Towards constraining warm dark matter with stellar streams through neural simulation-based inference

Joeri Hermans,[1]★ Nilanjan Banik,[2] Christoph Weniger,[3] Gianfranco Bertone,[3] and Gilles Louppe[1]
[1]*Montefiore Institute, University of Liège, Belgium*
[2]*Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA*
[3]*GRAPPA Institute, Institute for Theoretical Physics Amsterdam and Delta Institute for Theoretical Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

A statistical analysis of the observed perturbations in the density of stellar streams can in principle set stringent constraints on the mass function of dark matter subhaloes, which in turn can be used to constrain the mass of the dark matter particle. However, the likelihood of a stellar density with respect to the stream and subhaloes parameters involves solving an intractable inverse problem which rests on the integration of all possible forward realisations implicitly defined by the simulation model. In order to infer the subhalo abundance, previous analyses have relied on Approximate Bayesian Computation (ABC) together with domain-motivated but handcrafted summary statistics. Here, we introduce a likelihood-free Bayesian inference pipeline based on Amortised Approximate Likelihood Ratios (AALR), which automatically learns a mapping between the data and the simulator parameters and obviates the need to handcraft a possibly insufficient summary statistic. We apply the method to the simplified case where stellar streams are only perturbed by dark matter subhaloes, thus neglecting baryonic substructures, and describe several diagnostics that demonstrate the effectiveness of the new method and the statistical quality of the learned estimator.

Key words: methods: statistical, galaxy: structure, cosmology: dark matter

1 INTRODUCTION

Cold Dark Matter (CDM) models (Peebles 1982; Blumenthal et al. 1984) predict a hierarchical collapse in which large haloes form through the merging of smaller dark matter clumps (Moore et al. 1999; Avila-Reese et al. 1998; Zhao et al. 2003). This process is driven by CDM's scale-free halo mass function (Hofmann et al. 2001; Schneider et al. 2013) and depends on the initial conditions of the matter power spectrum, which in turn anticipates the existence of dark matter haloes down to 10^{-4} $M_\odot$ (Bertschinger 2006). Warm Dark Matter (WDM) models (Bond & Szalay 1983; Dodelson & Widrow 1994; Bode et al. 2001) on the other hand, in which the dark matter particle is much lighter, influence structure formation down to the scale of dwarf galaxies. While at large scales the collapse in WDM is hierarchical as well, it becomes strongly suppressed below the half-mode mass scale of the corresponding dark matter model, where the non-negligible velocity dispersion of dark matter particles prevents haloes to form and smooths the density field instead (Smith & Markovic 2011). Therefore, a powerful method of probing the particle nature of dark matter is to measure the abundances of the lowest mass subhaloes in our galaxy. While higher mass subhaloes will eventually initiate star formation and manifest themselves as dwarf galaxies, detecting low mass subhaloes ($\lesssim 10^9$ $M_\odot$) remains particularly hard since they either have very few faint stars or none at all.

Cold stellar streams that formed due to the tidal disruption of globular clusters by the Milky Way potential are a powerful probe for detecting and measuring the abundances of these low mass subhaloes (Ibata et al. 2002; Johnston et al. 2002; Yoon et al. 2011; Carlberg 2012; Erkal & Belokurov 2015a,b). When a subhalo flies past a stellar stream, it gravitationally perturbs the orbit of the stream stars around the point of closest approach, which leaves a visible imprint in the form of a region of low stellar density or a *gap*. Such gaps can be individually analyzed to infer the properties of a single subhalo perturber (Erkal & Belokurov 2015b). However, a stream is expected to encounter many subhalo impacts over its dynamical age, leading to complicated density structures that can be hard to separate into individual gaps. Therefore, a more pragmatic approach is to study the full stream density and statistically infer the subhalo abundance within the galactocentric radius of the stream (Bovy et al. 2017).

Stream-subhalo encounters are described by various quantities such as the impact parameter, the flyby velocity of the subhalo, mass and size of the subhalo, and the time and angle of the subhalo impact. While simulating stream-subhalo encounters and their effects on the stellar density through these parameters is relatively straightforward, the forward model does not easily lend itself to statistical inference. The reason for this is that the likelihood of a stellar density with respect to these parameters involves solving an intractable

★ E-mail: joeri.hermans@doct.uliege.be

inverse problem which rests on the integration of all possible forward realisations implicitly defined by the simulation model. It remains however possible to infer the underlying probabilities by relying on likelihood-free approximations (Cranmer et al. 2020). From this perspective, Bovy et al. (2017) applied Approximate Bayesian Computation (ABC) (Rubin 1984) to infer subhalo abundance using the power spectrum and bispectrum of the stream density as a summary statistic. With the same ABC technique, Banik et al. (2018); Banik et al. (2019b) applied the stream density power spectrum as a summary statistic to infer the particle mass of thermal relic dark matter.

It should be noted that ABC posteriors are *only* exact whenever the handcrafted summary statistic is *sufficient*, and the distance function chosen to express the similarity between observed and simulated data tends to 0, which in practice is never achievable. We address this issue by introducing a likelihood-free Bayesian inference pipeline based on amortised approximate likelihood ratios (AALR) (Hermans et al. 2019), which automatically learns a mapping between the data and the simulator parameters by solving a tractable minimization problem. Afterwards, the learned estimator is able to accurately approximate the posterior density function of arbitrary stellar streams supported by the simulation model. By automatically learning this relation from data, we obviate the need to handcraft a possibly insufficient summary statistic, therefore enabling domain-scientists to pivot from solving the intractable inverse problem to the more natural forward modeling. In addition, we describe several diagnostics to inspect the statistical quality of the learned estimators with respect to the simulation model. We demonstrate the effectiveness of this method by inferring the particle mass of dark matter within the stellar stream framework.

The paper is outlined as follows. In Section 2 we present the steps to forward model the stream-subhalo encounter simulations, and highlight our assumptions. Section 3 outlines the statistical formalism and the proposed methodology. Several diagnostics are discussed to probe the statistical quality. Section 4 evaluates the proposed methodology. We conclude in Section 5.

To support the reproducibility of this work, we document and provide all code on GitHub[1]. A tutorial demonstrating the technique on a toy problem is provided as well. Steps to obtain the simulated data and pretrained models are described there. Additionally, we annotate every result and figure with **</>**, which links to the code or Jupyter notebook used to generate it.

2 STREAM MODELING

We use the `streampepperdf` simulator[2] that is based within the `galpy` framework (Bovy 2015) to model stream-subhalo interactions. Baryonic structures in our galaxy, namely, the bar, spiral arms and the Giant Molecular clouds can induce stream density variations similar to those caused by subhalo impacts (Amorisco et al. 2016; Erkal et al. 2017; Pearson et al. 2017; Banik & Bovy 2019). However, owing to its retrograde orbit and a perigalacticon of ~ 14 kpc, the effect of the baryonic structures on the GD-1 stream (Grillmair & Dionatos 2006) is expected to be subdominant compared to that by a CDM like population of subhalos. Therefore, we have used the GD-1 stream for our analyses and ignored the effects from the baryonic structures. Since the location of GD-1's progenitor is not known, we adopt the model presented in Webb & Bovy (2019), which proposes that the progenitor cluster disrupted in its entirety approximately 500 Myr ago and resulted in the gap at the observed stream coordinate $\phi = -40°$. The dynamical age of the GD-1 stream is also unknown and so following the arguments in (Banik et al. 2019a), we consider all stream models in the range of 3-7 Gyr.

Our simulation model samples subhaloes in the sub-dwarf-galaxy mass range $[10^5 - 10^9]$ $M_\odot$, since density perturbations due to subhaloes less massive than 10^5 $M_\odot$ are below the level of noise in the current data. Warm Dark Matter (WDM) is modeled as a thermal relic candidate which is completely described by its particle mass. The implementation of the subhaloes follows the same procedure as in (Bovy et al. 2017; Banik et al. 2018; Banik et al. 2019a).

We summarize the salient steps of the forward model for completeness. The WDM mass function is modeled following Lovell et al. (2014):

$$\left(\frac{dn}{dM}\right)_{\rm WDM} = \left(1 + \gamma \frac{M_{\rm hm}}{M}\right)^{-\beta} \left(\frac{dn}{dM}\right)_{\rm CDM}, \quad (1)$$

where $\gamma = 2.7$, $\beta = 0.99$ and $\left(\frac{dn}{dM}\right)_{\rm CDM} \propto M^{-1.9}$. Here, $M_{\rm hm}$ is the half-mode mass that quantifies the scale below which the mass function is strongly suppressed. Both the CDM and WDM mass functions were obtained by fitting the subhaloes within a Milky Way like analogue from the Aquarius cosmological simulations (Springel et al. 2008). Being dark matter only simulations, these mass functions do not account for the disruption of subhaloes due to baryonic structures such as the disk, which has been shown to be capable of destroying around $\sim 10 - 50\%$ of the subhaloes within the galactocentric radius of the GD-1 stream and in the mass range $10^{6.5} - 10^{8.5}$ $M_\odot$ (D'Onghia et al. 2010; Sawala et al. 2017; Garrison-Kimmel et al. 2017; Kelley et al. 2019; Webb & Bovy 2020). Moreover, the disrupted fraction of WDM subhaloes is expected to be even higher due to their lower concentrations. In this paper we have ignored subhalo disruptions due to baryonic effects and defer a full analysis to a future publication.

For each simulated stream density, we consider the region $-34° < \phi < 10°$ in the observed coordinate frame, and normalize the stream density by dividing it by the mean density. The latter step is different from what was done in (Bovy et al. 2017; Banik et al. 2019a), where the authors normalize the stream density by dividing it by a 3rd order polynomial fit. We tested that both normalization procedures gave similar results. This was also demonstrated in Bovy et al. (2017), where they found that changing the order of the smoothing polynomial did not significantly affect the (ABC) posterior. Finally, noise is added to every simulated stream density by sampling a Gaussian realisation of the noise from the observed GD-1 data from de Boer et al. (2020).

