Sum of the masses of the Milky Way and M31: A likelihood-free inference approach

Pablo Lemos,[1,*] Niall Jeffrey,[2,1] Lorne Whiteway,[1] Ofer Lahav,[1] Noam Libeskind I,[3,4] and Yehuda Hoffman[5]

[1]*Department of Physics and Astronomy, University College London,*
Gower Street, London WC1E 6BT, United Kingdom
[2]*Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL,*
CNRS, Sorbonne Université, Université de Paris, Paris, France
[3]*Leibniz-Institut fr Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany*
[4]*University of Lyon, UCB Lyon-1/CNRS/IN2P3, IPN Lyon, France*
[5]*Racah Institute of Physics, Hebrew University, Jerusalem, 91904 Israel*

(Received 29 October 2020; accepted 18 December 2020; published 11 January 2021)

We use density estimation likelihood-free inference, Λ cold dark matter simulations of $\sim 2M$ galaxy pairs, and data from *Gaia* and the Hubble Space Telescope to infer the sum of the masses of the Milky Way and Andromeda (M31) galaxies, the two main components of the local group. This method overcomes most of the approximations of the traditional timing argument, makes the writing of a theoretical likelihood unnecessary, and allows the nonlinear modeling of observational errors that take into account correlations in the data and non-Gaussian distributions. We obtain an M_{200} mass estimate $M_{\mathrm{MW+M31}} = 4.6^{+2.3}_{-1.8} \times 10^{12}\ M_{\odot}$ (68% C.L.), in agreement with previous estimates both for the sum of the two masses and for the individual masses. This result is not only one of the most reliable estimates of the sum of the two masses to date, but is also an illustration of likelihood-free inference in a problem with only one parameter and only three data points.

DOI: 10.1103/PhysRevD.103.023009

I. INTRODUCTION

Likelihood-free inference (LFI) has emerged as a very promising technique for inferring parameters from data, particularly in cosmology. It provides parameter posterior probability estimation without requiring the calculation of an analytic likelihood (i.e., the probability of the data being observed given the parameters). LFI uses forward simulations in place of an analytic likelihood function. Writing a likelihood for cosmological observables can be extremely complex, often requiring the solution of Boltzmann equations, as well as approximations for highly nonlinear processes such as structure formation and baryonic feedback. While simulations have their own limitations and are computationally expensive, the quality and efficiency of cosmological simulations are constantly increasing, and they are likely to soon far surpass the accuracy or robustness of any likelihood function.

This is a rapidly growing topic in cosmology, due to the emergence of novel methods for likelihood-free inference [see e.g., [1,2]], with applications to data sets such as the joint light curve (JLA) and Pantheon supernova datasets [3,4], and the Dark Energy Survey science verification data [5], among others [6–8]. There are, therefore, many applications for which LFI could improve the robustness of parameter inference using cosmological data. In this work we perform a LFI-based parameter estimation of the sum of masses of the Milky Way and M31. The likelihood function for this problem requires significant simplifications, but forward simulations can be obtained easily.

The Milky Way and Andromeda are the main components of the Local Group, which includes tens of smaller galaxies. We define $M_{\mathrm{MW+M31}}$ as the sum of the MW and M31 masses. Estimating $M_{\mathrm{MW+M31}}$ remains an elusive and complex problem in astrophysics. As the mass of each of the Milky Way and M31 is known only to within a factor of 2, it is important to constrain the sum of their masses. The traditional approach is to use the so-called timing argument (TA) [9]. The timing argument estimates $M_{\mathrm{MW+M31}}$ using Newtonian dynamics integrated from the big bang. This integration is an extremely simplified version of a very complex problem. Therefore, alternative methods that do not rely on the same approximations become extremely useful.

In this work, we use the *multidark Planck* (MDPL) simulation [10–12], combined with data from the Hubble Space Telescope [HST, [13] and *Gaia* [14], to estimate $M_{\mathrm{MW+M31}}$. A similar data set was previously used in [15] to obtain a point estimate of $M_{\mathrm{MW+M31}}$ using Artificial Neural Networks (ANN) in conjunction with the TA. In contrast, our work uses density estimation likelihood-free inference

[*] pablo.lemos.18@ucl.ac.uk

TABLE I. Estimates of $M_{\mathrm{MW+M31}}$ from previous work. The third column shows the data used, with r in Mpc and v_r, v_t in km s^{-1}. The fourth column shows $M_{\mathrm{MW+M31}}$ in units of $10^{12}\ M_{\odot}$. Note that Gaussian approximations have been used to convert the reported confidence levels to 68% confidence levels in some cases.

Reference	Method	Assumed (r, v_r, v_t)	$M_{\mathrm{MW+M31}}$
Li & White (2008) [27]	TA calibrated on Sims	(0.784, 130, 0)	$5.27^{+2.48}_{-0.91}$
Gonzalez & Kravtsov (2014) [28]	Sims	(0.783, 109.3, 0)	$4.2^{+2.1}_{-1.2}$
McLeod *et al.* (2017) [15]	TA + Λ	(0.77 ± 0.04, 109.4 ± 4.4, 17 ± 17)	$4.7^{+0.7+2.9}_{-0.6-1.8}$
McLeod *et al.* (2017) [15]	Sims + ANN + Shear	(0.77 ± 0.04, 109.4 ± 4.4, 17 ± 17)	$4.9^{+0.8+1.7}_{-0.8-1.3}$
Phelps *et al.* (2013) [26]	Least Action	(0.79, 119, 0)	6.0 ± 0.5

[DELFI [2,16–18]], using the PYDELFI package [19], combined with more recent data. While the result is important on its own, this paper also illustrates the fundamental methodology of DELFI in a problem that is statistically simple but physically complex.

The structure of the paper is as follows: Sec. II reviews and describes previous estimates of $M_{\mathrm{MW+M31}}$. Section III describes the basics of LFI, and the particular techniques used in this work. Section IV and Sec. V describe the simulations and data, respectively, used in this work. Section VI shows our results, and conclusions are presented in Sec. VII.

II. PREVIOUS ESTIMATES

A first approach to estimating $M_{\mathrm{MW+M31}}$ from dynamics, known as TA, is based on the simple idea that MW and M31 are point masses approaching each other on a radial orbit that obeys

$$\ddot{r} = -\frac{GM_{\mathrm{MW+M31}}}{r^2} + \frac{1}{3}\Lambda r, \tag{1}$$

where $M_{\mathrm{MW+M31}}$ is the sum of the masses of the two galaxies [9,20], and where the Λ term, which represents a form of dark energy, was added in later studies [21,22]. It was also extended for modified gravity models [23]. Since we know the present-day distance r between MW and M31 and their relative radial velocity v_r, and if we assume the age of the universe and Λ, we can infer the mass $M_{\mathrm{MW+M31}}$. The analysis can be extended to cover the case of nonzero tangential v_t velocity [e.g., [15,24]]. The pros and cons of the timing argument are well known. It is a simple model which assumes only two point mass bodies; this ignores, for example, the tidal forces due to neighboring galaxy haloes in the local group and the extended cosmic web around it. While the timing argument model does not capture the complexity of the cosmic structure and resulting cosmic variance, it gives a somewhat surprisingly good estimate for $M_{\mathrm{MW+M31}}$. As shown below, it can also serve to test the sensitivity of the results to parameters such as the cosmological constant Λ and the Hubble constant H_0, in case simulations are not available for different values of these parameters.

A second approach is to consider the dynamics of all the galaxies in the local group using the least action principle [25], as, for example, was implemented in [26]. In this approach all members of the local group appear in the model, but as a result the derived masses are correlated, and the error bars should be interpreted accordingly. A third approach is to use N-body simulations of the local universe, assuming a cosmological model such as ΛCDM [15,27,28]. In this paper we apply this third method, but using the DELFI method (this provides significant improvements over other methods, as discussed in the following section).

Representative results for $M_{\mathrm{MW+M31}}$ from previous works are given in Table I. Throughout the paper we quote 68% credible intervals.

III. LIKELIHOOD-FREE INFERENCE

In Bayesian statistics we often face the following problem: given observed data D_{obs}, and a theoretical model I with a set of parameters θ, calculate the probability of the parameters given the data. In other words, we want to calculate the *posterior distribution* $\mathcal{P} \equiv p(\theta|D_{\mathrm{obs}}, I)$; here p is a probability (for a model with discrete parameters) or a probability density (for continuous parameters). We do so using Bayes' theorem:

$$p(\theta|D_{\mathrm{obs}}, I) = \frac{p(D_{\mathrm{obs}}|\theta, I)p(\theta|I)}{p(D_{\mathrm{obs}}|I)} \Leftrightarrow \mathcal{P} = \frac{\mathcal{L} \times \Pi}{\mathcal{Z}} \tag{2}$$

where $\mathcal{L}$ is called the likelihood, Π the prior, and $\mathcal{Z}$ the Bayesian evidence. The Bayesian evidence acts as an overall normalization in parameter estimation, and can therefore be ignored for this task. Thus, given a choice of prior distribution and a likelihood function, we can estimate the posterior distribution. However, obtaining a likelihood function is not always easy. The likelihood function provides a probability of measuring the data as a function of the parameter values, and often requires approximations both in the statistics and in the theoretical modeling.

