

The dynamical memory of tidal stellar streams: Joint inference of the Galactic potential and the progenitor of GD-1 with flow matching

Giuseppe Viterbo^{1,2}, Tobias Buck^{1,2}

¹ Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR), Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

e-mail: giuseppe.viterbo@iwr.uni-heidelberg.de

² Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Albert-Ueberle-Straße 2, D-69120 Heidelberg, Germany

e-mail: tobias.buck@iwr.uni-heidelberg.de

Received Month, XXXX; accepted Month Day, 7/7/2025

ABSTRACT

Context. Stellar streams offer one of the most sensitive probes of the Milky Way’s gravitational potential, as their phase-space morphology encodes both the tidal field of the host galaxy and the internal structure of their progenitors. In this work, we introduce a framework that leverages Flow Matching and Simulation-Based Inference (SBI) to jointly infer the parameters of the GD-1 progenitor and the global properties of the Milky Way potential.

Aims. Our aim is to move beyond traditional techniques (e.g. orbit-fitting and action-angle methods) by constructing a fully Bayesian, likelihood-free posterior over both host-galaxy parameters and progenitor properties, thereby capturing the intrinsic coupling between tidal stripping dynamics and the underlying potential.

Methods. To achieve this, we generate a large suite of mock GD-1-like streams using our differentiable N-body code *ODISSEO*, sampling self-consistent initial conditions from a Plummer sphere and evolving them in a flexible Milky Way potential model. We then apply conditional Flow Matching to learn the vector field that transports a base Gaussian distribution into the posterior $p(\theta | d)$, enabling efficient, amortized inference directly from stream phase-space data.

Results. We demonstrate that our method successfully recovers the true parameters of a fiducial GD-1 simulation, producing well-calibrated posteriors and accurately reproducing parameter degeneracies arising from progenitor–host interactions. Our results highlight the power of modern generative models for dynamical inference and provide a scalable pathway toward jointly constraining Galactic structure and the origins of stellar streams.

Conclusions. Flow Matching provides a powerful, flexible framework for Galactic Archaeology. Our approach enables joint inference on progenitor and Galactic parameters, capturing complex dependencies that are difficult to model with classical likelihood-based methods. This work paves the way for fully simulation-driven dynamical inference using Gaia and upcoming surveys.

Key words. Galaxies: evolution – Galaxies: formation – Galaxies: photometry – Methods: data analysis – Methods: statistical – Techniques: Simulation Based Inference

1. Introduction

Assuming a hierarchical Λ CDM accretion history, the Galaxy should be populated by tidal debris from accreted satellites, like dwarf galaxies and star clusters. As described in (Binney & Tremaine 2008), when the tidal forces acting on these objects are strong enough, stars get pulled out and end up on orbits which are slightly more/less energy than the progenitor’s orbit, forming the so-called leading/trailing arms. These narrow structures, referred to as stellar streams, have proven to be excellent tracers of fundamentals unknown in Galaxy evolution, like probing the dark matter halo in external galaxies (Walder et al. 2024) and our Galaxy (Küpper et al. 2015), charting dark matter subhalos in the Milky Way halo (Nibauer et al. 2024), or the potential property of dark matter itself (Mestre et al. 2024). In particular, the long, dynamically cold, GD-1 stream has been used to constrain the Milky Way potential with various techniques, like orbit fitting (Koposov et al. 2010; Malhan & Ibata 2019), backwards time integration (Price-Whelan et al. 2014; Palau et al. 2025),

action-angle modelling (Bovy et al. 2016), action-angle clustering (Reino et al. 2021) and particle-spray with stream-track (Bowden et al. 2015). More recently, new machine learning tools have been implemented to face this challenging problem, like in (Nibauer et al. 2022) and (Nibauer & Bonaca 2025), where they estimated the acceleration felt by the stars in the stream as a probe of the Milky Way potential, without relying on smooth analytic approximation and opening a new avenue for more realistic representation of the potential.

