CASBI: Chemical Abundance Simulation Based Inference

Giuseppe Viterbo

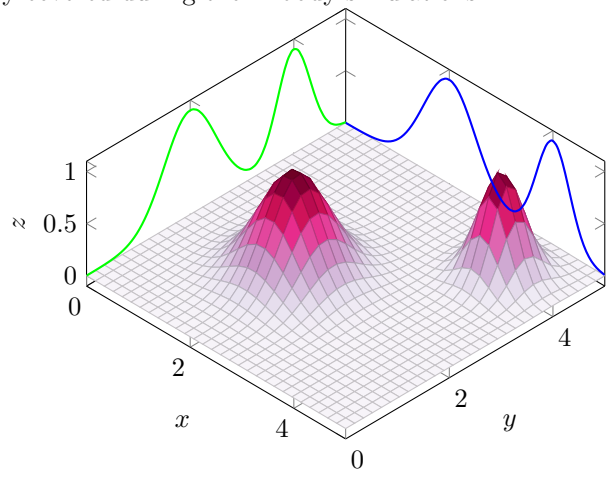
July 2, 2024

Contents

Chapter 1

Abstract

Galaxies evolve trough merging events and destroy lower-mass systems over their lifetimes. The contribution that those lower-mass system brings to the modern picture has been freezes in stellar halos by the long orbital timescales, making the relicts of this objects retain part of their initial progenitor orbit. But dynamical information is not enough to disentangle these components, and the complementary chemical information helps to characterized these building blocks. In fact, merging events tend to quench the star formation rate of these objects, making the chemical abundance plane (iron abundance against α element abundance) a distinct imprint that retains information on the conditions of formation of their stars, like the total mass and the age of the system until the merging event. These theoretical background allows us to attempt to decompose the stellar halo into its components, unraveling the merging history. In the modern era of large N-body galaxy simulations, we recast this problem into an SBI pipeline to recover the properties of this building block, (e.g. total stellar mass, infall time, ...) using the chemical abundance plane as observables. We therefore present CASBI (Chemical Abundance Simulation Based Inferece), a python package to recover the posterior probability of properties of building blocks of Milky Way like galaxy's halo. Moreover, CASBI incorporate conditional neural network architectures as generator to obtain observables from parameters, smoothly interpolating on region of the parameters space that weren't fully covered during the N-body simulations.



Chapter 2

Previous work

2.1 The reconstruction of the Assembly history of the Milky Way

Inferring the assembly history of the Milky Way is a challenging task, even in the era of the astrometric Gaia mission and its 6 dimensional phase space data, and the complementary chemical information obtained from the wide-field spectroscopic programs such as the GALAH survey [6], the H3 survey [3], APOGEE [20], RAVE [27], SEGUE [33], and LAMOST [4]. The dynamical times of the accreted objects are far longer than the age of the host galaxy, allowing the phase space to retain part of the information on the original orbit parameters. On the other hand, the chemical space is dependent on the star formation history, in particular type II SNe produce α-elements and iron with a almost constant ratio, while type Ia SNe produce more efficiently iron. Another factor that governs the chemical space is the total mass of the galaxy, since the more massive galaxies are more capable to resist the expulsion of metals due to feedback mechanism. The crossmatch between Gaia and spectroscopic data allowed for the discovery of the "Gaia-Sausage-Enceladus" (GSE) ([2], [12]), a massive accretion event whose remnant now dominates the observation of the inner stellar halo of our Galaxy. The GSE is described as 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 star of the Milky Way, as it is possible to see in Fig 2.1.

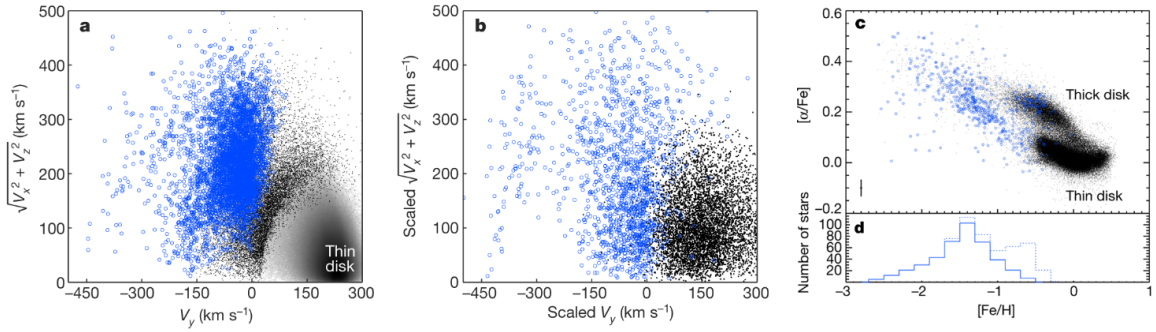


Figure 2.1: Left panel: Toomre diagram for Gaia DR2 data where in blue are stars selected to pick out the GSE structure. Middle panel: similar to the left panel but with simulated data of a minor merger, resulting in a less concentrated structure, mostly due to the fact that a more massive structure is more able to retain its original orbital properties. Right panel: the star in blue are the same of the left panel crossmateched with the APOGEE chemical information. This figure is from [12].

Robustly identify distinct structure is challenging, and disentangle the components in fully phase mix situations is nearly impossible. In order to characterize the assembly history [5] propose to use the "CARDs", the chemical abundance ratio distributions of the stars, obtained from a subsample of accreted object candidate from the FIRE-2 zoom-in cosmological simulations of MW-mass galaxies [31]. Although similar to CASBI on how to leverage N-body simulations, this method do not recovers posteriors for the parameters of the accreted objects but rather considers the host halo as a linear combinations of templates CARDs

$$\mathrm{CARD}_{\mathrm{halo,\ model}}(x_d) = \sum_i \sum_j A_{ij} \mathrm{CARD}_{\mathrm{temp},ij}(x_d | M_{\mathrm{sat},i}, t_{100,j}), \tag{2.1}$$

treating each coefficient A_{ij} as the fraction of mass contribution from the accretion event of the template satellite with mass $M_{\mathrm{sat},i}$ and quenching time $t_{100,j}$, and tries to recover those coefficients by maximize a loss that comperes the observed CARDs with the combination of the templates. An example of template constructed from dwarf galaxies is presented in Fig. 2.2. The template 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 [22]. This method and CASBI share two more aspect: 1. Both of these methods are meant to be used on simulations, and the integrations of observational data is not yet implemented, even though theoretically possible. 2. Both rely on the assumption that the chemical space of accreted and isolated dwarf galaxies is very similar, due to ram pressure quenching the star formation history of the accreted object and hence 'freezing' these abundance ratios at the infall time.

