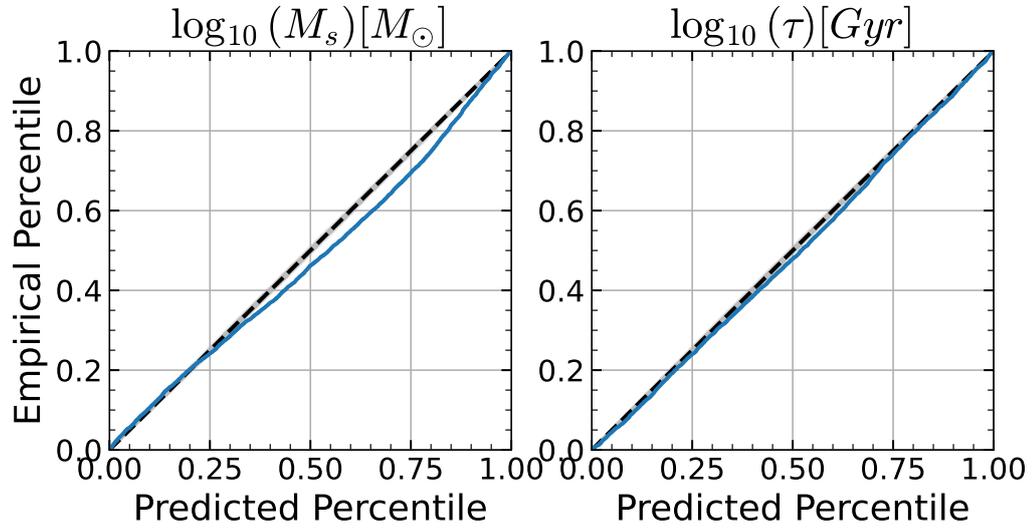
Figure C.2: Coverage for the marginal probabilities. The noise level is $\sigma_{Noise} = 0.02$
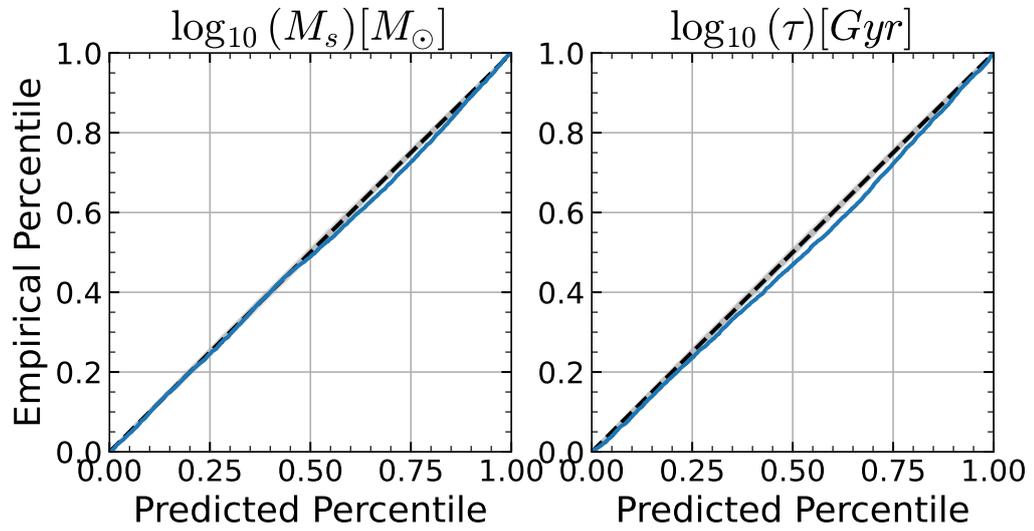
Figure C.3: Coverage for the marginal probabilities. The noise level is $\sigma_{Noise} = 0.04$
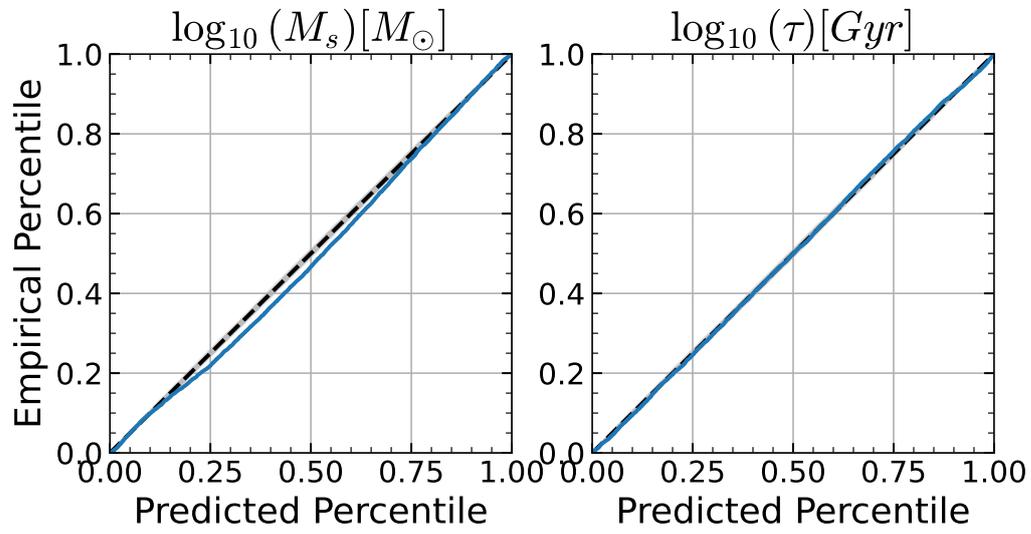
Figure C.4: Coverage for the marginal probabilities. The noise level is $\sigma_{Noise} = 0.06$
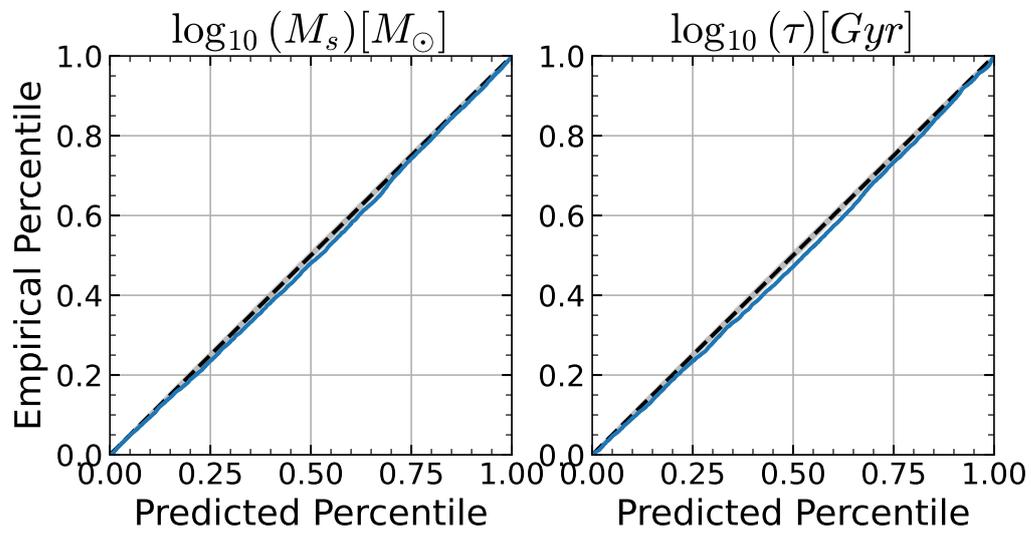
Figure C.5: Coverage for the marginal probabilities. The noise level is $\sigma_{Noise} = 0.13$

