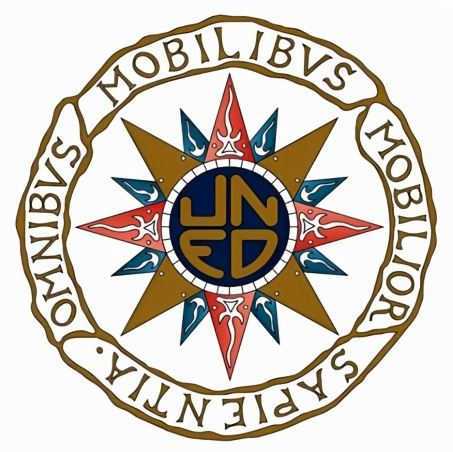
Universidad Nacional de Educación a Distancia (UNED)

Escuela Técnica Superior de Ingeniería Informática

Bayesian Inference on the Pleiades Open Cluster

Trabajo Fin de Máster
Máster en Investigación en Inteligencia Artificial

Autor: Francisco J. Palmero Moya
Directores: Javier Olivares Romero, Luis M. Sarro Baro



Septiembre 2023

«Y sucedió que íbamos por parte donde están las siete cabrillas, y en Dios y en mi ánima que como yo en mi niñez fui en mi tierra cabrerizo, que así como las vi, ¡me dio una gana de entretenerme con ellas un rato...!»

El ingenioso hidalgo Don Quijote de la Mancha,
Capítulo XLI, Miguel de Cervantes Saavedra

Contents

Preface

Undertaking the task of composing a master's thesis in the format of a scientific paper within the realm of Artificial Intelligence research is a journey that intertwines challenges and innovation. This preface aims to discuss some of the difficulties encountered during the process of adopting this format and to outline a practical approach that aims to address these challenges.

One of the primary challenges when writing a thesis in a paper format is striking a balance between the technical depth expected by the reviewers of the chosen journal and the diverse backgrounds of the readers, which may include academic supervisors who may not be as specialized in the subject matter. While the reviewers are well-versed in the technical nuances of the field, it is important to ensure that the thesis remains accessible to a wider audience.

To address this challenge, I have taken a two-fold approach. First, I have included a brief and formal discussion, Section 2, that provides the necessary context and fundamental concepts required for understanding the research. This section serves as a bridge between the technical nature of the research and the varying levels of expertise among readers. It is important to acknowledge that this section serves as a platform to establish with precision the mathematical terms harnessed throughout the thesis. This deliberate inclusion serves as a response to a common need within the scientific community, where papers often employ terms without prior definitions.

Furthermore, to avoid overwhelming the main narrative with detailed technical explanations, I have included appendices. These appendices contain in-depth discussions, derivations, and additional information that readers can explore at their own pace. This approach maintains the flow of the main thesis while providing interested readers with the opportunity to delve deeper into specific technical aspects.

It is pertinent to acknowledge that within the confines of a project, time emerges as an unyielding constraint, occasionally limiting the full optimization of certain challenges. It is a reality that despite fervent efforts, not all problems attain their zenith of resolution within the allocated timeframe. This thesis is not exempt from such limitations, as it is rooted in the practicality of real-world projects.

As you navigate through the pages of this thesis, my aim is that you will discover a synthesis of technical depth and clarity, *aurea mediocritas*.

Francisco José Palmero Moya
Delft, The Netherlands

Bayesian Inference on the Pleiades Open Cluster

F. J. Palmero[1]*, J. Olivares[1] and L. M. Sarro[1]

[1]Departamento de Inteligencia Artificial, UNED

Abstract

Context: *Age is one of the fundamental parameters of any astrophysical object, being it a galaxy, a star, or a planet. The main three stellar dating techniques in astrophysics are dynamical ages, isochrone ages, and the chemical abundance of the Lithium isotopes. However, these methods produce ages that differ by up to 50 % for stars younger than the Sun. The main culprit for these differences is the lack of a consistent and robust age calibration due to the fact that the only sufficiently precise age available is that of the Sun.*

Aims: *Our aim is to define and implement a Bayesian hierarchical model able to determine the ages of star clusters and associations through two age dating techniques: isochrones and Lithium abundance.*

Methods: *The resulting model combines existent photometric, parallax, and chemical abundance of Lithium data sets of stars belonging to stellar open clusters to infer its age distribution through modern and robust artificial intelligence methods. A Neural Network is trained given a grid of pre-calculated BT-Settl models to interpolate the spectral energy distributions of stars, working as a black-box interpolator in the model. The Bayesian hierarchical model not only facilitates simultaneous inference of star-level parameters but also offers an elegant framework for effectively pooling open cluster information and propagating uncertainty. Markov Chain Monte Carlo techniques allow us to sample the posterior distribution using the Hamiltonian Monte Carlo algorithm.*

Results: *Our model's robust performance on a synthetic dataset with known parameters, coupled with its successful age estimation of the Pleiades Open Cluster (116.8 $\pm$ 1.9 Myr), represents a significant advancement in the field by overcoming key challenges that have hindered previous attempts mixing artifical intelligence paradigms. The resulting model signifies a new methodology for age estimation that can be applied to a wide range of open clusters, with the Pleiades serving as the initial test benchmark.*

Key words: *stars: fundamental parameters, low-mass - methods: Hamiltonian Monte Carlo, Neural Networks*

1 Introduction

The age is unarguably one of the most fundamental parameters to characterize any astrophysical object, since nearly every aspect of astrophysics deals with how things evolve with time. In fact, the largest astrophysical missions of our times must rely on age estimates in order to conclusively interpret their measurements and answer the pressing astrophysical questions they were designed to solve.

As stated by Soderblom (2015): *«The clocks of the cosmos tick constantly, but too softly for us to hear, and therefore we cannot directly measure the ages of stars»*. We only know one accurate age, our only true anchor: the age of the Sun. It comes from radiometric measurements of the oldest Solar System material, which yields 4.5672 $\pm$ 0.6 Gyr (Amelin et al., 2002). In addition, we have a theoretical limit to the age of the universe at 13.799 $\pm$ 0.021 Gyr based on the Lambda cold dark matter concordance model (Ooba et al., 2018; Planck Collaboration, 2016). The ages of astrophysical objects significantly different from these two anchors must rely on strong assumptions based on not fully tested or calibrated theories. In spite of the tremendous progresses of the past decade for the models, theories and numerical simulations of stellar evolution, the estimated ages are highly method- and model-dependent (Barrado, 2016).

Soderblom (2010) presented a summary of most dating techniques and proposed their classification into five groups: fundamental, empirical, semi-empirical, statistical, and modelling. Two of the most used techniques are isochrone fitting and chemical abundance, which are briefly explained below for the case of Lithium abundance

Isochrones In stellar evolution, an isochrone is a curve on the HR-diagram, representing a population of stars of the same age but with different masses. Stars change their positions on the HR-diagram throughout their life. A few years after it was recognized that nuclear burning was the energy sources of the stars, it was discovered that the age of giants and globular clusters could be determined by looking at the evolution of the main sequence and the position in the HR-diagram. Isochrones can be used to date open clusters because their members all have roughly the same age.

Lithium abundance The most stable isotope of Li, ^{7}Li, is destroyed in stellar interiors at temper-

*Corresponding author: fpalmero6@alumno.uned.es
Submitted: September, 2023

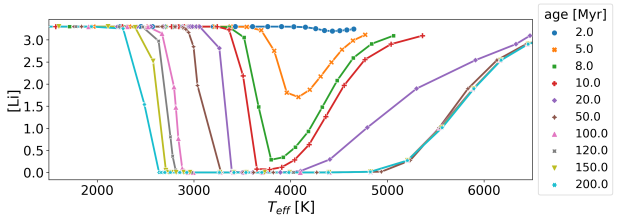


Figure 1: *Lithium abundance as a function of T_{eff} and age. The BT-Settl model (Allard, 2013) is used to generate the data.*

atures around 2.5 million Kelvins (Sestito & Randich, 2005). Due to convection, the surface abundance of Li can rapidly decrease up to a point of total depletion within a few million years. However, objects with masses below 0.06 $M_\odot$ never reach temperatures in their interiors high enough to burn Li. The presence or absence of Li is then highly sensitive to the mass and age of the object as observed in Figure 1. Thus, Li abundance can be used group stars with similar or the same age.

Star clusters are among the most powerful objects for use in calibrating stellar models owing to the common age, metallicity, and distance of their member stars. The nearest clusters are in many ways best suited for this type of work due to the high quality data. The vicinity ensures excellent astrometric, photometric, and spectroscopic measurements over a wide range of spectral types and stellar masses.
The Pleiades, also known as The Seven Sisters, is among the nearest star clusters to Earth, $\sim$ 134 pc (Galli, P. A. B. et al., 2017a). The Pleiades has long provided a benchmark for our understanding of stellar evolution and dynamics. The age estimate derived from isochrone fitting centre around 120 Myr (Heyl et al., 2022), which is in good agreement with age estimates based on the Lithium abundance (Galindo-Guil, F. J. et al., 2022).

In this study, we introduce a formalism for a Bayesian hierarchical model (BHM) that allows to simultaneously infer the low-level parameters of the individual stars, the middle hierarchy parameters of each age-dating technique, and the high hierarchy parameters of the cluster age. Conceptually, the inference problem can be viewed as a regression problem in which we aim to estimate the value of a continuous random variable (age) as a function of a number of other random variables (observations). BHMs provide us with a framework to incorporate domain probabilistic knowledge into the regression. Our goal is to give a first step towards a universal, autoconsistent, absolute and accurate method to estimate stellar ages over the entire time domain by constructing and applying a BHM that simultaneously infers the isochrone and Lithium abundance ages of several star clusters and associations. This is accomplished by the following sections: Section 2 provides a comprehensive overview of the fundamental concepts within the Bayesian framework, such as uncertainty propagation, priors and likelihood, and sampling techniques. Building upon this foundation, Section 3 delves into the intricate details of the BHM employed in this research, presenting a thorough description of its components (*dramatis personae*). In Section 4 the age estimation performance of the model is analyzed for synthetic data created by a generative model. Section 5 applies the model to estimate the age of the Pleiades working as a test benchmark for uor model.

2 Fundamentals

2.1 Probabilistic modelling

Parametric model A probabilistic model is an abstract description of a concrete system using mathematical concepts and language where variability is represented using probability distributions. The objective is to create a probability distribution that facilitates drawing inferences or making decisions based on data. If the probability distribution depends on some parameters, we have a parametric model[1]

Definition (Parametric model)**.** *Let Y be a continuous random variable for some data $y = \{y_i\}_{i=1}^N$. A model, or strictly a family of models, specifies the probability density function (pdf) of Y to be*

$$f_Y(y|\theta)$$

where $\theta = \{\theta_i\}_{i=1}^k$ belongs to the k-dimensional parameter space Ω_θ^k for some positive integer k.

Observations The observations y are a sample of the population, that is, they are a subset of the collection of all possible observations of a random variable. The model represents our hypothesis about how the data was generated. The parameters are introduced as an increase in degrees of freedom due to our limited knowledge of the system we are modelling (Appendix B). Therefore, we are compelled to make certain assumptions about $f_Y(y|\theta)$.

Exchangeability Under the assumption of exchangeability, $f_Y(y|\theta)$ is invariant to permutation of the indexes of y. The assumption is due to our lack of knowledge about how the indexes are relevant to $f_Y(y|\theta)$. If the observations are independent and

[1] Along the document we assume random variables as continuous (Appendix **??**)

identically distributed (iid), the model is symmetrical under indexes permutation since $f_Y(y|\theta)$ becomes

$$f_Y(y|\theta) = \prod_{i=1}^{N} f_Y(y_i|\theta) \tag{1}$$

Likelihood We wrote $f_Y(y|\theta)$ to emphasize that the pdf is a function of both y and θ, nevertheless the model pdf can be regarded as a function of θ for a fixed y. Our interest in this quantity, called likelihood, is motivated by the idea that it will be relatively larger for values of θ near that which generated the data. In fact, the likelihood is the probability of the data y given the parameters θ.

Prior There are two broad approaches to interpret probabilistic models: frequentist and Bayesian[2]. In the frequentist approach we treat θ as an unknown constant and therefore $\theta \in \Omega_\theta^k \subseteq \mathbb{R}^k$. In the second approach we think of the unknown θ that underlies our data as the outcome of a random variable Θ. Here probability represents a measure of uncertainty, where the random variable Θ is not necessarily the outcome of a random experiment[3]. Therefore, the pdf *a priori* (prior) $f_\Theta(\theta|I)$ represent our degree of belief in θ given the prior information I. I is simply whatever additional information that we may have about θ beyond data. Once the data have been observed, our beliefs about θ are contained in its pdf *a posteriori* (posterior) given the data

$$f_\Theta(\theta|y, I) = \frac{1}{Z} f_Y(y|\theta) f_\Theta(\theta|I) \tag{2}$$

where the normalization constant Z is the evidence, which has many uses in model evaluation and model averaging. Equation (2) is used for inference in this context.

Notation In order to simplify our notation, from now on we denote the posterior as $p(\theta|y)$, the prior as $p(\theta)$ and the likelihood as $p(y|\theta)$ or $\mathcal{L}(\theta)$.

Nuisance parameters There are some problems involving more than one parameter. It is dealing with such problems that the simple conceptual framework of the Bayesian approach reveals its principal advantages over other methods of inference. Although a problem can include several parameters of interest, conclusions will often be drawn for one, or only a few values of the parameters. In our case, the aim is to obtain the marginal posterior of the particular parameter of interest. Therefore, we first require the joint posterior distribution of all parameters, and then we integrate this distribution over those that are not of immediate interest to obtain the desired marginal distribution[4]. Parameters of this kind are often called nuisance parameters. Let $\theta = \{\theta_i\}_{i=1}^k \in \Omega_\theta^k$ be the parameters of a given model $p(y|\theta)$. If we split the parameter space into parameters of interest, $\phi \in \Omega_\phi^{l_\phi}$, and nuisance parameters, $\lambda \in \Omega_\lambda^{l_\lambda}$, such that $\Omega_\phi^{l_\phi} \oplus \Omega_\lambda^{l_\lambda} = \Omega_\theta^k$, the marginal posterior is

$$p(\phi|y) = \int_{\Omega_\lambda^{l_\lambda}} p(\theta|y) \mathrm{d}\lambda \tag{3}$$

Predictions To make inferences about an unknown observable, i.e., a future random variable Z, we follow a similar logic. The pdf of Z conditioned on observed data y is called posterior predictive pdf

$$\begin{aligned} p(z|y) &= \int_{\Omega_\theta^k} p(z, \theta|y)\, \mathrm{d}\theta \\ &= \int_{\Omega_\theta^k} p(z|\theta, y) p(\theta|y)\, \mathrm{d}\theta \\ &= \int_{\Omega_\theta^k} p(z|\theta) p(\theta|y)\, \mathrm{d}\theta \end{aligned} \tag{4}$$

The second and third lines display the posterior predictive as an average of conditional predictions over the posterior of θ. The last step follows from the assumed conditional independence of y and z given θ. On the other hand, it is also possible to make inferences about an unknown observable before the data y are considered

$$p(y) = \int_{\Omega_\theta^k} p(y, \theta)\, \mathrm{d}\theta = \int_{\Omega_\theta^k} p(\theta) p(y|\theta)\, \mathrm{d}\theta \tag{5}$$

This is often called the marginal distribution of y, but a more informative name is the prior predictive pdf.

2.2 Hierarchical modelling

Many statistical applications involve multiple parameters that can be regarded as related or connected in some way by the structure of the problem, implying that a model for these parameters should reflect their dependence. Therefore, the prior $p(\theta)$, as a joint probability of k random variables $\{\Theta_i\}_{i=1}^k$, is no longer

$$p(\theta) = \prod_{i=1}^{k} p(\theta_i) \tag{6}$$

since the parameters are not independent. Let $\mathcal{L}(\theta^1)$ be the likelihood of our model, where $\theta^1 \in \Omega_{\theta_1}^{l_1} \subset \Omega_\theta^k$.

[2] Appendix C.3
[3] Appendix C.1
[4] The marginal distribution of a subset of a collection of random variables is the probability of the variables contained in the subset. The process of compute the marginal distribution is called marginalization.

If θ^1 depends on other parameters $\theta^2 \in \Omega^{l_2}_{\theta_2} \subset \Omega^k_\theta$, such that $\Omega^{l_1}_{\theta_1} \oplus \Omega^{l_2}_{\theta_2} = \Omega^k_\theta$, the prior $p(\theta)$ can be factorized using the chain rule

$$p(\theta) = p(\theta^2)p(\theta^1|\theta^2) \tag{7}$$

In general, if we can group the parameters in l groups, such that there is a hierarchical dependency between them, the prior can be factorized

$$p(\theta) = p(\theta^l)\prod_{i=1}^{l} p(\theta^i|\theta^{i+1}) \tag{8}$$

where each $\theta^i \in \Omega^{l_i}_{\theta^i} \subset \Omega^k_\theta$ represents the subset of parameters in the i-th hierarchical level. This kind of models are known as Bayesian Hierarchical Models (BHM). The variables and conditional dependencies can be represented with a Probabilistic Graphical Model (PGM) via a directed acyclic[5] graph (DAG), where variables are represented by nodes, and edges represent dependence. In PGMs, stochastic variables are represented with circles and constants with squares. If the variable is known, as in the case of the data, it is represented with a filled symbol, otherwise with an empty symbol. If there is no line between two given elements, it indicates that they are assumed to be independent. Variables that repeat together, as in the case of the data, are grouped within a plate. The number of repetitions is indicated in one corner of the plate. For more details on PGMs see for example Koller and Friedman (2009).

2.3 Sampling the posterior distribution

The computing of the posterior distribution, key in any Bayesian model, is quite often unfeasible for large data sets or high-dimensional parametric spaces. The evidence,

$$Z \equiv \int_{\Omega^k_\theta} p(y|\theta)p(\theta)\ \mathrm{d}\theta, \tag{9}$$

makes the computation more difficult because the likelihood or the prior can have extremely complex structure, with multiple arbitrarily compact modes, unpredictably positioned in the (presumably high dimensional) parameter space Ω^k_θ. Another option is the use of a grid in the parametric space. The likelihood and the prior must be evaluated at each point in this grid and then multiplied. This approach is reasonable when the parametric space is of moderate dimension, but it requires the evaluation of the posterior q^k times, for a grid of q grid points in each dimension, and k the dimension of the parametric space. A different method and so far the only feasible approach for high dimensional parameter space is sampling the posterior using Markov chain Monte Carlo (MCMC) methods[6]. It is a Monte Carlo simulation that involves constructing a Markov chain with the desired distribution as its equilibrium distribution to obtain a sample from the distribution by recording states from the chain.