Another approach is presented in [8], which takes advantage of the mass-metallicity relation to decompose the metallicity distribution functions (MDF) of the host galaxy as a mixture of accreted halo's MDF, assumed gaussian for each of these building blocks. This decomposition rely on [15] that demonstrated that at the dwarf mass scale, not only the average metallicity vary with the mass, but the width of the MDF also varies, with the lowest mass dwarf having a wider spread of metallicities. The Likelihood that is used in this work for the [Fe/H] distribution, indicated as $\mathbf{z}$ is

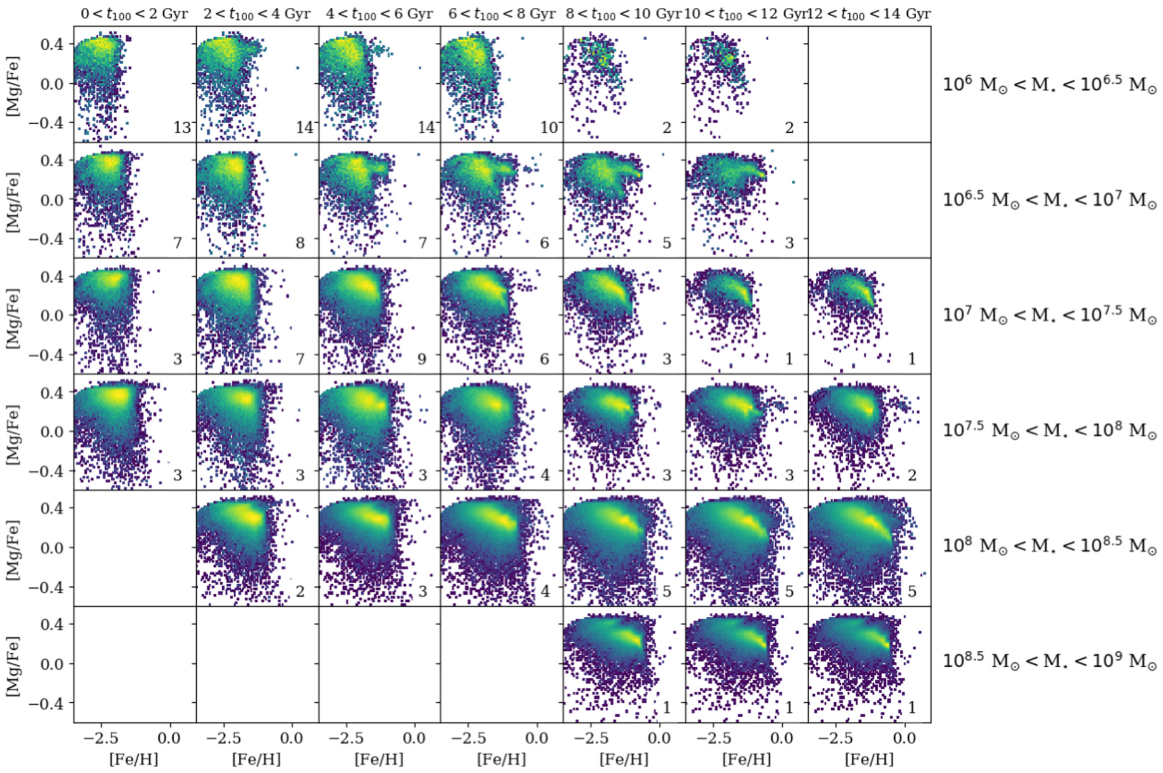


Figure 2.2: Template for accretion events constructed using dwarf galaxy in [5]. More massive dwarf galaxies have CARDs that extend to higher metallicities. At fixed stellar mass, galaxies that assemble more quickly (lower t_{100}) have more density at higher [Mg/Fe] than the component with a more extended star formation history.

then:

$$P(z|N, L_i, \mu_i, \sigma_i) = \frac{1}{\sum L_i} \sum_{i=1}^{N} L_i \mathcal{N}(z|\mu_i, \sigma_i), \tag{2.2}$$

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 [15]. Similarly to CASBI, this method has the problem of having a variable number of parameters, making it difficult to sample in practice, so to tackle this problem they decided to bin the luminosities L_j and count the number of contribution from each bin N_j. In order to perform the inference, they adopted a nested sampling scheme to obtain a posterior distribution for the number of galaxies in each luminosity bin, which can be considered a proxy for the star mass. The posterior probability for the Milky Way halo is reported in Fig. 2.3. The samples used in this posterior were obtained from different spectroscopic surveys after applying various cut to avoid contamination from thick disc stars. The cut were made based on parallax distance, radial distance, height with respect to the plane of the galaxy, and only stars with a retrograde orbit $v_\phi < 50$ were selected.

In CASBI we adopt the same superimposition of the components contribution, but we do not assume neither a prefix or an analytical form for the joint distribution of the chemical abundances, relaxing these assumption and relying only on the available samples from the N-body simulations.

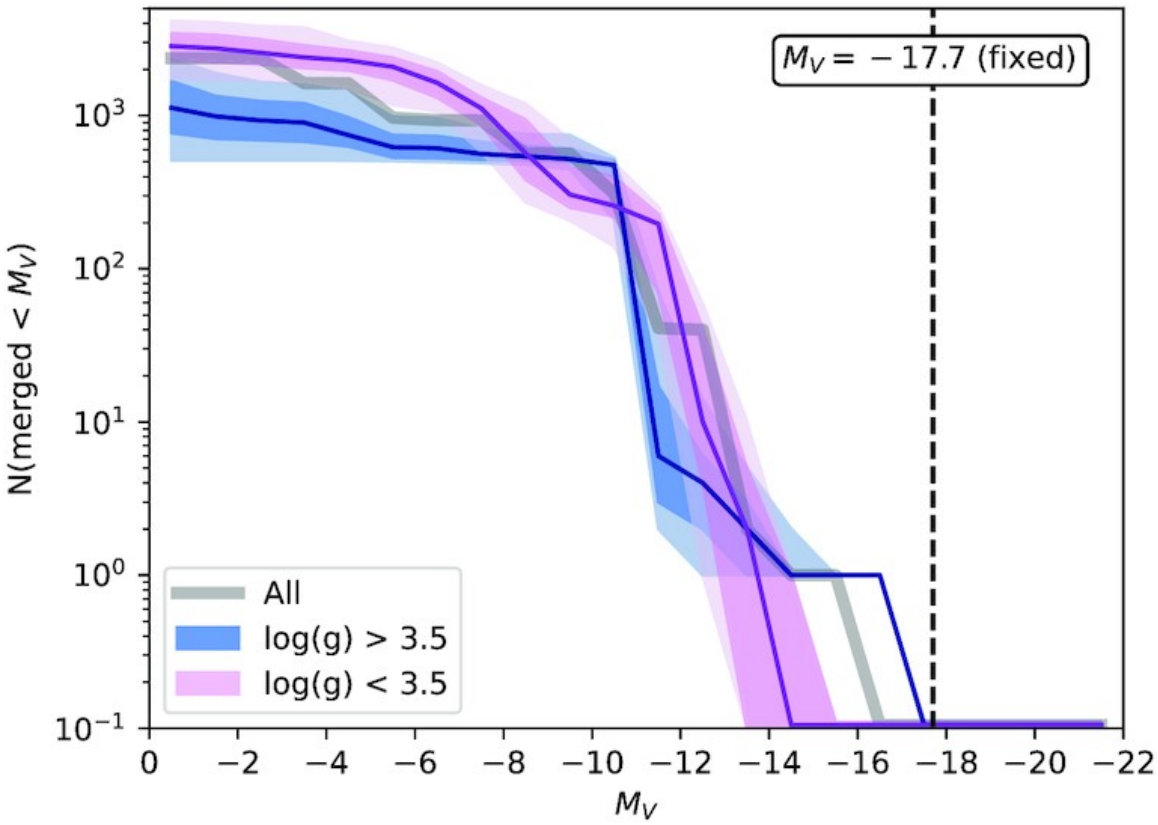


Figure 2.3: The estimated number of destroyed dwarf galaxy in the MW halo. The separation in two bin value of log(g) because dwarf stars and giants can have different metallicity biases.

Also a possible limitation of our method is that we do not take into consideration a distinction between destroyed and surviving accreted dwarf galaxy. This can lead, as described in [21] to a -0.3 dex offset of the mass metallicity [15] relation that might be necessary to take into account.

Chapter 3

CASBI: Chemical Abundance Simulation Based Inference

3.1 Simulation Based Inference

CASBI is a Simulation Based Inference (SBI) package to recover the properties of building blocks of Milky Way like galaxy's halo from observations of the chemical abundance plane. The SBI framework has existed along side the more traditional likelihood based inference methods for quite some years already, having its root in the Approximate Bayes Computation [26], and it has been used in a variety of fields, from cosmology to particle physics. The main difference between SBI and likelihood based methods, like MCMC, is that the former do not require the likelihood function to be known, but rather rely 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 $(\mathbf{x}, \boldsymbol{\theta})$.