3 METHOD

3.1 Statistical formalism

This work considers two inference scenarios. In the first we jointly infer the WDM mass $m_{\rm WDM}$ and the stream age $t_{\rm age}$. The second scenario solely considers $m_{\rm WDM}$ while marginalizing the stream age. Because our methodology generalizes to various domains, we ease

[1] Available at `https://git.io/JUvmj`.
[2] Available at `https://github.com/jobovy/streamgap-pepper` .

the discussion by simplifying the nomenclature into the following concepts:

Target parameters ϑ denote the main parameters of our simulation model. Depending on the inference scenario at hand, $\vartheta \triangleq (m_{\mathrm{WDM}},\ t_{\mathrm{age}})$ or $\vartheta \triangleq (m_{\mathrm{WDM}})$. Given the Bayesian perspective of this analysis, we define the priors over the WDM mass m_{WDM} and stream age t_{age} to be `uniform`$(1,\ 50)$ keV and `uniform`$(3,\ 7)$ billion years (Gyr) respectively. The upper bound of 50 keV is justified since it corresponds to a half-mode mass of $\sim 4\times10^4\ \mathrm{M}_\odot$, which is below the sensitivity of stellar streams given current observational uncertainties.

Observables x encapsulate the simulated stellar density of mock streams and the *observed* GD-1 density. An observable is encoded as a 66-dimensional vector along the linear angle ϕ between -34 and 10 degrees.

Nominal value ϑ^* or groundtruth used to simulate the observable x of a mock stream, i.e., $x \sim p(x|\vartheta^*)$.

Nuisance parameters η such as the impact angle and subhalo mass are not of direct interest, but their (random) effects must be accounted for to infer ϑ (Neyman & Scott 1948). However, this leaves us with the likelihood function $p(x|\vartheta,\eta)$. Given the Bayesian perspective of this work, we incorporate nuisance parameter uncertainty (Berger et al. 1999) by integration. The priors associated with the nuisance parameters are implicitly defined through the simulation model.

3.2 Motivation

Our multi-faceted simulation model induces an extensive space of possible execution paths, which, for example, correspond to randomly sampled dark matter haloes that impact the stellar stream throughout its evolution. The evaluation of the likelihood $p(x|\vartheta)$ of an observable x therefore involves amongst others the integration over a large variety of possible collision histories that are consistent with ϑ. Given the high-dimensional nature of this integral, directly evaluating data likelihoods is intractable.

A common Bayesian approach to address the intractability of the likelihood is to reduce the dimensionality of an observable x by means of a summary statistic $s(x)$. The reduction in dimensionality allows the posterior to be approximated numerically by collecting samples $\vartheta \sim p(\vartheta)$ for which observables produced by the forward model $s(x) \sim p(x|\vartheta)$ are similar, in terms of some distance, to the compressed representation of the observed data $s(x_o)$. This rejection sampling scheme is commonly referred to as *Approximate Bayesian Computation* (Rubin 1984) (ABC) and is, as the name indicates, *approximate*. Although the compression of x into a summary statistic makes the numerical approximation of the posterior tractable, it may reduce the statistical power of an analysis because the selected summary statistic often destroys relevant information. In fact, ABC is *only* exact whenever the summary statistic is *sufficient* and the distance function chosen to express the similarity between between $s(x)$ and $s(x_o)$ tends to 0. This is in practice never achievable because for a given simulation budget (i) a small acceptance threshold severely impacts the rate at which proposed samples are accepted, affecting the approximation of the posterior density function, and (ii) the *assumed* sufficiency of the summary statistic is virtually never thoroughly demonstrated in practice. Despite these shortcomings, ABC has been fruitfully applied in cosmology to constrain dark matter models within the context of stellar streams (Banik et al. 2018; Bovy 2019; Banik et al. 2019b), and more recently gravitational lensing (Gilman et al. 2020).

This work tackles the *intractability* of the likelihood from a different perspective. Instead of manually crafting a summary statistic and a distance function with a specific acceptance threshold, we propose to learn an amortised mapping from target parameters ϑ and observables x to posterior densities by solving a *tractable* minimization problem. The learned mapping has the potential to increase the statistical power of an analysis since the procedure, in contrast to ABC, *automatically* attempts to learn an internal sufficient summary statistic of the data. The automated procedure therefore enables domain-experts to solely focus on the forward modeling of the phenomena of interest, because the method does not require any consideration whether synthetic observables are compressible into low-dimensional summary statistics. Although the proposed method treats the simulation model as a black box, we would like to point out that it is possible to improve the efficiency of the minimization problem, provided that latent information can be extracted from the simulation model (Brehmer et al. 2018; Brehmer et al. 2019; Brehmer et al. 2020), albeit at some implementation cost.

3.3 Inference

The Bayesian paradigm finds model parameters compatible with observation by computing the *posterior*

$$p(\vartheta|x) = \frac{p(\vartheta)p(x|\vartheta)}{p(x)}. \tag{2}$$

Evaluating the posterior density for a given target parameter ϑ and an observable x in our setting is not possible because the likelihood $p(x|\vartheta)$ is per definition intractable. To enable the tractable evaluation of the posterior, we have to rely on likelihood-free surrogates for key components in Bayes' rule. Note that Equation 2 can be factorized into the product of the tractable prior probability and the intractable likelihood-to-evidence ratio $r(x|\vartheta)$:

$$p(\vartheta|x) = p(\vartheta)\frac{p(x|\vartheta)}{p(x)} = p(\vartheta)\frac{p(\vartheta,x)}{p(\vartheta)p(x)} = p(\vartheta)r(x|\vartheta). \tag{3}$$

Hermans et al. (2019) show that an amortised estimator $\hat{r}(x|\vartheta)$ of the intractable likelihood-to-evidence ratio can be obtained by training a discriminator $d(\vartheta,x)$ with inputs ϑ and x, to distinguish between samples from the joint $p(\vartheta,x)$ with class label 1 and samples from the product of marginals $p(\vartheta)p(x)$ with class label 0 using a discriminative criterion such as the binary cross entropy. Whenever the training criterion is minimized, the authors theoretically demonstrate that the optimal discriminator $d(\vartheta,x)$ models the Bayes optimal decision function

$$d(\vartheta,x) = \frac{p(\vartheta,x)}{p(\vartheta,x)+p(\vartheta)p(x)}. \tag{4}$$

Subsequently, given a model parameter ϑ and an observable x, we can use the discriminator as a density *ratio estimator* to compute the likelihood-to-evidence ratio

$$r(x|\vartheta) = \frac{1-d(\vartheta,x)}{d(\vartheta,x)} = \frac{p(\vartheta,x)}{p(\vartheta)p(x)} = \frac{p(x|\vartheta)}{p(x)}. \tag{5}$$

However, the computation of this formulation suffers from significant numerical issues in the saturating regime where the output of the discriminator tends to 0. Considering that $\log r(x|\vartheta) =$

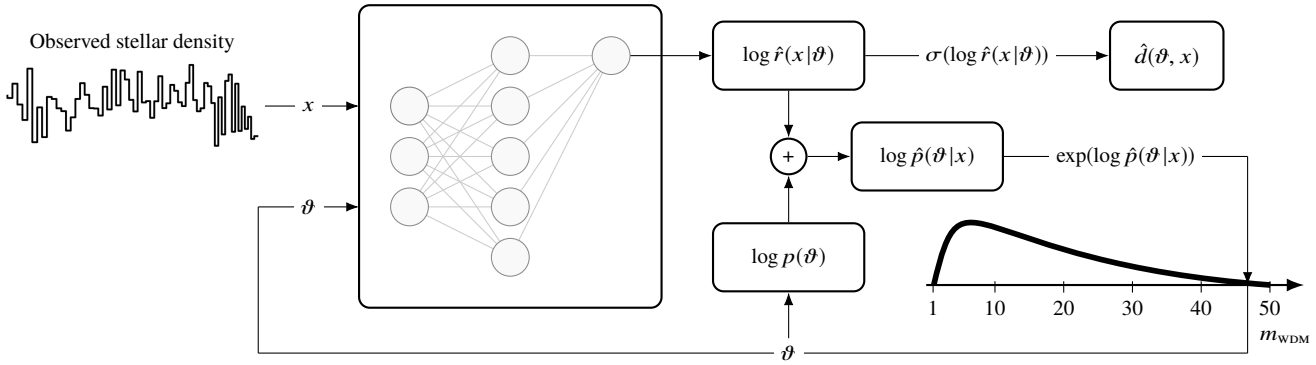

Figure 1. Graphical representation of the inference procedure after training the ratio estimator (neural network). The ratio estimator accepts a target parameter ϑ and an observable x as inputs, which are subsequently used to approximate the likelihood-to-evidence ratio $\hat{r}(x|\vartheta)$. The discriminator output $\hat{d}(\vartheta, x)$ — the sigmoidal projection $\sigma(\cdot)$ of $\log \hat{r}(x|\vartheta)$ — is only used during training. Given that the log prior probability of ϑ is a tractable quantity, we can easily approximate the log posterior probability $\log \hat{p}(\vartheta|x)$ by adding the approximated log likelihood-to-evidence ratio. Taking the exponent of the produced quantity results in a direct estimate of the posterior density. This procedure can be repeated for arbitrary target parameters ϑ supported by the prior. It should be noted that the neural network depicted here is an abstract representation. Our technique does not put any constraints on the architecture of the neural network. It is therefore possible to use of-the-shelf architectures of arbitrary complexity available in the Machine Learning literature.