Recent developments in generative models have boosted the adoption of the Simulation Based Inference (SBI) technique (Cranmer et al. 2020) as a valid alternative to the classical Bayesian approaches in various fields like Gravitational Wave (Dax et al. 2025), Galactic chemical enrichment (Buck et al. 2025; Gunes et al. 2025), Galactic Archaeology (Viterbo & Buck 2024; Sante et al. 2025), dark matter density profile in dwarf galaxies (Nguyen et al. 2023), and cosmology (Saoulis et al. 2025). In this work, we set out to jointly recover the gravitational potential of the Milky Way together with the parameters

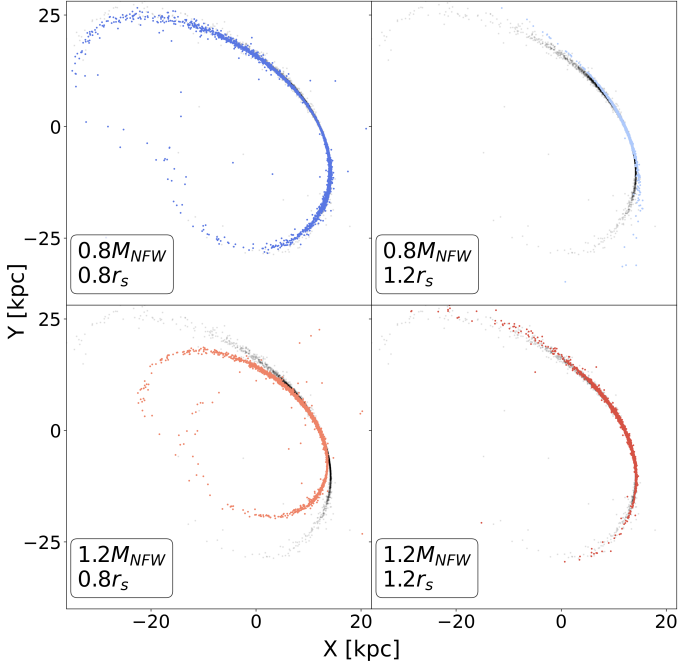


Fig. 1. Stream morphology in different Galactic potentials. The black scatter plot is the output of the simulation for the fiducial values of `BovyMWPotential2014` and the same in each sub-panel. We show the results for various combinations of 20% offset of the mass of the NFW halo and its scale radius as indicated in the legend of each sub-panel.

of the progenitor of the stellar stream, to compare and extend the work presented in (Alvey et al. 2024). This paper is structured as follows. In Sec. 2, we present the simulation choices, the N-body simulator `ODISSEO` used to create the training set, an introduction to the Flow Matching technique, and the details of the model architecture. In Sec. 3, we present the results of the inference for our reference GD1 simulation, and the test set results to validate the posterior calibration and accuracy. In Sec. 4, we summarize our findings and discuss limitations and future prospects.

2. Method

Our goal is to recover, in a statistically sound way, the physical parameters θ that govern the tidal stripping of a globular cluster evolving within a Milky Way-like gravitational field. Achieving this requires a set of modelling decisions: how we represent the host-galaxy potential (Sec. 2.1) and how we describe the progenitor system that seeds the stellar stream (Sec. 2.2). To infer θ , we adopt a fully Bayesian framework, more specifically, we adopt a Likelihood-Free approach (SBI; Sec. 2.4). SBI enables us to learn directly from simulations how the observed data d carry information about the underlying physical parameters. Specifically, we train a generative model¹ to approximate the posterior distribution $p(\theta | d)$ by automatically extracting informative summary statistics from the observations. The training dataset for this approach consists of pairs (d^i, θ^i) , where each synthetic observation d^i is produced through forward modelling: we draw parameters θ^i and pass them through a simulator S (`ODISSEO`, Sec. 2.3, yielding $d^i = S(\theta^i)$). The forward model must also incorporate observational caveats (e.g., uncertainty, selection effect), so that the forward process is equivalent to sam-

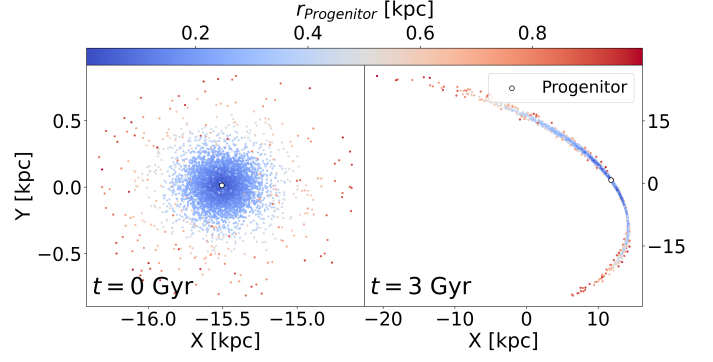


Fig. 2. Tidal stripping of $N=10^4$ stars from a Plummer sphere over a 3 Gyr evolution. The parameters to set this simulation were the fiducial `BovyMWPotential2014` for the Galaxy and $(M_{\text{Plummer}}, a_{\text{Plummer}}) = (10^{4.05} M_{\odot}, 100 \text{ pc})$ for the progenitor. The colorbar indicates the initial radial distance from the progenitor.