Recent advance of this technique was made possible by the use of machine learning models to emulate conditional probability distributions, a technique know as Neural Density Estimation (NDE) [23]. The NDE is achieved by training a Normalizing Flow architecture, a generative model that allows to obtain samples from a complex distribution $p(x)$ by constructing a series of **bijiective** 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 [14], implementing the transformations as Neural Network with parameters ϕ_i, in the end the models learns the following schema:

$$p(x) \sim x \equiv h_0 \overset{f^1_{\phi_1}}{\longleftrightarrow} h_1 \overset{f^2_{\phi_2}}{\longleftrightarrow} h_2 \dots \overset{f^K_{\phi_K}}{\longleftrightarrow} h_K \equiv z \sim \mathcal{N}(z; 0, \mathcal{I}), \tag{3.1}$$

by maximizing the negative log likelihood as loss function and using the change of variable formula as follows:

$$\begin{aligned} \log p(x) &= \log p(z) + \log \left| \det \left(\frac{\partial z}{\partial x} \right) \right| \\ &= \log p(z) + \sum_{i=1}^{K} \log \left| \det \left(\frac{\partial h_i}{\partial h_{i-1}} \right) \right| \\ &= \log p(z) + \sum_{i=1}^{K} \log \left| \det \left(\frac{\partial f^i_{\phi_i}(h_{i-1})}{\partial h_{i-1}} \right) \right|, \end{aligned} \tag{3.2}$$

where the last term is the sum of the log determinant of the Jacobian of the transformations $f^i_{\phi_i}$. 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 sum of log determinant tractable, the use of *Coupling layers* allows to split the input x along its dimensions and apply a transformation only to a subset of the dimensions, using the other as input for the transformation and keeping it fixed. The subset is than changed at each layers, allowing to have a permutation invariant transformation. The transformations $f^i_{\phi_i}$ are usually very simple invertible transformation like a translation and a scaling, or splines functions. The choice of the invertible function can affect the expressivity of the model, defined as the capability of approximate more complex multivariate distribution, at the cost of more parameters, computational time and inference time.

Following the discussion presented in [13], in Bayesian analysis we have the choice to approximate either the Posterior, the Likelihood or the Likelihood ratio, and this choice depend mostly on the problem that one wants to solve.

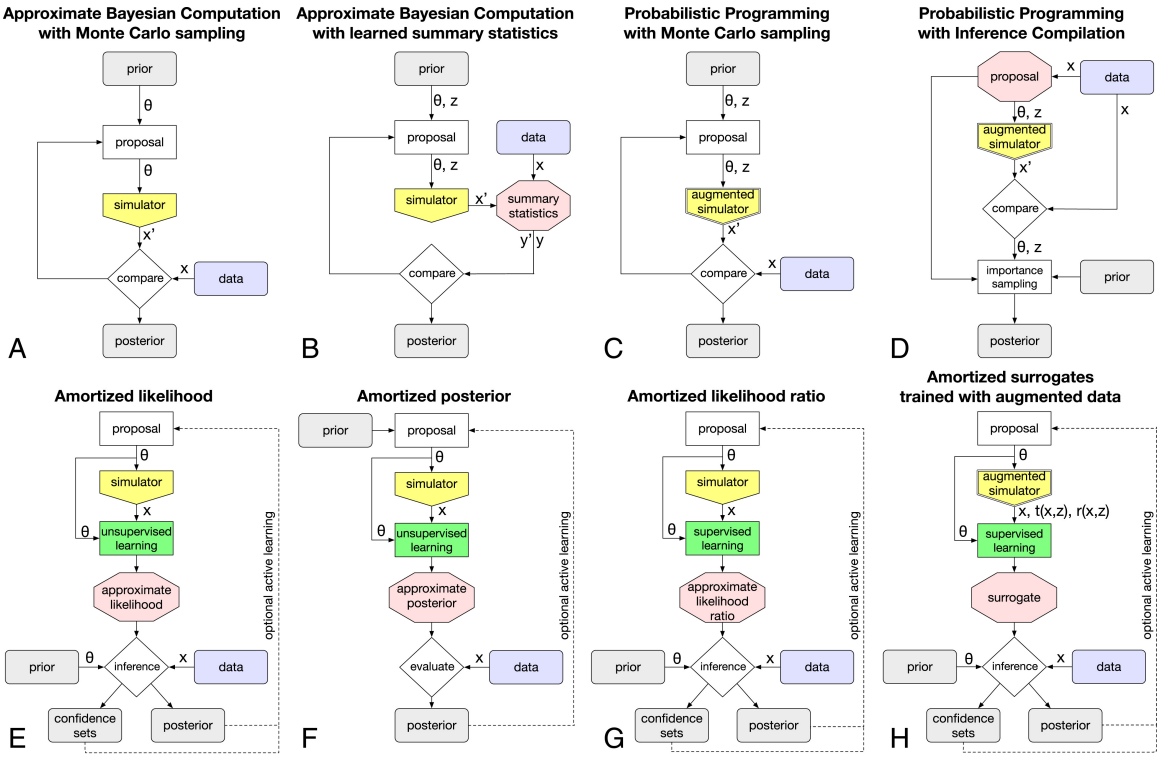


Figure 3.1: Different approaches to Simulation Based Inference, from [1].

In our case, due to the complexity of the Likelihood distribution of the chemical abundance space, we choose to approximate the Posterior distributions, and so we adopted the Neural Posterior Estimate (method **F** in Figure 3.1) that can be trained using the negative loglikelihood as loss function:

$$\begin{aligned} \mathcal{L}_{NPE}(\boldsymbol{\theta}) &= -\mathbb{E}_{\mathcal{D}_{train}} \log \hat{\mathcal{P}}(\theta_i | x_i) \\ &= -\mathbb{E}_{\mathcal{D}_{train}} \log \left(\frac{p(\theta)}{\tilde{p}(\theta)} q_\omega(\theta_i, x_i) \right), \end{aligned} \tag{3.3}$$

where our 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 $q_\omega(\theta_i, x_i)$, parametrized by the parameters ω.

Many excellent framework for handling SBI analysis are already available, and CASBI is build on top of the `ltu-ili` python package [13]. In particular, CASBI analysis were performed relying on the `sbi` backend [28] to train a *Neural Posterior Estimate*[1] of the parameters' posteriors. The preprocessing of the data is described in Section 3.2, the details of the training of the NPE is described in Section 3.4.

3.2 Simulator

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 [30]. The **NIHAO** is a set of 100 cosmological zoom-in hydrodynamical simulations with halos that range from dwarf ($M_{star} \sim 5 \times 10^9 M_\odot$) to Milky Way like ($M_{star} \sim 2 \times 10^{12} M_\odot$). In order to handle these simulations, in CASBI the preprocessing is done with the use of the functions available in `pynbody` [25]. In Fig. 3.2 we show face on samples of galaxies in the **NIHAO** simulations set.