Likelihood-free inference is an alternative method for calculating the posterior distribution; in this method we do not formally write down a likelihood function. Instead, we use *forward simulations* of the model to generate samples of the data and parameters. In the simplest version of LFI, we select only the forward simulations that are the most similar to the observed data, rejecting the rest. This method is known as approximate Bayesian computation [ABC, [29]]; it relies on choices of a distance metric (to measure similarity between simulated and observed data) and of a maximum distance parameter ϵ (used to accept or rejected simulations).

In this work, we will use a version of LFI called density estimation likelihood-free inference (DELFI). In this approach, we use all existing forward simulations to learn a conditional density distribution of the data[1] d given the parameters θ, using a density estimation algorithm. Examples of density estimation algorithms are kernel density estimation [KDE, [30–32]], mixture models, mixture density networks [33,34], and masked autoregressive flows [35]. We use the package PYDELFI, and estimate the likelihood function from the forward simulations using Gaussian mixture density networks (GMDN) and masked autoregressive flows (MAF). In this sense, the name likelihood-free inference is perhaps misleading: the inference is not likelihood-free, we simply avoid writing a likelihood and instead model it using forward simulations. A more accurate name for the method could therefore be explicit-likelihood-free. A more extended discussion of our choice of density estimation algorithms and conditional distribution is presented in the Appendix A.

DELFI has several advantages over the simpler ABC approach to LFI: it does not rely on a choice of a distance parameter ϵ (although admittedly the choice of basis in parameter space can change the implicit distance metric of the density estimator) and it uses all available forward simulations to build the conditional distribution, making it far more efficient.

While relatively new, likelihood-free inference has already been applied to several problems in astrophysics [e.g., [5,36–41]]. However, most applications involving LFI suffer from the curse of dimensionality: there can be hundreds, thousands, or even millions of observables (such as ~2000 multipoles in cosmic microwave background surveys, or 500 redshift and angular bins in cosmic shear analyses), and it is impossible to perform density estimation. Some form of data compression is therefore usually needed [42–44]. Similarly, due to the high dimensionality and complexity of these parameter spaces, efficient methods to generate the simulations (so as to minimize the number needed) have been developed [2,45,46]. However, our $M_{\mathrm{MW+M31}}$ problem has only three data points and one parameter of interest, making it an extremely simple application of the method from the statistical point of view; it illustrates all necessary techniques, and does not require data compression.

To summarize, the steps that we will follow are

(i) Generate a large number of simulations of systems similar to the one of interest. The simulations used in this work are described in Sec. IV.

(ii) Use a density estimator, in our case GMDN and MAF as part of the PYDELFI package, to obtain the sampling distribution for any data realization $p(d|\theta, I)$.

(iii) Evaluate this distribution at the observed data (which will be described in Sec. V) to obtain the likelihood function for our observed data realization: $p(d = D_{\mathrm{obs}}|\theta, I)$.

(iv) From a prior distribution and the likelihood, and using Bayes' theorem [Eq. (2)], get a posterior distribution $p(\theta|D_{\mathrm{obs}}, I)$.

IV. SIMULATIONS

We use the publicly available MDPL simulation. This is an N-body simulation of a periodic box with side length $L_{\mathrm{box}} = 1\ \mathrm{Gpc}\, h^{-1}$, populated with 2048^3 dark matter particles. Such a simulation achieves a mass resolution of $8.7 \times 10^9\ M_\odot\, h^{-1}$ and a Plummer equivalent gravitational softening of 7 kpc. The simulation was run with the ART code and assumed a ΛCDM power spectrum of fluctuations with $\Omega_\Lambda = 0.73$, $\Omega_b = 0.047$, $\Omega_m = 0.27$, $\sigma_8 = 8.2$ and $H_0 = 100\ h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, with $h = 0.7$.

Haloes are identified with the AHF halo finder [47]. AHF identifies local overdensities in the density field as possible halo centres. The local potential minimum is computed for each density peak and the particles that are gravitationally bound are identified. Haloes with more than 20 particles are recorded in the halo catalogue. The AHF catalogue includes some 11,960,882 dark matter haloes.

The halo catalogue is then searched for pairs of galaxies as follows. We first identify those haloes in the mass[2] range $5 \times 10^{10} < M/M_\odot h^{-1} < 5 \times 10^{13}$. For each such halo, we find all haloes (irrespective of mass) within ~4 Mpc and sort these by distance. If the closest of these is within the mass range $[5 \times 10^{10}, 5 \times 10^{13}]$ and is separated by more than 500 kpc but less than 1500 kpc, the two haloes are considered a potential pair. The partner is then examined to ensure that the pair is isolated (namely that there is no other halo with a mass $> 5 \times 10^{10}$ and closer to either pair member than the pair separation). If this is the case then the pair is kept for the analysis described here. In this way, 1,094,839 pairs are found (as opposed to the 30,190 pairs used in [15]).

We believe the criteria used to select the halo pairs in this work are not very restrictive and any cuts lie well outside

[1] Note that we use the letter d to refer to the data space used in DELFI to learn conditional distributions, and we use D_{obs} to refer to the observed data.

[2] Henceforth, all masses are defined as M_{200}, the total mass enclosed in the largest sphere surrounding them with an enclosed mean density over 200 times the critical value [27].

the realistic limits for the actual MW-M31 system. Though these selections are unrestrictive, we note that including more pairs does improve the density estimation, as DELFI can use all the available simulated data, and not only those that are close to the observation.

V. DATA

We use three observations to constrain $M_{\mathrm{MW+M31}}$: the distance to M31, and the radial and tangential components of its velocity. While the TA requires other observables, such as the age of the Universe t_0 and the cosmological constant Λ, in our approach these are already included in the simulations at fixed values. We discuss below how uncertainties in cosmological parameters affect the estimated mass.

For the distance to M31, we adopt the commonly used value $r = 770 \pm 40$ kpc [48–51]. For the velocity, we follow the results of [[52] henceforth VdM19]. The radial velocity in the galactocentric rest frame is $v_r = -109.4 \pm 4.4\ \mathrm{km\,s^{-1}}$ [53] from HST observations. The tangential velocity is slightly more cumbersome. VdM19 report the following value for components of the tangential velocity of M31 from a combination of *Gaia* DR2 and HST:

$$\mu = (10 \pm 11, -16 \pm 11)\ \mu\mathrm{as\,yr^{-1}}, \tag{3}$$

already in the galactocentric frame, i.e., after correcting for the solar reflex motion. VdM19 uses the distance to M31 to convert this to $\mathrm{kms^{-1}}$. They then use a method described in [54] to correct for the fact that taking the norm of the two components leads to a "bias" in the reported tangential velocity.[3] However, in this work, we take a different approach, illustrated in Fig. 1 for each simulation, we scatter the value of each component of the tangential velocity for the simulation according the observational error of each components shown in Eq. (3). By doing this, we are putting the observational errors in the simulated measurements, instead attempting to "debias" the tangential velocity summary statistic. We convert to $\mathrm{km\,s^{-1}}$ using the value of the distance for that sample, to account for the covariance between r and v_t. We take as the observed value the norm of the two observed components, $v_t = 72\ \mathrm{km\,s^{-1}}$. While this differs from the value reported in VdM19, this should not be a problem, as long as the way in which we calculate the tangential velocity in simulations and observations is consistent, and we use a summary statistic that extracts all the available information (which the norm of the components does). Furthermore, Appendix C shows that our results do not significantly change if we repeat the analysis using a purely radial motion ($v_t = 0$).

This approach takes into account both the non-Gaussian errors in v_t and the correlation of v_t with the errors in the distance measurement (these have not been accounted for in previous estimates of $M_{\mathrm{MW+M31}}$).

VI. RESULTS

A. Overview

Having discussed the method, the simulations, and the data, we have everything we need to perform LFI using DELFI. We have three data points $d = \{r, v_r, v_t\}$ and one parameter $\theta = M$. The process is illustrated in Figs. 2 and 3, and consists of the following:

(1) Left panel of Fig. 2: We generate a large number of forward simulations, as discussed in Sec. IV. Increasing the number of simulations will increase the accuracy of the density estimation, and of the resulting posterior. Reference [2] demonstrates an *active learning* scheme with PYDELFI, providing criteria to run new simulations based on discrepancies between the density estimates in the neural density estimator ensemble. However, due to our initial large number of simulations, we had no need to run such extra simulations on-the-fly.