Appendix D

TARP

In the following appendix I report the joint coverage plots (TARP) from the `ltu-ili` package for the test set with different noise levels σ_{Noise}.
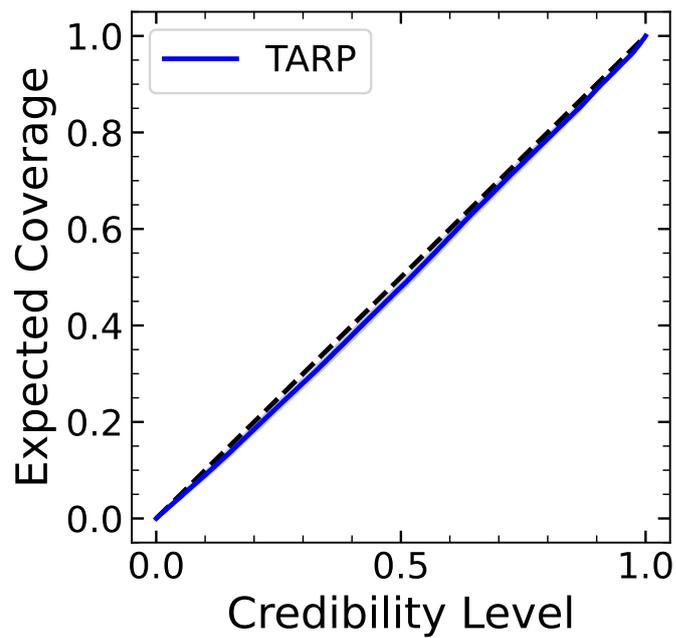
Figure D.1: TARP percentile-percentile plot. Approximation of coverage plot for the joint distribution. The noise level is $\sigma_{Noise} = 0.01$
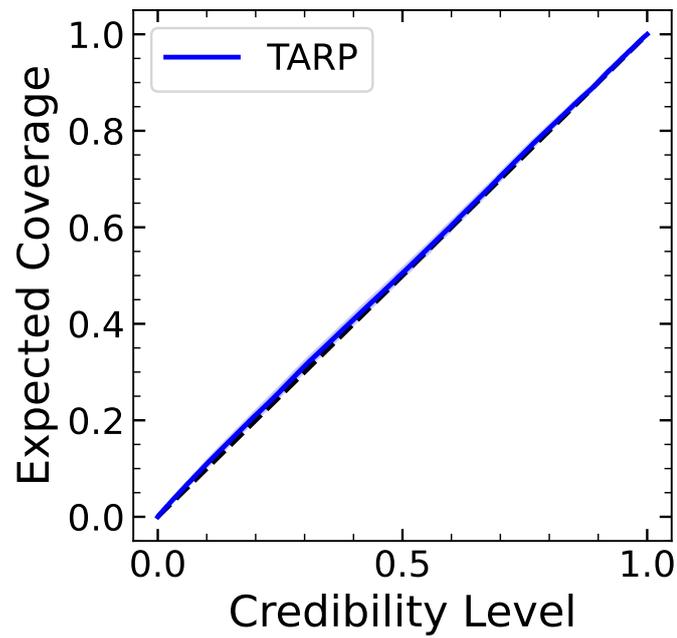
Figure D.2: TARP percentile-percentile plot. Approximation of coverage plot for the joint distribution. The noise level is $\sigma_{Noise} = 0.02$