[1] The `sbi` backed implement NPE using `nflows` [10]

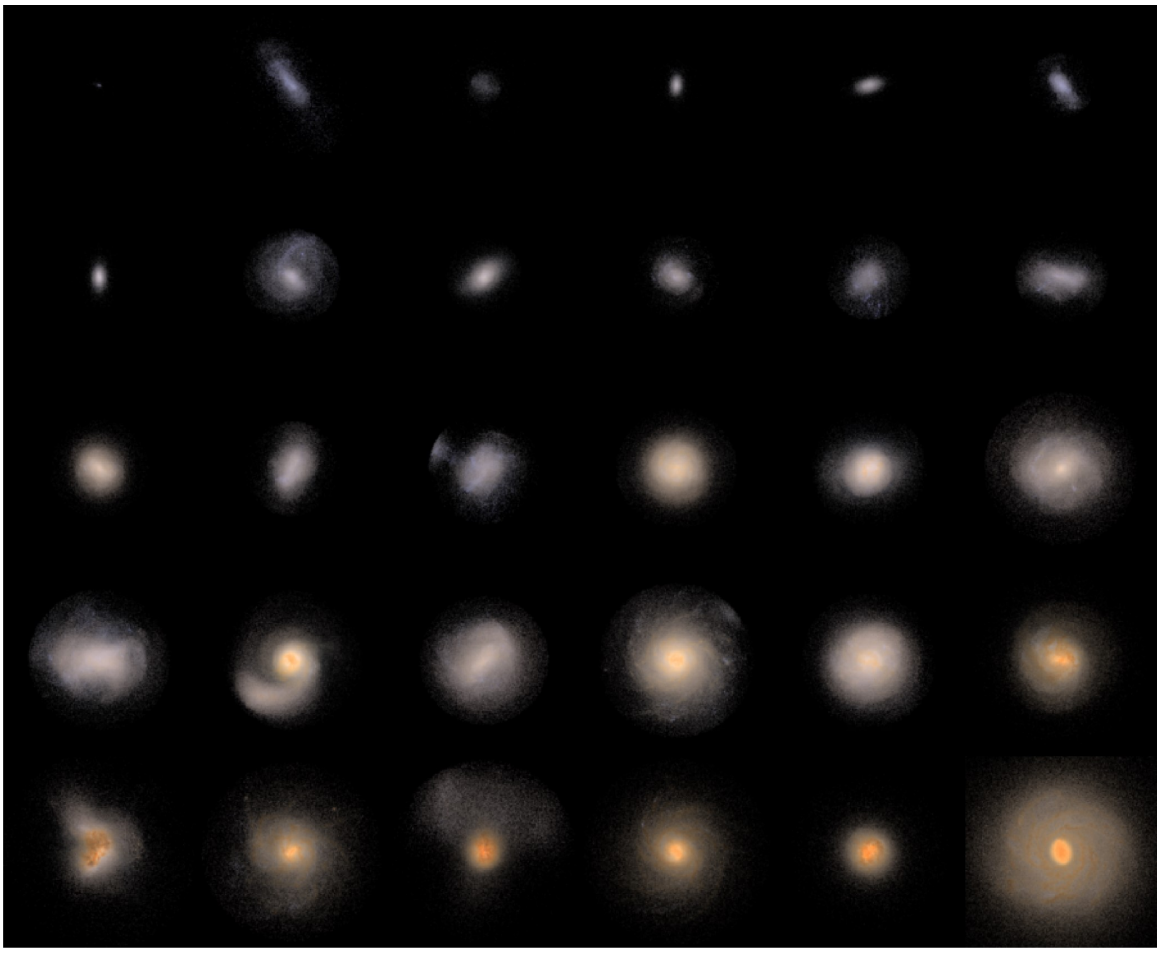


Figure 3.2: Face on **NIHAO** galaxies from [30].

Similarly to [5] and [8], we rely 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 we can treat each of the snapshot in this simulations as a possible building block of galactic halo.

The construction of the observables is done in by aggregating multiple subhalo into a single stellar halo. In order to create subhalo we construct 2D histogram, referred to as $\mathbf{x^i}$, by binning the chemical abundance plane $[[O/Fe], [Fe/H]]$[2] for each of the snapshot available in **NIHAO**. We have also filter the galaxies to have object 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 histogram have 64×64 pixels, and minimum and maximum values set after filtering all the stars that were outside the 0.01 percentile in either metallicity or α element abundance. Each of the x_i is uniquely identifiable trough the `Galaxy_name` attribute. 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 the N^j_{sub} is the number of accreted

[2]They are respectively proxy for α elements abundance and metallcity

objects present in the j-th galaxy halo. The actual choice of how to sample from the template library created from the **NIHAO** simulations can be adapted, we tested to randomly sample in 3.4 and to use a more physically informed approach by using a luminosity function and a total stellar mass budget in Section 3.5.

The goal of CASBI is to being able to recover $\theta^{\mathbf{i}}$ for each of the subhalos in the galactic halo from the observable $\mathbf{x^j}$= $\sum_i x_i^j$, and gaining insight on how many subhalos there are. Among all the possible parameters available from the simulations, we have decided to limit ourself to stellar mass M_{star} and age of the galaxy τ, also called infall time due to their equivalence in the assumption of quenched star formation after accretion.

3.3 Free Form Flow as a surrogate simulator

If it is possible to generate new data pairs at inference time, by sampling the prior and passing the samples to the simulator, and sequentially repeat the inference it is possible to achieve better accuracy, as it shown empirically in Fig. 3.3, at the cost of losing the **ammortize** property.

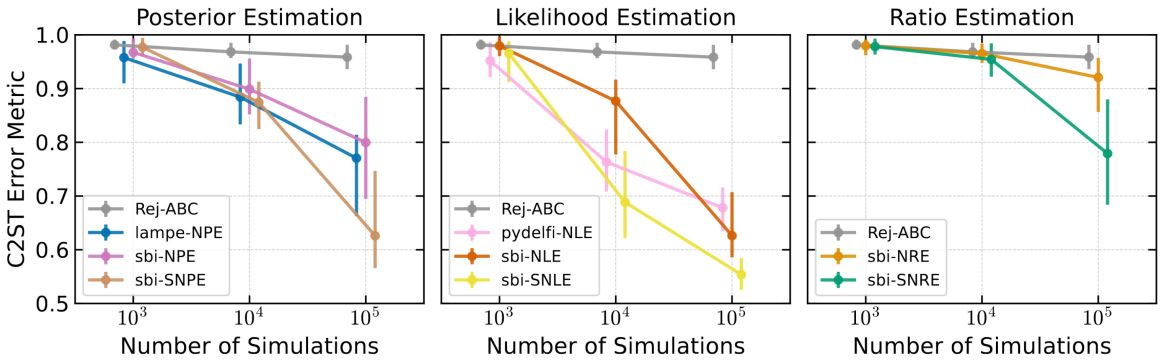


Figure 3.3: Different SBI method accuracy over 10 independent test with similar simulation budget. The error is defined in terms of the Classifier 2-sample Test (C2ST), with 0.5 being the optimal score. Figure credit [13].

Inspired by this result, we have decided to explorer 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. We have decided to implement a surrogate simulator that aims to learn how to sample new observation x from the Likelihood probability $p(x|\theta)$. The surrogate simulator that we adopted was Free Form Flow (FFF) [9], 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), like in [32], we decided to adopt FFF due to higher fidelity in the generation of new observation x, using the D-statistic of the two dimensional Kolmogorov-Smirnov Test as fidelity metric [19]. The architecture of the FFF is composed of a 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 map the latent space to the observation space. The major innovation 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 normalizing flow is calculating the Jacobian of the transformation, the authors propose to estimate its gradient using a pairs 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 3.2 as:

$$\mathcal{L}_{ML}^{f^{-1}} = \mathbb{E}_{x,v}[-\log p(f_\theta(x)) - v^T J_\theta \mathrm{SG}(J^{-1}v)], \tag{3.4}$$

  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 require the access 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:

$$\mathcal{L}_R = \frac{1}{2}\mathbb{E}_x[||x - g_\phi(f_\theta^{-1}(x))||^2]. \tag{3.5}$$

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

Combining both contributions from equations 3.4 and 3.5 leads to the following loss function:

$$\mathcal{L}_{FFF}^{g} = \mathcal{L}_{ML}^{g} + \beta\mathcal{L}_R, \tag{3.6}$$

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 we refer to the original paper [9].