(2) Right panel of Fig. 2: The observational errors are introduced as scatter in the forward simulations. More specifically, we displace the simulations by a number sampled from the error model presented in Sec. V. Note how in Fig. 2 this step does not affect the mass. This is because the mass in this problem is

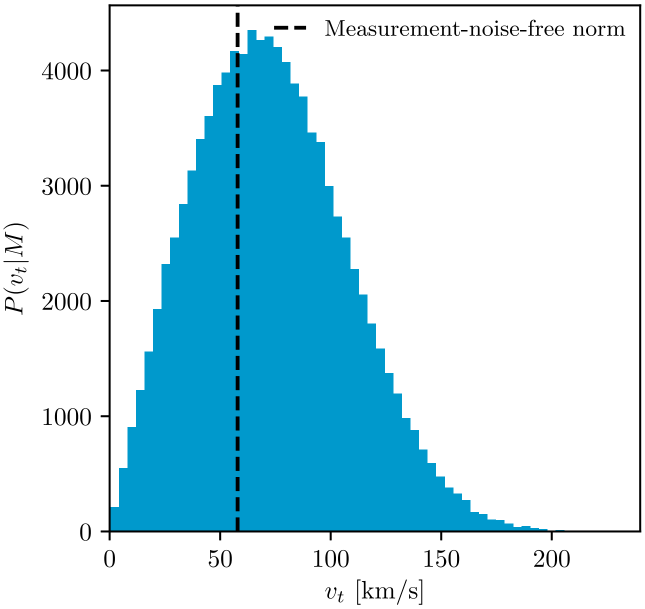

FIG. 1. An illustration of our non-Gaussian error modelling for the tangential velocity. The plot was obtained taking the components of the tangential velocity for a randomly chosen simulation (with fixed M), scattering a large number of times by the errors of Eq. (3), and calculating the norm for each sample. The black dashed line shows the norm of the components for the tangential velocity of the simulation without measurement error.

[3] We will discuss this supposed bias in a future publication.

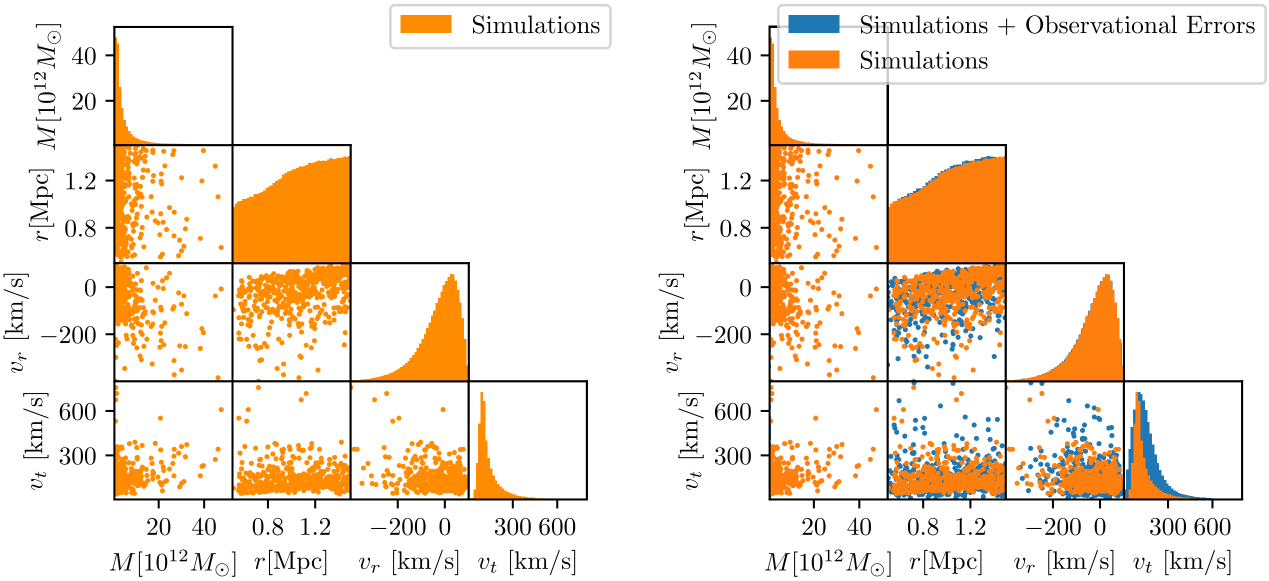

FIG. 2. Illustration of DELFI for the estimation of $M_{\mathrm{MW+M31}}$. The left panel is a scatter plot of the simulations described in Sec. IV. The right panel adds the observational errors by scattering the simulations.

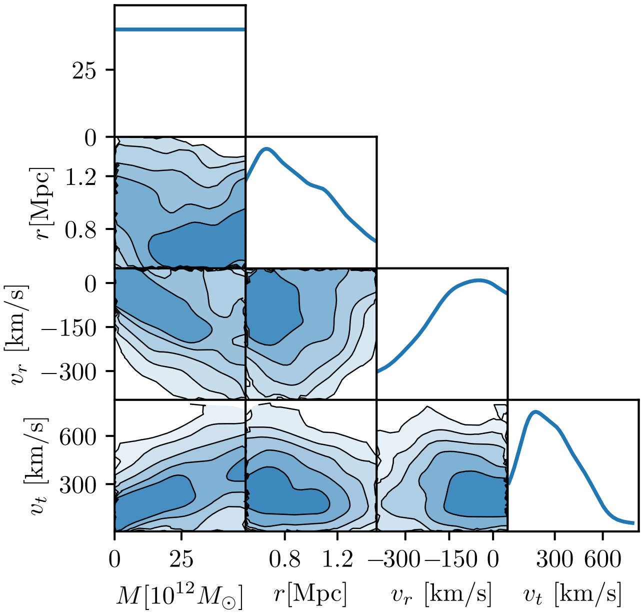

FIG. 3. Conditional distribution $p(d|\theta, I)$ density estimate from the points shown in the right panel of Fig. 2. The 2D subplots in the left column show the probabilities of the data $d = \{r, v_r, v_t\}$ conditional on the parameter $\theta = M$ (which has been given a uniform distribution), while the remaining 2D subplots have been marginalized over this uniform distribution. By evaluating this function at the observed data points $d = \{r, v_r, v_t\}$ we obtain the likelihood function. The points were sampled using the nested sampling [55] code POLYCHORD [56,57]. Note that the one dimensional distribution on M is flat by construction, as this is the parameter, and the distributions on r and v_r appear to be “cut” because of the selection criteria described in Sec. IV.

part of the parameters θ, not the data, as it is our goal to obtain a posterior distribution for the mass.

(3) We use density estimation to get the conditional density distribution $p(d|\theta, I)$, as shown in Fig. 3. While it might seem counterintuitive to learn the likelihood instead of directly learning the posterior, this allows us to then sample from a chosen prior, instead of being limited to the prior that is implicit in the simulations. This is discussed in more detail in Appendix A.

There are several algorithms that can be used to get a conditional probability distribution from samples. In this work we use GMDN and MAF[4] (as part of the PYDELFI package).

(4) Finally, we evaluate this conditional density distribution at the observed data

$$D_{\mathrm{obs}} = \{r = 0.77\ \mathrm{Mpc},\\ v_r = -109.3\ \mathrm{km\,s^{-1}},\ v_t = 72\ \mathrm{km\,s^{-1}}\}, \quad (4)$$

as discussed in Sec. V. This way, we get the likelihood function:

$$\mathcal{L} \equiv p(d = D_{\mathrm{obs}}|\theta, I). \quad (5)$$

[4] Note that while this work uses GMDN and MAF for density estimation, Fig. 3 uses KDE instead. This is because the plots were generated using the code ANESTHETIC [58], which uses KDE to plot smooth probability distributions. KDE is appropriate in this case, as ANESTHETIC only plots the one- and two-dimensional posterior distributions, whereas in this work we are trying to learn the full 4D distribution. ANESTHETIC uses the FASTKDE implementation [59,60].

Through this process we obtain a likelihood function, without ever having to write a theory or use a Gaussian approximation. While this process is limited by the number and accuracy of the available simulations, it has a big advantage over likelihood-based problems that use simplifying approximations to make the likelihood more tractable, or easier to compute. The calculation of $M_{\mathrm{MW+M31}}$ is a good example: the likelihood-based approach relies on the TA and data modeling approximations, which we know oversimplifies the problem. Instead, using DELFI, we can account for the complex nonlinear evolution of the system through our N-body simulation.

B. Density estimation validation

For the density estimation, with PYDELFI we use a combination of two GMDNs (with four and five Gaussian components) and two MAFs (with three and four components). The GMDNs have two layers with fifty components each, while the MAFs have thirty components on each of the two hidden layers. For a more robust density estimation we stack the results weighted by each density estimation's relative likelihood, as described in [2].