pling from the Likelihood $p(d | \theta)^2$. This simulation-inference amortized pipeline allows us to connect theoretical models of tidal stripping with the observable properties of stellar streams in a principled and scalable way.

2.1. The host: Model of the Milky Way

The gravitational potential of the Milky Way remains a subject of active debate. In the method outlined in the following section, we exploit the cold stellar stream GD-1 as a sensitive tracer of the Galaxy’s underlying density field. As illustrated in Fig. 1, even when the properties of the progenitor are fixed, tidal stripping can give rise to strikingly different present-day stream morphologies, depending on the assumed Galactic potential. We adopt the widely used `BovyMWPotential2014` model (Bovy 2014), which provides a smooth and flexible yet tractable three-component representation of the Milky Way.

This framework consists of:

1. A spherical dark-matter halo following a Navarro–Frenk–White profile (NFW) with density profile

$$\rho(r) = \frac{\rho_0}{\frac{r}{r_s} \left(1 + \frac{r}{r_s}\right)^2}, \quad (1)$$

where r is the spherical radii coordinates, r_s is the scale radius, and ρ_0 is the central density.

2. An axisymmetric disc described by a Miyamoto–Nagai (MN) with potential

$$\Phi(R, z) = -\frac{GM_{MN}}{\sqrt{R^2 + (a + \sqrt{z^2 + b^2})^2}}, \quad (2)$$

with R, z being the radii and height in cylindrical coordinates, M_{MN} is the total mass of the disc, and a, b are respectively the scale length and height.

3. A spherical bulge with an exponential cut-off with a density profile

$$\rho(r) = \rho_0 \left(\frac{r_1}{r_c}\right)^\alpha \exp(-(r/r_c)^2), \quad (3)$$

with r_c and α being respectively the cut-off radius and the power-law exponent.

¹ A Neural Network (NN) designed to learn conditional probability distributions.

² For this reason, the SBI technique is also referred to as Implicit Likelihood Inference.

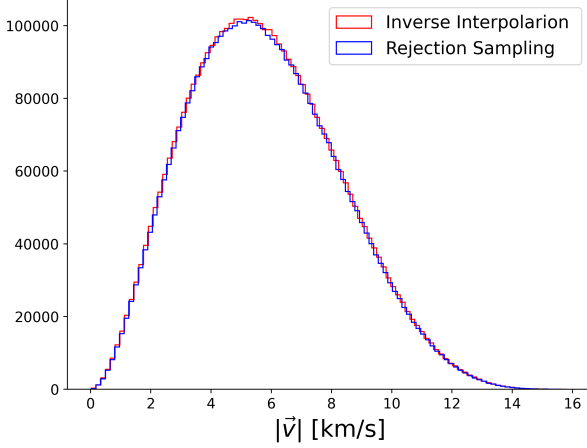


Fig. 3. Comparison between Rejection Sampling and Inverse Sampling of the interpolated inverse for the module of the velocity of particles sampled from a Plummer sphere.

2.2. The progenitor: Globular Cluster

The progenitor of GD-1 is widely attributed to be a disrupted globular cluster. Such systems are well approximated as spherically symmetric, self-gravitating stellar ensembles (Aarseth et al. 1974), for which a Plummer sphere provides an analytically convenient and physically motivated model. The Plummer potential captures both the central concentration and the finite spatial extent expected for low-mass, pressure-supported clusters, making it ideally suited for generating initial conditions in stream-formation simulations. A crucial aspect of modelling tidal disruption is that the internal phase-space distribution of stars within the progenitor directly shapes where and how stars are stripped. Stars located near the outer regions of the Plummer sphere—those with higher energies and larger orbital radii—reach the tidal boundary earlier and are therefore removed first as the cluster orbits within the Galactic potential. Conversely, tightly bound inner stars escape later and typically with different initial velocities (see e.g. Skúladóttir et al. 2025; Buder et al. 2025a,b, for a recent exploitation of this effect for Galactic Archaeology). We show an example of this process by evolving a Plummer sphere in the fiducial `BovyMWPotential2014` in Fig. 2.