There are different approaches for sampling using MCMC. For complicated models with many parameters, classical methods such as random-walk Metropolis (Metropolis et al., 1953) and Gibbs sampling (Geman & Geman, 1984) may require an unacceptably long time to converge to the target distribution. This is in large part due to the tendency of these methods to explore parameter space via inefficient random walks. Hamiltonian dynamics can be used to produce effective proposals for the Metropolis algorithm, thereby avoiding the slow exploration of the state space that results from the diffusive behavior of simple random-walk proposals. Though originating in physics, Hamiltonian dynamics can be applied to most problems with continuous state spaces by simply introducing momentum variables (Duane et al., 1987). In fact, it turns out that HMC is the most suited algorithm to overcome pathologies arising in BHM (Betancourt & Girolami, 2013). HMC's increased efficiency comes at a price. First, HMC requires the gradient of the log-posterior. Computing the gradient for a complex model is at best tedious and at worst impossible, but this requirement can be made less onerous by using automatic differentiation (Griewank & Walther, 2008). Additionally, HMC performance is highly sensitive to two user-specified parameters: a step size, ϵ, and a desired number of steps, L. In particular, if L is too small then the algorithm exhibits undesirable random-walk behavior, while if L is too large the algorithm wastes computation (Brooks et al., 2011a). Hoffman and Gelman (2011) proposed the No-U-Turn Sampler (NUTS), an extension to HMC that eliminates the need to set a number of steps and also derived a method for adapting the step size parameter on the fly (Appendix D).

[5] Note that parameters at each hierarchical level must be conditionally independent. Let $\theta^i \in \Omega^{l_i}_{\theta^i}$ be the set of parameters at hierarchical level l_i and S its power set, then for each $j = 1, ..., l_i$

$$p(\theta^i_j) = p(\theta^i_j|s),\ \forall s \in S \setminus \{\emptyset\}$$

On the other hand, if there is any cyclic dependency between hierarchical levels, the model is no longer hierarchical. □

[6] Variational Inference methods could be applied to compute an approximation of the posterior, nevertheless they provide a locally-optimal not a numerical approximation to the exact posterior (Kucukelbir et al., 2016). In fact, some MCMC approaches use Variational Inference to compute their starting points.

3 Bayesian hierarchical model

In this section, we will introduce the model that serves as the foundation for the subsequent analyses. The focus here is on presenting the model itself, deferring detailed arguments and discussions to the following section. The intent is to establish a comprehensive understanding of the model's structure and components before delving into its underlying justifications.

The variables used in the model can be divided in three groups: model parameters, deterministic and observed variables. The model parameters will be inferred. Some intermediate deterministic variables are introduced to facilitate the computation of the complete likelihood of the model. The observed variables correspond to the data provided. Figure 2 shows the graph of the proposed BHM following the notation explained in Subsection 2.2, whereas Table 1 present the variables in three different groups.

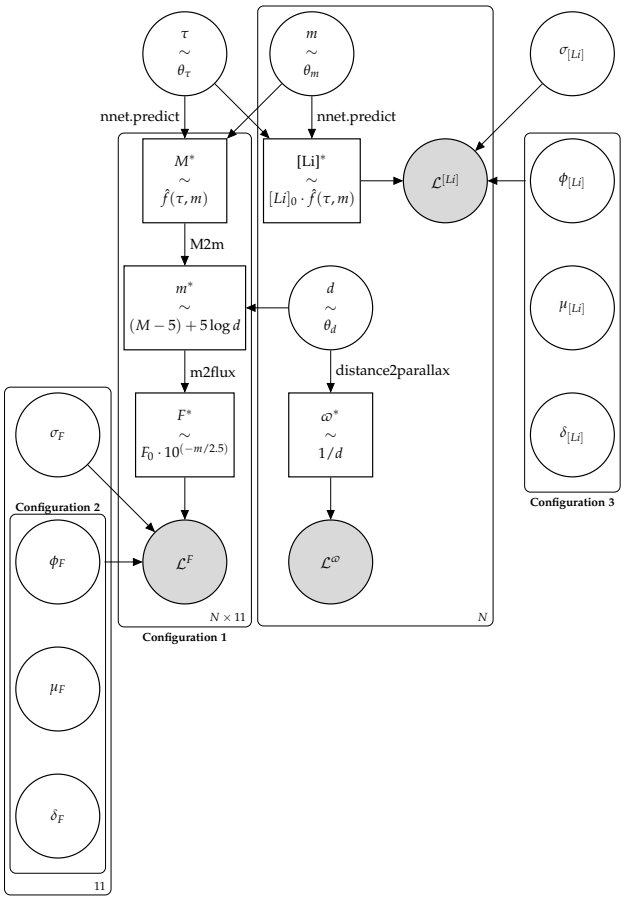


Figure 2: *Proposed BHM to determine the ages of star cluster and associations following plate notation described in Section 2. To symplify our notation, the hyperparameters corresponding to the mixture likelihoods have been groued into clusters showing only one edge that represents the dependency between each hyperparameter and the likelihood.*

3.1 Model variables

There is a core structure representing the generative model, i.e., a parametrized, quantitative description of a statistical procedure that could reasonably have generated the data.

High-level parameters Three main parameters describe the model: age, mass and distance. The most interesting parameter is the age of the cluster. Assuming that there are N stars of approximately the same age, we can group them into a global parameter, the age of the cluster denoted by θ_τ. Distance, $\theta_d = \{\theta_d^i\}_{i=1}^N$, and mass, $\theta_m = \{\theta_m^i\}_{i=1}^N$, refers to individual stars so that their dimension is equal to the number of stars in the dataset, N.

Neural Network as a black-box A physical model, called BT-Settl (Allard, 2013), is used to train a Neural Network (NN) and include its predictions as a deterministic relation, $\hat{f}(\tau, m)$, into our model. This relation is represented as a function `nnet.predict`. BT-Settl is a model that can simulate stellar atmospheres of small stars and brown dwarf. Our NN works as a "black-box" interpolator of the synthetic grid produced by BT-Settl. The inputs are age and mass, and the outputs are absolute magnitudes for each spectral band and Lithium abundance expressed as a ratio, relative to its initial Lithium abundance, $[\mathrm{Li}]_0$.

Intrinsic dispersion In general, it is customary to include an intrinsic dispersion in the model to account for over-simplistic assumptions or underestimated uncertainties (Appendix B). The process of fitting a NN to data involves optimizing its weights using a cost function J. It has been proven that a NN with one hidden layer with an arbitrary number of hidden units are capable of approximating any nonlinear function of the inputs to any desired optimal J^* (Hornik et al., 1989). In this sense, NN are a class of universal approximators. In practice, there are some additional sources of uncertainty which NNs do not take into account because they are just maximum likelihood estimators. In our case, we include an intrinsic dispersion for the likelihoods coming from the NN predictions, that is: observed Lithium abundance and flux. The intrinsic dispersion for Lithium abundance, $\sigma_{[\mathrm{Li}]}$, can be understood as a measure of uncertainties coming from imperfect predictions of the NN or due to BT-Settl model itself. It is a 1-dimensional parameter because we assume that it is the same for every observation. On the other hand, photometric observations comes from a measurement process that involves complex physical phenomena, such as extinction, that are not included into our model. The intrinsic dispersion for flux, σ_F, is added to the model

Table 1: *Model variables*

Parameters variables				
Notation	Description	Dims.	Units	Distribution
θ_τ	Open cluster age	1	Myr	$\mathcal{N}(\tau \mid \mu_\tau,\ \sigma_\tau)$ or $\mathcal{U}(\tau \mid a_\tau,\ b_\tau)$
θ_d	Distance to stars	N	pc	$\mathcal{N}(d \mid \mu_d,\ \sigma_d)$ or $\mathcal{U}(d \mid a_d,\ b_d)$
θ_m	Stars mass	N	$M_\odot$	$\mathcal{U}(m \mid a = 0.01,\ b = 1.5)$
ϕ, μ, δ	Mixture hyperparameters	It depends on the observed data and its likelihood		
$\sigma_{[\mathrm{Li}]}$	Lithium intrinsic dispersion	1	dex	HalfNormal with $\sigma = 1$
σ_F	Flux intrinsic disperesion	11	erg/s/cm^2	HalfNormal with $\sigma = 1$
Deterministic variables				
Notation	Description	Dims.	Units	Formula
ϖ^*	True parallax	N	mas	$\varpi = 1/d$, for d in kpc
$[\mathrm{Li}]^*$	True Li abundance	N	dex	$\hat{f}(\tau, m) \cdot [\mathrm{Li}]_0$
M^*	True absolute magnitude	$N \times 11$	1	$\hat{f}(\tau, m)$
m^*	True apparent magnitude	$N \times 11$	1	$m = (M - 5) + 5\log(d)$
F^*	True flux	$N \times 11$	erg/s/cm^2	$F = F_0 \cdot 10^{(-m/2.5)}$
Observed variables				
Notation	Description	Parameters		Distribution
$\mathcal{L}^{[\mathrm{Li}]}$	Lithium likelihood	$\theta_\tau, \theta_m, \sigma_{[\mathrm{Li}]}; \phi_{[\mathrm{Li}]}, \mu_{[\mathrm{Li}]}, \delta_{[\mathrm{Li}]}$		Normal or Mixture
$\mathcal{L}^{\varpi}$	Parallax likelihood	θ_d		Normal
$\mathcal{L}^{F}$	Flux likelihood	$\theta_\tau, \theta_m, \sigma_F; \phi_F, \mu_F, \delta_F$		Normal or Mixture

to account for these over-simplistic assumptions and any other underestimated uncertainty analogous to the Lithium case. It is an 11-dimensional parameter since we assume that each spectral band shares the same uncertainty process. The flux in the spectral band x is represented by F^x, therefore σ_{F^x} is a 1-dimensional parameter that quantifies flux intrinsic dispersion in spectral band x. Then, for flux and Lithium, if Δy_i represents the uncertainty of the observed value y_i, and σ the corresponding intrinsic dispersion, the likelihood of this data point has uncertainty

$$\Delta y_i + \sigma, \tag{10}$$

for all $i = 1, ..., N$.

Pruning outliers Sometimes models are very sensitive to outliers, that is, data points that are substantially farther from the expected relation (or not on the relation at all) because of unmodeled experimental uncertainty or unmodeled but rare sources of noise. It could even be due to a restricted model scope that does not apply to all the data. There is a standard procedure in astrophysics known as "sigma clipping", which identifies and removes outliers from a dataset. In our BHM, we follow a preferable approach modelling these outliers instead of removing them, i.e., our model is able to generate the outliers as well. Hogg et al. (2010) proposed a likelihood function defined as a mixture of two likelihoods: foreground component $\mathcal{L}_{\mathrm{fg}}$ and background component $\mathcal{L}_{\mathrm{bg}}$

$$\mathcal{L}(\theta, \phi, \mu, \delta) \equiv (1 - \phi)\mathcal{L}_{\mathrm{fg}}(\theta) + \phi\mathcal{L}_{\mathrm{bg}}(\mu, \delta) \tag{11}$$

where ϕ is the mixture weight, i.e., prior probability of the foreground component, (μ, δ) are the mean and standard deviation of the outlier distribution, and θ stands for any other parameter of the model. The terms foreground and background are just the notation used by Hogg et al. (2010) to represent the inliers distribution and outliers distribution, respectively. The outlier modelling approach is followed for Lithium abundance likelihood, $\mathcal{L}^{[\mathrm{Li}]}$, and flux likelihood, $\mathcal{L}^F$, where the parameters are

Pruning outliers parameters

Likelihood	Parameters	Dimensions
$\mathcal{L}^{[\mathrm{Li}]}$	$\phi_{[\mathrm{Li}]}, \mu_{[\mathrm{Li}]}$, and $\delta_{[\mathrm{Li}]}$	1
$\mathcal{L}^F$	ϕ_F, μ_F, and δ_F	11

They are optional parameters, and we decide whether to include them or not. In our approach, the use of a mixture likelihood, eliminates the need for pruning outliers by downweigthting their influence during inference, without resorting to arbitrary thresholds. In Section 4 we use this approach to study the robustness of our model for dealing with outliers.

Likelihoods Assuming a common initial Lithium abundance of 3.3 dex, we multiply the NN prediction for Lithium abundance by this quantity to obtain the true Lithium abundance, $[\text{Li}]^*$. The true Lithium abundance is then compared to observed Lithium abundance in the likelihood. The likelihood $\mathcal{L}^{[\text{Li}]}$ can be expressed as a product of individual likelihoods assuming independent and identically distributed (iid) observations

$$\mathcal{L}^{[\text{Li}]}\left(\theta_\tau, \theta_m, \sigma_{[\text{Li}]}\right) = \prod_{i=1}^{N} p_{[\text{Li}]}\left([\text{Li}]_i | \theta_\tau, \theta_m^i, \sigma_{[\text{Li}]}\right) \tag{12}$$

where the parameters for the mixture likelihood distribution must be included if they are used, we do not show them for simplicity. The $[\text{Li}]^*$ is an N-dimensional variable, but the likelihood is always a scalar.

On the other hand, the true absolute magnitude, M^*, given by the NN predictions is represented by an $N \times 11$ matrix, i.e., a row vector for each star where columns represent spectral bands. Since our observations cannot be absolute magnitudes[7], we convert these values to apparent magnitudes given the following relation

$$m = (M - 5) + 5\log(d) \tag{13}$$

where m is the apparent magnitude, M is the absolute magnitude, and d is the distance to the star measured in parsecs. Therefore, we compute the true apparent magnitude of our model, m^*, taking into account M^* and θ_d. This relation is represented as a function `M2m`. Distance to stars are never measured due to the vast scales involved in interstellar and intergalactic distances. They are inferred from observable quantities. Here we use parallax to compute distance given the following relation

$$d\ [\text{kpc}] = \frac{1}{\varpi\ [\text{mas}]} \tag{14}$$

Therefore the distance, θ_d, in kiloparsecs (kpc) is related to the true parallax, ϖ^*, in milliarseconds (mas) by means of the relation `distance2parallax`. Thus, the true parallax is N-dimensional. Then the true parallax is compared to the observed parallax. The total likelihood for parallax is again the product of individual likelihoods for the parallax measurement of each star

$$\mathcal{L}^{\varpi}(\theta_d) = \prod_{i=1}^{N} p_{\varpi}\left(\varpi_i \mid \theta_d^i\right) \tag{15}$$

Once we know the distance, we can compute the true apparent magnitude, m^*. We could compare m^* with the observed apparent magnitude, nevertheless it is not convenient in this case[8]. Therefore, although our observations are apparent magnitude, we introduced a *deus ex machina* in the form of a transformation to flux

$$F = F_0 \cdot 10^{-m/2.5} \tag{16}$$

where F_0 is the zero point. The true flux, F^*, is then related to m^* by `m2flux`. Finally, after the apparent magnitude observations are transformed to flux following the same relation (16), we can compute the likelihood assuming iid observations

$$\mathcal{L}^{F}(\theta_\tau, \theta_m, \sigma_F) = \prod_{i=1}^{N} p_F\left(F_i \mid \theta_\tau, \theta_m^i, \sigma_F\right) \tag{17}$$

where F_i is an 11-dimensional vector corresponding to the i-th data point which inputs are F_i^x for each spectral band x.

3.2 Software

In order to apply the theoretical framework presented in the previous section to real-world data and conduct meaningful Bayesian inference, a custom software implementation has been developed using the PyMC library (Salvatier et al., 2015) in Python. PyMC is a probabilistic programming library for Python that facilitates Bayesian statistical modeling through intuitive and flexible syntax, enabling users to define and conduct Bayesian inference for complex models. This subsection provides a detailed overview of the software architecture and the various modules that constitute it (Figure 3).

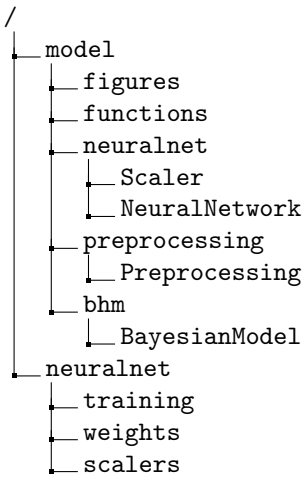


Figure 3: *Software structure*

The software core is the `bhm` module which contains the main class called `BayesianModel`. The class

[7] Absolute magnitude is an intrinsic property impossible to measure directly in practice.

[8] The likelihood for apparent magnitude do not follow a normal distribution. See Section 4

is instantiated by providing the data which has been previously preprocessed[9] using the `Preprocessing` class from `preprocessing` module. The model is then compiled given the priors definition for age and distance[10], and the options of whether to use a mixture likelihood or standard, for both photometric and Lithium observations. The mixture likelihood has been developed by ourselves since it is not available in the PyMC library. The custom implementation required two main functions: the log-likelihood[11] of the pdf and a random method to randomly generate samples[12]. Within the model compilation, the `Scaler` and `NeuralNetwork` classes from `neuralnet` module are called to provide the NN predictions. The `NeuralNetwork` class loads the optimal weights found during the training and their corresponding scalers if they were used. The model has a method called `summary` to provide a visual representation of its variables following the plate notation described in Section 2. Once the model is compiled, we are able to sample the posterior distribution with the `sample` method. The returned sample is stored in an InferenceData object. Finally, model evaluation is implemented by additional methods such as posterior predictive checks that call to the `figures` module for graphical representations most of them implemented in ArviZ (Kumar et al., 2019).

4 Methods

In this section, we delve into the practical aspects of our Bayesian model implementation and evaluation. Subsection 4.1 provides a comprehensive overview of the NN architecture employed as a fundamental component of our model, encompassing the training process, evaluation methods, and network topology. As we seek to demonstrate the efficacy of our model in practice, in Subsection 4.2 we will describe how the model evaluation is carried-out and test its performance in Subsection 4.3 by applying these techniques.

[9] Preprocessing steps are described in detail in Section 4.

[10] The mass prior is fixed given the constraints imposed by the NN as we will see in the next Section.

[11] The log-likelihood function is a logarithmic transformation of the likelihood function.

$$l(\theta) = \log(\mathcal{L}(\theta))$$

Since logarithms are strictly increasing functions, maximizing the likelihood is equivalent to maximizing the log-likelihood. However, for practical purposes it is more convenient to work with the log-likelihood function.

[12] The random method is only required for prior and posterior predictive.

4.1 Neural Network

In modern supervised learning, many Deep Neural Networks (DNN) are able to interpolate the data: the loss function can be driven to near zero on all samples simultaneously Cuchiero et al., 2020. In this work, we explicitly exploit this interpolation property for a grid of pre-calculated BT-Settl models to interpolate the spectral energy distribution of stars. BT-Settl is a stellar interior and stellar atmosphere model that attempts to describe the parameters of low-mass stars and brown dwarfs. Among others, those stars parameters are age, mass, Lithium abundance (as a ratio of its initial abundance) and absolute magnitudes for each spectral band. These parameters make up our dataset D. The SVO Theory Server provides data for 70 collections of theoretical spectra and observational templates, including the ones used in this document.