We hold back 10,000 simulations from the training set to be used for validating the likelihood that we have learned through the simulations. Each validation simulation has a "true" mass, position and velocity, and we can use these to estimate how well our likelihood works. The results from this validation are shown in Fig. 4. We also perform a quartile test, finding that 95.485% of the simulations fall within the 2σ predicted posterior, as expected.

C. Prior distribution

As previously discussed, we have the freedom to choose a suitable prior distribution (this is because we have used the simulations to learn a likelihood function, instead of directly learning the posterior distribution). The left panel of Fig. 8 shows four priors relevant to this study. Uninformative priors in this case could be either a flat prior or a logarithmic prior. In addition, in this problem Press-Schechter theory [61] supplies us with a physically-motivated prior. The Press-Schechter formalism predicts the number of virialized objects with a given mass. While this would be a fully correct prior only if the MW and M31 formed a single halo, it can provide a good prior distribution for the problem.[5] We calculate the Press-Schechter prior using the code COLOSSUS [62], [63] and the Tinker mass function [64]. Finally, the prior shown as a black dashed line is the one that would have been in use if we have learned the posterior directly from the simulations (when learning the posterior directly, we still have a prior, we simply lose the freedom to choose it). In this work, we adopt the Press-Schechter prior, which as shown in Fig. 8 is virtually equivalent to a flat prior in $\log M$. The effect of using different priors in our results will be discussed in Appendix B.

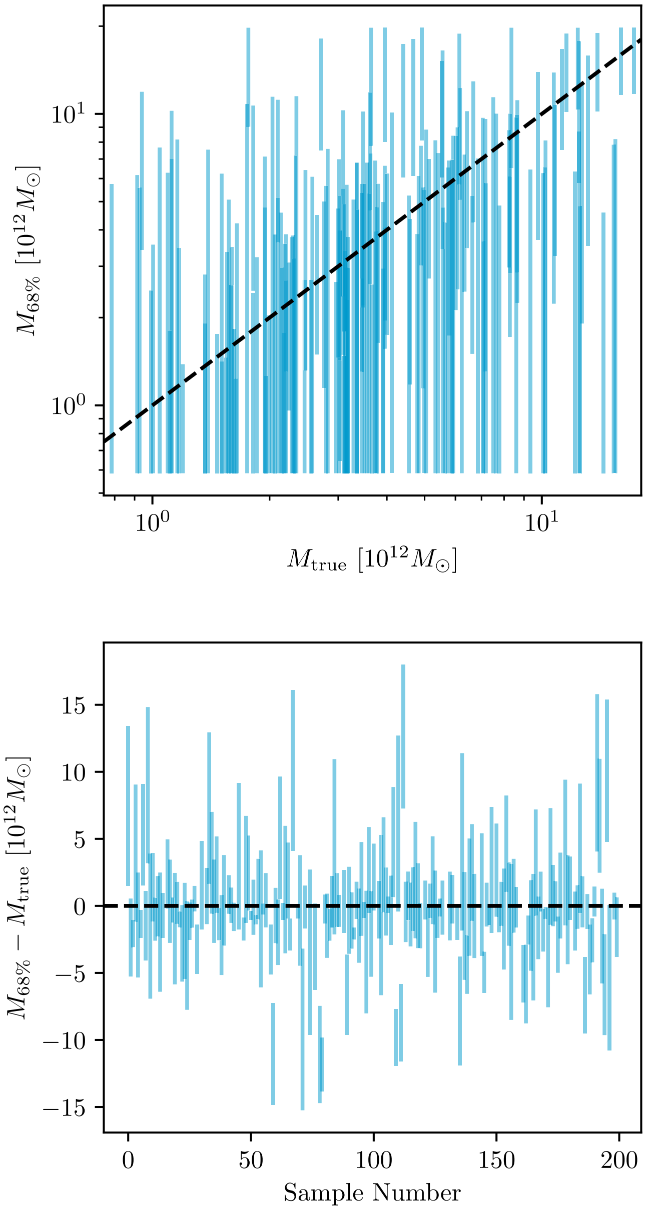

FIG. 4. Validation plot for our density estimation. We use 10,000 simulations that have not be used for training. For all these, we use r, v_r and v_t to estimate a mass, and compare with the true mass. The top figure plots predicted vs true mass, while the bottom plot shows the residuals. The bars show the 68% CL obtained using the method described in this paper.

Once we have obtained a likelihood function and prior, we can get a posterior using Bayes' theorem Eq. (2). We can describe this posterior by sampling from it using an algorithm such as Markov chain Monte Carlo (MCMC) or nested sampling [55]. However, in our case, a "brute-force" approach is more practical (because the posterior is only

[5] The local group is a bound system but not a virialized system, so placing it in the Press-Schechter mass function works as an approximation.

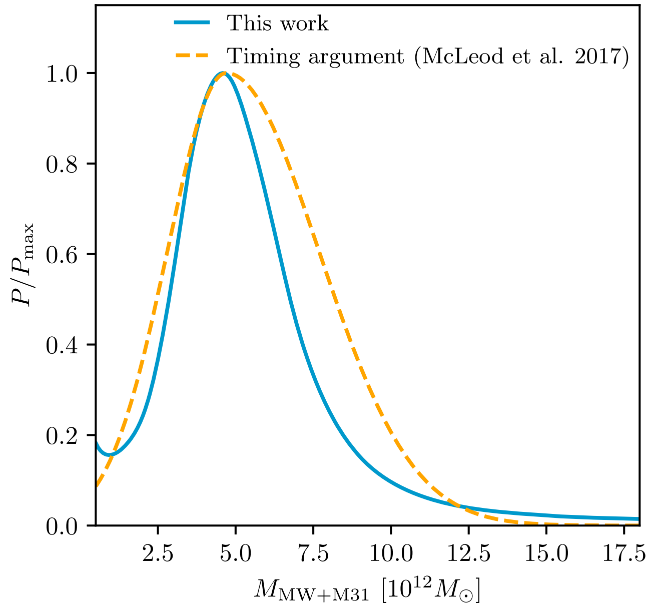

FIG. 5. The posterior on $M_{\rm MW+M31}$ obtained in this work (solid blue) compared to the timing argument result of [15] that includes Λ and the tangential velocity from [52] (dashed orange). While our peak is lower, the posteriors are fully consistent.

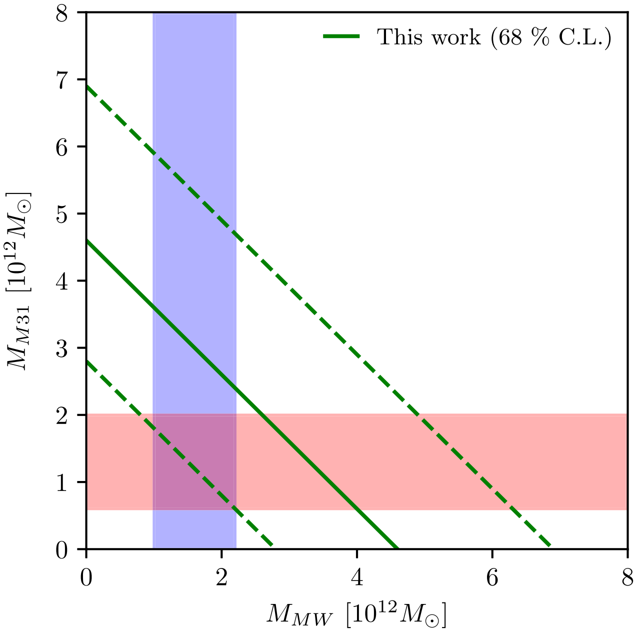

FIG. 6. A comparison of the estimates of the separate masses of M31, the MW, and their sum, the latter from this work. The plot shows the small discrepancy between separate estimates of the individual masses of the MW and M31 and this work.

one-dimensional): we simply calculate the posterior on a grid of mass values.

Our result using the Press-Schechter prior is shown in Fig. 5 (solid blue). Our peak and 68% confidence levels are $M_{\rm MW+M31} = 4.6^{+2.3}_{-1.8} \times 10^{12}\ M_\odot$, in good agreement with [15] (also shown in Fig. 5) but with improved error bars.

D. Results discussion

Our result is compared to previous results in Fig. 7. We see that all other estimates considered in this work are within the 68% confidence interval of our posterior in the mass, despite the different methods used. We notice that the least action result of [26]) obtains tighter constraints than our method; however, our result is the first one to fully account for the distribution of the observed errors in a robust (and Bayesian) manner. Other results use Gaussian approximations for observational errors, or neglect them completely, and therefore our result is the most accurate estimate of $M_{\rm MW+M31}$ to date. This framework also allows for more accurate estimates, in particular accounting for the presence of M33 and the LMC in the local group. This will be explored in future work.