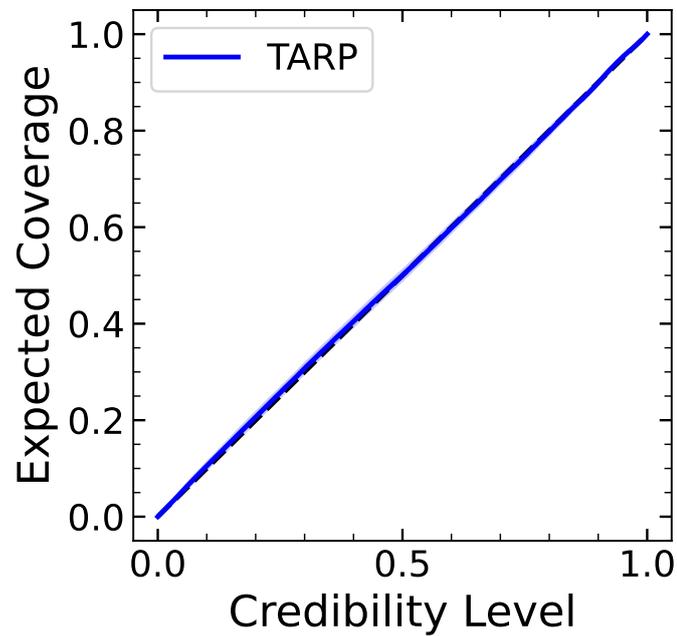
Figure D.3: TARP percentile-percentile plot. Approximation of coverage plot for the joint distribution. The noise level is $\sigma_{Noise} = 0.04$
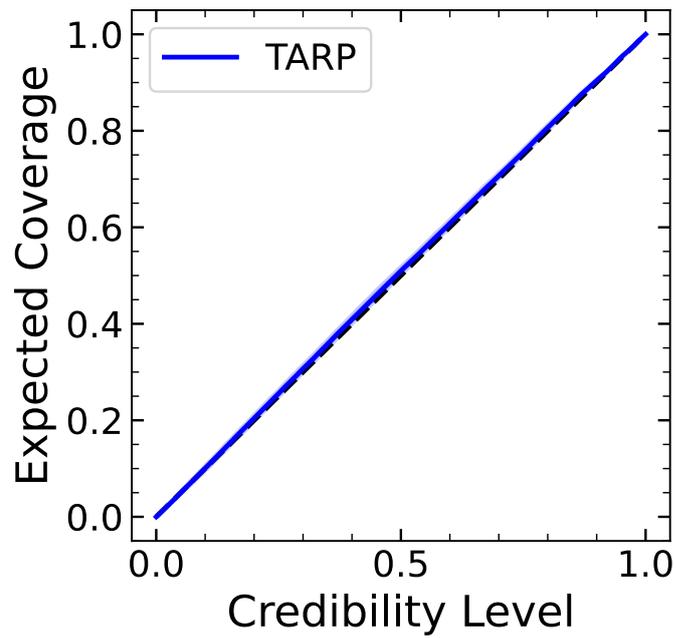
Figure D.4: TARP percentile-percentile plot. Approximation of coverage plot for the joint distribution. The noise level is $\sigma_{Noise} = 0.06$
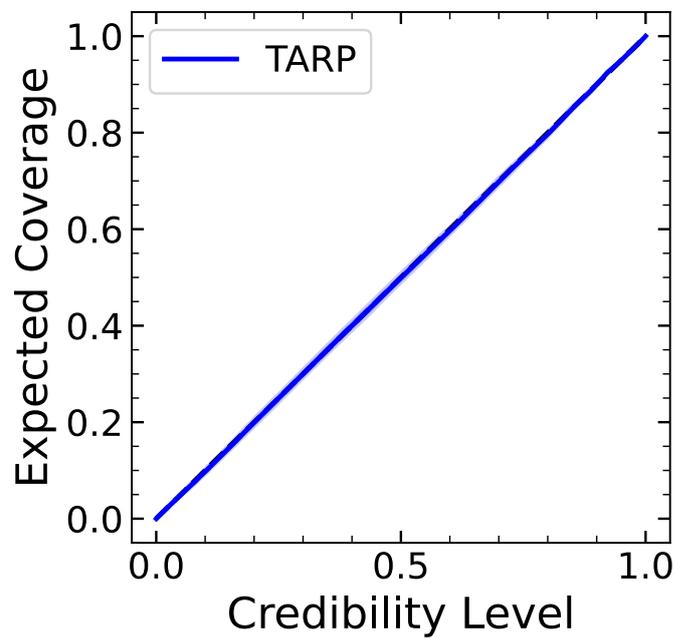
Figure D.5: TARP percentile-percentile plot. Approximation of coverage plot for the joint distribution. The noise level is $\sigma_{Noise} = 0.13$