$\text{logit}(d(\vartheta, x))$ for classifiers with a *sigmoidal* projection at the output, it is possible to directly obtain $\log r(x|\vartheta)$ from the classifier by extracting the quantity before the sigmoidal operation. This strategy ensures that the approximation of $\log r(\vartheta|x)$ is numerically stable. In addition, randomly shuffling ϑ in a batch $\vartheta, x \sim p(\vartheta, x)$ instead of drawing a new samples from the product of marginals significantly aids the convergence rate of the discriminator. After training, estimates of the posterior probability density function can be approximated for arbitrary (without retraining) target parameters ϑ and observables x by computing

$$\log p(\vartheta|x) \approx \log p(\vartheta) + \log \hat{r}(x|\vartheta), \tag{6}$$

provided that ϑ and x are supported by the prior $p(\vartheta)$ and the marginal model $p(x)$ respectively, thereby enabling consistent and fast likelihood-free posterior inference. Figure 1 provides a graphical overview. We refer the reader to Hermans et al. (2019) or our GitHub repository for implementation details.

The ratio estimator can likewise be adapted to compute a credible region (CR) at a desired level of uncertainty α by constructing a region Θ in the model parameter space which satisfies

$$\int_\Theta p(\vartheta) r(x|\vartheta)\, d\vartheta = 1 - \alpha. \tag{7}$$

Since many such regions Θ exist, we select the highest posterior density region, which is the smallest CR.

Although our analysis focuses on the Bayesian paradigm, it is possible use the ratio estimator in a frequentist setting (Cranmer et al. 2015; Brehmer et al. 2019). The likelihood-ratio $\lambda(x|\vartheta_0, \vartheta_1)$ between two hypotheses ϑ_0 and ϑ_1 can easily be computed with the ratio estimator as the denominators of $r(x|\vartheta_0)$ and $r(x|\vartheta_1)$ cancel out, i.e.,

$$\lambda(x|\vartheta_0, \vartheta_1) = \frac{p(x|\vartheta_0)}{p(x|\vartheta_1)} = \frac{r(x|\vartheta_0)}{r(x|\vartheta_1)}. \tag{8}$$

The same strategy applies to the likelihood-ratio (Cowan et al. 2011) test statistic for a specific observable x

$$-2 \log \lambda(\vartheta) = -2 \log \frac{p(x|\vartheta)}{p(x|\hat{\vartheta})}, \tag{9}$$

where the maximum likelihood estimate $\hat{\vartheta}$ is

$$\hat{\vartheta} = \arg\max_{\vartheta} r(x|\vartheta). \tag{10}$$

The test statistic can thus be expressed (Cranmer et al. 2015) as

$$-2 \log \lambda(\vartheta) = -2 \left[\log r(x|\vartheta) - \log r(x|\hat{\vartheta})\right]. \tag{11}$$

As a result of Wilks' theorem (Wilks 1938), we can directly convert the test statistic into a confidence level (CL) under the assumption that the statistic is χ^2_k-distributed with k degrees of freedom (in function of ϑ's dimensionality).

3.4 Diagnostics

Before making any scientific conclusion, it is crucial to verify the result of the involved statistical computation. This is especially challenging in the likelihood-free setting because evaluating the likelihood is intractable. The following subsections describe several diagnostics to assess the quality of the amortised ratio estimates. No additional training or fine-tuning is applied as this would change the statistical properties of the ratio estimator.

3.4.1 *Proper probability density*

A ratio estimator $\hat{r}(x|\vartheta)$ which correctly models the true likelihood-to-evidence ratio should satisfy

$$\int_\vartheta p(\vartheta) \hat{r}(x|\vartheta)\, d\vartheta \approx 1 \;\forall x. \tag{12}$$

The diagnostic should be applied to observables x of an evaluation dataset *and* real observables x_o. Passing the diagnostic on the evaluation dataset, while failing on x_o indicates that x_o is not supported by the marginal model $p(x)$, because ratio estimates in this regime are undefined and can therefore take on arbitrary values.

3.4.2 *Coverage*

Coverage quantifies the reliability of a statistical method to reconstruct the nominal value (Neyman 1937; Schall 2012; Strege et al. 2012; Prangle et al. 2013). The approximation of the ratio estimator

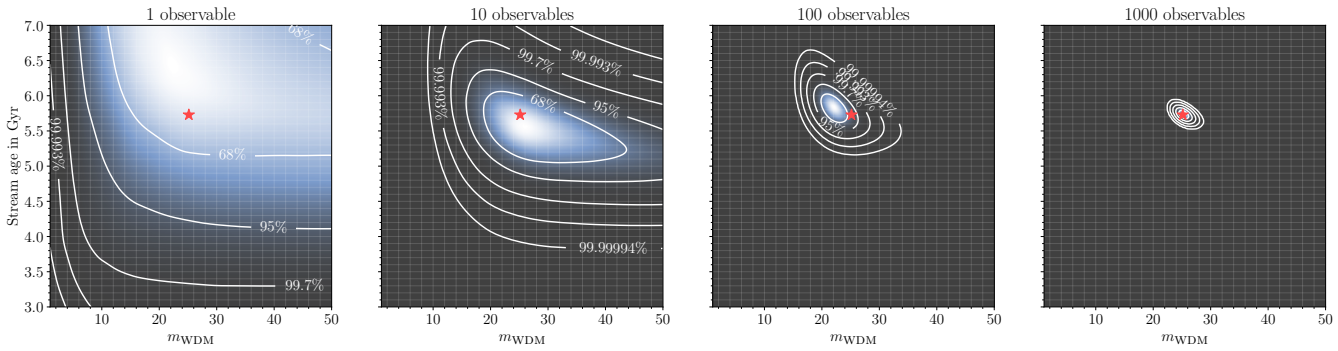

Figure 2. Demonstration of the mode convergence diagnostic described in Section 3.4.3. The figures show, from left to right, the posteriors for 1, 10, 100 and 1000 independent and identically distributed mock GD-1 observables. Every figure adopts the same nominal value or groundtruth, which is highlighted by the red star. As the amount of observables increases, the posteriors are becoming increasingly more tight around the nominal value. This indicates that the individual posteriors do not, in expectation, introduce significant bias for independent and identically distributed observables. </>

can thus be assessed by determining whether the empirical coverage probability matches the nominal coverage probability, which corresponds to the confidence level $1 - \alpha$. The empirical coverage probability is estimated using samples from a (large) presimulated evaluation dataset. This evaluation dataset consists of samples $\vartheta, x \sim p(\vartheta, x)$. For every pair (ϑ, x) in the evaluation dataset, we compute the corresponding credible or confidence interval. The fraction of samples for which the groundtruth was contained within the interval corresponds to the empirical coverage probability. If the empirical coverage probability $\geq 1 - \alpha$, then the ratio estimator passes the diagnostic. It is of course desirable that the empirical coverage probability of the ratio estimator converges to the confidence level. A substantially larger empirical coverage probability corresponds to intervals which are overly conservative. This implies that the ratio estimates are wrong, *but*, that in expectation the estimated posteriors are conservative, which is not an undesirable property. It should be noted that coverage can only be computed efficiently because our ratio estimator amortizes the estimation of the likelihood-to-evidence ratio. An equivalent study for ABC would have a significant computational cost.

3.4.3 *Convergence of the mode towards the nominal value*

The diagnostic is based on the idea that the maximum a posteriori (MAP) estimate converges towards the nominal value ϑ^* for an increasing number of independent and identically distributed observables $x \sim p(x|\vartheta^*)$. If the approximation of $\hat{r}(x|\vartheta)$ is correct, the MAP should in the limit coincide with the nominal value ϑ^*. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of i.i.d. observables. To compute the MAP, we need $p(\vartheta|\mathcal{X})$. As noted by Brehmer et al. (2019), Bayes' rule can be rewritten as

$$p(\vartheta|\mathcal{X}) = \frac{p(\vartheta) \prod_{x\in\mathcal{X}} p(x|\vartheta)}{\int p(\vartheta') \prod_{x\in\mathcal{X}} p(x|\vartheta')\, d\vartheta'} \tag{13}$$

$$= p(\vartheta) \left[\int p(\vartheta') \prod_{x\in\mathcal{X}} \frac{p(x|\vartheta')}{p(x|\vartheta)}\, d\vartheta' \right]^{-1} \tag{14}$$

$$\approx p(\vartheta) \left[\int p(\vartheta') \prod_{x\in\mathcal{X}} \frac{\hat{r}(x|\vartheta')}{\hat{r}(x|\vartheta)}\, d\vartheta' \right]^{-1}. \tag{15}$$

The integral can be estimated through Monte Carlo sampling. By checking whether the MAP concurs with the nominal value, we effectively probe the bias. Ideally, this diagnostic should be repeated for various groundtruths to inspect the behaviour of the ratio estimator over the complete model parameter space. In some settings however, the posterior may be multi-modal. In such scenarios the convergence of the mode(s) instead of the MAP should be assessed. A trial of the diagnostic is shown in Figure 2.

3.4.4 *Receiver operating characteristic*

We note that $\hat{r}(x|\vartheta)$ is only exact whenever

$$p(x)\frac{p(x|\vartheta)}{p(x)} = p(x)\hat{r}(x|\vartheta) = p(x|\vartheta), \tag{16}$$

is satisfied for all ϑ and x. Although $p(x)$ and $p(x|\vartheta)$ cannot be evaluated directly, it remains possible to sample from these densities. Given samples from the reweighted marginal model $p(x)\hat{r}(x|\vartheta)$, and from a specific likelihood $p(x|\vartheta)$, the idea is that $\hat{r}(x|\vartheta)$ can only be equivalent to $r(x|\vartheta)$ whenever a classifier tasked to distinguish between $p(x)\hat{r}(x|\vartheta)$ and $p(x|\vartheta)$, cannot extract any predictive features. The discriminative performance of this classifier can be assessed by means of a *Receiver Operating Characteristic* (ROC) curve. A diagonal ROC, which has an *Area Under Curve* (AUC) of 0.5, corresponds to a classifier which is insensitive. In that case, the ratio estimator passes the diagnostic. We emphasize that the ratio estimator can incorrectly pass the diagnostic whenever the classifier is not sufficiently expressive.