Preprocessing The original dataset of $|D| = 838$ is split into training $\mathcal{T}$ and validation $\mathcal{S}$ sets with a proportion of 80 % to 20%, respectively. The inputs x are age and mass, and the targets y are Lithium abundance and absolute magnitudes for each of the different spectral band[13]

$$x = (x_1, x_2) \equiv (\tau, m),$$
$$y = (y_1, y_2, ..., y_{12}) \equiv (G, G_{bp}, ..., \mathrm{Li})$$

There is a preprocessing transformation consisting on two different scalers before training. The first scaler applied is a Power-Transform implementing the Box-Cox method (Box & Cox, 1964) to make the data more Gaussian-like. Thus, for any data $x \in D$

$$x(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log x & \text{if } \lambda = 0 \end{cases} \tag{18}$$

where λ is a transformation parameter which is determined through maximum likelihood estimation. The second scaler is a min-max scaler which scales and transforms inputs to a specific range, in our case between 0 and 1. It is a linear mapping of the minimum and maximum values of the original data feature to the chosen range

$$x = \frac{x - x_m}{x_M - x_m} \tag{19}$$

where x_m and x_M represent the minimum and maximum[14] values for all $x \in D$, respectively. The values x_m and x_M, together with λ, are stored in `scalers` from Figure 3 for the scalers implementation in the Bayesian model.

[13] The spectral bands are: G, G_{bp}, G_{rp}, J, H, K_s, g, r, i, y, and z.

[14] The terms minimum and maximum in an N-dimensional data point is misleading but usually employed. They refer to the minimum or maximum over each feature $n = 1, ..., N$ in x.

Problem Setting We consider a supervised learning task where the model[15] is parametrized by $\omega \in \Omega \subseteq \mathbb{R}^p$. In general, the objective function can be expressed as an expectation over $z \in \mathcal{Z}$, a random variable indexing the samples of the training set $\mathcal{T}$

$$f(\omega) \triangleq \mathbb{E}_{z\in\mathcal{Z}}[\ell_z(\omega)] \tag{20}$$

where each ℓ_z is the loss function associated with the sample indexed by z. Our learning task is then a problem of finding a feasible set of parameters $w_* \in \Omega$ that minimizes f

$$w_* \in \underset{\omega\in\Omega}{\operatorname{argmin}} f(\omega) \tag{21}$$

The interpolation problem consist on finding a solution w_* that simultaneously minimizes all individual loss functions

$$\forall z \in \mathcal{Z}, \ell_z(w_*) = 0. \tag{22}$$

where, in some cases, it is more realistic to relax (22) to $\forall z \in \mathcal{Z}, \ell_z(w_*) = \epsilon$ for a small positive ϵ.

Neural Network Topology The network, shown in Figure 4, consists of an input layer with two nodes representing age and mass features. Subsequently, three hidden layers with 64 units each were incorporated, utilizing the rectified linear unit (ReLU) activation function and He normal initialization (He et al., 2015) for weight parameters. The bias are initialized at zero. The final output layer consists of two branches both with Glorot (Glorot & Bengio, 2010) weight initialization: one for photometry $y^{(M)}$ with 11 units and the other for Lithium $y^{(Li)}$ with a single sigmoid activation unit. The model can be considered a DNN due to the number of hidden layers and units. The rationale behind the increase in complexity comes from the fact that BT-Setll models are synthetic datasets without noise. The lack of noise in the training dataset avoid possible overfitting problems[16].

The ReLU activation function employed for the hidden layers mitigates the vanishing gradient problems found with other activation functions. In addition, ReLU allows an efficient computation due to its simplicity. He initialization, proposed by He et al. (2015), was specifically designed to address common issues that arises when using ReLU activation functions in DNN, allowing our DNN to be trained more efficiently and effectively. The sigmoid activation function is usually employed for classification problems, nevertheless, we have opted to define Lithium output unit with sigmoid activation in our regression problem because it effectively constraints the output values within the desired range of 0 to 1.

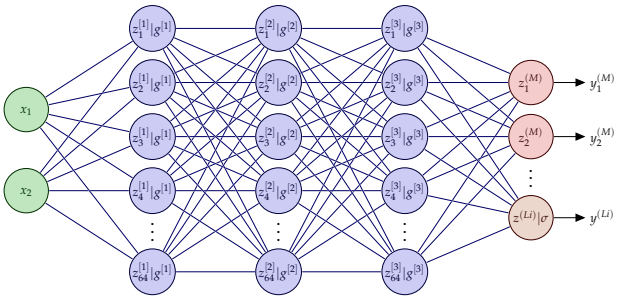


Figure 4: *Deep Neural Network for BT-Settl models interpolation. The inputs are age and mass, while the outputs are photometric values (for each spectral band) and Lithium abundance. The activation functions $g^{[i]}$ are ReLUs for each $i = 1, 2, 3$. Only the output unit corresponding to the Lithium has a non-linear activation function, the sigmoid activation σ.*

Learning-rate decay The model implements an algorithm for learning-rate decay called Adaptive Learning-rates for Interpolation with Gradients (ALI-G) (Berrada et al., 2020). ALI-G retains the two main advantages of Stochastic Gradient Descent (SGD), which are (i) a low computational cost per iteration and (ii) good generalization performance in practice. The steps are represented in Algorithm 1, where $\prod_\Omega(\omega)$ is the Euclidean projection of the vector $\omega \in \mathbb{R}^p$ on the set Ω.

Algorithm 1 ALI-G algorithm

Require: maximal learning-rate η, initial feasible $\omega_0 \in \Omega$, small constant $\delta > 0$

1: $t = 0$
2: **while** not converged **do**
3: Get $\ell_{z_t}(\omega_t)$, $\nabla\ell_{z_t}(\omega_t)$ with z_t draw iid
4: $\gamma_t = \min\left\{\frac{\ell_{z_t}(\omega_t)}{\left\|\nabla\ell_{z_t}(\omega_t)\right\|^2+\delta}, \eta\right\}$
5: $\omega_{t+1} = \prod_\Omega\left(\omega_t - \gamma_t\nabla\ell_{z_t}(\omega_t)\right)$
6: t = t + 1
7: **end while**

In our implementation, we use a maximum learning rate of $\eta = 0.1$ without Nesterov momentum aceleration (Botev et al., 2016).

Loss function The mean squared error (mse) is the loss function for photometry. In the case of Lithium, the mean absolute error (mae) loss function provides better results since Li values are between 0 and 1, and the squared value could underestimate the error. The total loss function of the model is then the sum of both. The implementation of other typical

[15]Please note that model here refers to the NN we are describing. In order to avoid any misunderstanding, we employ the term Bayesian model for the whole model, where the NN is only a component.

[16]Complex models are prone to overfitting when they capture and model the noise in the data, essentially memorizing random fluctuations instead of learning meaningful patterns.

regression loss functions such as Hubber loss (Huber, 1964) were useless. This ineffectuality arises from the inherent nature of the dataset, where the demand for robustness is absent.

Training The training process is executed over 10,000 epochs with a batch size of 16, and validation is performed using the root-mean-square error (rmse) metric over the validation dataset $\mathcal{S}$. The learning curves are shown in Figure 5 while the model performance on validation set $\mathcal{S}$ is showed in Table 2.

Table 2: *Neural Network Performance on evaluation set* $\mathcal{S}$

Target	loss	rmse
Li	0.0043 (mae)	0.0116
M	0.0102 (mse)	0.1009

Performance assessment Lithium predictions, as a function of two different color indexes[17] and age, are evaluated given the Figure 6. On the other hand, the photometric predictions are evaluated given the HR-diagrams in Figure 7. These figures have been carefully chosen to represent the physical meaning of the predictions, making possible an evaluation with an eye on its future implementation in the Bayesian model. Figure 6 (top panel) shows how the model fits the Lithium depletion lines in general, but fails to predict the right separation between depletion lines for two different isochrones of 70 Myr and 1000 Myr in a low-temperature range (represented in gray). In BT-Settl models, there is no significant differences between these depletion lines, while in the model predictions they could be considered significant. Looking at the possible implications in the Bayesian model, these significant differences will be more relevant in the age estimation in our model than according to BT-Settl. The culprit of this behavior could be the distribution of age points in the BT-Settle grid (Supplementary Figure 1). Indeed, the age values are concentrated at low-age, being the 75 % percentile only 600 Myr. Therefore, the low number of instances for age values greater than 600 Myr could reduce the model performance for those inputs values, nevertheless, there is not enough evidence to support the hypothesis as shown in Supplementary Figure 2. Apart from those discrepancies between BT-Settl models and our model predictions, we have not found any other evidence of a poor fit. As shown in Figure 7, the model fits the isochrones with enough quality to be included in our Bayesian model.

[17]Color indexes are closely related to the effective temperature of stars. The expected relation for Lithium and color index is the similar to Figure 1.

Implementation The software for training and evaluation was Keras library for Python. Those parameters which have not been mentioned in this subsection correspond to default values in Keras. The ALI-G algorithm employed is a Tensorflow implementation provided by the authors of Berrada et al. (2020) and can be found in their repository. The optimal weights w_* founds during training are stored in `weights` from Figure 3 for the NN implementation in the Bayesian model.

4.2 Model evaluation

In this subsection, we delve into the methodology employed to evaluate our Bayesian model. To ensure the robustness and reliability of our model, we conducted a comprehensive evaluation, considering both its overall performance and the individual contributions of specific model configurations. This previous step is essential for gaining insights into the model capabilities and limitations. The model evaluation is carried out as follows: a synthetic dataset with known parameters is generated in order to gauge the model ability to accurately recover information. These synthetic datasets serve as controlled experiments, enabling us to precisely measure the model performance given ground-thruth information. A figure-of-merit is required to measure the goodness of our inference. In our case, we follow a two-fold approach: first, the inference is qualitative and quantitative evaluated in terms of accuracy, that is, how close are our estimations from the ground-thruth; second, the goodness of our inference is measured based on posterior predictive checks which compare the empirical distribution of the data to the distribution described by the Bayesian model. In addition, since we are sampling the posterior distribution using MCMC, the chains' convergence must be tested. Finally, it is also important to assess the impact of the prior distribution on the posterior. A sensitivity analysis is performed investigating how the analysis compares for multiple prior distributions. The following methodology is employed for different model configurations.

Synthetic datasets Synthetic datasets are generated from prior predictive distribution, also called marginal distribution, defined in (5).

$$p(y) = \int_{\Omega_\theta^k} p(y,\theta)\,\mathrm{d}\theta = \int_{\Omega_\theta^k} p(\theta)p(y|\theta)\,\mathrm{d}\theta \qquad (23)$$

The marginal distribution is the probability of our data y marginalized over the model parameters θ. If we restrict the model parameters to a constrained parameters space Ω_θ^k, the marginal distribution will be highly biased since the marginalization is done

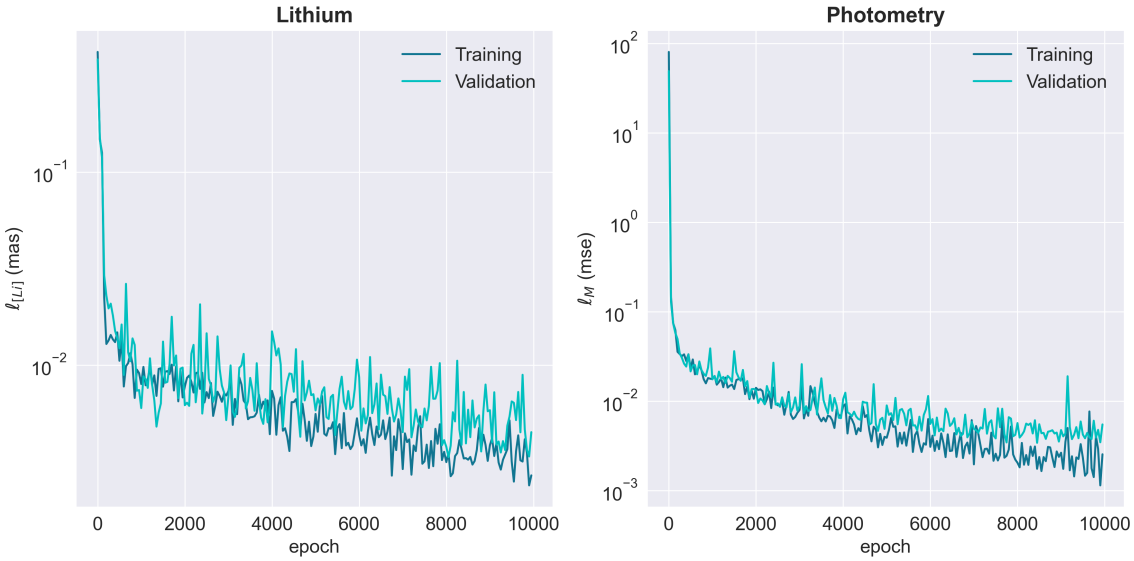


Figure 5: *Learning curves for Lithium and photometry outputs.*

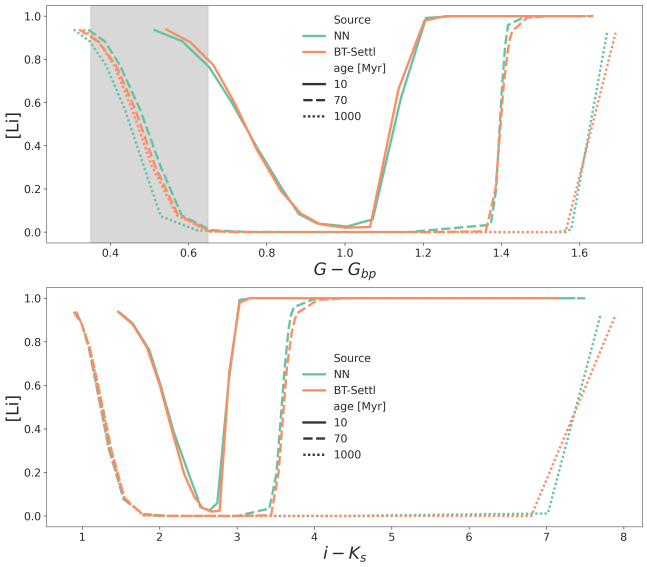


Figure 6: *Lithium predictions compared to BT-Settl models.*

over a limited parameters space. As an example, if we use a normal distribution for our prior,

$$p(\theta) \sim \mathcal{N}(\theta|\mu,\sigma) \tag{24}$$

with $\sigma < \epsilon$ for some ϵ, the marginalization over those values of θ distant from μ does not contribute to the marginal distributions since $\mathcal{N}(\theta|\mu,\sigma) \to 0$ when $\frac{1}{\sigma}\|\theta - \mu\| \to +\infty$. In addition, the main advantage of using the marginal distribution instead of simply a fixed value of θ is that it also incorporate uncertainty. Indeed, if we set our prior to a distribution with value θ_0 with zero uncertainty, such as

$$p(\theta) = \delta(\theta - \theta_0) \tag{25}$$

where δ is the Dirac delta function, the marginal distribution

$$p(y) = \int \delta(\theta - \theta_0) p(y|\theta)\, \mathrm{d}\theta = p(y|\theta_0) \tag{26}$$

becomes the likelihood evaluated at θ_0. However, if we relax our prior definition to also incorporate uncertainty

$$p(\theta) = \frac{1}{N}\sum_{i=0}^{N} \delta(\theta - \theta_i) \tag{27}$$

for some $N \in \mathbb{N}$, the marginal distribution becomes

$$p(y) = \int \frac{1}{N}\sum_{i=0}^{N} \delta(\theta - \theta_i) p(y|\theta)\, \mathrm{d}\theta \tag{28}$$

$$= \frac{1}{N}\sum_{i=0}^{N} p(y|\theta_i) \tag{29}$$

and any statistics $\langle g(y) \rangle$ of the data

$$\langle g(y) \rangle = \int g(y) p(y)\, \mathrm{d}y \tag{30}$$

$$= \frac{1}{N}\sum_{i=0}^{N} \int g(y) p(y|\theta_i)\, \mathrm{d}y \tag{31}$$

$$= \frac{1}{N}\sum_{i=0}^{N} \langle g(y|\theta_i) \rangle \tag{32}$$

incorporates the uncertainty coming from our uncertainty in θ. This is of great importance since the performance of our model in a Bayesian context is not only based on point-estimates but in pdf where uncertainty must be considered. Thus, our synthetic dataset consist on N samples randomly drawn according to $p(y)$.

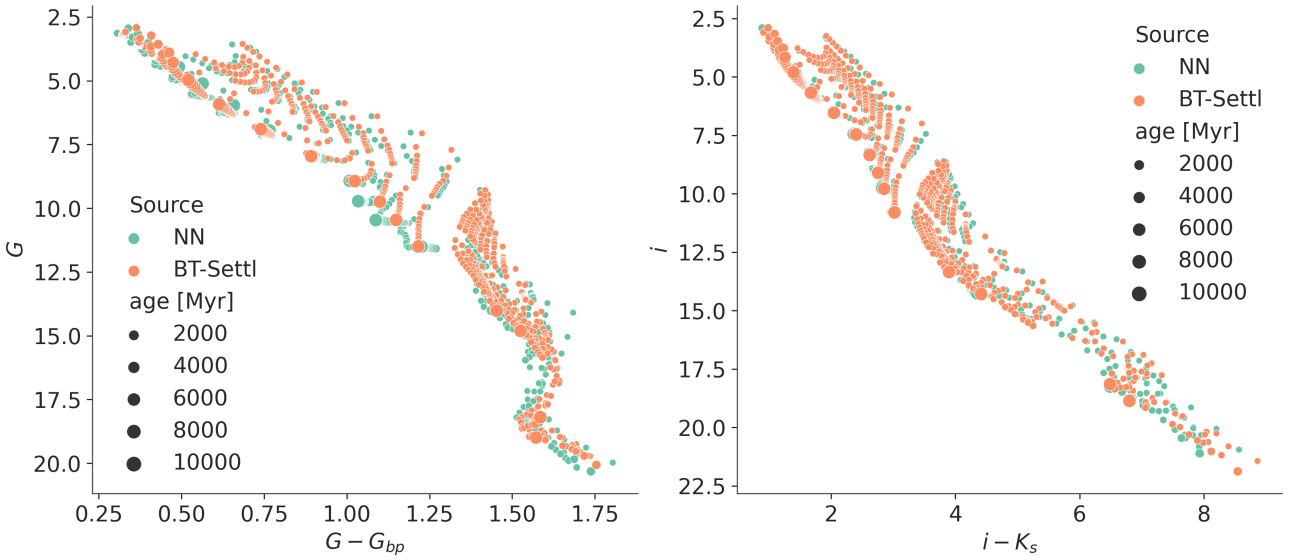


Figure 7: *Neural Network evaluation based on HR-diagram.*

Accuracy metrics As stated before, Bayesian models return a probability distribution instead of point-estimates. The accuracy metrics that represent how close are our inference estimation to the ground-thruth requires a specific definition that takes into account its probabilistic nature to define closeness. In our case, the metric is the Euclidean distance between each mean, each standard deviation and highest density intervals[18] (HDI). The idea is to quantitatively compare both distribution taking into account that we only have samples from the marginal distribution as our dataset and samples from the posterior distribution instead of its closed form. A visual representation is also showed to qualitatively compare distributions.