Bibliography

[Bel+18]   V Belokurov et al. "Co-Formation of the Disc and the Stellar Halo". In: *Monthly Notices of the Royal Astronomical Society* 478.1 (July 2018), pp. 611–619. ISSN: 0035-8711, 1365-2966. DOI: 10.1093/mnras/sty982. (Visited on 06/12/2024).

[BMB24]   Sven Buder et al. *Finding Accreted Stars in the Milky Way: Clues from NIHAO Simulations*. Apr. 2024. arXiv: 2404.13835 [astro-ph]. (Visited on 09/16/2024).

[BT08]   J. Binney et al. *Galactic Dynamics: Second Edition*. Princeton Series in Astrophysics. Princeton University Press, 2008. ISBN: 978-0-691-13027-9.

[Buc+23]   Tobias Buck et al. "The impact of early massive mergers on the chemical evolution of Milky Way-like galaxies: insights from NIHAO-UHD simulations". In: *Monthly Notices of the Royal Astronomical Society* 523.1 (July 2023), pp. 1565–1576. DOI: 10.1093/mnras/stad1503. arXiv: 2305.13759 [astro-ph.GA].

[Col+18]   Gaia Collaboration et al. "Gaia Data Release 2: Observational Hertzsprung-Russell Diagrams". In: *Astronomy & Astrophysics* 616 (Aug. 2018), A10. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201832843. arXiv: 1804.09378 [astro-ph]. (Visited on 09/26/2024).

[Con+19]   Charlie Conroy et al. "Mapping the Stellar Halo with the H3 Spectroscopic Survey". In: *The Astrophysical Journal* 883.1 (Sept. 2019), p. 107. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/ab38b8. (Visited on 06/12/2024).