In CASBI we have choose to use a Skip Connection Multi Layer Perceptron (SC-MLP) as both encoder and decoder, and we have followed the suggestions in Appendix B.1 of [9] to make this architecture conditional by concatenating to each layer the parameters θ sampled from our prior distributions. Even though the FFF architecture has good interpolation capability, returning average D-statics values lower than 0.3 when reconstructing the test set, the problem of sampling independent parameters values from the prior distributions make the net to extrapolate in regions of the conditional space where no data were shown, degrading the generate abundance halo. For example it could happen that a very massive M_{star} is sampled together with a very low infall time τ, which is a combination of parameters that is not physical and hence was not shown during training, so the generated halo is not realistic, namely it has a very high D-statistic value. Do to general poorer performance of SNPE using the FFF as a surrogate simulator, we have decided to not 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.

3.4 Two step Inference

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 have complete knowledge on the number of substructure that are present in the galactic halo. In the case of not fully phase mixed structure, the dynamical information could be use to help to disentangle this structure, and also to separate them from the host halo background. In CASBI we do not leverage on this information because it would require to construct stellar halo that have aggregated objects that are not dynamically biased. We leave this integration for future work. We decided to tackle this problem in the case of fully mixed remnants separating the inference in two steps, in the first we inferr 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 we 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 [8]. For each of the possible N_{sub} we 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, and we use 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, we 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 had 3 convolutional layers with 8, 16 and 32 filter, 3 maxplooling layers and 3 fully connected layers with 512, 256, 128 neurons. In this step we have not impose that the N_{sub} must be a discrete variable, and we have decide to just truncate the inferred value to the closet value. To the knowledge of the author no SBI framework has implemented a way of dealing with the inference of discrete random variables, so we leave a more precise implementation as a future work. We propose instead another method to obtain the number of substructure, by casting this inference as a classification problem. We use a SkipConnection CNN [3], considering the number of substructure as the label to assign to each x^j.

2. **Inference of $\theta^{\mathbf{j}}$**: Once we have the estimate $\tilde{N}_{sub}$, whether using dynamical information, the inference pipeline or the classification method, we can proceed to the inference of the parameters $\theta_{\mathbf{i}}^{\mathbf{j}}$. The prior for the parameters are assumed to be uniform between the minimum and maximum values available for the galaxies that we have filtered from the **NIHAO** simulations. We extract 10^5 random samples of $\tilde{N}_{sub}$ snapshots from the **NIHAO** simulations, and we construct the observable couples $(x^j, (\theta_1^j, \ldots, \theta_{\tilde{N}_{sub}}^j))$, with 20 % used as validation. We 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 training of the NPE is done using the `sbi` backend, using 4 `nsf` (neural spline flow) with 10 layers and 100 neurons each. Once again we use the same CNN architecture of the previous step as embedding for our observation x^j.

Even though highly modular, this two step inference has some limitations: the accuracy and calibration of the second step are heavily depend 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. We expected the pipeline to be able to recover most of the information from the most massive subhalos, due to

[3]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.

the degeneracy in the abundance plane of the less massive and components and the more distinct feature of the more massive one. The second problem is linked to the linear scaling of the parameter dimension as a function of the number of subhalos. In a realistic case we expect to have order of $\approx$100 subhalos, resulting in a parameter space of dimensionality 2*100 = 200. In order to reduce the impact of these two problems we rethink the inference pipeline, presented in the next section.

3.5 Realistic halo and 1 step Inference

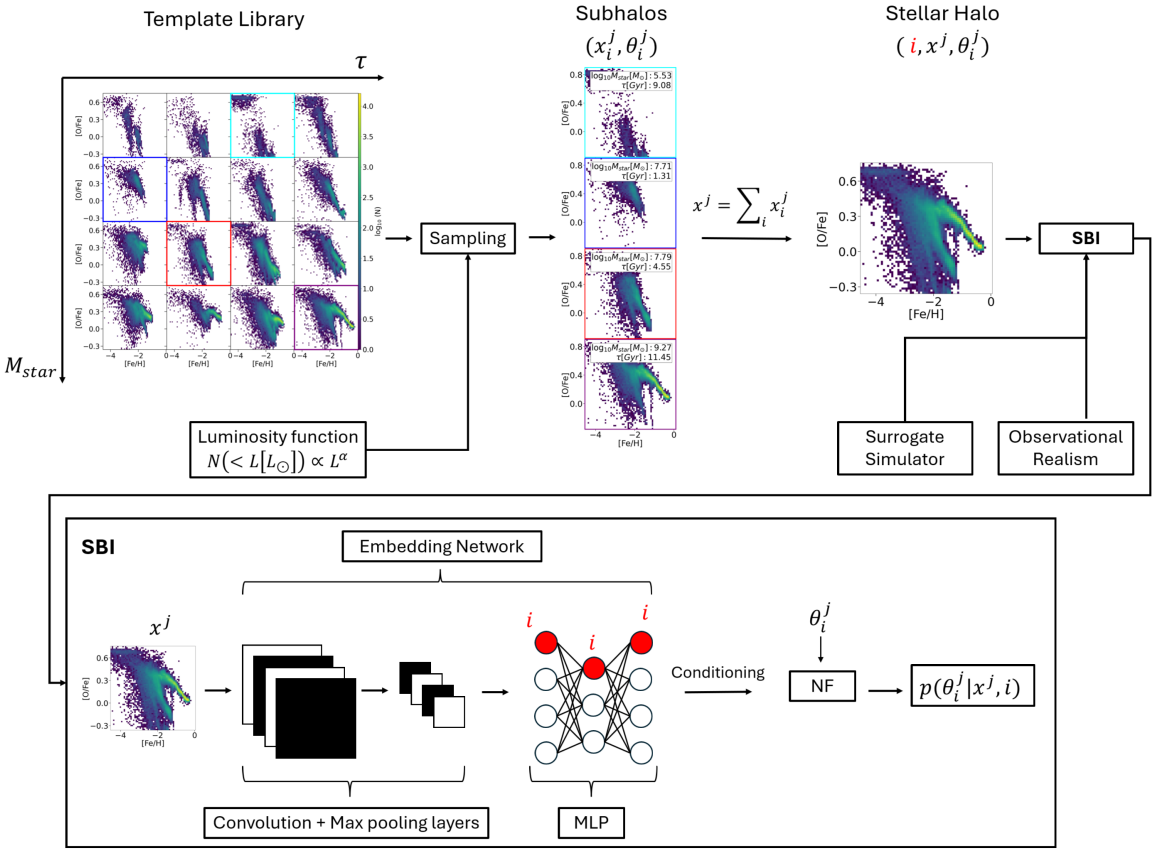


Figure 3.4: CASBI pipeline. In our analysis we have fixed the **Template Library** to be a subsample of the **NIHAO** simulations, as described in Section 3.2, but it can be swapped with user-choice simulations. The choice of the Template Library incorporate all the assumption that we make on the chemical 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** we sample non-repeated subhalos aiming to reproduce a **Luminosity function** $N(< L)$ that can be set by the user, for our analysis we have taken it from [16]. 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 our analysis for computational reasons). In our analysis we fixed the stellar mass budget of the halo accordingly to [7], but it can be set by the user accordingly to the estimated total stellar mass available in the observation that we want to infer on. The **SBI** pipeline can incorporate a **Surrogate Simulator** to perform the sequential version of NPE. In our analysis we do not adopt this component as described in Section 3.3. The **Observational Realism** encapsulate all that concerns bridging the gap between simulation and observation, i.e. uncertainties of spectroscopic surveys, selection functions, etc. In our analysis we haven't inject this component into the pipeline because we aimed to understand the limitations of this technique when no observational contamination is involved.