Posterior predictive checks Posterior predictive checks compare the empirical distribution of the data to the distribution described by the Bayesian model represented by the posterior predictive distribution defined in (4). The graphical representation of both distribution in an overlay density plot is usually employed but a poor choice because it is difficult to judge differences between the overlaid densities visually. In our case, we complement the overlaid comparison by plotting quantiles of the observed data against quantiles of the theoretical distribution, namely quantile-quantile (Q-Q) plots[19]. Q-Q plots are straightforward to interpret. If the data points follow a straight line (the diagonal line), it suggests that the dataset closely matches the theoretical distribution. Deviations from the line can indicate departures from the specified distribution. In addition, since we are dealing with a problem with physical interpretation, we also employed HR-diagrams to directly compare data points from our dataset with samples randomly drawn according to the posterior predictive distribution.

Convergence diagnostic In convergence diagnostics our aim is determining convergence of the underlying Markov chain to stationary state and convergence of Monte Carlo estimators to population quantities. The Gelman-Rubin (GR) diagnostic (Gelman & Rubin, 1992) appears to be the most popular method for assessing samples obtained from running MCMC algorithms (Roy, 2019). The GR diagnostic relies on multiple parallel chains to construct two estimators of the variance of the samples X, namely the within-chain variance estimate and the pooled variance estimate. The statistics $\hat{R}$ is then the ratio of the between-chain variance to the within-chain variance. Thus, if all chains have converged to the same distribution, $\hat{R}$ should be[20] close to 1. The effective sample size (ESS) for a MCMC-based estimator is also employed. ESS, in this context, quantifies how many iid samples are equivalent in information content to the actual MCMC sample. A higher ESS indicates that the samples are less correlated, and the MCMC estimate is more reliable. A lower ESS suggests that the samples are highly correlated, and one may need a larger MCMC sample size to obtain accurate estimates (Vehtari et al., 2021). Finally, the convergence of Monte Carlo estimators is evaluated in terms of Monte Carlo standard error for different quantiles.

[18]The highest density interval is a statistical concept that represents a range of values within which a probability distribution places the most likely values

[19]The Q-Q-plots are recommended by Eadie et al. (2023), one of the main references of this work for model evaluation.

[20]Gelman and Rubin (1992) argue that since the chains are started from an over-dispersed initial distribution, in finite samples, the numerator overestimates the target variance whereas the denominator underestimates it, making $\hat{R}$ larger than 1.

Sensitivity analysis A sensitivity analysis is performed investigating how the analysis compares for multiple priors with different age and distances prior distributions. The remaining parameters have been keep fixed for two main reasons: the focus on only two parameters reduces the analysis complexity and some parameters must be fixed to ensure convergence or inference within our model limits. Indeed, based on prior predictive checks and trial-error, we have defined some model parameters such as intrinsic dispersion, for both flux and Lithium, and mixture likelihood hyperparameters with those values that produces a more efficient sampling and low divergence rates. On the other hand, mass prior distribution must be bounded within the range of mass limits imposed by BT-Settl. The use of mass values outside the limits of BT-Settl grid is considered as extrapolation and our NN is an interpolator. Those values outside limits could produce catastrophic results. The fixed priors are defined in Table 1 while the hyperparameters are in Supplementary Table 1.

In general, three different priors setting are employed for each model configuration: one uninformative, and two weakly informative (right and wrong). The use of noninformative and weakly informative prior is due to the fact that our aim is to test the capacity of the model to recover the initial information without prior knowledge. In a synthetic dataset without noise and with known parameters, the possible inference with informative priors is irrelevant beyond an initial performance test for development.

4.3 Performance test

The synthetic datasets $\mathcal{S}$ are generated from a generative model

$$\theta_\tau \sim \mathcal{N}(\mu_\tau = 120, \sigma_\tau = 1) \tag{33}$$
$$\theta_d \sim \mathcal{N}(\mu_d = 135, \sigma_d = 1) \tag{34}$$

with $|\mathcal{S}| = 50$.

The inference is based on the priors

Table 3: *Prior settings for Bayesian model evaluation*

Models	Priors
Uninformative (U)	$\theta_\tau \sim \mathcal{U}(a_\tau = 10, b_\tau = 300)$ $\theta_d \sim \mathcal{U}(a_d = 10, b_d = 300)$
Informative (I)	$\theta_\tau \sim \mathcal{N}(\mu_\tau = 120, \sigma_\tau = 20)$ $\theta_d \sim \mathcal{N}(\mu_d = 135, \sigma_d = 20)$
Wrong (W)	$\theta_\tau \sim \mathcal{N}(\mu_\tau = 100, \sigma_\tau = 20)$ $\theta_d \sim \mathcal{N}(\mu_d = 120, \sigma_d = 20)$

The posterior sampling consist on 4 chains of 1000 draws each one. The presented results correspond to all samples from every chain, unless otherwise stated.

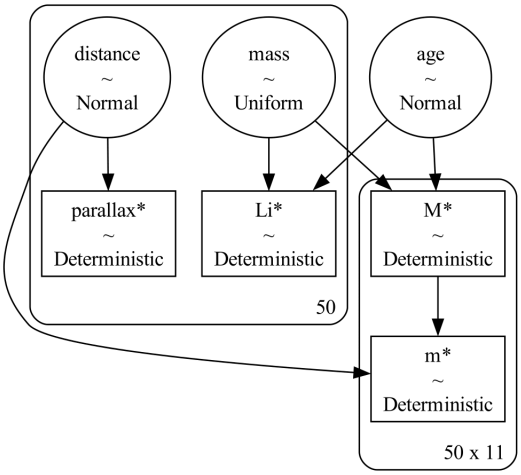


Figure 8: *Generative model for Bayesian model evaluation.*

The Bayesian model is then evaluated for different configurations with a special emphasize on age as the focus of our study. The different model configurations are showed in Figure 2. A deeper discussion is devoted to those model configurations more interesting for real-world datasets.

Conf. 1 The simplest model configuration with only photometric and parallax data corresponds to the use of a normal distribution for the flux likelihood.

Table 4: *Configuration 1.*
Performance on synthetic dataset $\mathcal{S}$

Models	$\hat{\mu}_\tau$	$\hat{\sigma}_\tau$	HDI (3%, 97%)
U	119.352	0.343	(118.735, 120.033)
I	119.351	0.355	(118.709, 120.047)
W	119.371	0.344	(118.751, 120.044)

Table 4 shows a small bias in the mean estimate $\hat{\mu}_\tau$, the model tends to predict a lower value for the mean age. The uncertainty represented by the standard deviation $\hat{\sigma}_\tau$ is also underestimated. The ground-thruth always lies between the HDI. The posterior predictive checks in all models look very similar and follow the expected relation. Q-Q plots are represented in Supplementary Figure 4, while posterior predictive checks based on HR-diagram are in Supplementary Figure 6 The convergence diagnostic are shown in Table 5. The GR $\hat{R}_\tau$-statistics is close to unity in all models. The informative prior shows the greatest bulk-ESS$_\tau$ and the wrong prior the worst. The mean Monte Carlo standard error is close to zero for every

model. Based on convergence diagnostic, the informative prior setting seems to provide slightly more reliable results.

Table 5: *Configuration 1. Convergence diagnostics*

Models	$\hat{R}_\tau$	bulk-ESS$_\tau$	mcse$_{\mu_\tau}$
U	1.001	3001	0.004
I	1.000	3450	0.004
W	1.001	2700	0.005

Conf. 2 The next model configuration increases the complexity of our model by including a mixture likelihood for flux, which has some associated hyperparameters. These hyperparameters are also evaluated, as part of the model.

Table 6: *Configuration 2. Performance on synthetic dataset $\mathcal{S}$*

Models	$\hat{\mu}_\tau$	$\hat{\sigma}_\tau$	HDI (3%, 97%)
U	120.208	0.375	(119.510, 120.927)
I	120.209	0.389	(119.442, 120.896)
W	120.212	0.395	(119.429, 120.896)

Table 7: *Configuration 2. Convergence diagnostics*

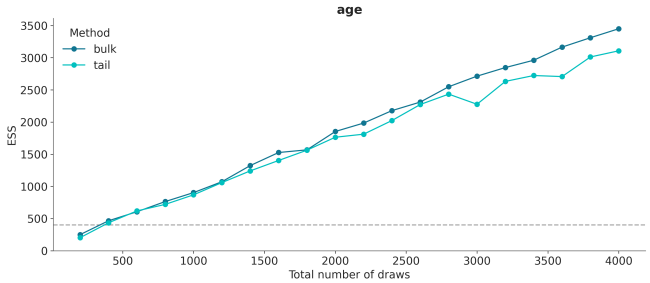
Models	$\hat{R}_\tau$	bulk-ESS$_\tau$	mcse$_{\mu_\tau}$
U	1.000	2241	0.006
I	1.000	1744	0.007
W	1.000	1585	0.007

Table 6 shows a small bias that tends to infer greater mean age than the ground-thruth. However, the bias is smaller compare to the first model configuration. The uncertainty is also underestimated, in this case, in the same proportion as previous model configuration. The ground-thruth always lies between the HDI. Q-Q plot for parallax is shown in Figure 10. There are some systematic differences around the interval (7.325, 7.35) [mass] between observed and predicted parallax, nevertheless they are small enough to be neglected. The convergence diagnostic are shown in Table 7. The GR $\hat{R}_\tau$-statistics is exactly the unity in all models. The uninformative prior shows the greatest bulk-ESS$_\tau$ and the wrong prior the worst. The mean Monte Carlo standard error is close to zero for every model. Based on convergence diagnostic, the uninformative prior setting seems to provide slightly more reliable results. The lower bulk-ESS$_\tau$ in this model could be result of the increase in complexity. The landscape becomes more complex and the MCMC algorithm requires more samples to provide reliable results. Indeed, if we compare ESS evolution as a function of the number of samples as in Figure 9, we find a greater slope for the simplest model.

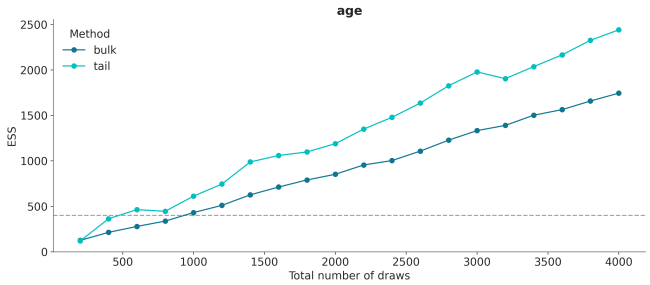
As we saw in Section 3, the use of a mixture likelihood (11) was proposed as a possible solution for inference when the dataset contains outliers. The expected weight for the outliers' distribution is then zero since our dataset does not contain outliers. The estimated background component weight $\hat{\phi}_F$ has mean value

$$\hat{\mu}_{\phi_{F_x}} = 0.001,\ \hat{\sigma}_{\phi_{F_x}} = 0.001,\ \ \forall x$$

where x is each of the different spectral bands. The Bayesian model successfully recovers the same parameter values that were used to generate the data.



(a) *Configuration 1. Simplest model.*



(b) *Configuration 2. Increase in complexity.*

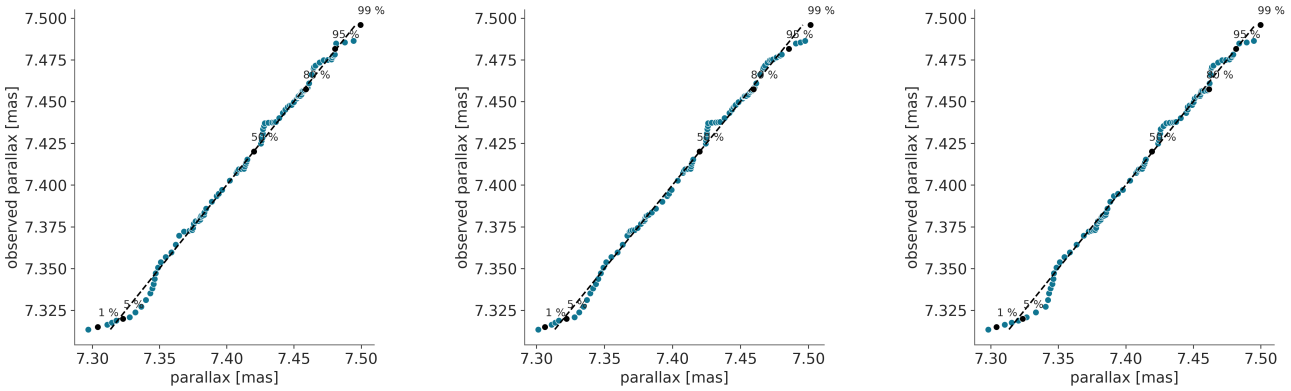
Figure 9: *ESS evolution as a function of total number of draws for two different model configuration varying in complexity.*

Conf. 3 The next model configuration also includes Lithium abundance in the dataset. The flux likelihood is normal in this case, while Lithium observations have a mixture likelihood.

Table 8: *Configuration 3. Performance on synthetic dataset $\mathcal{S}$*

Models	$\hat{\mu}_\tau$	$\hat{\sigma}_\tau$	HDI (3%, 97%)
U	119.894	0.478	(119.029, 120.779)
I	119.874	0.449	(119.070, 120.786)
W	119.898	0.449	(119.078, 120.747)

The estimated age mean and standard deviation are slightly underestimated as shown in Table 8. The



(a) *Uninformative* **(b)** *Weakly informative* **(c)** *Wrong*

Figure 10: *Q-Q plot of model configuration 2 for different priors settings.*

Table 9: *Configuration 3. Convergence diagnostics*

Models	$\hat{R}_\tau$	bulk-ESS$_\tau$	mcse$_{\mu_\tau}$
U	1.04	3051	0.006
I	1.01	5071	0.004
W	1.02	4828	0.005

ground-thruth always lies between the HDI. The GR $\hat{R}_\tau$-statistics is worse than previous models, while the ESS is higher. In particular, the informative prior setting provides high quality results due to its low $\hat{R}_\tau$, high ESS and standard mcse. Therefore, informative prior seems to be the most reliable model based on convergence diagnostic. Q-Q plots of Lithium for different priors settings are shown in Figure 11. As we can see, both distributions are almost identical based on Q-Q plots. The model successfully recover the Lithium distribution using posterior predictive checks. On the other hand, HR-diagram for posterior predictive checks are showed in Figure 12. The choice of right or (weakly) wrong priors seems to be irrelevant in terms of accuracy or posterior predictive checks, but helps in the quality of the sampling. Indeed, the high ESS implies that the inference requires fewer samples to obtain reliable results.

Analogous to the previous model with mixture likelihood for flux, in this model configuration the expected weight for the background component of Lithium is zero. In fact, the posterior for $\hat{\phi}_{[\mathrm{Li}]}$ has

$$\hat{\mu}_{\phi_{[\mathrm{Li}]}} = 0.004,\ \hat{\sigma}_{\phi_{[\mathrm{Li}]}} = 0.004$$

for every prior setting in this model configuration.

Conf. 4 Finally, the most complex model is evaluated. The model include Lithium and photometric observations, both with mixture likelihood. The estimated age mean is slightly overestimated as shown in Table 10, while the estimated standard deviation is underestimated. The ground-truth always lies between the HDI.

Table 10: *Configuration 4. Performance on synthetic dataset $\mathcal{S}$*

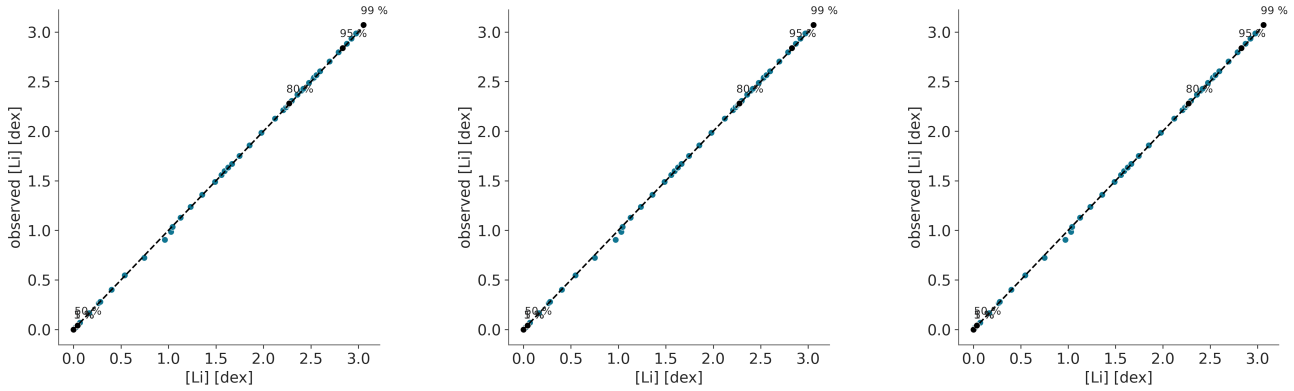
Models	$\hat{\mu}_\tau$	$\hat{\sigma}_\tau$	HDI (3%, 97%)
U	120.154	0.576	(119.035, 121.224)
I	120.240	0.427	(119.408, 121.028)
W	120.185	0.558	(119.103, 121.177)

The main disadvantage of complex models is the increase in the required number of samples to obtain reliable results. Table 11 shows the poor convergence performance while Figure 13 provides a visual representation given by ESS evolution as a function of the number of draws.

Table 11: *Configuration 4. Convergence diagnostics*

Models	$\hat{R}_\tau$	bulk-ESS$_\tau$	mcse$_{\mu_\tau}$
U	1.05	1485	0.006
I	1.03	1269	0.004
W	1.05	1554	0.005

The increase in complexity does not provide any significant advantage compare to previous models given our figure-of-merits. However, a further study is required to evaluate the performance in-depth. The number of samples has been keep fixed in our model evaluation to 4 chains of 1000 draws each one, boosting simpler models convergence performance. The low number of samples allows an evaluation of multiple models in a reasonable amount of time, in detrimental of those models that requires a greater number of samples to provide reliable results. However, since our computational resources are limited, we seek a trade-off between model performance, assets in accuracy and convergence metrics, and effi-



(a) *Uninformative* **(b)** *Weakly informative* **(c)** *Wrong*

Figure 11: *Q-Q plot of model configuration 3 for different priors settings.*

ciency. Our model implementation allows the use of different configurations that can adapt to the available computational resources, while providing high quality results as we have seen. The assessment outlined in this section highlights the primary pros and cons of each model configuration, serving as a point of reference when making the ultimate decision on which model configuration to employ in real-world scenarios.