The simulation was run using one particular set of cosmological parameters but in reality these parameters are uncertain and we should marginalize over them. This is infeasible for us, as we have a pre-run set of simulations with fixed cosmological parameters, but we can estimate the size of the effect by reference to the timing argument (TA).

The TA uses the same observational constraints as does this work, and like this work is based on modeling/simulating the trajectories of galaxies similar to those in the MW + M31 system; as a result the TA should have similar sensitivities to cosmological parameters as this work. The TA sensitivities can be estimated by numerically differentiating the mass estimation algorithm described in [22]. We parametrize this algorithm using h and Ω_Λ (from which Λ and the age of the universe may be derived, the latter assuming $\Omega_m + \Omega_\Lambda = 1$). We find $\partial M_{\rm MW+M31}/\partial\Omega_\Lambda = -2.4 \times 10^{12}\ M_\odot$ and $\partial M_{\rm MW+M31}/\partial h = 7.4 \times 10^{12}\ M_\odot$. Multiplying these sensitivities by uncertainties on cosmological parameters ($\Delta\Omega_\Lambda = 0.006$ and $\Delta h = 0.004$ [65]) yields uncertainties on the mass estimate that are immaterial compared to the uncertainty implied by the posterior width, and hence will be ignored. This conclusion continues to hold even if we assume a larger uncertainty on h reflecting the current tension between early- and late-Universe measurements of this parameter. For example, a change in h of 0.066 induces a change in the TA $M_{\rm MW+M31}$ of $0.49 \times 10^{12}\ M_\odot$ (in agreement with [23]); adding this in quadrature to the uncertainty implied by the posterior width yields only a marginal increase in total uncertainty (from $2.3 \times 10^{12}\ M_\odot$ to $2.35 \times 10^{12}\ M_\odot$). This calculation illustrates that in a simulation-based approach it is important to have a benchmark analytical model, to gauge if parameters not explored by the simulations are relevant and if extra simulations are needed.

Finally, we can compare our results with separate estimates of the masses of the Milky Way and M31.

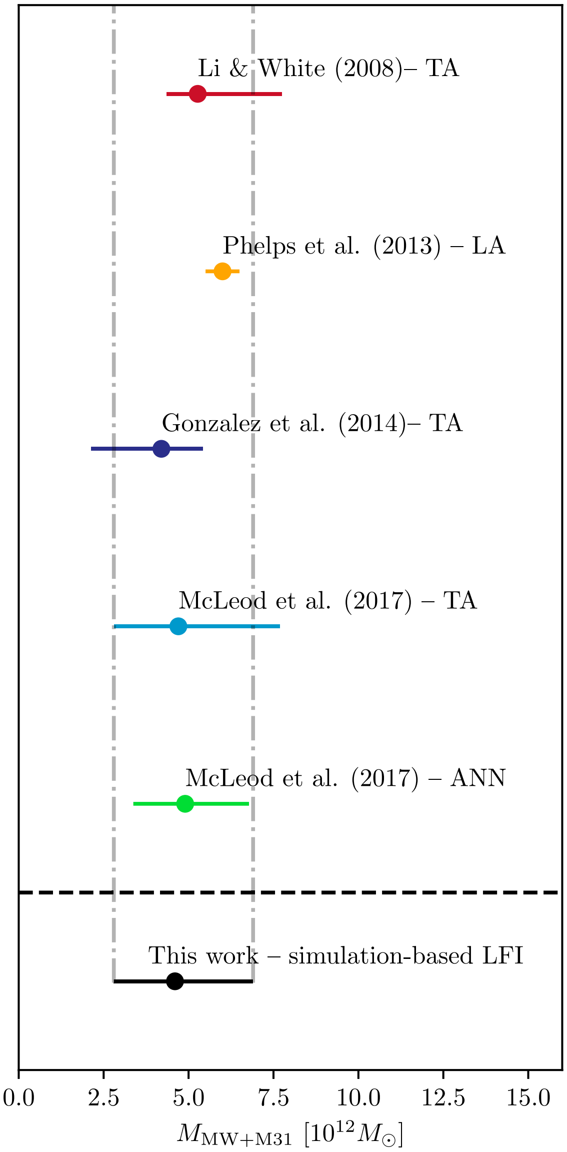

FIG. 7. Comparison of this work with previous estimates of the $M_{\mathrm{MW+M31}}$, shown as best fit and 68% confidence intervals. The result of this work is shown at the bottom; it is the first to account fully for the observational errors, and to not rely on the approximation of the TA.

There are several values in the literature for the separate masses of each galaxy, in some cases discrepant. Given this discrepancy, we take a number of estimates of each mass obtained through different methods, and assume that the true value is contained within the ranges of the different estimates, as done in [66]. While conservative, this method should provide with ranges that contain the true mass of each galaxy, and allow us to combine estimates in tension. Through this method, we get the following:

(i) $M_{\mathrm{MW}} \in (1.0, 2.2) \times 10^{12}\ M_\odot$, from [67–72]

(ii) $M_{\mathrm{M31}} \in (0.6, 2.0) \times 10^{12}\ M_\odot$, from [67,73–75]

Combining these two measurements yields $M_{\mathrm{MW+M31}} \in (1.6, 4.2) \times 10^{12}\ M_\odot$. This is slightly lower than our result, but still in agreement, as illustrated in Fig. 6. We can see in Fig. 7 that all estimates of the sum of the masses based on the relative distance and velocity of the bodies (TA, ANN and our approach) obtain slightly larger values than the sum of the separate masses. A possible explanation for this could be the fact that all these approaches ignore the effect of other bodies such as the LMC and M33 in the observed velocities, which could bias the sum of the masses to higher values [76]. The effect of the LMC and M33 in our posterior mass will be explored in future work.

VII. CONCLUSIONS

In this work we have used density estimation likelihood-free inference with forward-modelling to estimate the posterior distribution for sum of the masses of the Milky Way and M31 using observations of the distance and velocity to M31. We obtain a mass $M_{\mathrm{MW+M31}} = 4.6^{+2.3}_{-1.8} \times 10^{12}\ M_\odot$ (M_{200}). Our method overcomes the several approximations of the traditional timing argument, accounts for non-Gaussian sources of observational measurement error, and uses a physically motivated prior; this makes it the most reliable estimate of $M_{\mathrm{MW+M31}}$ mass to date.

The sensitivity analysis performed in this study illustrates that in any simulation-based approach it is important to have a benchmark analytical (or semianalytical) model, to assess how to cover the parameter space of required simulations.

This works serves not only to obtain state-of-the-art estimates of the $M_{\mathrm{MW+M31}}$; by applying likelihood-free inference to a problem that is physically rich and complex yet statistically simple (thanks to its low dimensionality), we can illustrate how the method works, what different choices need to be made, and what challenges need to be tackled. The ability to robustly infer $M_{\mathrm{MW+M31}}$ without requiring an analytic theory or a likelihood demonstrates the potential of likelihood-free inference methods in astronomy and cosmology.

ACKNOWLEDGMENTS

We thank Andrey Kravtsov, Roeland van der Marel, Ekta Patel and Vasily Belokurov. We are also grateful to David Benisty for participating in discussions that led to the idea of this paper. P. L. and O. L. acknowledge STFC Consolidated Grant No. ST/R000476/1. N. I. L. acknowledges financial support of the Project IDEXLYON at the University of Lyon under the Investments for the Future Program (Grant No. ANR-16-IDEX-0005). Y. H. has been partially supported by the Israel Science Foundation Grant No. ISF 1358/18.

APPENDIX A: DENSITY ESTIMATORS

One of the key elements of DELFI is the estimation of a probability distribution from samples. This corresponds to going from Fig. 2 (right panel) to Fig. 3. The density estimation problem arises in many fields (for example image analysis [77,78]) and several algorithms have been developed to address it. In this section, we review some of the most popular density estimation methods in the context of LFI. For an overview of neural density estimation in the context of LFI we recommend [2].

Density estimation algorithms that rely on there being samples near the point of interest, such as spline or kernel density estimation (KDE), struggle in high dimensional spaces due to the sparsity of the sampling. They are very useful, however, for estimating low dimensional PDFs, which is why they are often used for plotting marginalized posterior distributions. Public codes such as GetDist [79], CHAINCONSUMER [80] or ANESTHETIC [58] use KDE to generate plots of marginalized posterior distributions.