These variations imprint distinct escape conditions along the orbit, influencing the width, density variations, and energy–angle structure of the resulting stream. Consequently, even subtle differences in the progenitor’s density profile or velocity dispersion propagate into observable differences in the stream morphology, making an accurate generative model of the progenitor essential for robust inference. In the following section, we describe in detail how we sample particles from a Plummer sphere to initialise the progenitor’s six-dimensional phase-space distribution. This sampling procedure forms one of the first steps of our forward model and ensures that the tidal stripping points—and thus the emergent stream structure—are physically consistent when generating the training data used for our SBI pipeline.

2.2.1. Differentiable Initial Condition

The initial positions and velocities of the progenitor depend inherently on the two parameters of the Plummer model: the total mass M_{Plummer} and its scale radius a_{Plummer} .

To sample star particles from a Plummer sphere, we need to sample (x, v) as follows:

1. Given the radial mass profile of the Plummer model:

$$M(r) = M_{\text{Plummer}} \left(\frac{r}{a_{\text{Plummer}}} \right)^3 \left(1 + \frac{r^2}{a_{\text{Plummer}}^2} \right)^{-3/2} \quad (4)$$

we apply inverse sampling, using $F(r) = M(r)/M_{\text{Plummer}}$ as cumulative distribution function, to obtain r by simply evaluating:

$$r = \sqrt{\frac{a_{\text{Plummer}}^2}{u^{-2/3} - 1}} \quad \text{where } u \sim \mathcal{U}(0, 1) \quad (5)$$

2. Given the radial distance r , and assuming spherical symmetry, we generate (x, y, z) by sampling a random direction on the unit sphere
3. Using the Plummer potential

$$\Phi_{\text{Plummer}}(r) = -\frac{GM_{\text{Plummer}}}{\sqrt{r^2 + a_{\text{Plummer}}^2}}, \quad (6)$$

We can find for each r what is the associated escape velocity $v_e(r)$ simply by setting the total energy of the particle to be 0, obtaining $v_e(r) = \sqrt{-2\Phi(r)}$. This is the maximum velocity that a gravitationally bound particle can have. In order to obtain the velocity, we can use the distribution function, which for the Plummer sphere is the closed form

$$g(v)dv \propto (-E)^{7/2} v^2 dv \quad (7)$$

where $E = -v_e^2 + \frac{1}{2}v^2$. We can then define $q(v) = v/v_e$ so that the previous expression can be rewritten as the unnormalized probability density function

$$g(q) = (1 - q)^{7/2} q^2 \quad \text{with } 0 \leq q \leq 1. \quad (8)$$

In this case, inverse sampling is not possible because the cumulative distribution function $G(q) = \int_0^q (1 - s)^{7/2} s^2 ds$ does not have an analytic inverse F , such that $F(G(q)) = q$. However, since $G(q)$ can be evaluated numerically, we approximate its inverse $\tilde{F}(u)$, with $u \sim \mathcal{U}(0, 1)$, by interpolating over a set of points $(G(q'), q')$, where q' are equally spaced values between 0 and 1. In practice, this amounts to swapping the input and output of G and constructing an interpolation of the inverse, which can then be sampled at the values of u to obtain our final samples for q . We report in Fig. 3 a comparison plot of sampling 10^5 particles showing how close it matches the classical rejection sampling approach presented in Aarseth et al. (1974). An advantage over the classical rejection sampling approach is that the inverse interpolation sampling is differentiable, allowing for gradient based approaches, which we could leverage easily with our differentiable simulator, as shown in (Viterbo & Buck 2025). In this work however, we do not yet make use of the differentiability.

4. Once we have sampled q we can rescale them, using v_e , to obtain the velocity module v and finally, assuming spherical symmetry, we generate (v_x, v_y, v_z) by sampling random direction on the unit sphere.