3.4.5 *Alternative diagnostics*

Our list of diagnostics is not exhaustive. Some diagnostics are specific to our ratio estimator and can only be computed efficiently because ratio estimates are amortised. In fact, the development of diagnostics for the simulation-based inference literature is an active area of research. For more recent methodologies we refer the reader to Talts et al. (2018) and Dalmasso et al. (2020).

3.5 Overview of the proposed recipe

(i) Simulate a train and test dataset by sampling from the joint $p(\vartheta, x)$. This is done by drawing samples $\vartheta \sim p(\vartheta)$ and conditioning the simulation model on ϑ to generate observables $x \sim p(x|\vartheta)$. These simulations can be parallelised arbitrarily because the samples are drawn independently. The effective number of simulations depends on the problem at hand. In practice additional simulations were

added whenever the ratio estimators did not pass the coverage diagnostic, or, if we found over-fitting to be a significant issue during training.

(ii) Train several discriminators $d(\vartheta, x)$ on the previously simulated dataset. This has several uses. First, the ratio estimators can be ensembled to reduce the variance of the approximation. Secondly, as there is only a single true likelihood-to-evidence ratio $r(x|\vartheta)$, the variability of ratio estimates within the ensemble can be used to quickly assess the convergence. A significant deviation in the ratio estimates is indicative of a ill-tuned optimization procedure.

(iii) Probe the trained ratio estimators for flaws with the diagnostics. Afterwards, apply the diagnostic described in Section 3.4.1 to the observable(s) x_o.

(iv) Compute the posterior $\hat{p}(\vartheta|x_o) = p(\vartheta)\hat{r}(x_o|\vartheta)$ and the desired credible or confidence intervals.

4 EXPERIMENTS AND RESULTS

We demonstrate the usage of our technique on various synthetic realisations of GD-1. Diagnostics are applied to probe the statistical quality of the approximated posteriors under the specified simulation model. By comparing our technique against ABC, we highlight the gain in statistical power our technique can bring to the scientific community. We compute *preliminary* constraints on $m_{\rm WDM}$ based on observations of GD-1 by *Gaia* proper motions (Gaia Collaboration & Brown 2018; Gaia Collaboration & Prusti 2016) and *Pan-STARRS* photometry. It should be noted these constraints only hold under the assumed simulation model. An analysis of (simulation) model misspecification is outside the scope of this work.

4.1 Setup

Simulations We follow the simulation formalism described in Section 2 and the priors defined in Section 3.1. 10 million pairs $(\vartheta, x) \sim p(\vartheta, x)$ are drawn from the simulation model for training, and 100,000 for testing. The simulations in the training dataset are reused in our ABC analyses.

Ratio estimator training All architectures are trained with identical hyperparameter settings. No exhaustive hyperparameter optimization or architecture-search was conducted. Options such as weight-decay and batch-normalization (BN) (Ioffe & Szegedy 2015) were evaluated to reduce over-fitting. All ratio estimators use SELU (Klambauer et al. 2017) activations and were trained using the ADAMW (Loshchilov & Hutter 2017) optimizer for 50 epochs with a batch-size of 4096. We found that larger batch-sizes, for our setting, generalized better. Appendix A investigates the influence of the batch-size on the approximations in detail. We empirically found SELU and ELU activations to be preferable over RELU-like activations, because the approximation of the posterior density function was generally smoother. This work considers 3 main architectures; (i) a simple feedforward MLP, and variants to RESNET (He et al. 2016) such as (ii) RESNET-18 and (iii) RESNET-50. Both use 1 dimensional convolutions without dilations since the usage of dilated convolutions did not yield any significant improvements in terms of test loss. Because our methodology treats ϑ as an input feature, we cannot easily condition the convolutional layers of the RESNET-based architectures on ϑ. This would require conditional convolutions (Yang et al. 2019) or hypernetworks (Ha et al. 2016) to generate specialized kernels for a given ϑ. To retain the simplicity of our architecture, we inject the dependency on ϑ in the fully connected trunk of the convolutional ratio estimators. Other architectural considerations were not explored. Appendix B lists the hyperparameter settings.

Approximate Bayesian Computation Instead of using the stream density power as summary statistics as in Bovy et al. (2017); Banik et al. (2019a), we construct a summary statistics based on the stream density itself. We divide the synthetic observable x (with $n = 66$ bins) by the observable of interest x_o to obtain the bin-wise stellar density ratio $d = x/x_o$. Our summary statistic and distance function are jointly expressed as

$$s(x) = \frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})^2, \tag{17}$$

where $\bar{d}$ is the mean stellar density ratio. Ideally, if both observables match perfectly, then $s(x) = 0$. The acceptance threshold is tuned such that for any given observable of interest x_o, the number of accepted posterior samples is 0.1% of the simulation budget, therefore yielding the smallest threshold with respect to the specified acceptance rate. This corresponds to approximately 10,000 posterior samples. Our goal is to highlight generic aspects of ABC with respect to the proposed method in terms of tuning of the analyses, and its statistical quality *for the given simulation budget*. We emphasize that several scheduling and threshold strategies for ABC exist in the literature, see e.g. (Lintusaari et al. 2017; Prangle 2017). We opted here for a method that is based on the same number of simulations used for training the neural network. The threshold was chosen heuristically to obtain sufficiently smooth posteriors across the entire parameter space, and was not tuned depending on the WDM mass and stream age. This is different from the targeted convergence check and simulation strategy in previous ABC-based streams analyses (Bovy et al. 2017; Banik et al. 2018; Bovy 2019; Banik et al. 2019b). We cannot exclude that the ABC results shown here could further improve by significantly increasing the number of simulations beyond what was needed for the neural network training. This is beyond the scope of the current work.

4.2 Statistical quality

We now assess the statistical properties of the trained ratio estimators. For every architecture, we select the weights which achieved the smallest test-loss.

Proper probability density The computational cost of the integration does not allow us to do an exhaustive analysis. Instead, we apply the diagnostic to 1000 randomly sampled observables. As before, we repeat the experiment 10 times. The following results were obtained: MLP (1.023 ± 0.11), MLP-BN (1.037 ± 0.09), RESNET-18 (1.00 ± 0.02), RESNET-18-BN (0.973 ± 0.03), RESNET-50 (0.993 ± 0.03), and RESNET-50-BN (1.001 ± 0.04) </>. Although the average integrated area under the approximated posterior density functions approaches 1 for all ratio estimator architectures, the results suggest that the approximations of the RESNET-based architectures are more robust. A more careful analysis of the integrated areas, presented in Figure 3, confirms this. Interestingly, the integrated areas for RESNET architectures *with* batch-normalization have a larger spread compared to their counterparts without batch-normalization. Our evaluations on GD-1 will therefore focus on RESNET-based architectures without batch-normalization.

Empirical coverage probability						
Architecture	68% CR	95% CR	99.7% CR	68% CL	95% CL	99.7% CL
$\hat{r}(x\|\vartheta)$ with $\vartheta \triangleq (m_{\rm WDM})$						
MLP	$0.685_{\pm 0.004}$	$0.954_{\pm 0.002}$	$0.997_{\pm 0.001}$	$0.750_{\pm 0.004}$	$0.968_{\pm 0.002}$	$0.999_{\pm 0.000}$
MLP-BN	$0.687_{\pm 0.006}$	$0.951_{\pm 0.002}$	$0.997_{\pm 0.000}$	$0.760_{\pm 0.003}$	$0.970_{\pm 0.002}$	$0.999_{\pm 0.000}$
RESNET-18	$0.667_{\pm 0.004}$	$0.943_{\pm 0.002}$	$0.996_{\pm 0.001}$	$0.721_{\pm 0.005}$	$0.960_{\pm 0.002}$	$0.997_{\pm 0.000}$
RESNET-18-BN	$0.672_{\pm 0.004}$	$0.945_{\pm 0.001}$	$0.996_{\pm 0.001}$	$0.736_{\pm 0.003}$	$0.961_{\pm 0.002}$	$0.998_{\pm 0.000}$
RESNET-50	$0.671_{\pm 0.005}$	$0.947_{\pm 0.003}$	$0.996_{\pm 0.001}$	$0.726_{\pm 0.005}$	$0.963_{\pm 0.000}$	$0.998_{\pm 0.001}$
RESNET-50-BN	$0.678_{\pm 0.004}$	$0.949_{\pm 0.004}$	$0.996_{\pm 0.001}$	$0.743_{\pm 0.002}$	$0.966_{\pm 0.001}$	$0.998_{\pm 0.000}$
$\hat{r}(x\|\vartheta)$ with $\vartheta \triangleq (m_{\rm WDM},\ t_{\rm age})$						
MLP	$0.685_{\pm 0.005}$	$0.953_{\pm 0.002}$	$0.998_{\pm 0.000}$	$0.752_{\pm 0.003}$	$0.968_{\pm 0.001}$	$0.999_{\pm 0.000}$
MLP-BN	$0.685_{\pm 0.004}$	$0.952_{\pm 0.003}$	$0.997_{\pm 0.000}$	$0.758_{\pm 0.003}$	$0.970_{\pm 0.002}$	$0.999_{\pm 0.000}$
RESNET-18	$0.666_{\pm 0.005}$	$0.945_{\pm 0.002}$	$0.995_{\pm 0.001}$	$0.724_{\pm 0.005}$	$0.961_{\pm 0.002}$	$0.998_{\pm 0.000}$
RESNET-18-BN	$0.671_{\pm 0.003}$	$0.945_{\pm 0.003}$	$0.996_{\pm 0.001}$	$0.736_{\pm 0.004}$	$0.961_{\pm 0.002}$	$0.998_{\pm 0.000}$
RESNET-50	$0.674_{\pm 0.006}$	$0.944_{\pm 0.002}$	$0.996_{\pm 0.001}$	$0.740_{\pm 0.004}$	$0.970_{\pm 0.002}$	$0.999_{\pm 0.000}$
RESNET-50-BN	$0.677_{\pm 0.004}$	$0.947_{\pm 0.003}$	$0.997_{\pm 0.000}$	$0.738_{\pm 0.004}$	$0.970_{\pm 0.002}$	$0.999_{\pm 0.000}$