A mixture model (MM) represents a PDF p as a weighted sum of component distributions:

$$p(\mathbf{y}) = \sum_{c=1}^{N} \alpha_c \mathcal{D}(\mathbf{y}; \Phi_c). \tag{A1}$$

Here N is the number of components in the mixture while $\mathcal{D}$ is some family of distributions described by parameters Φ; the weights $\{\alpha_c\}$ and parameters $\{\Phi_c\}$ are fit to observed or training data. A common choice is the Gaussian mixture model (GMM), in which each component distribution is Gaussian: $\mathcal{D}(\mathbf{y}; \Phi_c) = \mathcal{N}(\mathbf{y}; \mu_c, \sigma_c)$.

GMMs can successfully represent a large number of PDFs. In addition, they have the advantage that the weights and parameters $\{\alpha_c, \mu_c, \sigma_c\}$ can be easily fit to the data using the expectation-maximization algorithm [81]. There are, however, some issues with GMMs: they are sensitive to the choice of N, and they have problems fitting certain features (such as the sharp edges that can arise when flat priors are used).

In the context of LFI we are interested in modeling a *conditional* distribution $p(\mathbf{y}|\mathbf{x})$ (for example in our case p is the conditional likelihood, for which $\mathbf{y} = d$ and $\mathbf{x} = \theta$). Such conditional distributions can be modeled by mixture density networks (MDNs) [33,34]. As with MMs, they model the PDF as a weighted sum of component distributions, but now the weights and parameters describing the components are themselves (possibly nonlinear) functions of $\mathbf{x}$:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{c=1}^{N} \alpha_c(\mathbf{x}) \mathcal{D}_c(\mathbf{y}; \Phi_c(\mathbf{x})). \tag{A2}$$

Again, a common choice is the Gaussian MDN (GMDN), in which each component is Gaussian: $\mathcal{D}(\mathbf{y}; \Phi_c(\mathbf{x})) = \mathcal{N}(\mathbf{y}; \mu_c(\mathbf{x}), \sigma_c(\mathbf{x}))$.

The functions $\{\alpha_c(\mathbf{x}), \mu_c(\mathbf{x}), \sigma_c(\mathbf{x})\}$ can be modeled by a neural network with a set of weights; these weights are then fit to the data. As with GMMs, GMDNs require specification of the number N of mixture components to be used; however, this dependence is much smaller than in the case of GMMs (as GMDNs can fit complex distributions using only a small number of components).

We finish by describing masked autoregressive flows (MAFs), which have recently emerged as a powerful density estimation method [35,45]. They do not rely on a choice of number of components, and have the advantage of providing simple tests of the goodness of fit to the samples.

Here is the motivation for *masking* as a strategy for density estimation. Consider training a neural network *NN* to mimic; one can imagine training a parrot, for example. The trainer speaks (=input signal), and rewards the bird if its output matches this training input. If *NN* has sufficient complexity then it will learn to mimic the input. Now repeat the process but with a bird with covered (=masked) ears. The bird cannot hear the input, but nevertheless receives the training reward if its output matches the input. Now the bird can only "play the percentages"; it learns the optimal strategy, which is to output a weighted average of the input signals (weighted by their frequency of usage by the trainer). In this way the masked parrot learns the probability distribution of the input signal i.e., has become a density estimator.

That was the one-dimensional case. The two dimensional case needs two parrots. The first is trained on signal x_1, which it cannot hear, with the result that it learns $p(x_1)$. The second is trained on x_2, which it cannot hear, but it is allowed to hear x_1. As a result it learns $p(x_2|x_1)$. Thus between them they learn $p(x_1)p(x_2|x_1) = p(x_1, x_2)$ as desired. The multidimensional case is similar. This strategy is called an *autoregressive autoencoder*. Note that it treats the coordinates asymmetrically.

The conditional distributions $p(x_i|x_1, \ldots, x_{i-1})$ learned by *NN* are typically modeled as Gaussian. Consider generating samples from the estimated probability distribution p; for each sample we need we a set of n random unit normals (i.e., a draw from $N(0, I)$), which we transform to get samples from p—call this transform T. The details of T come from the means and standard deviations of the conditional distributions, which can be obtained from n evaluations of *NN*. However, importantly, the inverse mapping T^{-1} can be found with just *one* evaluation of *NN*. This is the idea of the masked autoencoder for distribution estimation (MADE) algorithm [82].

Now apply T^{-1} to the training data D. The resulting "pulled-back" data $T^{-1}(D)$ will ideally be a set of samples from $N(0, I)$ and its deviation from this ideal gives a direct measure of how imperfect is our modeling of p. We can then use the pulled-back data as the training data for yet another MADE process, and so on through several iterations. Between iterations we permute the coordinate axes, thereby symmetrising how we treat them. With sufficient iterations, the multiply-pulled-back training data approaches $N(0, I)$; at this point the algorithm has an easy-to-evaluate mapping between p and $N(0, I)$, which suffices for doing calculations. This is the masked autoregressive flows (MAF) algorithm [35,45].

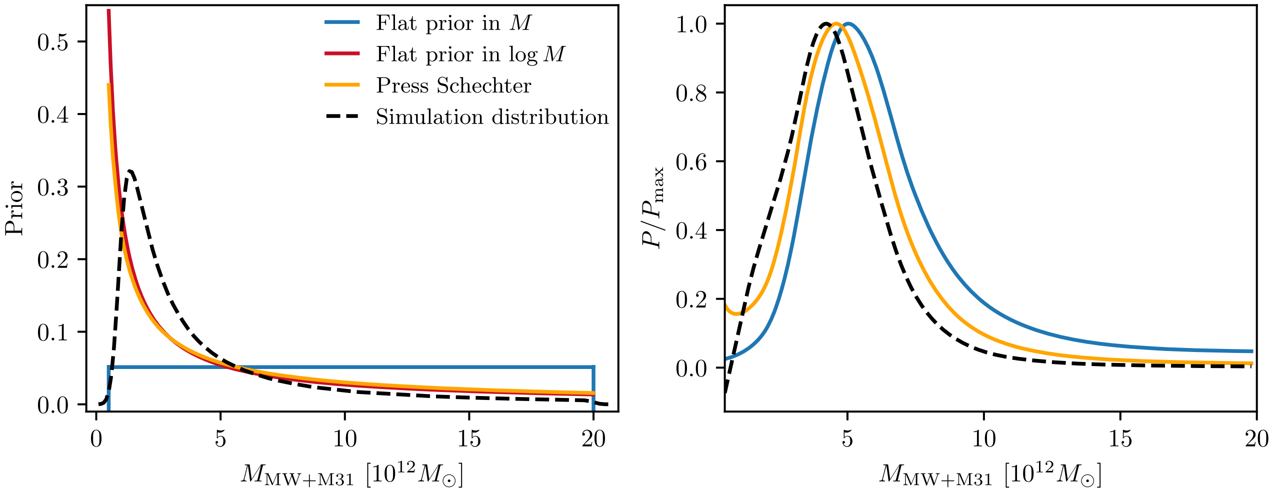

FIG. 8. On the left, a comparison of four possible priors: a flat prior in the mass (blue), a flat prior in the logarithm of the mass (red), a physically motivated Press-Schechter prior (orange), and the prior from the distribution of the simulations (dashed black). On the right, the corresponding posterior obtained from each prior. Note that we did not show the posterior for the flat prior in the logarithm of the mass, as this is equivalent to the Press-Schechter prior.

APPENDIX B: DEPENDENCE ON PRIORS

In this Appendix, we explore how the posterior distribution of $M_{\rm MW+M31}$ (as shown in Fig. 5 and discussed in Sec. VI) depends on our choice of prior. We consider the four different priors illustrated in Fig. 8:

(i) A flat prior in the mass;
(ii) A logarithmic prior in the mass;
(iii) A prior based on the Press-Schechter distribution (as adopted in this work);
(iv) A prior distribution matching the distribution of masses in the simulation.

In our case the second and third choices are virtually the same (as shown in the left panel of Fig. 8), and so we omit the ‘logarithmic in mass’ prior when examining how the priors affect our results. The posteriors obtained when using the remaining three priors are shown in the right panel of Fig. 8, and we see that our result is essentially independent of our choice of prior, be it the Press-Schechter prior, a flat prior on the mass ($M_{\rm LG}(\text{Flat prior}) = 5.0^{+2.7}_{-1.7} \times 10^{12}\ M_\odot$) or the prior from the simulation distribution. ($M_{\rm LG}(\text{Simulation Distribution}) = 4.3 \pm 1.7 \times 10^{12}\ M_\odot$)

APPENDIX C: DEPENDENCE ON TANGENTIAL VELOCITY

Imagine a 2-dimensional velocity vector $\mathbf{V} = (V_x, V_y)$. If V_x, V_y are uncorrelated, normally distributed with zero mean and equal variance σ, then the overall speed $V = \sqrt{(V_x^2 + V_y^2)}$ will be characterized by the Rayleigh distribution [83], with mean $\bar{V} = \sigma\sqrt{\pi/2}$, rather than naively zero. Similarly, nonzero measurements of V_x, V_y with error bars will result in a distribution function with a mean that is *not* $\bar{V} = \sqrt{(V_x^2 + V_y^2)}$. In our analysis we pay attention to this via the forward modeling approach, starting with simulated (V_x, V_y) and propagating their impact on the final posterior for $M_{\rm MW+M31}$.