In order to avoid the need for a two step inference and still retaining the possibility to access to the information on how many subhalos populate a given abundance plane, we 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 on $(x, \theta) = (i, x^j, \theta_i^j)$ pairs, where $x^j = \sum_i x_i^j$ and the x_i^j are ordered accordingly to their stellar mass M_{star}. In this way the j-th stellar halo abundance plane is shown as many times as the number of subhalos present in it, and in order to guide the model into inferring the right parameters θ_i^j the embedding is conditioned on the integer i-th by concatenating it to each input of the fully connected layers of the CNN used to embed the observations, and it is concatenate also before passing the embedded information to the Normalizing Flow. This conditioning can be appreciated also in the lower part of Fig. 3.4, in which the CASBI SBI pipeline shows the integer i as a red node concatenated to each layer of the embedding network.

In order to create more realistic mock galaxy halo we have decided to adopt a sampling scheme for the subhalos that is based on the luminosity function described in [16]. The luminosity function described the subhalo distribution in a range of luminosities that spans from $M_V = -2$ all the way to the luminosity of the Large Magellanic cloud:

$$\frac{dN}{dM_V} = 10 \times 10^{0.1(M_V+5)} \tag{3.7}$$

we can then manipulate this equation to express it as a function of the Luminosity L:

$$\frac{dN}{dL} = \frac{dN}{dM_V} \times \frac{dM_V}{dL} = 10^{0.1(M_{V,\odot}-2.5\log_{10}(L)+5)+1} \times (-2.5L^{-1}) \sim L^{-1.25} \tag{3.8}$$

which in the end can be integrated to obtain the number of subhalos with luminosity lower the L that we are going to adopt for sampling stellar halo:

$$N(< L) = K \times L^{1+\alpha}, \tag{3.9}$$

where K represent a constant and $\alpha = -1.25$ is the single power law exponent obtained by [16]. Other work based not only on SDSS observations like [16] but also on ΛCDM N-body simulation set $\alpha = -1.9 \pm 0.2$ ([29]). We fix this value to -1.25 and we leave the analysis of the impact of this choice as a future work. Assuming $L_\odot = M_\odot$, we normalize equation 3.9 after setting the support to be the interval of masses that we have available in our catalogue of NIHAO simulations ($10^5 M_\odot < M_{star} < M_{star}^{halo}$) and we sample from this distributions using an inverse scheme, where M_{star}^{halo} is the mass budget for our mock halo of $M = 1.4 \pm 0.2 \times 10^9 M_\odot$ based on [7]. After obtaining the analytic samples we take the first and second Nearest Neighbors (NN) that are within a 10% of the analytic sampled mass as subhalo for our mock halos and we reduce the total mass budget by the mass of the NN that we have used. The choice of the mass budget can be adapted and comes from observation that do not take into account 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 we make sure to sample non repeated subhalo within the same mock halo and we avoid repetitions of the same combinations of subhalos between mock subhalos both within training and test set and across these two sets.

In Fig. 3.4 we show 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 be a different suite of simulated galaxies (e.g. [24]), the sampling scheme can incorporate different luminosity function and stellar halo budget, the NPE and embedding network architecture and hyperparameter 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 [9], GRUMPY [17]) can be implemented to allow for the Sequential version of the NPE.

3.6 Calibration

This section is highly inspired by [13]. In posterior estimation we aim to maximize the constraining of θ given the observable x_0 and whether the uncertainties are calibrated to our training data. These criteria are naturally adversarial and the this problem can be interpreted as another instance of the bias-variance trade off. It is possible to confront various NPE accuracy 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 concentrate more probability mass around the true value, which translate directly to a higher constraining power. Moreover it is possible to gauge the overall constrain power against a *ground truth* posterior, namely long-run MCMC output. A classical metric to confront posterior samples is the C2ST, which is defined as the accuracy of a classifier to distinguish between true and infered posterior samples, with a C2ST value of 0.5 implying that the two sampled distributions are the same.

The calibration of the model uncertainties can be obtain using the Probability Integral Transformation (PIT), defined as the cumulative density function of our posterior given x_0:

$$\text{PIT}(\theta|x_0) = \int_{-\infty}^{\theta} \hat{\mathcal{P}}(\theta|x_0)d\theta. \tag{3.10}$$

Due to the poor scaling of the PIT to higher dimension, it is better to construct and estimate PIT value as:

$$\text{PIT}(\theta|x_0) = \mathbb{E}_{\hat{\theta}\sim\hat{\mathcal{P}}(\theta|x)}[\Theta(\hat{\theta} - \theta)], \tag{3.11}$$

where Θ is the Heaviside step function. The PIT counts the number of time 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 distributed uniformly in range [0, 1] for each of the test set samples. Usually the PIT distribution is studied using percentile-percentile (P-P) plots, comparing the CDF of the PIT value to the CDF of a uniform distribution. This tool can be used to constrain over / sunder dispersions. In Fig. 3.5 we show an example of uncalibrated posteriors approximation and the corresponding P-P plot [4]. Since as the dimensionality of θ increase the proper coverage requires exponentially more samples, we decide to rely on the PIT of each component θ_i of the marginal posterior, and we expect that the value

$$\text{PIT}(\theta_i|x_0) = \mathbb{E}_{\hat{\theta}\sim\hat{\mathcal{P}}(\theta_i|x)}[\Theta(\hat{\theta}_i - \theta_i)], \tag{3.12}$$

has the same properties as the PIT obtain in equation 3.11 if the model is globally consistent on the test set.

Lastly, as an approximation to check multivariate posterior coverage, we use the Test of Accuracy with Random Points (TARP) [18]. 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 method are already integrated in the `ltu-ili` package.

[4] The plot is not accurate, it is created to give an idea of the general relation between the posterior behavior and how it is reflected in the P-P plot.

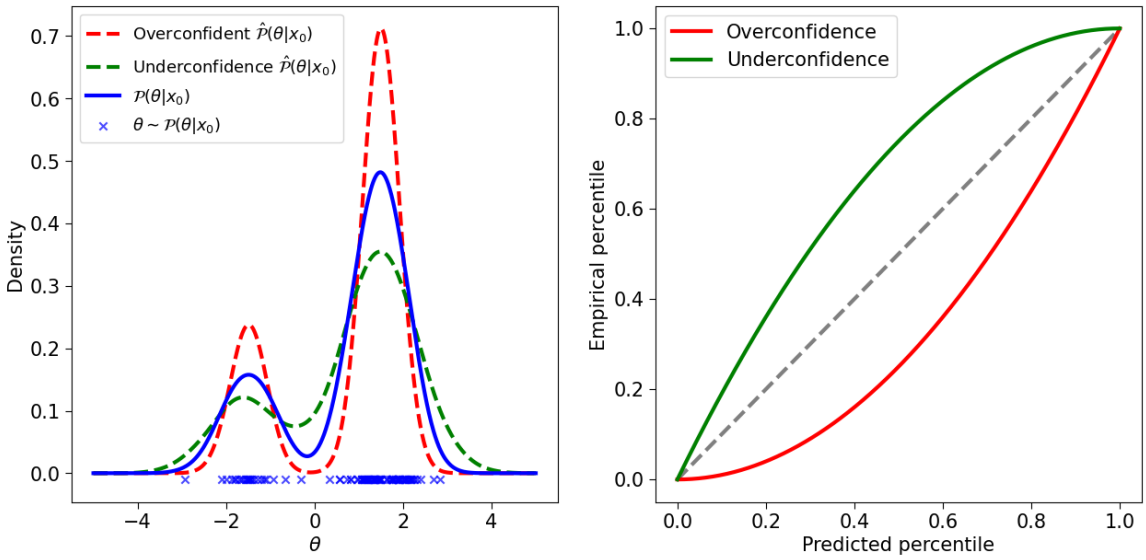


Figure 3.5: Examples of SBI uncalibrated posterior. Left panel: over (red) and under (green) posterior obtained from the same data θ, for which the true posterior (blue) is shown. Right panel: P-P plot for the two uncalibrated posterior. The goal of the calibration is to obtain a distribution that is underconfident and as close as possible to the diagonal. This figure is inspired by [11].