Table 1. Results of the overage diagnostic. Architectures with the BN suffix make use of Batch Normalization. For all ratio estimator architectures, we report Bayesian credible regions and frequentist confidence intervals. Although credible regions do not necessarily have a frequentist interpretation, they are in fact much closer to the nominal coverage probability compared to the confidence intervals. On the contrary, the confidence intervals have coverage, but are slightly conservative. Our analyses will therefore focus on constraints based on confidence intervals. </>

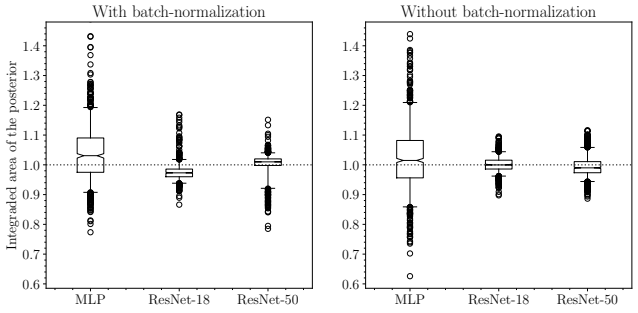

Figure 3. Result of the proper probability density diagnostic. As expected, high-capacity models (RESNET) have tighter approximations compared to the MLP architectures. An interesting discrepancy between the usage of with and without batch normalization is observed. *(Left)* With batch-normalization. *(Right)* Without batch-normalization. </>

Coverage Table 1 summarizes the empirical coverage probability of the ratio estimators. For every ratio estimator, we compute the credible and confidence intervals as described in Section 3.3. For both paradigms, we evaluate the interval construction on 10,000 observables, which is repeated 10 times. The empirical coverage probability of a ratio estimator is therefore based on approximately 100,000 observables in total. We empirically find that MLP-based architectures have coverage under both Bayesian credible and frequentist confidence intervals. This is not the case for RESNET-based architectures. It is noteworthy that the empirical coverage probability of the credible regions are much closer to the nominal coverage probabilities compared to their frequentist counterparts. Additional statistical power could therefore be extracted if the credible regions could be tuned to sufficiently cover the groundtruth at a given nominal coverage probability.

Receiver operating characteristic We now directly probe the correctness of the approximated likelihood-to-evidence ratios. Every ratio estimator is evaluated on 10 uniformly sampled test-hypotheses. 10,000 observables are drawn from every test-hypothesis. For every test-hypothesis, we repeat the computation of the area under curve 10 times to account for the stochastic training of the classifier tasked to distinguish between samples from the reweighted marginal model and samples from the test-hypothesis. Figure 4 summarizes the results. In general, we find that all ratio estimators are unable to perfectly approximate $r(x|\vartheta)$. This result is not unexpected, because the coverage diagnostic indicates that the confidence intervals are conservative, which implies that our estimates of the *true* likelihood-to-evidence ratio must be wrong. Incorrect, but conservative estimates of the posterior are not a significant issue because we mainly seek to constrain $m_{\rm WDM}$.

We additionally find that the quality of the ratio estimates degrades for larger values of $m_{\rm WDM}$ across all architectures. Several strategies could be applied to address this. First, more expressive architectures could be explored which potentially make more efficient use of the available data. Second, by using additional observables could be simulated to aid the approximation of the underlying densities. In our specific case, a straightforward application of this strategy would be to simulate additional observables for $\vartheta \gtrapprox 20$ keV. We would like to emphasize that increasing the size of the training dataset by simulating additional observables at specific target parameters ϑ *should not be done*, because this implicitly changes the prior and therefore the underlying marginal model. Instead, additional observables should only be simulated by sampling from the joint $p(\vartheta, x)$.

4.3 Evaluation

The performance of both methods is assessed on various randomly sampled GD-1 mock simulations with distinct nominal target parameters. A compact overview of the computed posteriors is shown in Figure 5. A full overview can be found in Appendix D. We find the proposed methodology to be preferable over the ABC analysis regarding the reconstruction of the nominal target parameters, and with respect to our stronger, statistically tested, confidence intervals.

In conjunction with the foregoing statistical validation of the ratio estimators, these results highlight the fact that ABC requires a carefully crafted summary statistic; a problem that is absent, or effectively automatised, in the proposed method. As mentioned earlier, an ABC posterior is only exact whenever the summary statistic

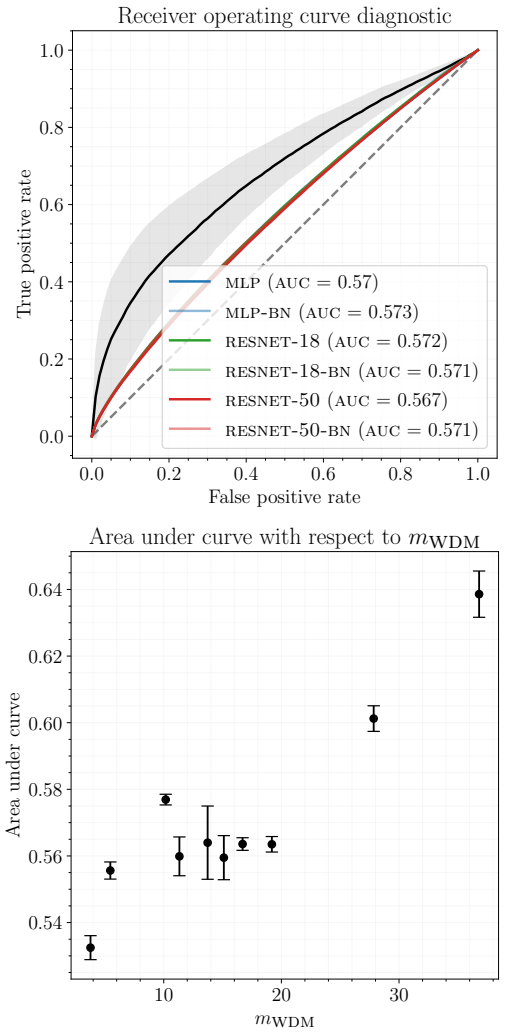

Figure 4. Summary of the receiver operating curve diagnostic. *(Top)* Area Under Curve (AUC) for all test-hypotheses. A baseline measurement, indicated by the black line, does not reweigh the marginal model. Although the ratio estimators perform significantly better compared to the baseline, the diagnostic indicates that all ratio estimators do not perfectly approximate the likelihood-to-evidence ratio (since AUC $\neq$ 0.5). This is not necessarily an issue, because the coverage diagnostic demonstrates that the confidence intervals are conservative. *(Bottom)* Average AUC of the test-hypotheses under consideration. Larger values of $m_{\rm WDM}$ are associated with a degraded quality of the ratio estimates. </>

is sufficient *and* the acceptance threshold tends to 0. If these conditions are not met, the posterior is possibly inaccurate or biased. The necessity of a sufficient summary statistic underlines an important issue with ABC in practice; the *assumed* sufficiency. Determining the statistical validity of an ABC analysis is computationally demanding and often not feasible. Our method does not suffer from this issue, because the estimation of the posterior density is amortised.

4.4 Towards constraining $m_{\rm WDM}$ with GD-1

We now apply our methodology to obtain a *preliminary* constraint on $m_{\rm WDM}$, based on the observed stellar density along the GD-1 stream. The posteriors in this section are computed using the previously trained and statistically validated RESNET-50 ratio estimator.

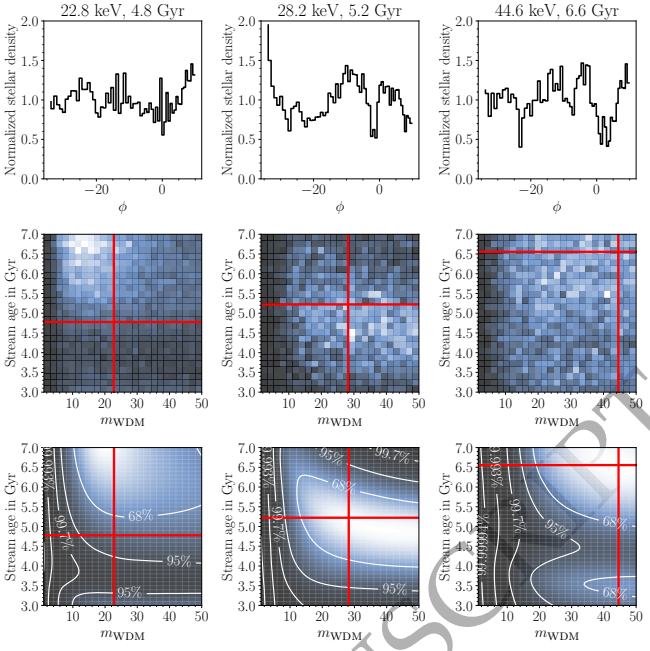

Figure 5. Compact summary of comparisons against ABC. All comparisons are listed in Appendix D. Every column relates to a single mock simulation. The rows show, from top to bottom, the observable, the approximate posterior ABC, and our method respectively. The red cross indicates the groundtruth. ABC and our method are in agreement for most mock simulations. </>

We would like to remind the reader that the coverage diagnostic indicates that the derived confidence intervals are slightly conservative. Our results suggest a strong preference for CDM over WDM. The posteriors and credible intervals at various confidence levels are shown in Figure 6. We find the integrated area under the approximated posterior to be (0.96 ± 0.011 </>). After marginalizing the stream age, the proposed methodology yields $m_{\rm WDM} \geq 17.5$ keV (95% CR) and $m_{\rm WDM} \geq 10.5$ keV (99.7% CR). No significant constraints can be put on the age of GD-1, although an older stream is preferred. A frequentist perspective based on likelihood ratio limits finds $m_{\rm WDM} \geq 13.15$ keV (95% CL) and $m_{\rm WDM} \geq 7.85$ keV (99.7% CL) after marginalizing the stream age. Assuming the posterior approximated by ABC is exact, we find $m_{\rm WDM} \geq 10.8$ keV (95% CL) and $m_{\rm WDM} \geq 3.5$ keV (99.7% CL). We emphasize that *our simulation model does not account for baryonic effects, disturbances caused by massive ($\gtrsim 10^9\ M_\odot$) subhaloes, and effects induced by variations in the Milky Way potential.*

However, our results are promising. We expect that the proposed method will enable an optimal discrimination between dark matter and baryonic effects (provided the latter can be convincingly modeled). It thus constitutes a powerful probe for constraining the mass of thermal or sterile neutrino dark matter (Dodelson & Widrow 1994; Shi & Fuller 1999; Abazajian et al. 2001; Asaka & Shaposhnikov 2005; Boyarsky et al. 2009) (although a discrimination between such WDM models might be challenging).