While distance and radial velocity have been measured in numerous occasions using different methods, observations of the tangential velocity are far more scarce [52–54]. Their work treats the effect of converting measurements of two components of the tangential velocity into a modulus in

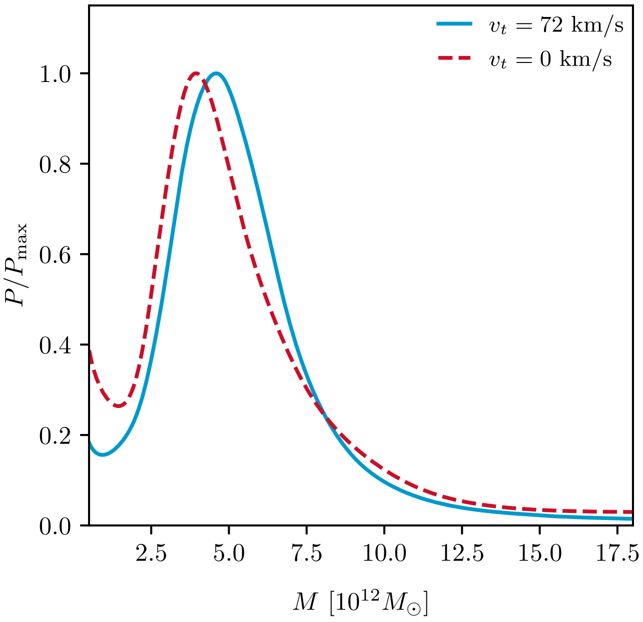

FIG. 9. The posterior on $M_{\rm MW+M31}$ for different values of the tangential velocity: $v_t = 72\ {\rm km\,s^{-1}}$ as used in this work (solid blue), and a purely radial motion $v_t = 0$ (dashed red).

a novel way. We check the robustness of our approach in this section. We do so by comparing the main result of the paper to the case of no tangential velocity. As shown in Fig. 9, our posterior on the mass does not depend strongly on the tangential velocity. Therefore, we are confident on the accuracy of our posterior in the mass.

[1] F. Leclercq, Bayesian optimization for likelihood-free cosmological inference, Phys. Rev. D **98**, 063511 (2018).

[2] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, Fast likelihood-free cosmology with neural density estimators and active learning, Mon. Not. R. Astron. Soc. **488**, 5093 (2019).

[3] M. Betoule, R. Kessler, J. Guy, J. Mosher, D. Hardin *et al.* Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples, Astron. Astrophys. **568**, A22 (2014).

[4] Y.-C. Wang, Y.-B. Xie, T.-J. Zhang, H.-C. Huang, T. Zhang, and K. Liu, Likelihood-free cosmological constraints with artificial neural networks: An application on hubble parameters and SN Ia, arXiv:2005.10628.

[5] N. Jeffrey, J. Alsing, and F. Lanusse, Likelihood-free inference with neural compression of DES SV weak lensing map statistics, arXiv:2009.08459 [MNRAS (to be published)], https://doi.org/10.1093/mnras/staa3594.

[6] J. Brehmer, S. Mishra-Sharma, J. Hermans, G. Louppe, and K. Cranmer, Mining for dark matter substructure: Inferring subhalo population properties from strong lenses with machine learning, Astrophys. J. **886**, 49 (2019).

[7] D. K. Ramanah, R. Wojtak, and N. Arendse, Simulation-based inference of dynamical galaxy cluster masses with 3D convolutional neural networks, arXiv:2009.03340 [MNRAS (to be published)], https://doi.org/10.1093/mnras/staa3922.

[8] L. Tortorelli, M. Fagioli, J. Herbel, A. Amara, T. Kacprzak, and A. Refregier, Measurement of the B-band galaxy luminosity function with approximate Bayesian computation, J. Cosmol. Astropart. Phys. 09 (2020) 048.

[9] F. D. Kahn and L. Woltjer, Intergalactic matter and the Galaxy, Astrophys. J. **130**, 705 (1959).

[10] www.cosmosim.org

[11] F. Prada, A. A. Klypin, A. J. Cuesta, J. E. Betancort-Rijo, and J. Primack, Halo concentrations in the standard Λ cold dark matter cosmology, Mon. Not. R. Astron. Soc. **423**, 3018 (2012).

[12] K. Riebe, A. M. Partl, H. Enke, J. Forero-Romero, S. Gottlöber, A. Klypin, G. Lemson, F. Prada, J. R. Primack, M. Steinmetz, and V. Turchaninov, The MultiDark database: Release of the Bolshoi and MultiDark cosmological simulations, Astron. Nachr. **334**, 691 (2013).

[13] G. Meylan, J. P. Madrid, and D. Macchetto, Hubble space telescope science metrics, Publ. Astron. Soc. Pac. **116**, 790 (2004).

[14] Gaia Collaboration, The Gaia mission, Astron. Astrophys. **595**, A1 (2016).

[15] M. McLeod, N. Libeskind, O. Lahav, and Y. Hoffman, Estimating the mass of the local group using machine learning applied to numerical simulations, J. Cosmol. Astropart. Phys. 12 (2017) 034.

[16] F. V. Bonassi, L. You, and M. West, Bayesian Learning from marginal data in bionetwork models, Stat. Appl. Genetics Mol. Biol. **10**, 49 (2011).

[17] Y. Fan, D. J. Nott, and S. A. Sisson, Approximate Bayesian computation via regression density estimation, arXiv:1212.1479.

[18] G. Papamakarios and I. Murray, Fast ϵ-free inference of simulation models with Bayesian conditional density estimation, arXiv:1605.06376.

[19] https://github.com/justinalsing/pydelfi

[20] D. Lynden-Bell, The dynamical age of the local group of galaxies, Observatory **101**, 111 (1981), https://ui.adsabs.harvard.edu/abs/1981Obs...101..111L/abstract.

[21] J. Binney and S. Tremaine, *Galactic Dynamics* (Princeton University Press, 1987).

[22] C. Partridge, O. Lahav, and Y. Hoffman, Weighing the local group in the presence of dark energy, Mon. Not. R. Astron. Soc. **436**, L45 (2013).

[23] M. McLeod and O. Lahav, The two body problem in the presence of dark energy and modified gravity: Application to the local group, J. Cosmol. Astropart. Phys. 09 (2020) 056.

[24] D. Benisty, E. I. Guendelman, and O. Lahav, Milky Way and Andromeda past-encounters in different gravity models: The impact on the estimated local group mass, arXiv:1904.03153 [Phys. Rev. D (to be published)].

[25] P. J. E. Peebles, Orbits of the nearby Galaxies, Astrophys. J. **429**, 43 (1994).

[26] S. Phelps, A. Nusser, and V. Desjacques, The mass of the Milky Way and M31 using the method of least action, Astrophys. J. **775**, 102 (2013).

[27] Y.-S. Li and S. D. M. White, Masses for the local group and the Milky Way, Mon. Not. R. Astron. Soc. **384**, 1459 (2008).

[28] R. E. Gonzalez, A. V. Kravtsov, and N. Y. Gnedin, On the mass of the local group, Astrophys. J. **793**, 91 (2014).

[29] D. B. Rubin, Bayesianly justifiable and relevant frequency calculations for the applied statistician, Ann. Stat. **12**, 1151 (1984).

[30] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Stat. **27**, 832 (1956).

[31] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. **33**, 1065 (1962).

[32] J. S. Simonoff, *Smoothing Methods in Statistics*, Springer Series in Statistics (Springer, New York, 1996).

[33] C. M. Bishop, Mixture density networks, Technical Report, 1994, https://research.aston.ac.uk/en/publications/mixture-density-networks.

[34] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[35] G. Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation, in *NIPS'17: Proceedings of the 31st International Conference on Neural Information* (2017), pp. 2338–2347.

[36] E. Cameron and A. N. Pettitt, Approximate Bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift, Mon. Not. R. Astron. Soc. **425**, 44 (2012).

[37] A. Weyant, C. Schafer, and W. Michael Wood-Vasey, Likelihood-free cosmological inference with Type Ia Supernovae: Approximate Bayesian computation for a complete treatment of uncertainty, Astrophys. J. **764**, 116 (2013).

[38] J. Akeret, A. Refregier, A. Amara, S. Seehars, and C. Hasner, Approximate Bayesian computation for forward modeling in cosmology, J. Cosmol. Astropart. Phys. 08 (2015) 043.