5 Results

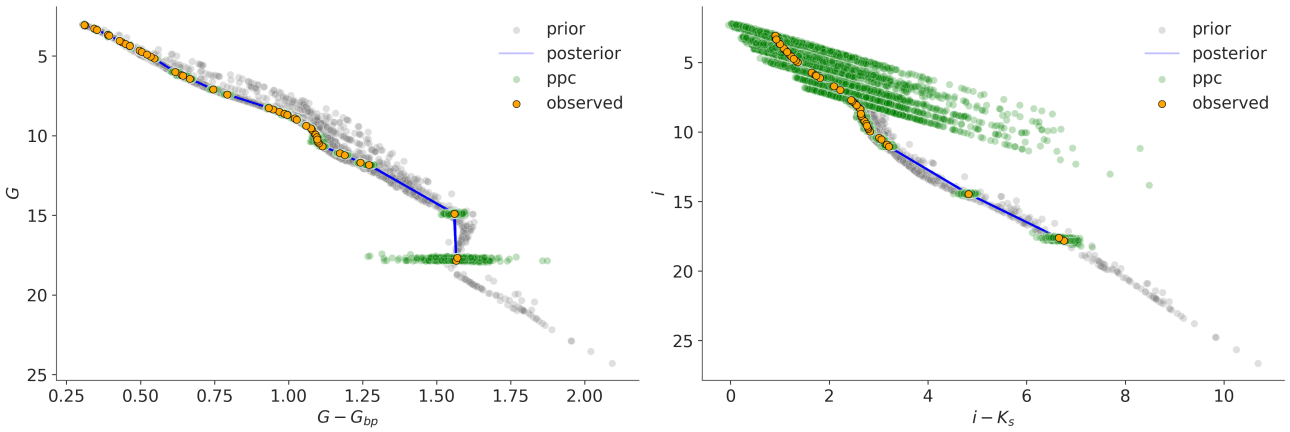
In this section, we present the outcomes of applying the Bayesian model we developed to infer the age of open clusters based on Lithium abundance, photometric data, and parallax measurements. To rigorously assess the model's performance and its practical utility, we chose the iconic Pleiades open cluster as our test bench. Its popularity comes from its unique combination of properties. It is young (125 ± 8 Myr, Stauffer et al., 1998), close to the sun ($134.4^{+2.9}_{-2.8}$ pc, Galli, P. A. B. et al., 2017b), massive ($870 \pm 35\ M_{\odot}$, Converse & Stahler, 2008), has low extinction ($A_v = 0.12$, Guthrie, 1987), and an almost solar metallicity (Takeda et al., 2016). Its age, often considered one of the fundamental calibrations for stellar evolution models, has been a subject of extensive investigation over the years. By comparing our age estimates with established values from the literature, we assess the model's ability to accurately infer open cluster ages in real-world scenarios with a methodology that combines multiple sources of information, providing a more robust and precise estimate.

In the preceding Section 4.3, we detailed the formulation of our Bayesian model and the extensive testing it underwent on synthetic datasets with known parameters. These tests provided crucial insights into the model's performance, its sensitivity to various prior settings, and its capacity to handle uncertainties inherent in astronomical observations. With this solid foundation, we now turn our attention to the real astronomical data of the Pleiades cluster.

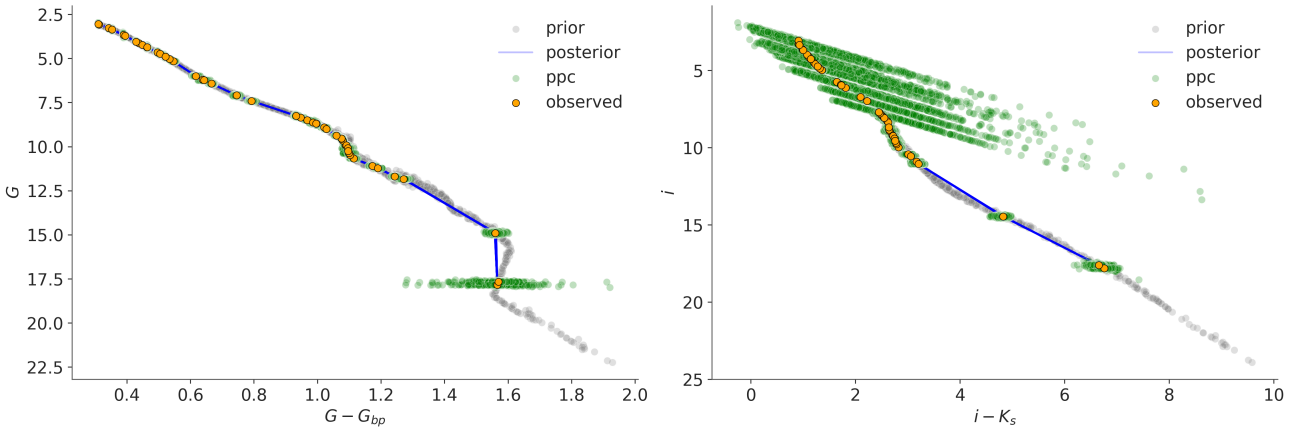
Dataset Our dataset comprises measurements of Lithium abundance, photometric, and parallax measurements[21] for the members of the Pleiades cluster. These members represent the common membership shared between those identified by Olivares et al. (2018, 2021) and those identified by Meingast et al. (2021). The corresponding spectroscopic surveys are Gaia-ESO, Pan-Starrs, and 2MASS. On the other hand, the Lithium abundance measurements correspond to those reported by Bouvier et al. (2018) which were derived from equivalent widths, effective temperatures, and growth curves.

Preprocessing The dataset consist on 932 members overall, however, there are some mssing values for photometry and Lithium abundance. In the case of photometry, these missing values correspond to the actual data (i.e., the mean value) or its uncertainties (i.e., the standard deviation). There are a broad range of techniques to deal with missing values. The most commonly used are data imputation and data deletion. In data imputation, missing values are replaced by estimates coming from mean, median, regression based on other features, etc. On the other hand, data deletion implies removing entire rows with missing values (list-wise deletion) or ignore missing values when computing statistics (pair-wise deletion). In our approach, we have applied pair-wise deletion when the missing value corresponds to data, and data imputation for uncertainty. The pair-wise deletion consist on assigning a unity likelihood when the value is missing. Therefore, if the input F_j^x corresponds to a missing value, its contribution to the

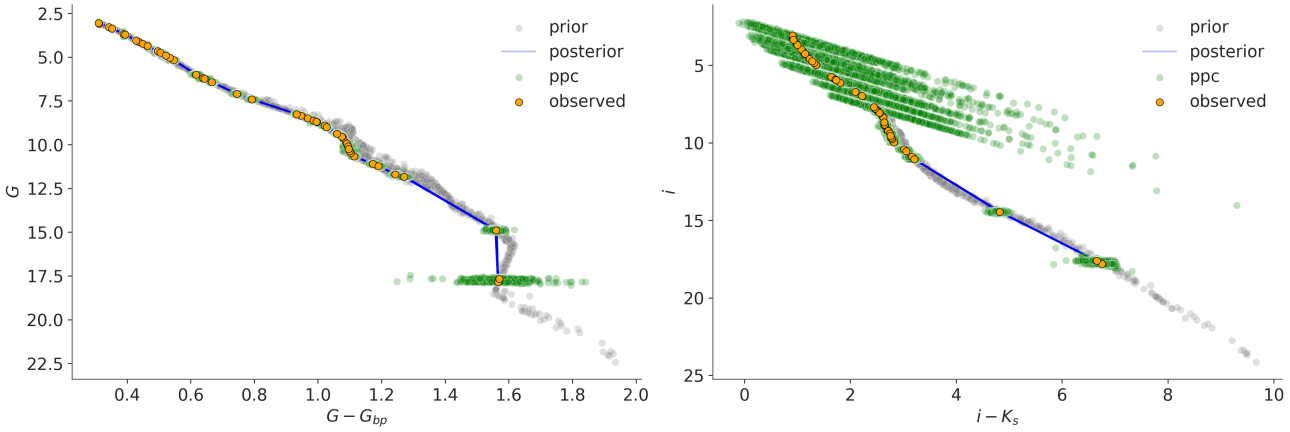
[21] A general description of these measure techniques is presented in Appendix A



(a) *Uninformative*

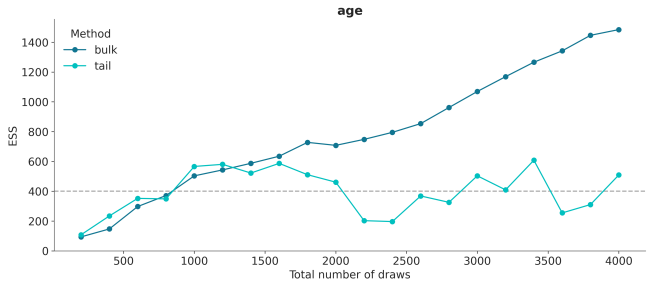


(b) *Weakly informative*

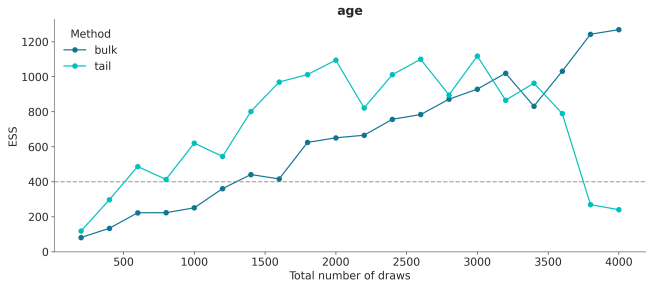


(c) *Wrong*

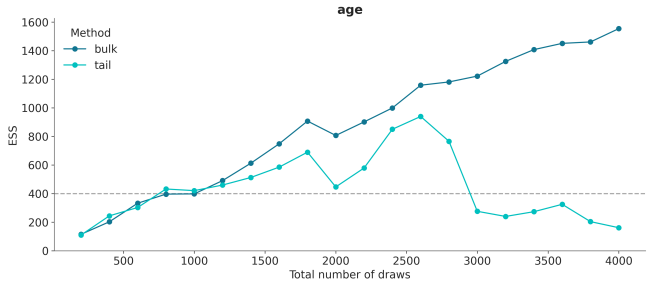
Figure 12: *Posterior predictive checks based on HR-diagram of model configuration 3 for different priors settings.*



(a) *Uninformative*



(b) *Weakly informative*



(c) *Weakly informative*

Figure 13: *ESS evolution as a function of total number of draws for different prior setting in model configuration 4.*

log-likelihood would be null

$$l(\theta) = \sum_{i=1}^{N} (1 - \delta_{i,j}) \log \mathcal{L}_i^{F_x}(\theta) \tag{35}$$

where $\delta_{i,j}$ is the Kronecker delta[22]. On the other hand, the missing uncertainties are replaced by the maximum uncertainty value of its spectral band. This approach takes a pessimistic position, since we are assuming that any missing uncertainty is the greatest value it can take based on our observations. Supplementary Figures 7 and 8 show the results after preprocessing and before preprocessing, respectively. Finally, those stars having absolute magnitude lower than our BT-Settl limits are also deleted to avoid possible problems coming from NN predictions. In the case of Lithium, every missing value is missing both: actual data and uncertainty. Then, list-wise deletion is applied in this case, resulting in only 88 observations after deletion. The possible use of imputation for data values is far from desirable in the given scientific context. The use of pair-wise deletion does not delete an entire row if there is one missing value, as list-wise deletion. This increase the amount of data by allowing those members with missing values in some spectral bands, but not in all, contribute to the inference. The choice of maximum values for uncertainty imputation is the most conservative approach, finding a trade-off between the use of as much data as possible and maintaining a cautious perspective on data quality and reliability.

[22]Kronecker delta is defined as

$$\delta_{i,j} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases} \tag{36}$$

Configuration The Bayesian model configuration corresponds to the model configuration 3. The use of normal likelihood for flux and mixture likelihood for Lithium abundances seems to find a trade-off between model complexity and reliable results based on previous analysis. The increase in complexity due to our larger dataset makes the computation harder, requiring more time to acquire reliable results. As we have keep fixed the number of draws and chains, the chosen model configuration 3 showed faster convergence in our diagnostic making it more suitable for our problem.

Priors setting The priors for distance and age employed corresponds to informative priors

$$\theta_\tau \sim \mathcal{N}(\mu_\tau = 120,\ \sigma_\tau = 5)$$
$$\theta_d \sim \mathcal{N}(\mu_d = 135,\ \sigma_d = 5)$$

Inference results The posterior distribution for age is shown in Figure 14. The inferred age was 116.8 ± 1.9 Myr.

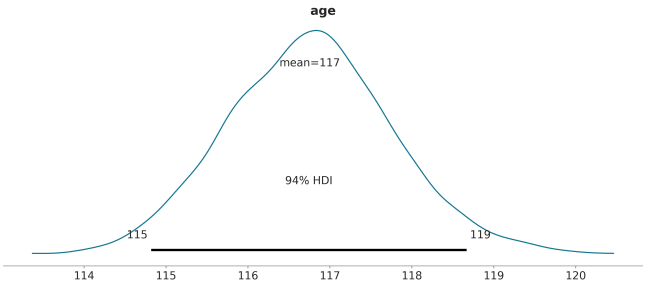


Figure 14: *Age posterior distribution for the Pleiades Open Cluster*

The convergence diagnostic shows an GR $\hat{R}$-statistics close to unity, nevertheless the ESS is less than previous models.

Despite the low ESS value, an examination of the ESS trend depicted in Figure 15 does not reveal any concerning patterns that would lead us to believe that extending the sample length would address the issue of the low ESS value.

Table 12: *Results. Convergence diagnostics*

Model	$\hat{R}_\tau$	bulk-ESS$_\tau$	mcse$_{\mu_\tau}$
I	1.01	476	0.047

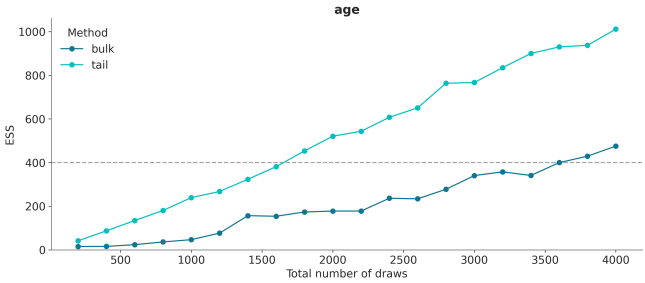


Figure 15: *ESS evolution as a function of total number of draws for Pleiades inference.*

Furthermore, the favorable estimate of the $\hat{R}$-statistics, indicating good convergence, may be attributed to the convergence of each chain towards the same distribution, as evident in Figure 16. Finally, the posterior distribution is compared to the observations on Figure 17.

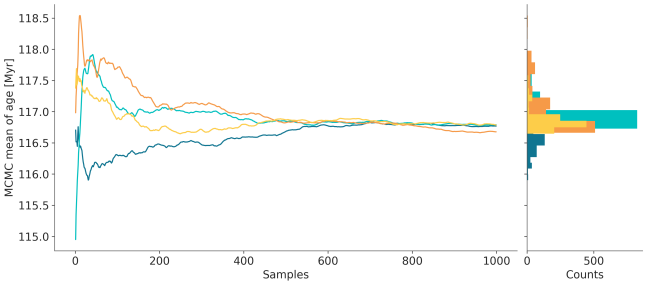


Figure 16: *Age posterior mean for each chain as a function of the number of samples*

6 Conclusions

A Bayesian hierarchical model for estimating the ages of open clusters based on chemical abundance of Lithium, photometric, and parallax measurements has been developed. Our model successfully combines artificial intelligence methodologies, including neural networks and Bayesian hierarchical models, resulting in several notable advantages. The incorporation of a neural network into our model is particularly advantageous due to its inherent ability as a universal interpolator. This neural network seamlessly interpolates data within a physical model grid, such as BT-Settl, enabling us to bridge gaps and address intricacies in our understanding of stellar properties. Furthermore, our approach leverages the benefits of a Bayesian hierarchical model, providing several key advantages. Bayesian statistics offers a rigorous framework for incorporating uncertainty into our age estimations, making our results more robust and reliable. Additionally, hierarchical modelling allows us to capture complex dependencies and correlations among different data sources, such as chemical abundance of Lithium, photometric, and parallax measurements, leading to a more comprehensive and accurate inference of open cluster ages.

Sampling the posterior distribution defined by a complex model such as ours can be challenging, particularly when aiming to employ the most effective samplers like Hamiltonian Monte Carlo, which necessitate the computation of gradients. However, our implementation is designed to facilitate this sampling process, enabling the use of advanced samplers and further enhancing the model's computational tractability. This critical enhancement represents a substantial step forward in the field of astrophysical modeling, offering a solution to the sampling challenges faced by many Bayesian models and providing a distinct advantage over counterparts that are unable to harness the power of advanced samplers like Hamiltonian Monte Carlo.

Moreover, our approach exhibits a remarkable level of flexibility that extends beyond the confines of any particular physical model. While the BT-Settl grid, used to interpolate the spectral energy distributions of stars, may impose limitations based on its theoretical constraints, our modular implementation allows for the seamless integration of alternative neural network interpolators. This means that our model can readily adapt to different datasets and physical limits, making it a versatile tool for age estimation across a wide range of stellar environments. This adaptability ensures that our model remains at the forefront of stellar dating research, capable of accommodating emerging data sources and evolving understanding of stellar phenomena.

The robustness and reliability of our model have been validated through extensive testing. We assessed its capabilities using synthetic datasets with known parameters, demonstrating its remarkable ability to accurately recover information. Additionally, we applied our model to a real-world dataset, such as the well-studied Pleiades open cluster, where it delivered excellent results. This dual success in handling both synthetic and real data underscores the model's versatility and its potential as a valuable tool for advancing our understanding of star cluster ages in various astrophysical contexts.

While our research has made significant strides in advancing the field of astrophysical modeling, it is essential to acknowledge the areas that remain open for further investigation. For instance, one promising

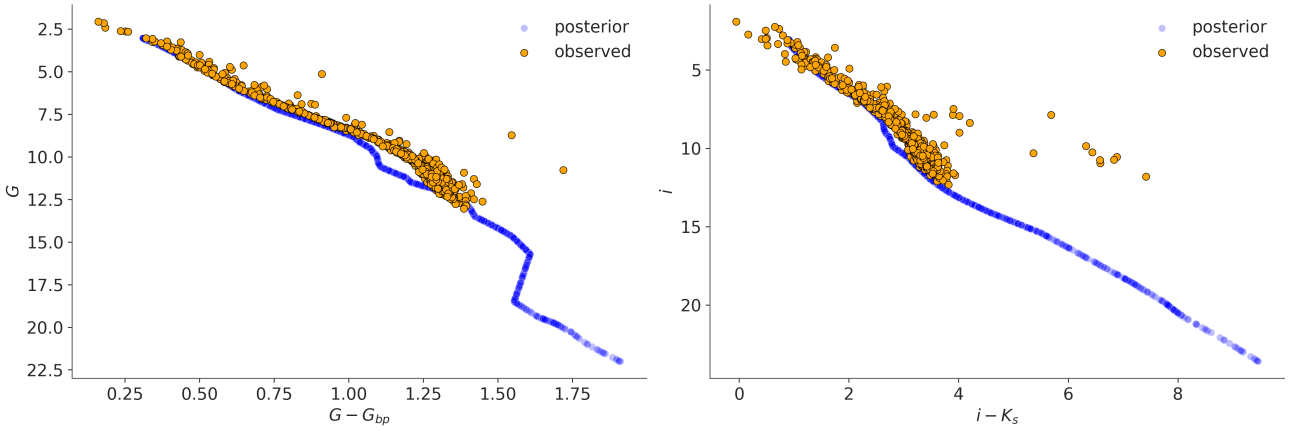


Figure 17: *Posterior distribution compare to observations for the Pleiades*

avenue for future research could involve extending our model to incorporate additional parameters, such as the detection and characterization of binary stars within open clusters. By detecting and accounting for binary stars in our model, we can improve the accuracy of age estimations for these stellar populations, as their presence and interactions can impact the inferred parameters. Finally, although the model evaluation was done based on different prior settings, the test benchmark on the Pleiades only employed one (informative) prior setting. The knowledge incorporation into the prior helps in the sampling process, nevertheless, a more rigorous test would involve different prior settings. Particularly, the replicability of our results with noninformative priors would be a confirmation of our model performance.