5 SUMMARY AND DISCUSSION

This work proposes a general recipe for the usage of neural simulation-based inference in the natural sciences. Although the procedure generalizes to many domains, we apply our methodology

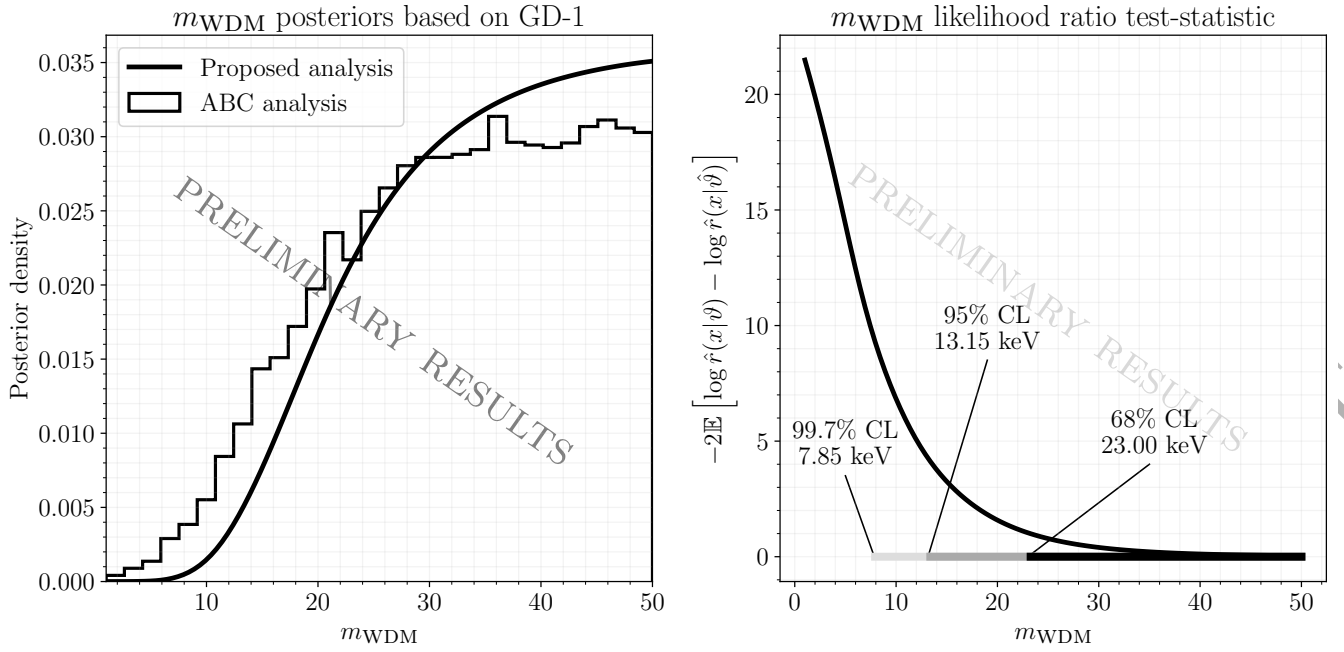

Figure 6. Age-marginalized results based on the observed stellar density variations of GD-1. *The results shown here illustrate the power of the proposed methodology, but should be considered as preliminary, since e.g. baryonic effects are not yet fully included in the simulation model. (Left)* Direct comparison of the reference ABC and the proposed analysis. Both posteriors indicate a preference for CDM over WDM within the assumed simulation model. We find that the proposed method is able to put stronger constraints on m_{WDM}. *(Right)* Likelihood ratio test-statistic used to derive the lower limit confidence intervals. </>

in the stellar stream framework to determine the nature of the dark matter particle. We summarize our findings as follows:

- Bayesian inference based on Amortised Approximate Likelihood Ratios (AALR) is a powerful and convenient analysis framework to study the statistical properties of density variations of stellar streams. In Figure 5 we demonstrate that (at least in the absence of the uncertainties from the baryonic effects), GD-1-like streams could be used to simultaneously constrain the mass of thermal relic dark matter and the age of the stream.
- AALR, in contrast to ABC, does not require handcrafted summary statistics and tuned acceptance thresholds. Our out-of-the-box AALR analysis are expected to be at least as good as any ABC implementation, and to often significantly outperform ABC, as evident in Figure 6.
- The amortised posterior estimation in AALR allows for a variety of diagnostics, including coverage and bias tests, which are computationally demanding and often infeasible for ABC. We explicitly demonstrate that posteriors approximated by AALR are unbiased and that the corresponding confidence intervals have coverage, as show in in see Figure 2 and Table 1 respectively.

Finally, our preliminary results for GD-1 are promising as they indicate that AALR is an excellent and versatile method to probe the nature of dark matter with stellar streams. At face value, we can probe WDM masses up to 17.5 keV (95% credible lower limit for a GD-1-like stream). *We emphasize however that our simulation codes do not account for baryonic effects, which are expected to significantly impact the results.* In upcoming analyses we plan to use the improved statistical power achieved through AALR to obtain more statistically robust and tighter constraints on the particle mass of dark matter. However, we do expect some loss in sensitivity when including baryonic effects, because we expect the task of discriminating between CDM and WDM impacted streams to be harder for AALR.

ACKNOWLEDGEMENTS

All authors would like to specifically thank Jo Bovy for providing us with the the base of the simulation model, and the development of `galpy`. Both were critical to this work. Joeri Hermans would like to thank the National Fund for Scientific Research for his FRIA scholarship. Gilles Louppe is recipient of the ULiège - NRB Chair on Big data and is thankful for the support of NRB. Finally, all authors would like to thank the developers of the *Jupyter* (Kluyver et al. 2016) project, *Matplotlib* (Hunter 2007), *Numpy* (Van Der Walt et al. 2011), *PyTorch* (Paszke et al. 2019), and *Scikit-Learn* (Pedregosa et al. 2011) for enabling this work.

DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

Abazajian K., Fuller G. M., Patel M., 2001, Phys. Rev. D, D64, 023501
Amorisco N. C., Gòmez F. A., Vegetti S., White S. D. M., 2016, MNRAS, 463, L17
Asaka T., Shaposhnikov M., 2005, Phys. Rev. B, 620, 17
Avila-Reese V., Firmani C., Hernández X., 1998, ApJ, 505, 37
Banik N., Bovy J., 2019, MNRAS, 484, 2009
Banik N., Bertone G., Bovy J., Bozorgnia N., 2018, J. Cosmology Astropart. Phys., 7, 061

Banik N., Bovy J., Bertone G., Erkal D., de Boer T. J. L., 2019a, arXiv e-prints, p. arXiv:1911.02662

Banik N., Bovy J., Bertone G., Erkal D., de Boer T. J. L., 2019b, arXiv e-prints, p. arXiv:1911.02663

Berger J. O., Liseo B., Wolpert R. L., et al., 1999, Statistical science, 14, 1

Bertschinger E., 2006, Phys. Rev. D, 74, 063509

Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517

Bode P., Ostriker J. P., Turok N., 2001, ApJ, 556, 93

Bond J. R., Szalay A. S., 1983, ApJ, 274, 443

Bovy J., 2015, The Astrophysical Journal Supplement Series, 216, 29

Bovy J., 2019, in Essig R., Feng J., Zurek K., eds, , Vol. 56, Illuminating Dark Matter. pp 9–18

Bovy J., Erkal D., Sanders J. L., 2017, MNRAS, 466, 628

Boyarsky A., Ruchayskiy O., Shaposhnikov M., 2009, Annual Review of Nuclear and Particle Science, 59, 191

Brehmer J., Cranmer K., Louppe G., Pavez J., 2018, Phys. Rev. Lett., 121, 111801

Brehmer J., Mishra-Sharma S., Hermans J., Louppe G., Cranmer K., 2019, ApJ, 886, 49

Brehmer J., Louppe G., Pavez J., Cranmer K., 2020, Proceedings of the National Academy of Sciences, 117, 5242

Carlberg R. G., 2012, ApJ, 748, 20

Cowan G., Cranmer K., Gross E., Vitells O., 2011, The European Physical Journal C, 71

Cranmer K., Pavez J., Louppe G., 2015, arXiv preprint arXiv:1506.02169

Cranmer K., Brehmer J., Louppe G., 2020, Proceedings of the National Academy of Sciences

D'Onghia E., Springel V., Hernquist L., Keres D., 2010, ApJ, 709, 1138

Dalmasso N., Lee A., Izbicki R., Pospisil T., Kim I., Lin C.-A., 2020, in International Conference on Artificial Intelligence and Statistics. pp 3349–3361 (arXiv:1905.11505)