[Cui+12]   Xiang-Qun Cui et al. "The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)". In: *Research in Astronomy and Astrophysics* 12.9 (Sept. 2012), pp. 1197–1242. ISSN: 1674-4527. DOI: 10.1088/1674-4527/12/9/003. (Visited on 06/12/2024).

[Cun+22]   Emily C. Cunningham et al. "Reading the CARDs: The Imprint of Accretion History in the Chemical Abundances of the Milky Way's Stellar Halo". In: *The Astrophysical Journal* 934.2 (Aug. 2022), p. 172. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/ac78ea. (Visited on 03/06/2024).

[Dax+23]   Maximilian Dax et al. *Flow Matching for Scalable Simulation-Based Inference*. Oct. 2023. arXiv: 2305.17161 [cs]. (Visited on 03/28/2024).

[DB24]   Alis J. Deason et al. "Galactic Archaeology with Gaia". In: *New Astronomy Reviews* 99 (Dec. 2024), p. 101706. ISSN: 13876473. DOI: 10.1016/j.newar.2024.101706. arXiv: 2402.12443 [astro-ph]. (Visited on 09/25/2024).

[DBS19]   Alis J. Deason et al. "The Total Stellar Halo Mass of the Milky Way". In: *Monthly Notices of the Royal Astronomical Society* 490.3 (Dec. 2019), pp. 3426–3439. ISSN: 0035-8711, 1365-2966. DOI: 10.1093/mnras/stz2793. arXiv: 1908.02763 [astro-ph]. (Visited on 06/14/2024).

[De +15]   G. M. De Silva et al. "The GALAH Survey: Scientific Motivation". In: *Monthly Notices of the Royal Astronomical Society* 449.3 (May 2015), pp. 2604–2617. ISSN: 1365-2966, 0035-8711. DOI: 10.1093/mnras/stv327. (Visited on 06/12/2024).

[Dea+23]   Alis J Deason et al. "Unravelling the Mass Spectrum of Destroyed Dwarf Galaxies with the Metallicity Distribution Function". In: *Monthly Notices of the Royal Astronomical Society* 520.4 (Feb. 2023), pp. 6091–6103. ISSN: 0035-8711, 1365-2966. DOI: 10.1093/mnras/stad535. (Visited on 03/06/2024).

[Dra+24]   Felix Draxler et al. *Free-Form Flows: Make Any Architecture a Normalizing Flow*. Apr. 2024. arXiv: 2310.16624 [cs, stat]. (Visited on 05/08/2024).

[Fal+]   Maciej Falkiewicz et al. "Calibrating Neural Simulation-Based Inference with Differentiable Coverage Probability". In: ().

[Hel+18] Amina Helmi et al. "The Merger That Led to the Formation of the Milky Way's Inner Stellar Halo and Thick Disk". In: *Nature* 563.7729 (Nov. 2018), pp. 85–88. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0625-x. (Visited on 06/12/2024).

[Ho+24] Matthew Ho et al. *LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology.* Feb. 2024. arXiv: 2402.05137 [astro-ph]. (Visited on 03/06/2024).

[KD18] Diederik P. Kingma et al. *Glow: Generative Flow with Invertible 1x1 Convolutions.* July 2018. arXiv: 1807.03039 [cs, stat]. (Visited on 03/13/2024).

[Kir+11] Evan N. Kirby et al. "MULTI-ELEMENT ABUNDANCE MEASUREMENTS FROM MEDIUM-RESOLUTION SPECTRA. III. METALLICITY DISTRIBUTIONS OF MILKY WAY DWARF SATELLITE GALAXIES". In: *The Astrophysical Journal* 727.2 (Feb. 2011), p. 78. ISSN: 0004-637X, 1538-4357. DOI: 10.1088/0004-637X/727/2/78. (Visited on 06/26/2024).

[Kir+13] Evan N. Kirby et al. "THE UNIVERSAL STELLAR MASS-STELLAR METALLICITY RELATION FOR DWARF GALAXIES". In: *The Astrophysical Journal* 779.2 (Nov. 2013), p. 102. ISSN: 0004-637X, 1538-4357. DOI: 10.1088/0004-637X/779/2/102. (Visited on 07/03/2024).

[KK09] Steffen R. Knollmann et al. "AHF: AMIGA'S HALO FINDER". In: *The Astrophysical Journal Supplement Series* 182.2 (June 2009), pp. 608–624. ISSN: 0067-0049, 1538-4365. DOI: 10.1088/0067-0049/182/2/608. (Visited on 09/27/2024).