As we conclude this study, it is essential to recognize that the avenues for further exploration and discovery are numerous and promising. The insights gained from this research serve as a foundation upon which future astrophysicists can build, deepening our understanding of the cosmos. It is our hope that this work serves as an initial step towards a universal, auto-consistent, absolute and accurate method to estimate stellar ages over the entire time domain that overcome the current pathologies in the field. With continued dedication and collaboration within the astrophysical community, we are poised to unlock even more profound insights into the age and evolution of stars, enriching our understanding of the universe.

Code availability

The core code for this research project is available on GitHub at the following repository: franciscopalmeromoya/bayesian-ages-open-clusters

References

Soderblom, D. (2015). Stellar clocks. *Nature*, *517*(7536), 557–558.

Amelin, Y., Krot, A. N., Hutcheon, I. D., & Ulyanov, A. A. (2002). Lead isotopic ages of chondrules and calcium-aluminum-rich inclusions. *Science*, *297*(5587), 1678–1683.

Ooba, J., Ratra, B., & Sugiyama, N. (2018). Planck 2015 constraints on the non-flat lambda-cdm inflation model. *The Astrophysical Journal*, *864*(1), 80.

Planck Collaboration. (2016). Planck 2015 results - xxiv. cosmology from sunyaev-zeldovich cluster counts. *Astronomy and Astrophysics*, *594*, A24.

Barrado, D. (2016). Clusters: Age scales for stellar physics (E. Moraux, Y. Lebreton, & C. Charbonnel, Eds.). *EAS Publications Series*, *80-81*, 115–175.

Soderblom, D. (2010). The ages of stars. *Annual Review of Astronomy and Astrophysics*, *48*(1), 581–629.

Sestito & Randich. (2005). Time scales of li evolution: A homogeneous analysis of open clusters from zams to late-ms. *A & A*, *442*(2), 615–627. https://doi.org/10.1051/0004-6361:20053482

Allard, F. (2013). The bt-settl model atmospheres for stars, brown dwarfs and planets. *Proceedings of the International Astronomical Union*, *8*(S299), 271–272.

Galli, P. A. B., Moraux, E., Bouy, H., Bouvier, J., Olivares, J., & Teixeira, R. (2017a). A revised moving cluster distance to the pleiades open cluster. *A&A*, *598*, A48.

Heyl, J., Caiazzo, I., & Richer, H. B. (2022). Reconstructing the pleiades with gaia edr3. *The Astrophysical Journal*, *926*(2), 132.

Galindo-Guil, F. J., Barrado, D., Bouy, H., Olivares, J., Bayo, A., Morales-Calderón, M., Huélamo, N., Sarro, L. M., Rivière-Marichalar, P., Stoev, H., Montesinos, B., & Stauffer, J. R. (2022). Lithium depletion boundary, stellar associations, and gaia. *A&A*, *664*, A70.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques - adaptive computation and machine learning*. The MIT Press.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2016). Automatic differentiation variational inference.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, *195*(2), 216–222.

Betancourt, M. J., & Girolami, M. (2013). Hamiltonian monte carlo for hierarchical models.

Griewank, A., & Walther, A. (2008). *Evaluating derivatives* (Second). Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9780898717761

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (Eds.). (2011a). *Handbook of markov chain monte carlo*. Chapman; Hall/CRC. https://doi.org/10.1201/b10905

Hoffman, M. D., & Gelman, A. (2011). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/https://doi.org/10.1016/0893-6080(89)90020-8

Hogg, D. W., Bovy, J., & Lang, D. (2010). Data analysis recipes: Fitting a model to data.

Salvatier, J., Wiecki, T., & Fonnesbeck, C. (2015). Probabilistic programming in python using pymc.

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). Arviz a unified library for exploratory analysis of bayesian models in python. *Journal of Open Source Software*, *4*(33), 1143. https://doi.org/10.21105/joss.01143

Cuchiero, C., Larsson, M., & Teichmann, J. (2020). Deep neural networks, generic universal interpolation, and controlled odes.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. Retrieved September 7, 2023, from http://www.jstor.org/stable/2984418

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics*. https://api.semanticscholar.org/CorpusID:5575601

Berrada, L., Zisserman, A., & Kumar, M. P. (2020). Training neural networks for and by interpolation.

Botev, A., Lever, G., & Barber, D. (2016). Nesterov's accelerated gradient and momentum as approximations to regularised update descent.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Eadie, G. M., Speagle, J. S., Cisewski-Kehe, J., Foreman-Mackey, D., Huppenkothen, D., Jones, D. E., Springford, A., & Tak, H. (2023). Practical guidance for bayesian inference in astronomy.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Roy, V. (2019). Convergence diagnostics for markov chain monte carlo.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Stauffer, J. R., Schultz, G., & Kirkpatrick, J. D. (1998). Keck spectra of pleiades brown dwarf candidates and a precise determination of the lithium depletion edge in the pleiades*. *The Astrophysical Journal*, *499*(2), L199. https://doi.org/10.1086/311379

Galli, P. A. B., Moraux, E., Bouy, H., Bouvier, J., Olivares, J., & Teixeira, R. (2017b). A revised moving cluster distance to the pleiades open cluster. *A&A*, *598*, A48. https://doi.org/10.1051/0004-6361/201629239

Converse, J. M., & Stahler, S. W. (2008). The distribution of stellar mass in the pleiades. *The Astrophysical Journal*, *678*(1), 431. https://doi.org/10.1086/529431

Guthrie, B. N. G. (1987). Interstellar extinction in the pleiades. *Quarterly Journal of the Royal Astronomical Society*, *28*(3), 289–293. http://inis.iaea.org/search/search.aspx?orig_q=RN:19020379

Takeda, Y., Hashimoto, O., & Honda, S. (2016). Photospheric carbon and oxygen abundances of F–G type stars in the Pleiades cluster*. *Publications of the Astronomical Society of Japan*, *69*(1), 1. https://doi.org/10.1093/pasj/psw105

Olivares, J., Sarro, L. M., Moraux, E., Berihuete, A., Bouy, H., Hernández-Jiménez, S., Bertin, E., Galli, P. A. B., Huelamo, N., Bouvier, J., & Barrado, D. (2018). The seven sisters DANCe. IV. Bayesian hierarchical model. *A&A*, *617*, Article A15, A15. https://doi.org/10.1051/0004-6361/201730972

Olivares, J., Bouy, H., Sarro, L. M., Moraux, E., Berihuete, A., Galli, P. A. B., & Miret-Roig, N. (2021). Miec: A Bayesian hierarchical model for the analysis of nearby young open clusters. *A&A*, *649*, Article A159, A159. https://doi.org/10.1051/0004-6361/202140282

Meingast, S., Alves, J., & Rottensteiner, A. (2021). Extended stellar systems in the solar neighborhood. V. Discovery of coronae of nearby star clusters. *A&A*, *645*, Article A84, A84. https://doi.org/10.1051/0004-6361/202038610

Bouvier, J., Barrado, D., Moraux, E., Stauffer, J., Rebull, L., Hillenbrand, L., Bayo, A., Boisse, I., Bouy, H., DiFolco, E., Lillo-Box, J., & Morales Calderón, M. (2018). The lithium-rotation connection in the 125 Myr-old Pleiades cluster. *A&A*, *613*, Article A63, A63. https://doi.org/10.1051/0004-6361/201731881

Copi, I., & Cohen, C. (2005). *Introduction to logic*. Pearson/Prentice Hall.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Bernardo, J. M., & Smith, A. (2000). *Bayesian Theory*. John Wiley; Sons Ltd.

Jaynes, E., & Justice, J. H. (1986). Bayesian methods: General background. In *Maximum entropy and bayesian methods in applied statistics: Proceedings of the fourth maximum entropy workshop university of calgary, 1984* (pp. 1–25). Cambridge University Press. https://doi.org/10.1017/CBO9780511569678.003

Stigler, S. M. (1986). *The history of statistics : The measurement of uncertainty before 1900*. Cambridge, Mass. : Belknap Press of Harvard University Press, 1986.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109. Retrieved August 9, 2023, from http://www.jstor.org/stable/2334940

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (Eds.). (2011b). *Handbook of markov chain monte carlo*. Chapman; Hall CRC.

Appendices

These appendices contain in-depth discussions, derivations, and additional information that readers can explore at their own pace. This approach maintains the flow of the main thesis while providing interested readers with the opportunity to delve deeper into specific technical aspects.

A Astrophysics

A.1 The life of a star

Stellar evolution, the gradual transformation of stars over time, is a complex phenomenon driven by numerous factors, including a star's initial mass and composition. Stars are born out of the gravitational collapse of cool, dense molecular clouds. Over the course of millions of years, these protostars settle down into a state of equilibrium, at this point, Hydrogen is converted into Helium in the core and the star is born onto the main sequence. The star will continue to burn hydrogen into helium and will remain a main sequence star for about 90 % of its life.

There are a number of physical characteristics of stars that provide important information on the lives of stars. Two quantities, mass and age, are probably most fundamental. The progress of a star's life is predestined by its mass, because ultimately the mass determines how much energy the star can produce and how quickly it will do so. The age of a star tells you how far along it is in its evolution. However, both of these quantities are hard to measure directly, instead we try to find a way to classify stars based upon a simple observation.

A.2 Observations

Accurate dating methods are indispensable for unraveling the mysteries of stellar evolution. As we saw in Section 1, two of the most used techniques are isochrone fitting and chemical abundance, which were briefly explained. However, these dating techniques relies on accurate astrophysical measurements which are indispensable for the correct interpretation of stellar properties and evolution. In this context, photometry and astrometry play crucial roles in gathering essential data.

Photometry Photometry is a branch of astronomy that involves measuring the brightness of celestial objects, typically stars, using various filters to capture light at specific wavelengths[23]. It is a fundamental technique for understanding the properties of stars, galaxies, and other astronomical objects. The primary goal of photometry is to quantify the amount of light emitted or reflected by an object, usually expressed in terms of its apparent magnitude.

Astrometry Astrometry is another branch of astronomy focused on precisely measuring the positions and motions of celestial objects in the sky. Astrometry provides critical data for tracking the orbits of planets, asteroids, and stars, as well as studying the structure and dynamics of our galaxy and the universe at large. It involves accurately determining the coordinates (right ascension and declination) of celestial objects and their proper motions over time.

The combination of both techniques enables the precise measurement of absolute magnitudes, which is essential for understanding the intrinsic luminosity of celestial objects. By accurately determining the brightness of stars through photometry and their precise positions through astrometry, allows the derivation of absolute magnitudes given the relation 13. On the other hand, to determine the abundance of Lithium in stars, astronomers often rely on a combination of photometry and spectroscopy. In spectroscopy, the process involves capturing a star's light spectrum through a spectrograph, where it is separated into its constituent colors. Subsequently, astronomers identify specific absorption lines within the spectrum, including the Lithium absorption line, and then proceed to measure the equivalent width of the Lithium line, quantifying its depth and width. Finally, to determine the Lithium abundance in the star, astronomers compare this measured equivalent width with a theoretical model designed for stars sharing similar characteristics such as temperature, luminosity, and metallicity. The abundance of Lithium is then calculated based on this comparison.

[23]Please, see spectral band entry in Glossary

B Uncertainties

As we saw in Appendix C.1, uncertainty is a measure of the lack of confidence in the outcome of an event or the value of a quantity, often due to incomplete information or variability. Traditional approaches in statistics fail to distinguish inherently different sources of uncertainty, often referred to as aleatoric and epistemic uncertainty.

Aleatoric Aleatoric uncertainty refers to the notion of randomness, that is, the variability in the outcome of an experiment which is due to inherently random effects.

Epistemic Epistemic uncertainty refers to uncertainty caused by a lack of knowledge, i.e., to the epistemic state of the agent. As opposed to aleatoric uncertainty, epistemic uncertainty can in principle be reduced on the basis of additional information.

C Plausible reasoning and Probability theory

In this section, we will delve into the formal concepts of probability theory and explore the mathematical equivalence between the frequentist and Bayesian interpretations of probability in some areas. While these interpretations are often discussed in terms of their philosophical underpinnings, we will focus on their common mathematical foundations and demonstrate how they yield equivalent results in key areas such as random variables, measurable spaces, and more.

C.1 Logic, uncertainty and probability

As is generally credited to the *Organon* of Aristotle (fourth century BC) deductive reasoning (*apodeixis*) can be analyzed ultimately into the repeated application of two strong syllogism: *modus ponens* and *modus tollens* (Copi & Cohen, 2005). This kind of reasoning is powerful but relies heavily on having complete and accurate information about the problem at hand. However, in most real-world situations, we often lack complete information, and uncertainties are prevalent. The absence of perfect information can stem from various factors, such as incomplete data, measurement errors, hidden variables, or complex interactions between multiple factors. In such cases, we resort to probabilistic reasoning and use weaker syllogisms or inductive reasoning (*apagoge*). One of such syllogisms could be:

$A \Rightarrow B$
B is true.

---

therefore, A becomes more plausible. $\therefore$

Note that their conclusions do not follow necessary from the premises, they can yield no more than a degree of probability. The reasoning is valid even when we have a whole set of propositions $\{A_i\}_{i=1}^{N}$ or when we go to the continuum limit. According to Jaynes (2003) the relation is simply: Aristotelian deductive logic is the limiting of plausible reasoning, as we become more and more certain of its conclusion. Setting aside any philosophical interpretation of plausible reasoning, for us probabilities define a particular scale on which degrees of plausibility, taking into account uncertainties, can be measure[24].

C.2 Bayesian framework

Thomas Bayes was a British clergyman and amateur mathematician, who died in 1761. Among his papers was found a curious manuscript, published after his death in 1763. It gives the basis of "Bayesian Statistics". In almost his first published work (1774), Laplace rediscovered Bayes' principle in greater clarity and generality. The basic theorem appears today as almost trivially simple; yet it is by far the most important principle underlying scientic inference (Jaynes & Justice, 1986). Let A, B, and C denote three propositions and let AB stand for the proposition "both A and B are true". Since AB and BA are the same proposition, requiring consistency to the basic product and sum rules of probability theory we find what is always called "Bayes'

[24]A rigorous development of the alignment between probability and quantitative degrees of plausibility can be located in (Bernardo & Smith, 2000)

Theorem" today, although Bayes never wrote it (Stigler, 1986):

$$\Pr(A|BC) = \Pr(A|C)\frac{\Pr(B|AC)}{\Pr(B|C)} \tag{37}$$

In (37) we have a mathematical representation of the process of learning. $\Pr(A|C)$ is our "prior probability" for A, when we know only C. $\Pr(A|BC)$ is its "posterior probability", updated as a result of acquiring new information B.

C.3 Bayesian vs. frequentist

In the Bayesian interpretation of probability, the probabilities represent quantitavively degrees of belief or uncertainty. Thus, given a set of statements $S = \{A_i\}_{i=1}^{N}$ the probability of A_i is a function $\Pr : S \to [0,1] \in \mathbb{R}$ such that $\Pr(A_i)$ represents our degrees of belief. Out of all possible monotonic functions which could, in principle, serve to represent our degrees of belief equally well, we choose this particular one, not because it is more correct but because it is more convenient, i.e., they obey the simplest rules of combination: the product and sum rules. This situation is analogous to that in thermodynamics, where out of all possible empirical temperature scales, which are monotonic functions of each other, we finally decide to use the Kelvin scale because it is more convenient.

Nevertheless, for historical reasons the mathematical development of the probability theory foundations have been developed in terms of random experiments. Therefore, probability is usually defined in terms of the Kolmogorov axioms where we have a probability space $(\Omega, \mathcal{F}, P)$ that provides a formal model of a random process or experiment. Thus, we have a sample space, Ω, which is the set of all possible outcomes; an event space, which is a set of events[25], $\mathcal{F}$; and a probability function, P, which assigns to each event in the event space a probability, which is a number between 0 and 1. This is precisely why the term random variable could be misleading, since the randomness does not come from the output of an experiment, but from degrees of belief.

On the other hand, Cox's theorem, named after the physicist Richard Threlkeld Cox, is a derivation of the laws of probability theory from a certain set of postulates that justifies the so-called "logical" interpretation of probability, as the laws of probability derived by Cox's theorem are applicable to any proposition. This is the interpretation of probability we are closer to, as we are following Jaynes (2003).

The probabilistic model we have proposed yields a probability density function for the parameters of the model. The inferred parameters come from Bayesian inference, i.e., the use of Bayes theorem to update our prior knowledge about them. Thus, we had to assign a prior probability to them representing our degrees of belief based on scientific literature.

D More about Markov chain Monte Carlo

Metropolis chain Monte Carlo (MCMC) was invented at Los Alamos, one of the few places where computers were available at the time. Metropolis et al. (1953)[26] simulated a liquid in equilibrium with its gas phase without the need to simulate the exact dynamics; they only needed to simulate some Markov chain having the same equilibrium distribution. Simulations following the scheme of (Metropolis et al., 1953) are said to use the Metropolis algorithm. Hastings (1970) generalized the Metropolis algorithm, and simulations following this scheme are said to use the Metropolis-Hasting algorithm. A special case of Metropolis-Hasting algorithm was introduced by Geman and Geman (1984), apparently without knoledge of earlier work. Simulations following their scheme are said to use the Gibbs sampler. In this section, we introduce some topics required to understand MCMC mainly based on Brooks et al. (2011b).

Markov chains A discrete-time Markov chain is a sequence of random variables $\{X_i\}_{i=1}^{N}$, namely a stochastic process, with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states, i.e., if the conditional distribution of X_{n+1} given $\{X_i\}_{i=1}^{n}$ depends on X_n only. The set in which the X_i take values is called the state space of the Markov chain.

A Markov chain has stationary transition probabilities if the conditional distribution of X_{n+1} given X_n does not depend on n. This is the main kind of Markov chain of interest in MCMC, nevertheless some kinds

[25] An event being a set of outcomes in the sample space

[26] The fifth author was E. Teller, the "father of the hydrogen bomb".

of adaptive MCMC have nonstationary transition probabilities. The joint distribution of Markov chain is determined by

- The marginal distribution of X_1, called the initial distribution.
- The conditional distribution of X_{n+1} given X_n, called the transition probability distribution.