Dodelson S., Widrow L. M., 1994, Phys. Rev. Lett., 72, 17

Erkal D., Belokurov V., 2015a, MNRAS, 450, 1136

Erkal D., Belokurov V., 2015b, MNRAS, 454, 3542

Erkal D., Koposov S. E., Belokurov V., 2017, MNRAS, 470, 60

Gaia Collaboration Brown 2018, Astronomy and Astrophysics, 616, A1

Gaia Collaboration Prusti 2016, A&A, 595, A1

Garrison-Kimmel S., et al., 2017, MNRAS, 471, 1709

Gilman D., Birrer S., Nierenberg A., Treu T., Du X., Benson A., 2020, MNRAS, 491, 6077

Grillmair C. J., Dionatos O., 2006, ApJ, 643, L17

Ha D., Dai A., Le Q. V., 2016, arXiv preprint arXiv:1609.09106

He K., Zhang X., Ren S., Sun J., 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778

Hermans J., Begy V., Louppe G., 2019, arXiv e-prints, p. arXiv:1903.04057

Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R., 2012, arXiv preprint arXiv:1207.0580

Hoffer E., Hubara I., Soudry D., 2017, in Advances in Neural Information Processing Systems. pp 1731–1741

Hofmann S., Schwarz D. J., Stoecker H., 2001, Phys. Rev. D, 64, 083507

Hunter J. D., 2007, Computing in Science & Engineering, 9, 90

Ibata R., Lewis G., Irwin M., Quinn T., 2002, MNRAS, 332, 915

Ioffe S., Szegedy C., 2015, arXiv preprint arXiv:1502.03167

Johnston K. V., Spergel D. N., Haydn C., 2002, ApJ, 570, 656

Kelley T., Bullock J. S., Garrison-Kimmel S., Boylan-Kolchin M., Pawlowski M. S., Graus A. S., 2019, MNRAS, 487, 4409

Keskar N. S., Mudigere D., Nocedal J., Smelyanskiy M., Tang P. T. P., 2017, 5th International Conference on Learning Representations

Klambauer G., Unterthiner T., Mayr A., Hochreiter S., 2017, in Advances in neural information processing systems. pp 971–980

Kluyver T., et al., 2016, in Loizides F., Schmidt B., eds, Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp 87 – 90

Lintusaari J., Gutmann M. U., Dutta R., Kaski S., Corander J., 2017, Syst. Biol., 66, e66

Loshchilov I., Hutter F., 2017, arXiv preprint arXiv:1711.05101

Lovell M. R., Frenk C. S., Eke V. R., Jenkins A., Gao L., Theuns T., 2014, MNRAS, 439, 300

Masters D., Luschi C., 2018, arXiv preprint arXiv:1804.07612

Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, ApJ, 524, 9

Neyman J., 1937, Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236, 333

Neyman J., Scott E. L., 1948, Econometrica, 16, 1

Paszke A., et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, , Vol. 394, Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp 8024–8035, doi:10.1016/s0140-6736(19)32614-5

Pearson S., Price-Whelan A. M., Johnston K. V., 2017, Nature Astronomy, 1, 633

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Peebles P. J. E., 1982, ApJ, 263, L1

Prangle D., 2017, Bayesian Anal., 12, 289

Prangle D., Blum M. G. B., Popovic G., Sisson S. A., 2013, arXiv e-prints, p. arXiv:1301.3166

Rubin D. B., 1984, The Annals of Statistics, pp 1151–1172

Sawala T., Pihajoki P., Johansson P. H., Frenk C. S., Navarro J. F., Oman K. A., White S. D. M., 2017, MNRAS, 467, 4383

Schall R., 2012, Biometrical journal, 54, 537

Schneider A., Smith R. E., Reed D., 2013, MNRAS, 433, 1573

Shi X., Fuller G. M., 1999, Phys. Rev. Lett., 82, 2832

Smith R. E., Markovic K., 2011, Phys. Rev. D, 84, 063507

Smith S. L., Kindermans P.-J., Ying C., Le Q. V., 2017, arXiv preprint arXiv:1711.00489

Springel V., et al., 2008, MNRAS, 391, 1685

Strege C., Trotta R., Bertone G., Peter A. H., Scott P., 2012, Phys. Rev. D, 86, 023507

Talts S., Betancourt M., Simpson D., Vehtari A., Gelman A., 2018, arXiv preprint arXiv:1804.06788

Van Der Walt S., Colbert S. C., Varoquaux G., 2011, Computing in Science & Engineering, 13, 22

Webb J. J., Bovy J., 2019, MNRAS, 485, 5929–5938

Webb J. J., Bovy J., 2020, arXiv e-prints, p. arXiv:2006.06695

Wilks S. S., 1938, The annals of mathematical statistics, 9, 60

Yang B., Bender G., Le Q. V., Ngiam J., 2019, in Advances in Neural Information Processing Systems. pp 1307–1318

Yoon J. H., Johnston K. V., Hogg D. W., 2011, ApJ, 731, 58

Zhao D., Mo H., Jing Y., Boerner G., 2003, MNRAS, 339, 12

de Boer T. J. L., Erkal D., Gieles M., 2020, MNRAS, 494, 5315

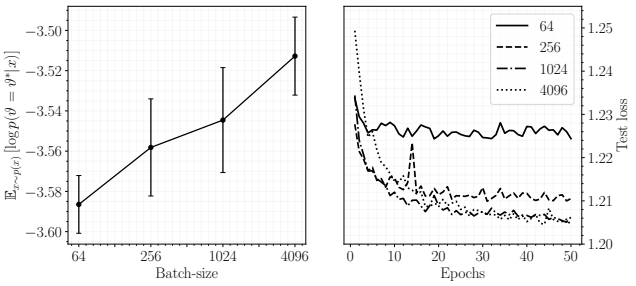

Figure A1. *(Left)* The expected log posterior probability of the nominal ϑ^* for $\vartheta^*, x \sim p(\vartheta, x)$. Under the assumption that $p(\vartheta)\hat{r}(x|\vartheta)$ is a proper probability density, this quantity captures the ability of the approximated posteriors to reconstruct the nominal value. *(Right)* Average test-loss curve for various batch-sizes. For a given learning rate, there is a clear inverse relation between the test-loss and expected log posterior probability of the nominal value. Larger batch-sizes should therefore be preferred. </>

APPENDIX A: INFLUENCE OF THE BATCH-SIZE DURING TRAINING ON THE APPROXIMATED POSTERIORS

To investigate the effect of the batch-size on the approximated posteriors, we train several ratio estimators based on the MLP-BN (batch normalization) architecture with batch-sizes 64, 256, 1024 and 4096. At 95% CR, the empirical coverage probabilities of these ratio estimators are 0.961 ± 0.004, 0.954 ± 0.004, 0.952 ± 0.008 and 0.952 ± 0.006 respectively </>. Figure A1 shows the test-loss curves and $\mathbb{E}_{\vartheta,x\sim p(\vartheta,x)}[\log \hat{p}(\vartheta = \vartheta|x)]$ for every batch-size setting. Under the assumption that $p(\vartheta)\hat{r}(x|\vartheta)$ is a proper probability density, Equation A captures the ability of $\hat{r}(x|\vartheta)$ to reconstruct the groundtruth. As indicated by Figure A1, there is a clear negative correlation between the test-loss and the expected log posterior probability of the nominal value. Although not entirely unexpected, this suggests that larger batch-sizes have the potential to further reduce the test-loss at a given learning rate. Practitioners should therefore analyze the behaviour of their optimization procedure with respect to the batch-size as well.

The observations made here are in line with the machine learning literature (Hoffer et al. 2017), although others (Keskar et al. 2017; Masters & Luschi 2018) have suggested that smaller batch-sizes lead to models which generalize to a greater degree. This especially seems to be the case whenever the testing loss surface differs from the training loss surface (Keskar et al. 2017). Unlike typical deep learning applications with a fixed dataset, this issue can easily be addressed within the likelihood-free setting, because the similarity of these loss surfaces can be ensured by continuously drawing new samples from the simulation model. For a given learning rate, larger batch-sizes should therefore be preferred (Smith et al. 2017). Alternatively, this could also be explained due to the fact that larger batch-sizes provide more empirical evidence (less stochasticity) to approximate the ratio.

APPENDIX B: HYPERPARAMETERS

The same hyperparameters are used across all architectures. We did not explore specific settings for every individual architecture, demonstrating the robustness of our technique. A learning rate of 0.0001 with a batch-size of 4096 and a weight-decay factor of 0.1 was used during training. The ratio estimators do not use dropout (Hinton et al. 2012). The remaining hyperparameters (e.g., of Batch Normalization) were set to the *PyTorch* defaults. </>.