[KM22] Andrey Kravtsov et al. "GRUMPY: A Simple Framework for Realistic Forward-Modelling of Dwarf Galaxies". In: *Monthly Notices of the Royal Astronomical Society* 514.2 (June 2022), pp. 2667–2691. ISSN: 0035-8711, 1365-2966. DOI: 10.1093/mnras/stac1439. arXiv: 2106.09724 [astro-ph]. (Visited on 05/22/2024).

[Kop+08] S. Koposov et al. "The Luminosity Function of the Milky Way Satellites". In: *The Astrophysical Journal* 686.1 (Oct. 2008), pp. 279–291. ISSN: 0004-637X, 1538-4357. DOI: 10.1086/589911. (Visited on 06/17/2024).

[Kyl20]   Gilles Louppe Kyle Cranmer Johann Brehmer. *The Frontier of Simulation-Based Infer-ence.* https://www.pnas.org/doi/10.1073/pnas.1912789117. 2020. DOI: 10.1073/pnas.1912789117. (Visited on 05/23/2024).

[Lem+23]   Pablo Lemos et al. "Sampling-Based Accuracy Testing of Posterior Estimators for General Inference". In: (2023).

[LO18]   David Lopez-Paz et al. *Revisiting Classifier Two-Sample Tests.* Mar. 2018. arXiv: 1610.06545 [stat]. (Visited on 10/05/2024).

[LRH]   Raul H C Lopes et al. "The Two-Dimensional Kolmogorov-Smirnov Test". In: ().

[Maj+17]   Steven R. Majewski et al. "The Apache Point Observatory Galactic Evolution Experiment (APOGEE)". In: *The Astronomical Journal* 154.3 (Sept. 2017), p. 94. ISSN: 0004-6256, 1538-3881. DOI: 10.3847/1538-3881/aa784d. (Visited on 06/12/2024).

[Mye+22]   G. C. Myeong et al. "Milky Way's Eccentric Constituents with $Gaia$, APOGEE $\&$ GALAH". In: *The Astrophysical Journal* 938.1 (Oct. 2022), p. 21. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/ac8d68. arXiv: 2206.07744 [astro-ph]. (Visited on 10/04/2024).

[Nai+20]   Rohan P. Naidu et al. "Evidence from the H3 Survey That the Stellar Halo Is Entirely Comprised of Substructure". In: *The Astrophysical Journal* 901.1 (Sept. 2020), p. 48. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/abaef4. (Visited on 06/26/2024).

[Pan+21]   Nondh Panithanpaisal et al. "The Galaxy Progenitors of Stellar Streams around Milky Way–Mass Galaxies in the FIRE Cosmological Simulations". In: *The Astrophysical Journal* 920.1 (Oct. 2021), p. 10. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/ac1109. (Visited on 06/26/2024).

[Pap19]   George Papamakarios. *Neural Density Estimation and Likelihood-free Inference.* Oct. 2019. arXiv: 1910.13233 [cs, stat]. (Visited on 06/12/2024).

[Pil+23]   Annalisa Pillepich et al. *Milky Way and Andromeda Analogs from the TNG50 Simulation.* Mar. 2023. arXiv: 2303.16217 [astro-ph]. (Visited on 06/28/2024).

[Pon+13]   Andrew Pontzen et al. "Pynbody: N-Body/SPH Analysis for Python". In: *Astrophysics Source Code Library* (May 2013), ascl:1305.002. (Visited on 06/11/2024).

[Rub84]   Donald B. Rubin. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician". In: *The Annals of Statistics* 12.4 (Dec. 1984), pp. 1151–1172. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176346785. (Visited on 06/12/2024).

[Ste+06]   M. Steinmetz et al. "The Radial Velocity Experiment (RAVE): First Data Release". In: *The Astronomical Journal* 132.4 (Oct. 2006), pp. 1645–1668. ISSN: 0004-6256, 1538-3881. DOI: 10.1086/506564. (Visited on 06/12/2024).

[Tol+08]   Erik J. Tollerud et al. "Hundreds of Milky Way Satellites? Luminosity Bias in the Satellite Luminosity Function". In: *The Astrophysical Journal* 688.1 (Nov. 2008), pp. 277–289. ISSN: 0004-637X, 1538-4357. DOI: 10.1086/592102. (Visited on 06/17/2024).