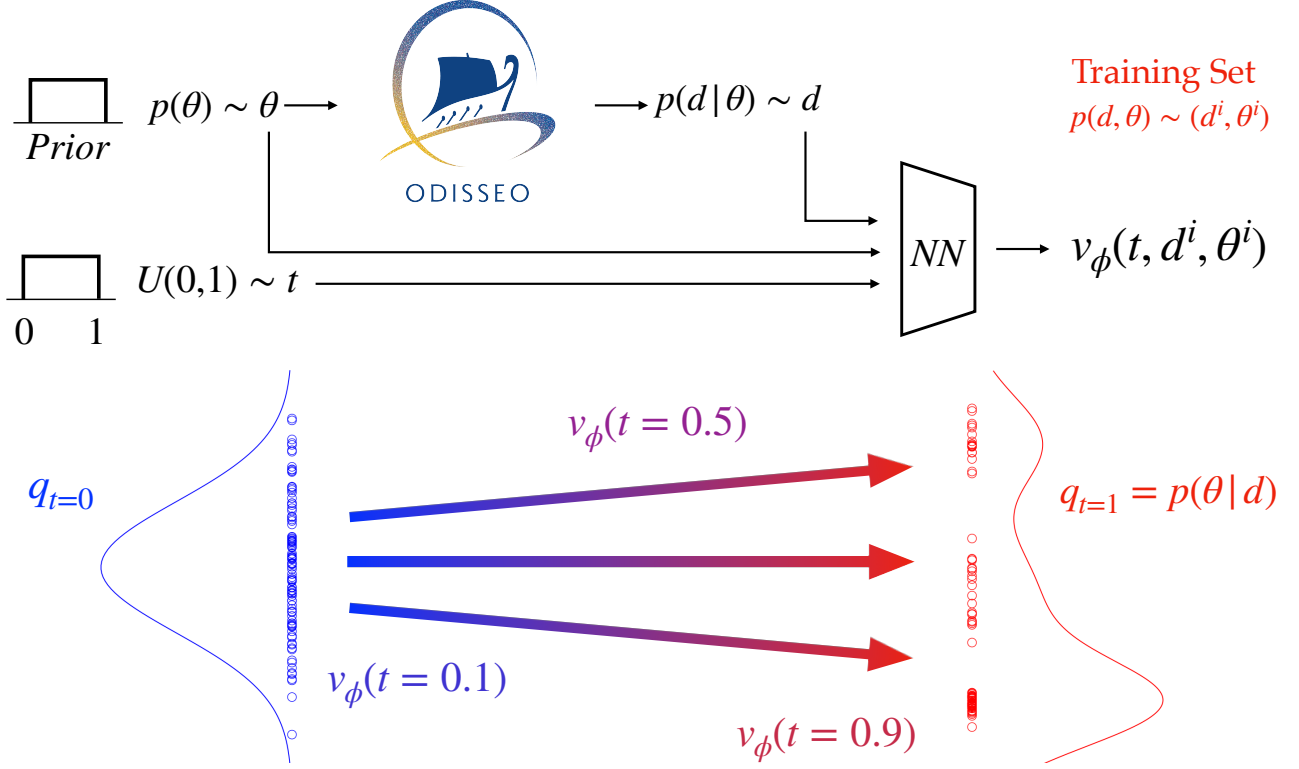


Fig. 4. Schematic of flow matching for posterior estimation with **ODISSEO**. The training set is generated by sampling $i = 0, \dots, N$ parameters θ^i from the prior $p(\theta)$, and then forward model using **ODISSEO** to obtain the observation $d^i \sim p(d | \theta)$. Following the flow described in Sec. 2.4.1, we sample $t \sim U(0, 1)$ to train a NN to approximate the vector field $v_\phi(t, d^i, \theta^i)$. In the lower section, we report a simplified Flow matching objective for a 1D case. The NN is called for different t to regress the vector field that governs the ODE to transform the sampling distribution $q_{t=0} = \mathcal{N}(0, I)$ into the posterior distribution $q_{t=1} = p(\theta | d)$. Note that in this schematic we refer to θ_1 described in Sec. 2.4.1 as θ .