Markov chains of interest in MCMC have uncountable state space, and then we must think of the initial distribution as an unconditional probability distribution and the transition probability distribution as an unconditional probability distribution.

Stationarity A stochastic process is stationary if for every positive integer k the distribution of the k-tuple $(X_{n+1}, ..., X_{n+k})$ does not depend on n. A Markov chain is stationary if it is a stationary stochastic process. Since the conditional distribution of $(X_{n+2}, ..., X_{n+k})$ given X_{n+1} does not depend on n, it follows that a Markov chain is stationary if and only if the marginal distribution of X_n does not depend on n.

An initial distribution is said to be stationary or invariant or equilibrium for some transition probability if the Markov chain specify by this initial distribution and transition probability is stationary. Having an equilibrium distribution is an important property of a Markov chain transition probability. In fact, MCMC samples the equilibrium distribution meaning that all Markov chains used in MCMC have equilibrium distribution.

Reversibility A transition probability distribution is reversible with respect to an initial distribution if, for the Markov chain $\{X_i\}_{i=1}^N$ they specify, the distribution of pairs (X_i, X_{i+1}) is exchangeable for every $i = 1, ..., N$.

A Markov chain is reversible if its transition probability is reversible with respect to its initial distribution. Reversibility implies stationarity, but not vice versa. A reversible Markov chain has the same laws running forward or backward in time, that is, for any i and k the distributions of $(X_{i+1}, ..., X_{i+k})$ and $(X_{i+k}, ..., Xi + 1)$ are the same. Hence the name.

Convergence There is a great deal of theory about convergence of Markov chains. Unfortunately, none of it can be applied to get useful convergence information for most MCMC applications.

A Markov chain can appear to have converged to its equilibrium distribution when it has not. This happens when parts of the state space are poorly connected by the Markov chain dynamics: it takes many iterations to get from one part to another. Some algorithms such as Hamiltonian Monte Carlo are designed to solve this problems.

Burn-in Burn-in is a colloquial term that describes the practice of throwing away some iterations at the beginning of an MCMC run. This notion says that you start at somewhere, them you run the Markov chain for B staper (the burn-in period) during which you throw away all the data. After the burn-in you run normally, using each iterate in your MCMC calculations. It is just a method of finding a good starting point, and some Markov chains do not need burn-in.

D.1 Metropolis-Hasting

The simplest algorithm is the Metropolis-Hasting (M-H) algorithm. Suppose that the specified distribution, i.e., the desired stationary distribution of the MCMC sampler we are constructing has unnormalized density f. This means that f is a positive constant times a probability density. Thus f is a nonnegative-valued function that integrates (for continuous state) or sums (for discrete state) to a value that is infinite and nonzero. The M-H update does the following.

- If the current state is θ, propose a move to a θ', having conditional probability density given θ denoted q
- Compute the Hasting ratio

$$r(\theta|\theta') = \frac{f(\theta')q(\theta'|\theta)}{f(\theta)q(\theta|\theta')} \tag{38}$$

- Accept the proposed move $\theta \to \theta'$ with probability

$$a(\theta|\theta') = \min\left[1, r(\theta|\theta')\right] \tag{39}$$

that is, the state after the update is θ' (accepted) with probability $a(\theta|\theta')$, and the state after the update is θ (rejected) with probability $1 - a(\theta|\theta')$.

The last step is often called Metropolis rejection. Note that the Hasting ration is undefined if $f(\theta) = 0$, thus we must always arrange that $f(\theta) > 0$ in the initial state. Since $r(\theta|\theta') = 0$ if $h(\theta') = 0$, the M-H update can never move to a new θ' having $f(\theta') = 0$.

The special case of M-H update when $q(\theta|\theta') = q(\theta'|\theta)$ for al θ, θ' is the Metropolis update. The Hasting ratio from (38) simplifies to

$$r(\theta|\theta') = \frac{f(\theta')}{f(\theta)} \tag{40}$$

and is called the Metropolis ratio or the odds ratio.Thus Metropolis updates save a little time in calculating $r(\theta|\theta')$ but otherwise have no advantages over M-H update.

One of the main reason of the wide use of M-H algorithm in Bayesian inference is due to the fact that $f(\theta)$ is not necessarily normalized. The posterior distribution (2) is normalized thanks to the marginal distribution (5) which is quite often unfeasible given the size of the data. Nevertheless, in the M-H algorithm we do not need to compute Z, since it is a possitive constant that cancel out in the Hasting (or Metropolis) ratio.

D.2 Hamiltonian Monte Carlo

In this section, we briefly discuss theoretical and practical aspects of Hamiltonian Monte Carlo (HMC) based on Brooks et al. (2011b, Chapter 5), particularly the No-U-Turn Sampler (NUTS) method proposed by Hoffman and Gelman (2011), an MCMC algorithm that closely resembles HMC. The main reason we focus on HMC is that it turns out to be the most suited algorithm to overcome pathologies arising in BHM (Betancourt & Girolami, 2013). As stated in the main text, the slow exploration of the parameter space in classical MCMC is mainly due to the endemic behaviour of the random walks. If Metropolis updates are done using a simple random-walk proposal distribution, in some cases the new proposed states with high probability of acceptance tend to lie close to the current state giving rise to a correlation between consecutive samples. Thus, the number of iterations needed to reach a state independent of the current state, crucial property in any Markov chain, could be high. In HMC, a new state is proposed by computing a trajectory according to Hamiltonian dynamics. A state proposed in this way can be distant from the current state but nevertheless have a high probability of acceptance. This bypasses the slow exploration of the state space that occurs in classical MCMC.
Hamiltonian dynamics operates on a d-dimensional position vector, q, and a d-dimensional momentum vector, p, so that the full state space has $2d$ dimensions. The system is described by a function of q and p known as the Hamiltonian, $H(q, p)$. The partial derivatives of the Hamiltonian determine how $q(t)$ and $p(t)$ change over time, t, according to Hamilton's equations

$$\frac{\mathrm{d}q_i(t)}{\mathrm{d}t} = \frac{\partial H(q,p)}{\partial p_i} \tag{41}$$

$$\frac{\mathrm{d}p_i(t)}{\mathrm{d}t} = -\frac{\partial H(q,p)}{\partial q_i} \tag{42}$$

where the Hamiltonian is the sum of potential energy, U, and kinetic energy, K, usually written in HMC as

$$H(q,p) = U(q) + K(p) \tag{43}$$

The distribution we wish to sample can be related to a potential energy function via the concept of a canonical distribution from statistical mechanics. The canonical ensemble assigns a probability to each distinct state, (q, p), given by the following equation

$$P(q,p) = \frac{1}{Z}\exp\left[-\beta H(q,p)\right] \tag{44}$$

$$= \frac{1}{Z}\exp\left[-\beta U(q)\right]\exp\left[-\beta K(p)\right] \tag{45}$$

where β is a free parameter[27], and Z is a normalization constant known as partition function. In Bayesian statistics, the posterior distribution for the model parameters, θ, is the usual focus of interest, and hence these parameters will take the role of the position, q. We can express the posterior distribution as a canonical distribution (with β = 1) using a potential energy function defined as

$$U(q) = \log\left[p(q|y)p(q)\right] \tag{46}$$

[27]In physics, it represents the inverse of the temperature in a convenient unit system.

where $p(q)$ is the prior, and $p(q|y)$ is the likelihood given data y. HMC samples from the canonical distribution for q and p defined by equation (44), in which q has the distribution of interest, as specified using the potential energy function (46). On the other hand, we can choose the distribution of the momentum variables, p, which are independent of q, as we wish, specifying the distribution via the kinetic function, $K(p)$. Common practice is to use a quadratic kinetic energy

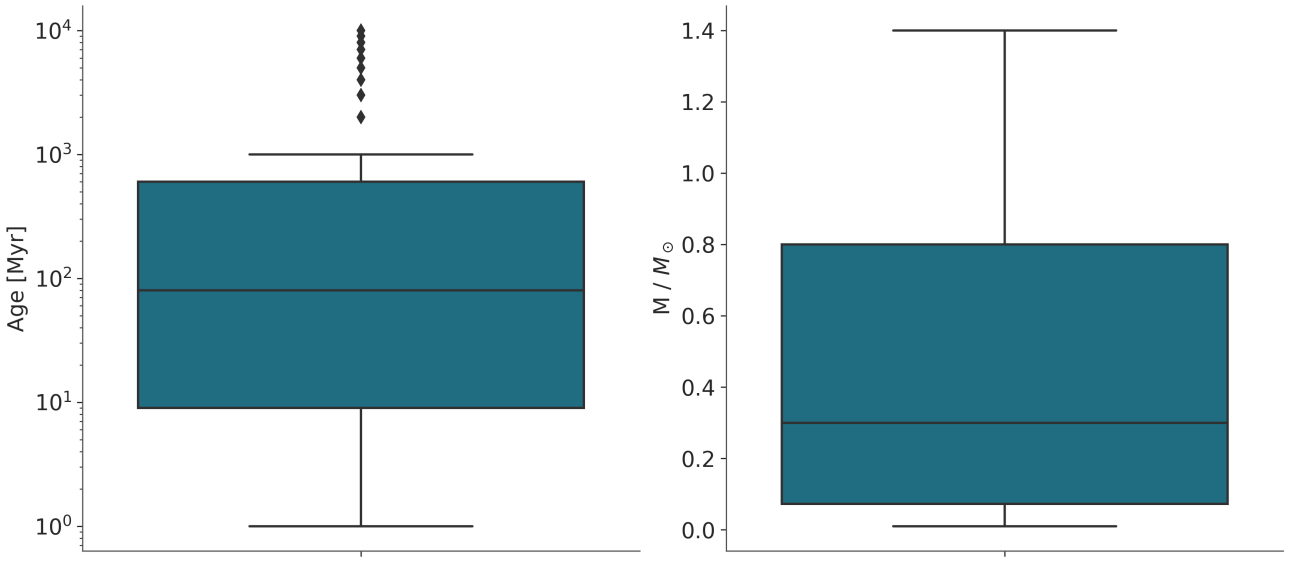
$$K(p) = \sum_{i=1}^{k} \frac{p_i^2}{2m_i} \tag{47}$$

Several properties of Hamiltonian dynamics, such as reversibility, Hamiltonian and volume conservation, are crucial to its use in constructing MCMC updates, the main problem comes from the computer implementation. Hamilton's equations must be approximated by discretizing time, using some small stepsize, ϵ. The proposed method by Brooks et al. (2011b) because of its results is the leapfrog method. Then, the Metropolis update using Hamiltonian dynamics works as follows: starting with the current state, (q, p), Hamiltonian dynamics is simulated for L steps using the leapfrog method, with stepsize of ϵ. Therefore, L and ϵ are parameters of the algorithm, which need to be tuned to obtain good performance.

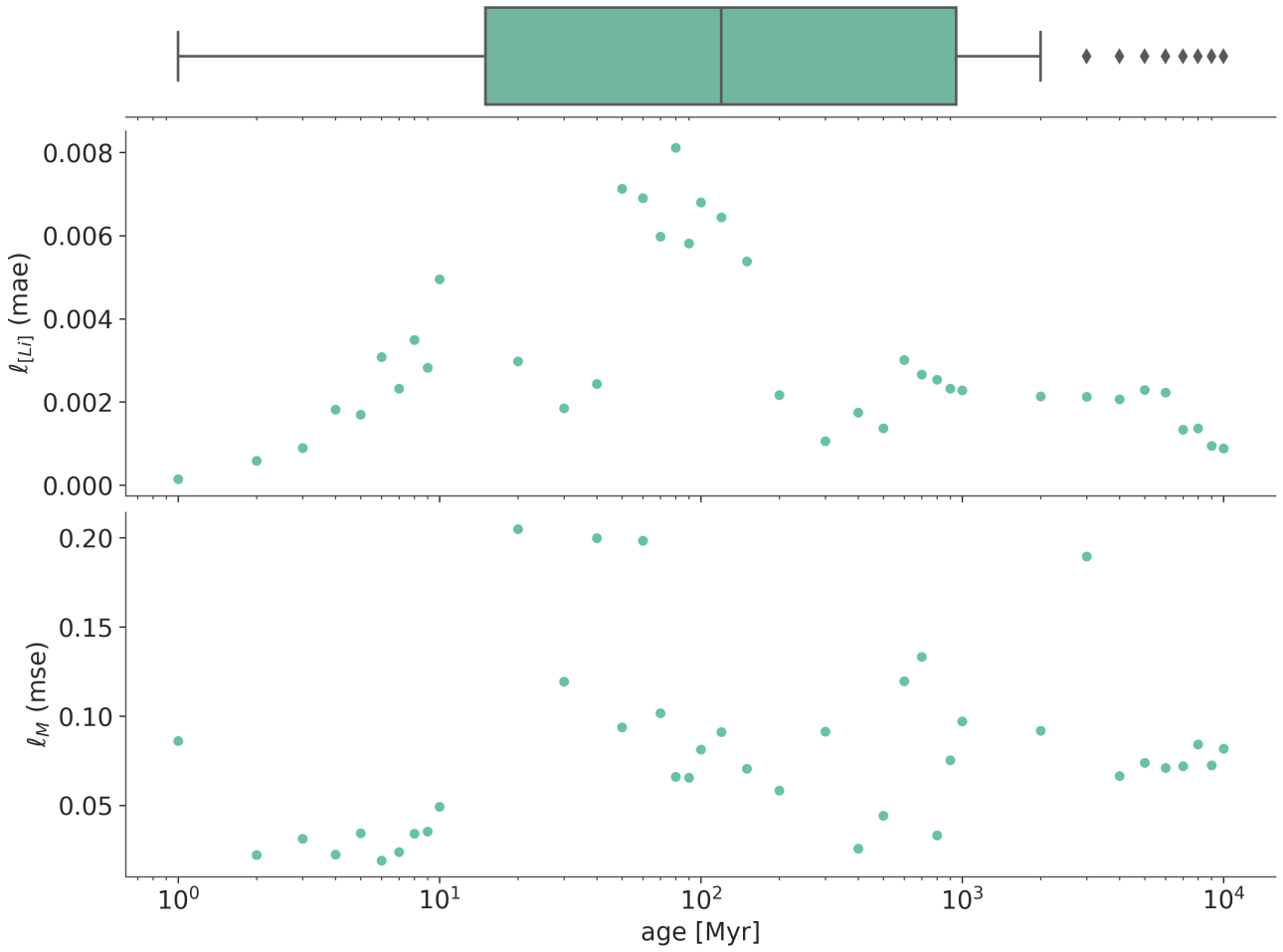
No-U-Turn Sampler HMC performance is highly sensitive to these parameters. In particular, if L is too small then the algorithm exhibits undesirable random walk behavior, while if L is too large the algorithm waste computation. The No-U-Turn Sampler (NUTS) algorithm proposed by Hoffman and Gelman (2011) is an extension to HMC that eliminates the need to set a number of steps L. They also developed a method to for automatically adapting stepsize parameter ϵ.

The algorithm is too complex for the scope of the section, but the main idea is that NUTS algorithm detects situations where the dot product between the difference in positions and the current momentum state becomes negative or drops below a threshold. This criterion effectively identifies when the trajectory starts to "U-turn" and curve back on itself, indicating that the algorithm should terminate the trajectory to prevent inefficient exploration and ensure more efficient sampling of the target distribution. If the termination criterion is not met, the algorithm uses a binary tree doubling procedure to decide whether to continue in the same direction or reverse direction. This procedure involves creating a binary tree of sub-trajectories and checking which side of the tree to explore based on the termination criterion. Additionally, NUTS adapts the step size and trajectory length based on the acceptance rate of proposals, making it possible to run NUTS with no hand tunning at all.

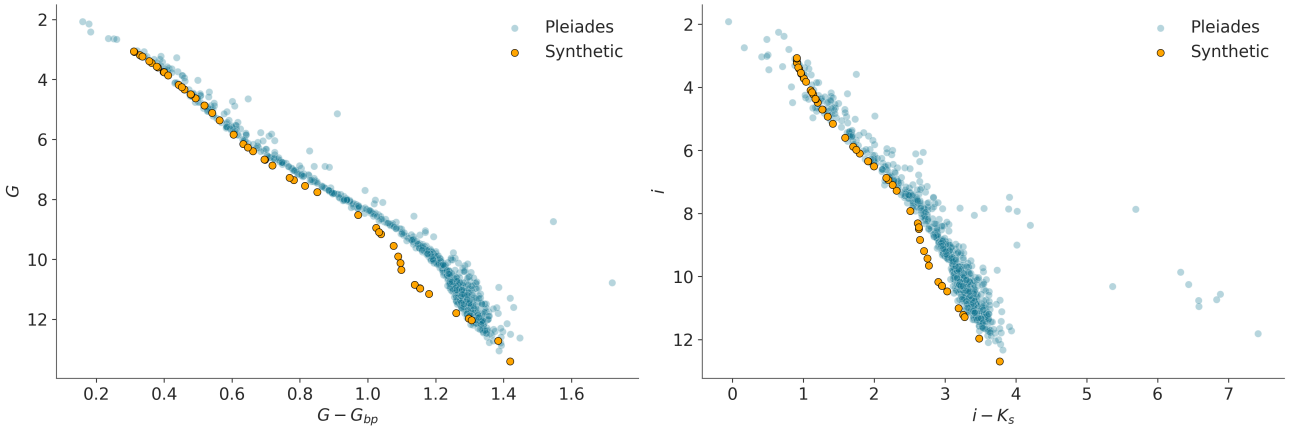
E Supplementary Figures



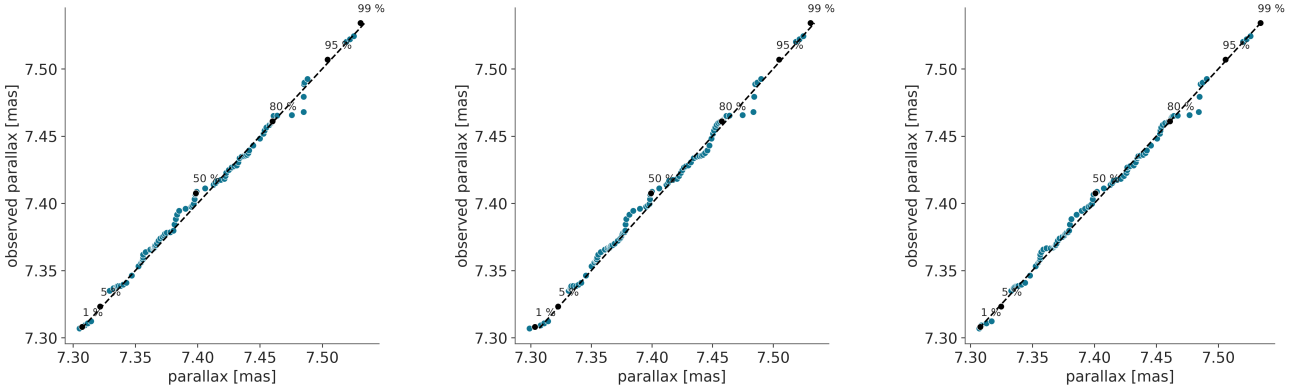
Supplementary Figure 1: *Age and mass distribution in BT-Settl models (Allard, 2013) grid. The age grid has a higher density of values between 10 and 1000, while for bigger values they are much more spread.*