Chapter 4

Analysis

4.1 NIHAO UHD

Chapter 5

Conclusion

5.1 Future work

5.1.1 GRUMPY

5.1.2 Number halos has a free parameter: hierarchical sbi

5.1.3 True test: GAIA

Bibliography

[1] The frontier of simulation-based inference. https://www.pnas.org/doi/10.1073/pnas.1912789117.

[2] V. Belokurov, D. Erkal, N. W. Evans, S. E. Koposov, and A. J. Deason. Co-formation of the disc and the stellar halo. *Monthly Notices of the Royal Astronomical Society*, 478(1):611–619, July 2018.

[3] C. Conroy, A. Bonaca, P. Cargile, B. D. Johnson, N. Caldwell, R. P. Naidu, D. Zaritsky, D. Fabricant, S. Moran, J. Rhee, A. Szentgyorgyi, P. Berlind, M. L. Calkins, S. Kattner, and C. Ly. Mapping the Stellar Halo with the H3 Spectroscopic Survey. *The Astrophysical Journal*, 883(1):107, Sept. 2019.

[4] X.-Q. Cui, Y.-H. Zhao, Y.-Q. Chu, G.-P. Li, Q. Li, L.-P. Zhang, H.-J. Su, Z.-Q. Yao, Y.-N. Wang, X.-Z. Xing, X.-N. Li, Y.-T. Zhu, G. Wang, B.-Z. Gu, A.-L. Luo, X.-Q. Xu, Z.-C. Zhang, G.-R. Liu, H.-T. Zhang, D.-H. Yang, S.-Y. Cao, H.-Y. Chen, J.-J. Chen, K.-X. Chen, Y. Chen, J.-R. Chu, L. Feng, X.-F. Gong, Y.-H. Hou, H.-Z. Hu, N.-S. Hu, Z.-W. Hu, L. Jia, F.-H. Jiang, X. Jiang, Z.-B. Jiang, G. Jin, A.-H. Li, Y. Li, Y.-P. Li, G.-Q. Liu, Z.-G. Liu, W.-Z. Lu, Y.-D. Mao, L. Men, Y.-J. Qi, Z.-X. Qi, H.-M. Shi, Z.-H. Tang, Q.-S. Tao, D.-Q. Wang, D. Wang, G.-M. Wang, H. Wang, J.-N. Wang, J. Wang, J.-L. Wang, J.-P. Wang, L. Wang, S.-Q. Wang, Y. Wang, Y.-F. Wang, L.-Z. Xu, Y. Xu, S.-H. Yang, Y. Yu, H. Yuan, X.-Y. Yuan, C. Zhai, J. Zhang, Y.-X. Zhang, Y. Zhang, M. Zhao, F. Zhou, G.-H. Zhou, J. Zhu, and S.-C. Zou. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). *Research in Astronomy and Astrophysics*, 12(9):1197–1242, Sept. 2012.

[5] E. C. Cunningham, R. E. Sanderson, K. V. Johnston, N. Panithanpaisal, M. K. Ness, A. Wetzel, S. R. Loebman, I. Escala, D. Horta, and C.-A. Faucher-Giguère. Reading the CARDs: The Imprint of Accretion History in the Chemical Abundances of the Milky Way's Stellar Halo. *The Astrophysical Journal*, 934(2):172, Aug. 2022.

[6] G. M. De Silva, K. C. Freeman, J. Bland-Hawthorn, S. Martell, E. W. De Boer, M. Asplund, S. Keller, S. Sharma, D. B. Zucker, T. Zwitter, B. Anguiano, C. Bacigalupo, D. Bayliss, M. A. Beavis, M. Bergemann, S. Campbell, R. Cannon, D. Carollo, L. Casagrande, A. R. Casey, G. Da Costa, V. D'Orazi, A. Dotter, L. Duong, A. Heger, M. J. Ireland, P. R. Kafle, J. Kos, J. Lattanzio, G. F. Lewis, J. Lin, K. Lind, U. Munari, D. M. Nataf, S. O'Toole, Q. Parker, W. Reid, K. J. Schlesinger, A. Sheinis, J. D. Simpson, D. Stello, Y.-S. Ting, G. Traven, F. Watson, R. Wittenmyer, D. Yong, and M. Žerjal. The GALAH survey: Scientific motivation. *Monthly Notices of the Royal Astronomical Society*, 449(3):2604–2617, May 2015.

[7] A. J. Deason, V. Belokurov, and J. L. Sanders. The total stellar halo mass of the Milky Way. *Monthly Notices of the Royal Astronomical Society*, 490(3):3426–3439, Dec. 2019.

[8] A. J. Deason, S. E. Koposov, A. Fattahi, and R. J. J. Grand. Unravelling the mass spectrum of destroyed dwarf galaxies with the metallicity distribution function. *Monthly Notices of the Royal Astronomical Society*, 520(4):6091–6103, Feb. 2023.

[9] F. Draxler, P. Sorrenson, L. Zimmermann, A. Rousselot, and U. Köthe. Free-form Flows: Make Any Architecture a Normalizing Flow, Apr. 2024.

[10] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Nflows: Normalizing flows in PyTorch. Zenodo, Nov. 2020.

[11] M. Falkiewicz, N. Takeishi, I. Shekhzadeh, A. Wehenkel, A. Delaunoy, G. Louppe, and A. Kalousis. Calibrating Neural Simulation-Based Inference with Differentiable Coverage Probability.

[12] A. Helmi, C. Babusiaux, H. H. Koppelman, D. Massari, J. Veljanoski, and A. G. A. Brown. The merger that led to the formation of the Milky Way's inner stellar halo and thick disk. *Nature*, 563(7729):85–88, Nov. 2018.

[13] M. Ho, D. J. Bartlett, N. Chartier, C. Cuesta-Lazaro, S. Ding, A. Lapel, P. Lemos, C. C. Lovell, T. L. Makinen, C. Modi, V. Pandya, S. Pandey, L. A. Perez, B. Wandelt, and G. L. Bryan. LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology, Feb. 2024.

[14] D. P. Kingma and P. Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions, July 2018.

[15] E. N. Kirby, G. A. Lanfranchi, J. D. Simon, J. G. Cohen, and P. Guhathakurta. MULTI-ELEMENT ABUNDANCE MEASUREMENTS FROM MEDIUM-RESOLUTION SPECTRA. III. METALLICITY DISTRIBUTIONS OF MILKY WAY DWARF SATELLITE GALAXIES. *The Astrophysical Journal*, 727(2):78, Feb. 2011.

[16] S. Koposov, V. Belokurov, N. W. Evans, P. C. Hewett, M. J. Irwin, G. Gilmore, D. B. Zucker, H.-W. Rix, M. Fellhauer, E. F. Bell, and E. V. Glushkova. The Luminosity Function of the Milky Way Satellites. *The Astrophysical Journal*, 686(1):279–291, Oct. 2008.

[17] A. Kravtsov and V. Manwadkar. GRUMPY: A simple framework for realistic forward-modelling of dwarf galaxies. *Monthly Notices of the Royal Astronomical Society*, 514(2):2667–2691, June 2022.

[18] P. Lemos, A. Coogan, Y. Hezaveh, and L. Perreault-Levasseur. Sampling-Based Accuracy Testing of Posterior Estimators for General Inference.

[19] R. H. C. Lopes, I. Reid, and P. R. Hobson. The two-dimensional Kolmogorov-Smirnov test.