2.3. N-body simulation with **ODISSEO**

Although fast particle-spray algorithms exist (e.g. Chen et al. 2025; Fardal et al. 2015; Nibauer et al. 2024), we model tidal disruption directly using the direct **ODISSEO** N-body integrator. This self-consistent approach captures the ejection dynamics due to close encounters that particle-spray techniques approximate³, and it produces forward simulations well suited for a simulation-based inference pipeline. As presented in (Viterbo & Buck 2025), **ODISSEO** is a N-body code developed to study particle systems in which external potentials play a crucial role. The code is implemented in a purely functional style in Jax (Bradbury et al. 2018), which ensures that the full simulation is trivially parallelizable on GPU, and the high-level Python interface allows for easy setup, prototyping, community-driven development, and maintenance. In addition, **ODISSEO** natively supports just-in-time compilation and execution on CPUs, GPUs, and TPUs, ensuring both flexibility and computational efficiency. We have used **ODISSEO** as our simulator to quickly generate the training set of mock GD1 simulations that are needed to train the model used in our SBI approach. Additionally, as described in (Viterbo & Buck 2025), the simulation can be differentiated by Automatic Differentiability (AD), allowing for gradient descent methods or variational inference for parameter estimation. In particular, the gradient can be pulled through the time integration, and also through the initial condition sampling, as described in 2.2. In this work, we do not use the differentiability.

³ To keep the computational cost tractable, we nevertheless adopt a small Plummer softening length, as described in the following section.

Parameter	Prior range	True value
Host Parameters		
$M_{\text{vir}} [M_\odot]$	$[0.5 M_{\text{vir}}^{\text{true}}, 2.0 M_{\text{NFW}}^{\text{true}}]$	4.37×10^{11}
$r_{\text{NFW}} [\text{kpc}]$	$[0.5 r_s^{\text{true}}, 2.0 r_s^{\text{true}}]$	16.0
$M_{\text{MN}} [M_\odot]$	$[0.5 M_{\text{MN}}^{\text{true}}, 2.0 M_{\text{MN}}^{\text{true}}]$	6.82×10^{10}
$a_{\text{MN}} [\text{kpc}]$	$[0.5 a_{\text{MN}}^{\text{true}}, 2.0 a_{\text{MN}}^{\text{true}}]$	3.0
Progenitor Parameters		
$t_{\text{end}} [\text{Gyr}]$	$[0.5, 5.0]$	3.0
$M_{\text{Plummer}} [M_\odot]$	$[10^{3.0}, 10^{4.5}]$	1.12×10^4
$a_{\text{Plummer}} [\text{kpc}]$	$[0.5 a_{\text{Plummer}}^{\text{true}}, 2.0 a_{\text{Plummer}}^{\text{true}}]$	0.008
$x^c [\text{kpc}]$	$[10.0, 14.0]$	11.8
$y^c [\text{kpc}]$	$[0.1, 2.5]$	0.79
$z^c [\text{kpc}]$	$[6.0, 8.0]$	6.4
$v_x^c [\text{km s}^{-1}]$	$[90.0, 115.0]$	109.5
$v_y^c [\text{km s}^{-1}]$	$[-280.0, -230.0]$	-254.5
$v_z^c [\text{km s}^{-1}]$	$[-120.0, -80.0]$	-90.3

Table 1. Prior ranges and true values for model parameters. The prior ranges for t_{end} , M_{Plummer} , x_c^0 , y_c^0 are taken from (Alvey et al. 2024).

2.3.1. Stellar streams simulation

In this section, we describe the general steps that were used to generate simulations of a stellar stream, which go as follows:

1. *Cluster trajectory.* Given the present day position and velocity $(\mathbf{x}^c, \mathbf{v}^c)$ for the progenitor of the star cluster, we trace the trajectory back to t_{end} , as a single particle with mass $M_{Plummer}$ in the chosen gravitational potential, to the initial phase space position $(\mathbf{x}_0^c, \mathbf{v}_0^c)$.
2. *Populate with star particles.* We draw $N=1000$ star particle's position and velocity $(\hat{x}_i, \hat{v}_i)_{i=0,\dots,N}$ from a Plummer potential centered on the origin, and then we shift them in phase space by $(\mathbf{x}_0^c, \mathbf{v}_0^c)$.
3. *Stream evolution.* We evolve the star particles forward in time for a total integration time of t_{end} using the 5th order explicit Runge–Kutta method Tsit5 ODE solver available in `diffraX` (Kidger 2022).

All the simulations have been carried out using `ODISSEO`⁴. We decided to treat the particles as phase-space tracers, so we used a Plummer softening of 0.1 pc to avoid the formation of dynamical binaries.