[39] C. H. Hahn, M. Vakili, K. Walsh, A. P. Hearin, D. W. Hogg, and D. Campbell, Approximate Bayesian computation in large-scale structure: constraining the galaxy–halo connection, Mon. Not. R. Astron. Soc. **469**, 2791 (2017).

[40] A. Peel, C.-A. Lin, F. Lanusse, A. Leonard, J.-L. Starck, and M. Kilbinger, Cosmological constraints with weak lensing peak counts and second-order statistics in a large-field survey, Astron. Astrophys. **599**, A79 (2017).

[41] T. Kacprzak, J. Herbel, A. Amara, and A. Rfrgier, Accelerating approximate Bayesian computation with quantile regression: Application to cosmological redshift distributions, J. Cosmol. Astropart. Phys. 02 (2018) 042.

[42] J. Alsing, B. Wandelt, and S. Feeney, Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology, Mon. Not. R. Astron. Soc. **477**, 2874 (2018).

[43] J. Alsing and B. Wandelt, Generalized massive optimal data compression, Mon. Not. R. Astron. Soc. **476**, L60 (2018).

[44] A. F. Heavens, E. Sellentin, and A. H Jaffe, Extreme data compression while searching for new physics, Mon. Not. R. Astron. Soc. **498**, 3440 (2020).

[45] G. Papamakarios, D. Sterratt, and I. Murray, Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019), pp. 837–848, http://proceedings.mlr.press/v89/papamakarios19a.html.

[46] J.-M. Lueckmann, G. Bassetto, T. Karaletsos, and J. H Macke, Likelihood-free inference with emulator networks, in *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference* (PMLR, 2019), pp 32–53, http://proceedings.mlr.press/v96/lueckmann19a.html.

[47] A. Knebe, S. R. Knollmann, S. I. Muldrew, F. R. Pearce, and M. A. Aragon-Calvo *et al.* Haloes gone MAD: The halo-finder comparison project, Mon. Not. R. Astron. Soc. **415**, 2293 (2011).

[48] S. Holland, The distance to the M31 globular cluster system, Astron. J. **115**, 1916 (1998).

[49] Y. C. Joshi, A. K. Pandey, D. Narasimha, R. Sagar, and Y. Giraud-Hiraud, Identification of 13 cepheids and 333 other variables in M 31, Astron. Astrophys. **402**, 113 (2003).

[50] I. Ribas, C. Jordi, F. Vilardell, E. L. Fitzpatrick, R. W. Hilditch, and E. F. Guinan, First determination of the distance and fundamental properties of an eclipsing binary in the andromeda galaxy, Astrophys. J. Lett. **635**, L37 (2005).

[51] A. McConnachie and M. Irwin, Structural parameters for the m31 dwarf spheroidals, Mon. Not. R. Astron. Soc. **365**, 1263 (2006).

[52] R. P. van der Marel, M. A. Fardal, S. T. Sohn, E. Patel, G. Besla, A. del Pino, J. Sahlmann, and L. L. Watkins, First Gaia dynamics of the Andromeda system: DR2 proper motions, orbits, and rotation of M31 and M33, Astrophys. J. **872**, 24 (2019).

[53] R. P. van der Marel, G. Besla, T. J. Cox, S. T. Sohn, and J. Anderson, The M31 velocity vector. III. Future Milky Way M31-M33 orbital evolution, merging, and fate of the sun, Astrophys. J. **753**, 9 (2012).

[54] R. P. van der Marel and P. Guhathakurta, M31 transverse velocity and local group mass from satellite kinematics, Astrophys. J. **678**, 187 (2008).

[55] J. Skilling *et al.*, Nested sampling for general Bayesian computation, Bayesian Anal. **1**, 833 (2006), https://projecteuclid.org/euclid.ba/1340370944.

[56] W. J. Handley, M. P. Hobson, and A. N. Lasenby, POLYCHORD: Nested sampling for cosmology, Mon. Not. R. Astron. Soc. **450**, L61 (2015).

[57] W. J. Handley, M. P. Hobson, and A. N. Lasenby, POLYCHORD: Next-generation nested sampling, Mon. Not. R. Astron. Soc. **453**, 4385 (2015).

[58] W. Handley, anesthetic: Nested sampling visualisation, J. Open Source Softw. **4**, 1414 (2019).

[59] T. O'Brien, K. Kashinath, N. Cavanaugh, W. Collins, and J. O'Brien, A fast and objective multidimensional kernel density estimation method: Fastkde, Computational Statistics and Data Analysis **101**, 148 (2016).

[60] T. O'Brien, W. Collins, S. Rauscher, and T. Ringler, Reducing the computational cost of the ecf using a nufft: A fast and objective probability density estimation method, Computational Statistics and Data Analysis **79**, 222 (2014).

[61] W. H. Press and P. Schechter, Formation of Galaxies and clusters of Galaxies by self-similar gravitational condensation, Astrophys. J. **187**, 425 (1974).

[62] B. Diemer, COLOSSUS: A PYTHON toolkit for cosmology, large-scale structure, and dark matter halos, Astrophys. J. Suppl. Ser. **239**, 35 (2018).

[63] https://bdiemer.bitbucket.io/colossus/index.html

[64] J. L. Tinker, A. V. Kravtsov, A. Klypin, K. Abazajian, M. S. Warren, G. Yepes, S. Gottlober, and D. E. Holz, Toward a halo mass function for precision cosmology: The limits of universality, Astrophys. J. **688**, 709 (2008).

[65] Planck Collaboration, Planck 2018 results—VI. Cosmological parameters, Astron. Astrophys. **641**, A6 (2020).

[66] N. I. Libeskind *et al.*, The HESTIA project: Simulations of the local group, Mon. Not. R. Astron. Soc. **498**, 2968 (2020).

[67] J. D. Diaz, S. E. Koposov, M. Irwin, V. Belokurov, and N. W. Evans, Balancing mass and momentum in the local group, Mon. Not. R. Astron. Soc. **443**, 1688 (2014).

[68] D. Zaritsky and H. Courtois, A dynamics-free lower bound on the mass of our Galaxy, Mon. Not. R. Astron. Soc. **465**, 3724 (2017).

[69] K. Hattori, M. Valluri, E. F. Bell, and I. U. Roederer, Old, metal-poor extreme velocity stars in the solar neighborhood, Astrophys. J. **866**, 121 (2018).

[70] L. Posti and A. Helmi, Mass and shape of the Milky Way's dark matter halo with globular clusters from Gaia and hubble, Astron. Astrophys. **621**, A56 (2019).

[71] L. L. Watkins, R. P. van der Marel, S. T. Sohn, and N. W. Evans, Evidence for an intermediate-mass Milky Way from Gaia DR2 halo globular cluster motions, Astrophys. J. **873**, 118 (2019).

[72] E. V. Karukes, M. Benito, F. Iocco, R. Trotta, and A. Geringer-Sameth, A robust estimate of the Milky Way mass from rotation curve data, J. Cosmol. Astropart. Phys. 05 (2020) 033.

[73] E. Corbelli, S. Lorenzoni, R. Walterbos, R. Braun, and D. Thilker, A wide-field H I mosaic of messier 31. II. The disk warp, rotation, and the dark matter halo, Astron. Astrophys. **511**, A89 (2010).

[74] A. Tamm, E. Tempel, P. Tenjes, O. Tihhonova, and T. Tuvikene, Stellar mass map and dark matter distribution in M 31, Astron. Astrophys. **546**, A4 (2012).

[75] P. R. Kafle, S. Sharma, G. F. Lewis, A. S. G. Robotham, and S. P. Driver, The need for speed: escape velocity and dynamical mass measurements of the Andromeda galaxy, Mon. Not. R. Astron. Soc. **475**, 4043 (2018).

[76] J. Peñarrubia, F. A. Gómez, G. Besla, D. Erkal, and Y.-Z. Ma, A timing constraint on the (total) mass of the large magellanic cloud, Mon. Not. R. Astron. Soc. **456**, L54 (2016).

[77] L. Theis and M. Bethge, Generative image modeling using spatial lstms, in *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems* (2015), pp. 1927–1935.

[78] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications, arXiv: 1701.05517.

[79] A. Lewis, GetDist: A PYTHON package for analysing Monte Carlo samples, 2019, https://getdist.readthedocs.io.

[80] S. R. Hinton, ChainConsumer, J. Open Source Softw. **1**, 00045 (2016).

[81] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B **39**, 1 (1977).

[82] M. Germain, K. Gregor, I. Murray, and H. Larochelle, Made: Masked autoencoder for distribution estimation, in *International Conference on Machine Learning* (PMLR, 2015), pp. 881–889, http://proceedings.mlr.press/v37/germain15.html.

[83] L. Rayleigh, The problem of the random walk, Nature (London) **72**, 318 (1905).