[Wag+21]   Sebastian Wagner-Carena et al. "Hierarchical Inference with Bayesian Neural Networks: An Application to Strong Gravitational Lensing". In: *The Astrophysical Journal* 909.2 (Mar. 2021), p. 187. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/abdf59. (Visited on 09/28/2024).

[Wan+15]   Liang Wang et al. "NIHAO Project I: Reproducing the Inefficiency of Galaxy Formation across Cosmic Time with a Large Sample of Cosmological Hydrodynamical Simulations". In: *Monthly Notices of the Royal Astronomical Society* 454.1 (Nov. 2015), pp. 83–94. ISSN: 0035-8711, 1365-2966. DOI: 10.1093/mnras/stv1937. arXiv: 1503.04818 [astro-ph]. (Visited on 03/06/2024).

[WB23]   Luca Wolf et al. *GalacticFlow: Learning a Generalized Representation of Galaxies with Normalizing Flows*. Dec. 2023. arXiv: 2312.06017 [astro-ph]. (Visited on 03/06/2024).

[Wet+16]   Andrew R. Wetzel et al. "RECONCILING DWARF GALAXIES WITH ΛCDM COSMOLOGY: SIMULATING A REALISTIC POPULATION OF SATELLITES AROUND A MILKY WAY–MASS GALAXY". In: *The Astrophysical Journal Letters* 827.2 (Aug. 2016), p. L23. ISSN: 2041-8205, 2041-8213. DOI: 10.3847/2041-8205/827/2/L23. (Visited on 06/12/2024).

[WSQ04]   James Wadsley et al. "Gasoline: An Adaptable Implementation of TreeSPH". In: *New Astronomy* 9.2 (Feb. 2004), pp. 137–158. ISSN: 13841076. DOI: 10.1016/j.newast.2003.08.004. arXiv: astro-ph/0303521. (Visited on 10/01/2024).

[Yan+09]   Brian Yanny et al. "SEGUE: A SPECTROSCOPIC SURVEY OF 240,000 STARS WITH $g = 14\text{-}20$". In: *The Astronomical Journal* 137.5 (May 2009), pp. 4377–4399. ISSN: 0004-6256, 1538-3881. DOI: 10.1088/0004-6256/137/5/4377. (Visited on 06/12/2024).

Acknowledgments

I would like to express my special thanks to Dr. Tobias Buck for his help in the creation of this project during my Erasmus in the Astro-Ai lab and also to Dr. William Oliver who helped shape the final version of CASBI. I learned so much during this experience and I am thankful to be able to continue working with them. I express my gratitude to Dr. Giuliano Iorio and Prof. Antonino Milone, who inspired me to pursue an academic career and guided me in the writing of this work. I would like to thank my family, Antonio, Vito and Luisa, for their constant support and love. Special thanks to my colleagues Marco, Tommaso and Giacomo of the "Three Blue One Brown" team, I will always carry the sleepless nights, broken gits, dead clusters and endless laughter in my heart. Honorable mention to Giacomo, more of a brother than a friend, for pushing me far beyond what I thought I deserved, I will never again be as lucky as the day I decided to put Mollymauk as my lock screen. Deep thanks to Gabriele, representative, colleague, friend of 1001 misfortunes in our academic journey, without him this thesis would have had more typos than content, and I hope to be able to work with him again. A big and loving hug to my bachelor friends Antonella, Lorenzo, Michele, Carmen, Giovanni, Giancarlo, and Mattia, thank you for putting up with me and supporting me after all this time. I am extremely grateful for the friendship, the experience and all the fond memories that I have with Davide, Valerio, and Tommaso, you really thought me how to "paddle your own canoe". I thank my group of firends Lisa, Chiara, Giorgia, Nicholas, Leonardo, Sofia, Alice, Gioele, and Alberto for the laughter and the lighthearted moments that we have shared. A huge thank you to my Dungeon and Dragons group Carlo, Giovanni, Jacopo, Antonio, Valerio, and Luca, you heard me do silly voices and allowed me to express all seven of my personalities at once, the moments spent with you meant much more than you could imagine, and I hope to get

71

back behind the screen as soon as possible.

Last but certainly not least, I would like to deeply thank Paola, my lifetime partner in crime, my biggest fan, my dearest friend, for giving me the strength to carry on, the love to make all of this possible, and the inspiration to become a better person. I am really proud to have you still by my side and I thank you for everything.