[20] S. R. Majewski, R. P. Schiavon, P. M. Frinchaboy, C. A. Prieto, R. Barkhouser, D. Bizyaev, B. Blank, S. Brunner, A. Burton, R. Carrera, S. D. Chojnowski, K. Cunha, C. Epstein, G. Fitzgerald, A. E. G. Pérez, F. R. Hearty, C. Henderson, J. A. Holtzman, J. A. Johnson, C. R. Lam, J. E. Lawler, P. Maseman, S. Mészáros, M. Nelson, D. C. Nguyen, D. L. Nidever, M. Pinsonneault, M. Shetrone, S. Smee, V. V. Smith, T. Stolberg, M. F. Skrutskie, E. Walker, J. C. Wilson, G. Zasowski, F. Anders, S. Basu, S. Beland, M. R. Blanton, J. Bovy, J. R. Brownstein, J. Carlberg, W. Chaplin, C. Chiappini, D. J. Eisenstein, Y. Elsworth, D. Feuillet, S. W. Fleming, J. Galbraith-Frew, R. A. García, D. A. García-Hernández, B. A. Gillespie, L. Girardi,

J. E. Gunn, S. Hasselquist, M. R. Hayden, S. Hekker, I. Ivans, K. Kinemuchi, M. Klaene, S. Mahadevan, S. Mathur, B. Mosser, D. Muna, J. A. Munn, R. C. Nichol, R. W. O'Connell, J. K. Parejko, A. C. Robin, H. Rocha-Pinto, M. Schultheis, A. M. Serenelli, N. Shane, V. S. Aguirre, J. S. Sobeck, B. Thompson, N. W. Troup, D. H. Weinberg, and O. Zamora. The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *The Astronomical Journal*, 154(3):94, Sept. 2017.

[21] R. P. Naidu, C. Conroy, A. Bonaca, B. D. Johnson, Y.-S. Ting, N. Caldwell, D. Zaritsky, and P. A. Cargile. Evidence from the H3 Survey That the Stellar Halo Is Entirely Comprised of Substructure. *The Astrophysical Journal*, 901(1):48, Sept. 2020.

[22] N. Panithanpaisal, R. E. Sanderson, A. Wetzel, E. C. Cunningham, J. Bailin, and C.-A. Faucher-Giguère. The Galaxy Progenitors of Stellar Streams around Milky Way–mass Galaxies in the FIRE Cosmological Simulations. *The Astrophysical Journal*, 920(1):10, Oct. 2021.

[23] G. Papamakarios. Neural Density Estimation and Likelihood-free Inference, Oct. 2019.

[24] A. Pillepich, D. Sotillo-Ramos, R. Ramesh, D. Nelson, C. Engler, V. Rodriguez-Gomez, M. Fournier, M. Donnari, V. Springel, and L. Hernquist. Milky Way and Andromeda analogs from the TNG50 simulation, Mar. 2023.

[25] A. Pontzen, R. Roškar, G. Stinson, and R. Woods. Pynbody: N-Body/SPH analysis for python. *Astrophysics Source Code Library*, page ascl:1305.002, May 2013.

[26] D. B. Rubin. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151–1172, Dec. 1984.

[27] M. Steinmetz, T. Zwitter, A. Siebert, F. G. Watson, K. C. Freeman, U. Munari, R. Campbell, M. Williams, G. M. Seabroke, R. F. G. Wyse, Q. A. Parker, O. Bienaymé, S. Roeser, B. K. Gibson, G. Gilmore, E. K. Grebel, A. Helmi, J. F. Navarro, D. Burton, C. J. P. Cass, J. A. Dawe, K. Fiegert, M. Hartley, K. S. Russell, W. Saunders, H. Enke, J. Bailin, J. Binney, J. Bland-Hawthorn, C. Boeche, W. Dehnen, D. J. Eisenstein, N. W. Evans, M. Fiorucci, J. P. Fulbright, O. Gerhard, U. Jauregi, A. Kelz, L. Mijović, I. Minchev, G. Parmentier, J. Peñarrubia, A. C. Quillen, M. A. Read, G. Ruchti, R.-D. Scholz, A. Siviero, M. C. Smith, R. Sordo, L. Veltz, S. Vidrih, R. Von Berlepsch, B. J. Boyle, and E. Schilbach. The Radial Velocity Experiment (RAVE): First Data Release. *The Astronomical Journal*, 132(4):1645–1668, Oct. 2006.

[28] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke. Sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, Aug. 2020.

[29] E. J. Tollerud, J. S. Bullock, L. E. Strigari, and B. Willman. Hundreds of Milky Way Satellites? Luminosity Bias in the Satellite Luminosity Function. *The Astrophysical Journal*, 688(1):277–289, Nov. 2008.

[30] L. Wang, A. A. Dutton, G. S. Stinson, A. V. Macciò, C. Penzo, X. Kang, B. W. Keller, and J. Wadsley. 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, Nov. 2015.

[31] A. R. Wetzel, P. F. Hopkins, J.-h. Kim, C.-A. Faucher-Giguère, D. Kereš, and E. Quataert. RECONCILING DWARF GALAXIES WITH ΛCDM COSMOLOGY: SIMULATING A REALISTIC POPULATION OF SATELLITES AROUND A MILKY WAY–MASS GALAXY. *The Astrophysical Journal Letters*, 827(2):L23, Aug. 2016.

[32] L. Wolf and T. Buck. GalacticFlow: Learning a Generalized Representation of Galaxies with Normalizing Flows, Dec. 2023.

[33] B. Yanny, C. Rockosi, H. J. Newberg, G. R. Knapp, J. K. Adelman-McCarthy, B. Alcorn, S. Allam, C. A. Prieto, D. An, K. S. J. Anderson, S. Anderson, C. A. L. Bailer-Jones, S. Bastian, T. C. Beers, E. Bell, V. Belokurov, D. Bizyaev, N. Blythe, J. J. Bochanski, W. N. Boroski, J. Brinchmann, J. Brinkmann, H. Brewington, L. Carey, K. M. Cudworth, M. Evans, N. W. Evans, E. Gates, B. T. Gänsicke, B. Gillespie, G. Gilmore, A. N. Gomez-Moran, E. K. Grebel, J. Greenwell, J. E. Gunn, C. Jordan, W. Jordan, P. Harding, H. Harris, J. S. Hendry, D. Holder, I. I. Ivans, Ž. Ivezič, S. Jester, J. A. Johnson, S. M. Kent, S. Kleinman, A. Kniazev, J. Krzesinski, R. Kron, N. Kuropatkin, S. Lebedeva, Y. S. Lee, R. F. Leger, S. Lépine, S. Levine, H. Lin, D. C. Long, C. Loomis, R. Lupton, O. Malanushenko, V. Malanushenko, B. Margon, D. Martinez-Delgado, P. McGehee, D. Monet, H. L. Morrison, J. A. Munn, E. H. Neilsen, A. Nitta, J. E. Norris, D. Oravetz, R. Owen, N. Padmanabhan, K. Pan, R. S. Peterson, J. R. Pier, J. Platson, P. R. Fiorentin, G. T. Richards, H.-W. Rix, D. J. Schlegel, D. P. Schneider, M. R. Schreiber, A. Schwope, V. Sibley, A. Simmons, S. A. Snedden, J. A. Smith, L. Stark, F. Stauffer, M. Steinmetz, C. Stoughton, M. SubbaRao, A. Szalay, P. Szkody, A. R. Thakar, S. Thirupathi, D. Tucker, A. Uomoto, D. V. Berk, S. Vidrih, Y. Wadadekar, S. Watters, R. Wilhelm, R. F. G. Wyse, J. Yarger, and D. Zucker. SEGUE: A SPECTROSCOPIC SURVEY OF 240,000 STARS WITH g = 14-20. *The Astronomical Journal*, 137(5):4377–4399, May 2009.