Empirical coverage probability			
Architecture	68% CR	95% CR	99.7% CR
$\hat{r}(x\|\vartheta)$ with $\vartheta \triangleq (m_{\rm WDM})$			
MLP	0.704 ± 0.004	0.972 ± 0.002	0.999 ± 0.000
MLP-BN	0.706 ± 0.003	0.970 ± 0.001	0.999 ± 0.000
RESNET-18	0.687 ± 0.004	0.955 ± 0.002	0.998 ± 0.000
RESNET-18-BN	0.693 ± 0.004	0.966 ± 0.002	0.999 ± 0.000
RESNET-50	0.689 ± 0.006	0.967 ± 0.001	0.998 ± 0.000
RESNET-50-BN	0.698 ± 0.004	0.969 ± 0.001	0.999 ± 0.000
$\hat{r}(x\|\vartheta)$ with $\vartheta \triangleq (m_{\rm WDM},\ t_{\rm age})$			
MLP	0.704 ± 0.004	0.973 ± 0.001	0.999 ± 0.000
MLP-BN	0.709 ± 0.004	0.970 ± 0.001	0.999 ± 0.000
RESNET-18	0.688 ± 0.005	0.965 ± 0.002	0.998 ± 0.000
RESNET-18-BN	0.692 ± 0.006	0.967 ± 0.002	0.999 ± 0.000
RESNET-50	0.694 ± 0.005	0.968 ± 0.002	0.999 ± 0.000
RESNET-50-BN	0.695 ± 0.006	0.968 ± 0.001	0.999 ± 0.000

Table C1. Summary of the coverage diagnostic with an artificially lowered highest density level. In doing so, we make the credible intervals more conservative such that the procedure has coverage at the specified confidence levels. </>

APPENDIX C: NEYMAN CONSTRUCTION WITH RATIO ESTIMATORS AND BAYESIAN CREDIBLE REGIONS

As indicated by Table 1, the method responsible for computing the Bayesian credible regions closely approximates the nominal coverage probability, even though credible regions do not necessarily have a frequentist interpretation. For most ratio estimators however, the method does not sufficiently cover the groundtruth. The credible regions in question are derived from the intersection between the posterior density and the highest density such that the area below the intersection is approximately $1 - \alpha$. Credible regions can therefore be made more conservative by artificially lowering the highest density level until they have coverage at some given confidence level. We achieve this by introducing a bias term α_b such that the integrated area under the credible region Θ is $1 - \alpha - \alpha_b$. Using the same ratio estimators as in Table 1, we repeat the experiment with α_b 0.002, 0.02 and 0.02 across all architectures for 68% CR, 95% CR and 97.7% CR respectively </>. The results are shown in Table C1. As expected, the credible regions with the additional bias term have coverage.

APPENDIX D: COMPARISONS AGAINST APPROXIMATE BAYESIAN COMPUTATION

See next page.

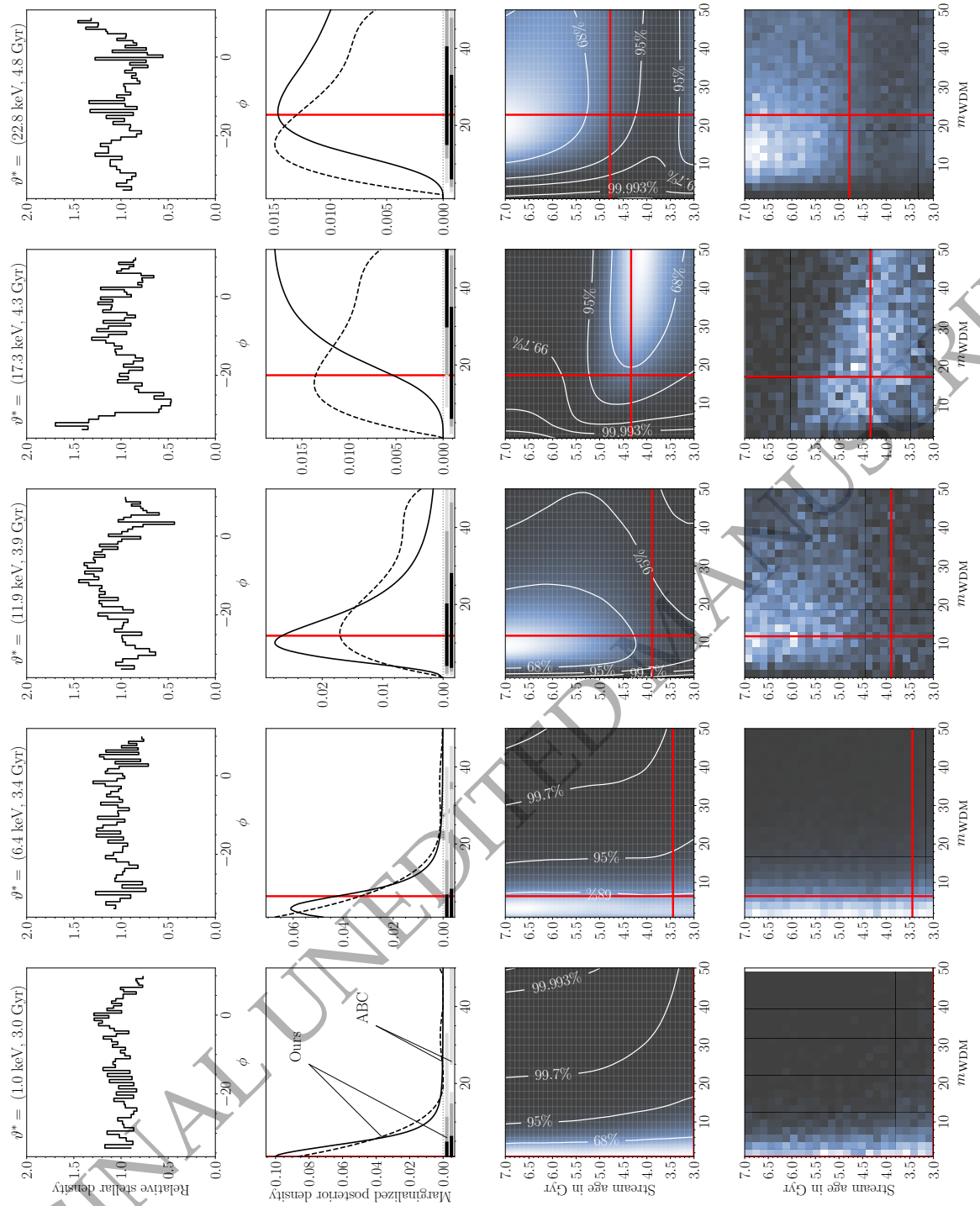

Figure D1. Direct comparison of ABC against the proposed method. The top row shows the observable. The second row the marginalized posteriors for both methods. Finally, row 3 and 4 show the joint posterior for our method and ABC respectively. The nominal target parameter is indicated by the red line. It is visually apparent that the proposed methodology produces stronger constraints of the groundtruth compared to ABC. </>

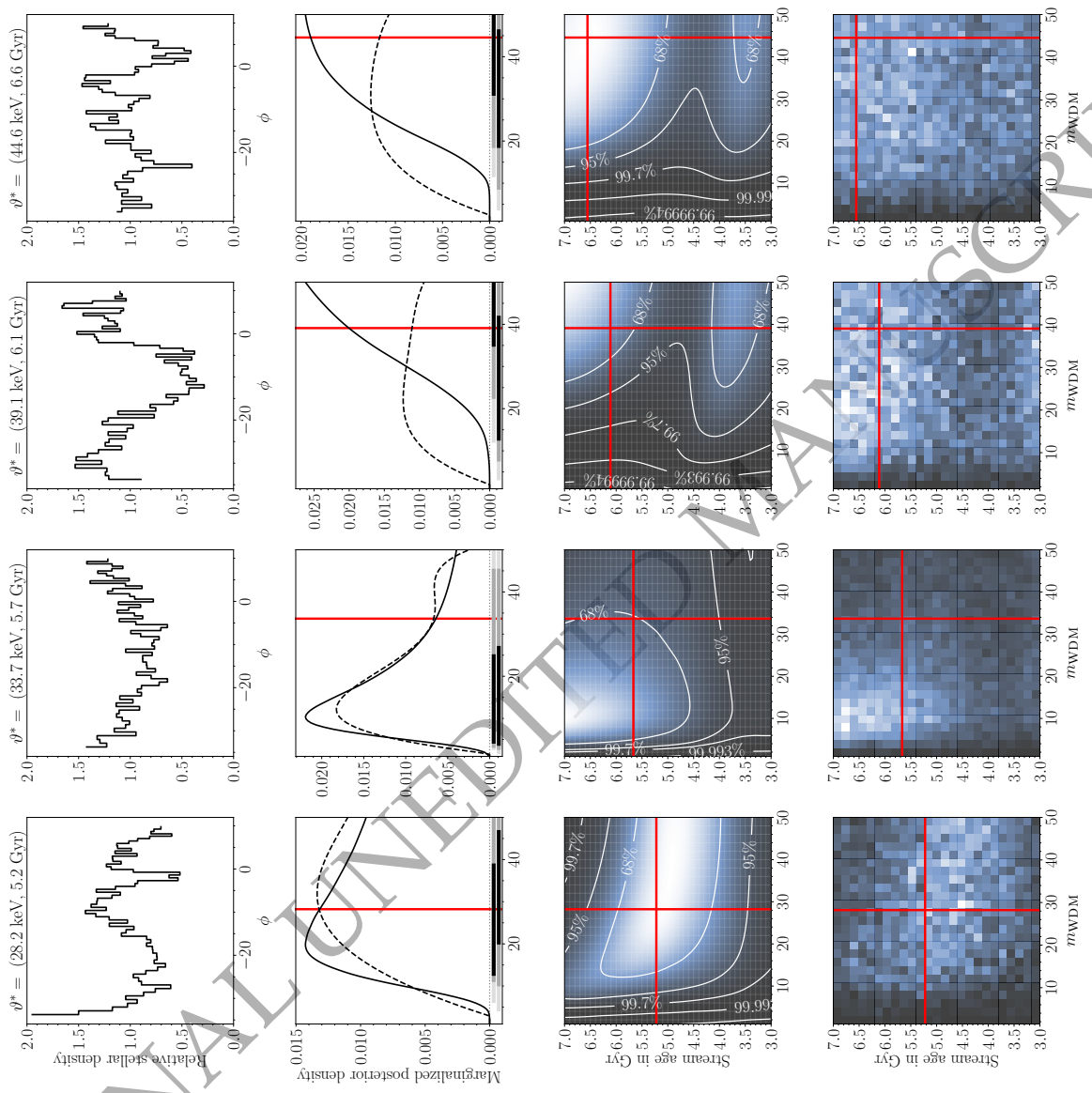

Figure D2. Direct comparison of ABC against the proposed method. Refer to Figure D1 for the initial results. </>