Supplementary Figure 2: *Loss values for NN interpolation of BT-Settl models as a function of age. The higher concentration of age inputs at low values does not imply a poor performance for greater age values.*

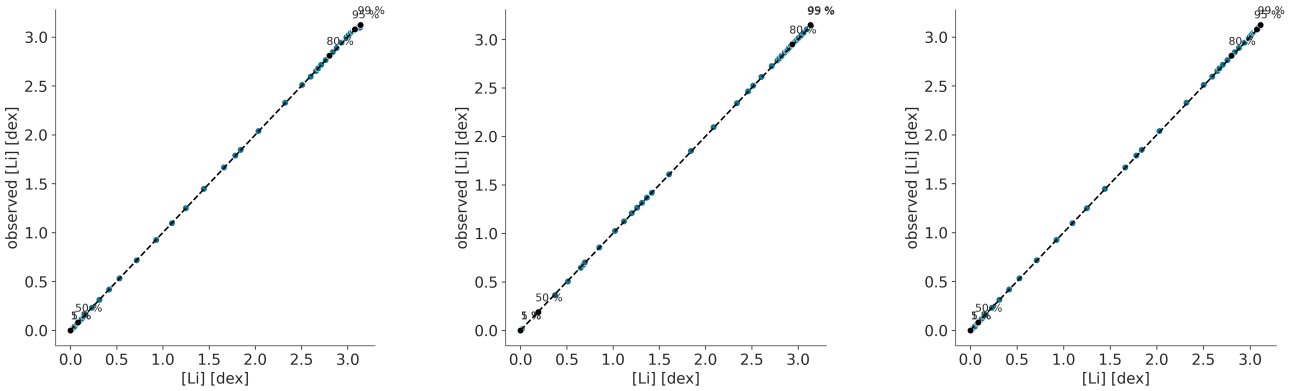


Supplementary Figure 3: *Synthetic dataset compare to Pleiades dataset based on HR-diagram representation.*



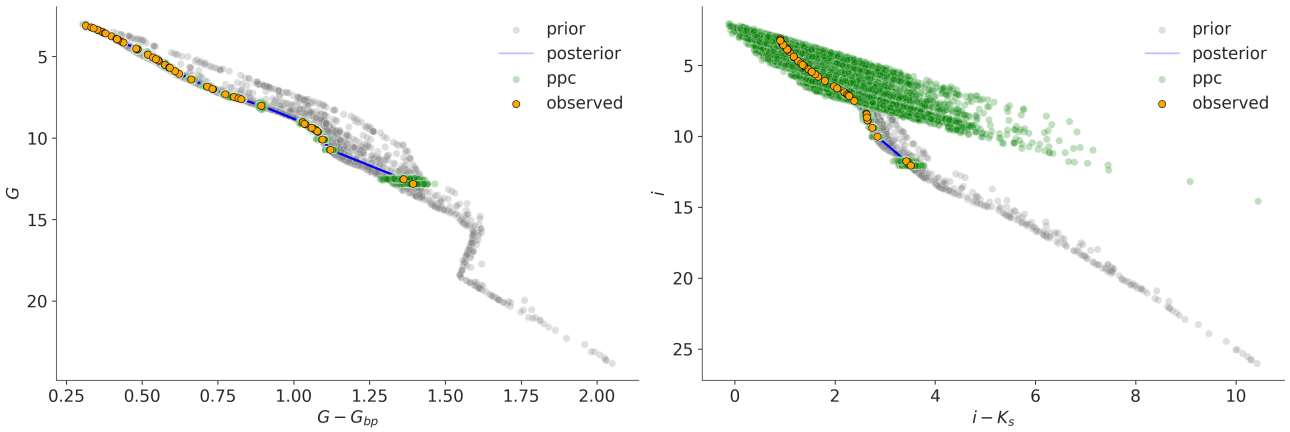
(a) *Uninformative* **(b)** *Weakly informative* **(c)** *Wrong*

Supplementary Figure 4: *Q-Q plot of model configuration 1 for different priors settings.*

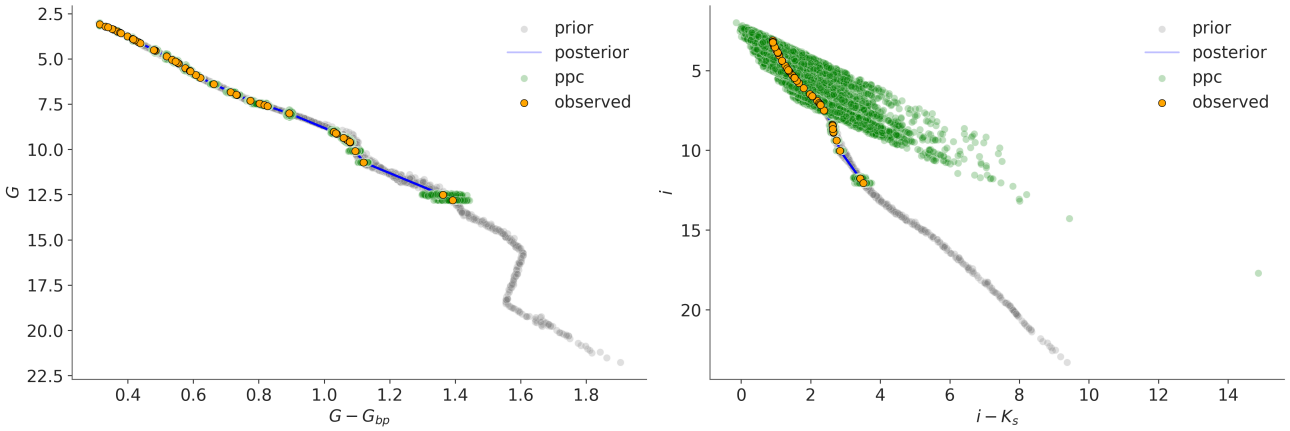


(a) *Uninformative* **(b)** *Weakly informative* **(c)** *Wrong*

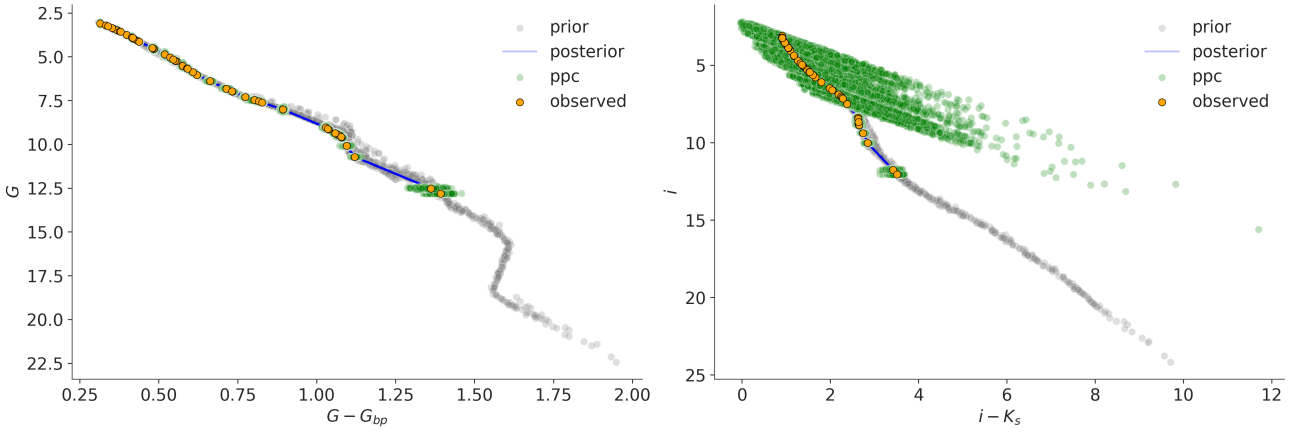
Supplementary Figure 5: *Q-Q plot of model configuration 4 for different priors settings.*



(a) *Uninformative*

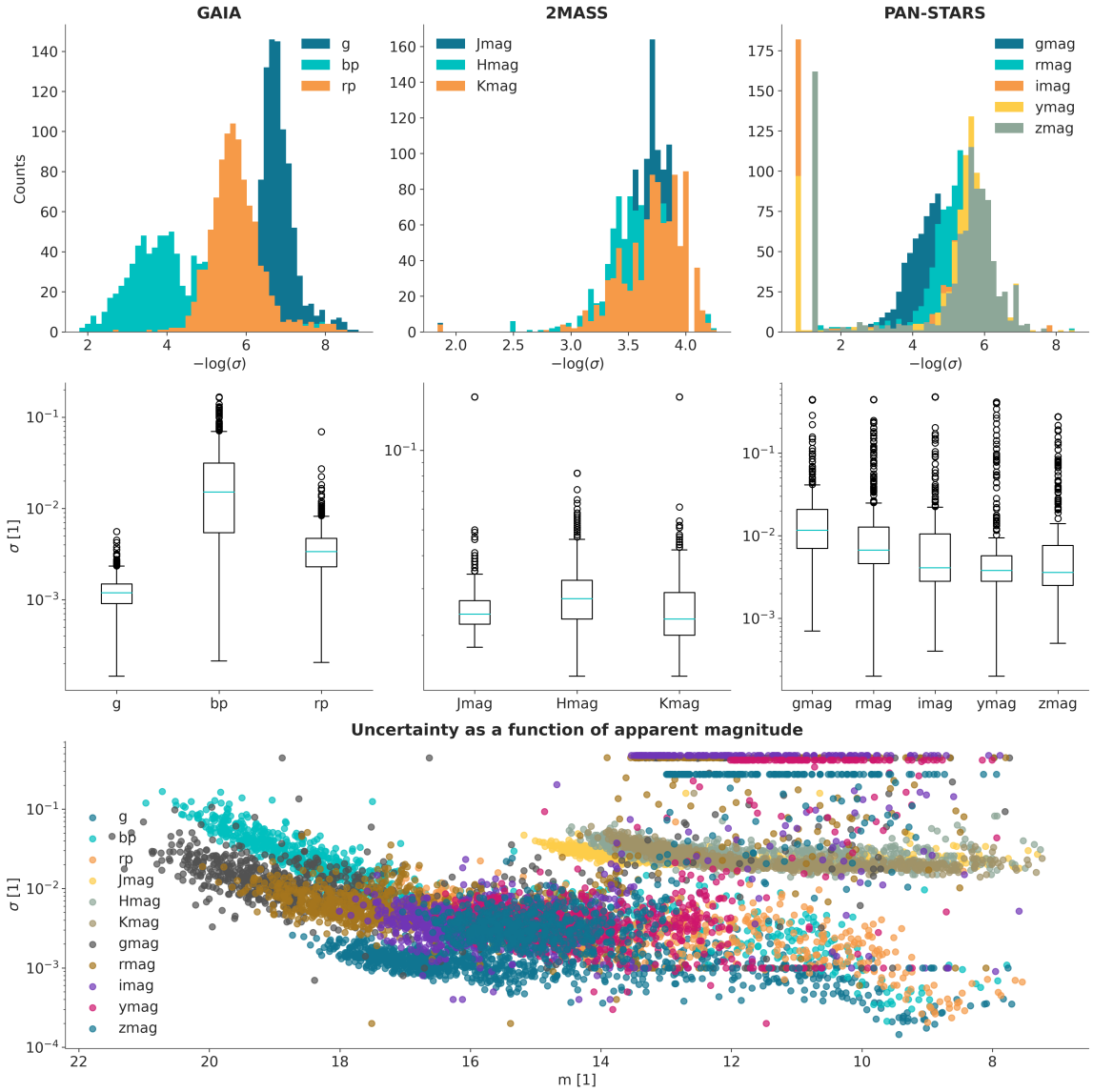


(b) *Weakly informative*

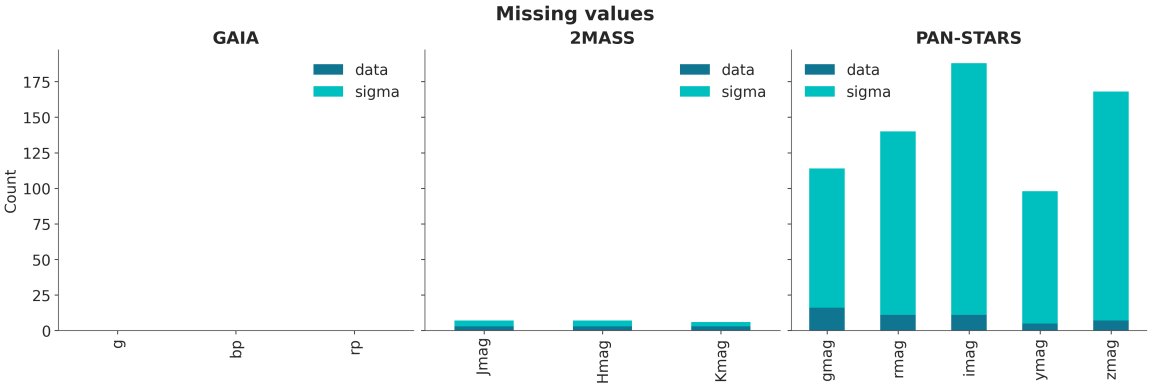


(c) *Wrong*

Supplementary Figure 6: *Posterior predictive checks based on HR-diagram of model configuration 1 for different priors settings.*



Supplementary Figure 7: *Photometric observations with their corresponding uncertainties after preprocessing. Missing values in uncertainty for each spectral band have been replaced by the maximum uncertainty in its spectral band.*



Supplementary Figure 8: *Missing values in our dataset. They can be found in both their actual value and their corresponding uncertainty. Gaia is free of missing values, while Pan-Starrs contains many missing values in uncertainty.*

F Supplementary Tables

Supplementary Table 1: *Mixture likelihood hyperparameters*

Notation	Description	Dims.	Units	Distribution
	Lithium observations			
$\phi_{[\mathrm{Li}]}$	Weight of outliers' distribution	1	1	$\mathcal{U}(a = 0,\ b = 1)$
$\mu_{[\mathrm{Li}]}$	Mean outliers' distribution	1	dex	$\mathcal{U}(a = -10,\ b = 10)$
$\delta_{[\mathrm{Li}]}$	Variance outliers' distribution	1	dex	$\mathcal{U}(a = 0,\ b = 10)$
	Flux observations			
ϕ_F	Weight of outliers' distribution	11	1	$\mathcal{U}(a = 0,\ b = 1)$
μ_F	Mean outliers' distribution	11	$\mathrm{erg/s/cm^2}$	`m2flux` $\mathcal{U}(a = -10,\ b = 50)$
δ_F	Variance outliers' distribution	11	$\mathrm{erg/s/cm^2}$	$\mathcal{U}(a = 0,\ b = 10)$

Glossary

absolute magnitude In astronomy, absolute magnitude refers to an intrinsic magnitude defined to directly compare luminosity by hypothetically placing all objects at a standard reference distance of 10 parsec from the observer, without extinction (or dimming) of its light due to absorption by interstellar matter and cosmic dust. 6, 7, 8, 9, 24

absorption line Absorption lines are dark lines in a spectrum resulting from the absorption of specific wavelengths of light by atoms or molecules, indicating their presence and identifying their elemental or molecular composition. 24

apparent magnitude Apparent magnitude is a measure used in astronomy to quantify the brightness of a celestial object as it appears to an observer on Earth. 7, 8, 24

binary star Binary stars are a pair of stars that are gravitationally bound to each other and orbit around a common center of mass. 21

brown dwarf A brown dwarf is a type of celestial object that falls in between the characteristics of a star and a planet. Brown dwarfs are not massive enough to sustain the nuclear fusion reactions that power stars like our Sun, but they are more massive than typical planets. 6, 9

color index Color index is a measure of an astronomical objects color, obtained by comparing its brightness in different wavelengths or filters, often used to classify and study celestial bodies such as stars and galaxies. 11

DAG A Directed Acyclic Graph (DAG) is a data structure that consists of nodes (or vertices) connected by directed edges, where the edges have a specific direction from one node to another and there are no cycles (loops) in the graph. 5

effective temperature Effective temperature is the temperature of an idealized black body that emits the same amount of radiation as a given star or object, representing its overall thermal energy output. 17

equivalent width The equivalent width is a measure of the width of a spectral line that has the same integrated flux as the actual line, often used in astronomy and spectroscopy. 17

extinction In astronomy, extinction refers to the process by which light from a celestial object, such as a star or a galaxy, is absorbed and dispersed as it passes through interstellar or intergalactic dust, gas, and other material. 6, 17

growth curve Growth curves in astrophysics are graphs that depict how the intensity of an astronomical signal or feature changes with time or another relevant parameter, used to study various cosmic phenomena like stellar evolution, transient object detection (e.g., supernovae), and variability in the light of variable stars. 17

HR-diagram The Hertzsprung-Russell (HR) Diagram is a fundamental tool in astronomy that graphically represents the relationship between the luminosities (brightnesses) and effective surface temperatures of stars. It is named after the astronomers Ejnar Hertzsprung and Henry Norris Russell, who independently developed the concept. 2, 11, 13, 14, 16, 18, 31, 32

isotope An isotope is a variant of a chemical element that has the same number of protons in its atomic nucleus (thus belonging to the same element), but a different number of neutrons, resulting in a slightly different atomic mass. 2

main sequence In astronomy, the main sequence is a prominent and crucial phase in the life cycle of a star, representing the phase when a star is primarily engaged in nuclear fusion and maintaining stability through the balance of gravitational forces and energy production. The specific location of a star on the main sequence depends on its mass. Higher-mass stars are hotter and more luminous, while lower-mass stars are cooler and less luminous. 2, 24

metallicity Metallicity in astrophysics refers to the abundance of elements heavier than hydrogen and helium in an astronomical object. 17

plate In the context of probabilistic graphical models, such as Bayesian networks, a "plate" is a graphical notation used to represent repeated structures or repetitions of variables. Plate notation is particularly useful for representing situations where a group of variables share the same relationships and characteristics and appear multiple times in the model. 5, 6, 9

power set In mathematics, the power set of a set is the collection of all possible subsets of that set, including the empty set and the set itself. 5

spectral band A spectral band refers to a specific range of wavelengths or frequencies within the electromagnetic spectrum. Spectral bands are utilized to study and manipulate specific ranges of electromagnetic radiation for various purposes, such as provide information about the composition, temperature, and other properties of objects in space. 6, 7, 8, 9, 10, 15, 19, 24, 33

zero point In photometric measurements, the "zero point" refers to a reference level or calibration point against which all other measurements are made. It is a crucial concept in photometry, which is the branch of science that deals with the measurement of light intensity or brightness, typically in the context of celestial objects or other sources of light. 8