2.4. Inference

We aim to infer jointly the parameters that describe the progenitor (θ_{prog}) of a fiducial GD1 stream simulation and the parameters of the Milky Way potential (θ_{host}) in which the tidal stripping of its progenitor has happened. In practice, we aim to both reproduce the results obtained by Alvey et al. (2024) and to extend the inference to also be able to use the stellar stream as a tracer for the gravitational potential of the host galaxy. We have decided to face this challenging task by training a Neural Density Estimator (NDE) ((Cranmer et al. 2020)) of the posterior distribution $p(\theta \mid d_{obs})$, where $\theta = (\theta_{prog}, \theta_{host})$ and $d_{obs} = (\phi_1, \phi_2, r, v_{\phi_1} \cos \phi_2, v_{\phi_2}, v_r)$ is the phase space of all the stars in the GD1 stellar stream projected on the plane of the stream, like in Alvey et al. (2024) and Koposov et al. (2010)⁵. Since we modelled the progenitor of the GD1 with a Plummer sphere (Sec. 2.2), the parameters are going to be the total time of integration, total mass and the scale radius of the Plummer sphere, its present day position and velocity so that $\theta_{prog} = (t_{end}, M_{Plummer}, a_{Plummer}, \mathbf{x}^c, \mathbf{v}^c)$. Since we wanted to test the constraining power over the amplitude and shape of the host potential, we vary only the Navarro-Frenk-White halo mass and scale radius and the Miyamoto-Nagai mass and scale length, while fixing all the other parameters to the default fiducial parameters in `Galax` (Starkman et al. 2024) for `MWPotential2014`, to contain the computational cost of sampling on a small but still informative parameter space, leaving for future work extension to models with higher degrees of freedom like triaxial NFW profiles or time evolving potentials. The final host potential parameters are then $\theta_{host} = (M_{NFW}, r_{NFW}, M_{MN}, a_{MN})$. To train a NDE we need to generate couples of parameters and observation $(\theta, d) \sim p(\theta, d)$, by letting $\theta \sim p(\theta)$, with $p(\theta)$ being the prior over the parameters, and $d = S(\theta) \sim p(d \mid \theta)$, with S being the `ODISSEO` simulator which implicitly define the Likelihood $p(d \mid \theta)$. Moreover, we applied to the observation the same observational window and noise level reported in Tab. 2 in (Alvey et al. 2024), leaving the background contamination and selection functions effect for future work. The prior choice is reported in Tab. 1.

⁴ The data are publicly available at <https://zenodo.org/records/17711491>

⁵ These are the standard set of co-ordinates used in the literature: (ϕ_1, ϕ_2) are two angles coordinate, the corresponding proper motion (v_{ϕ_1}, v_{ϕ_2}) , and radial distance and velocities (r, v_r) .

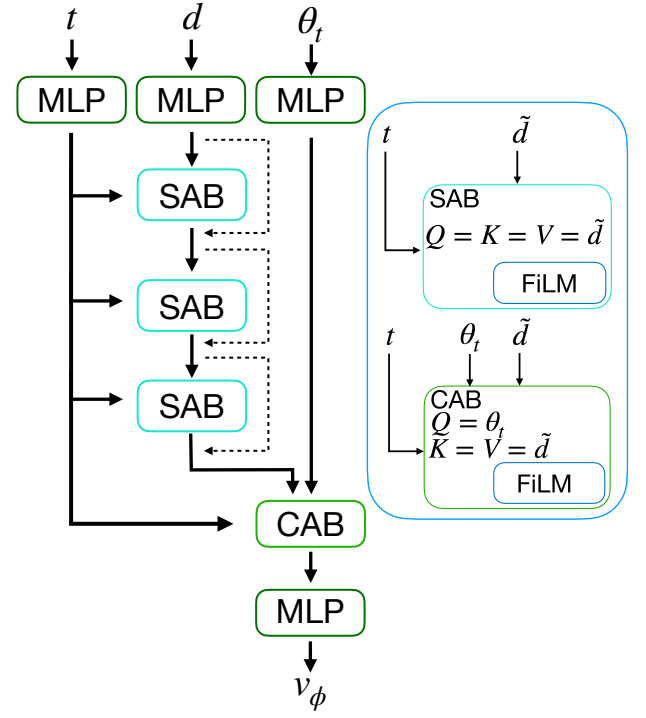


Fig. 5. Flow matching SetTransformer. We indicate with $\tilde{d}$ both the observation d and the intermediate output of the layers. To embed each of the stars in the observation d , the parameter state θ_t at the ODE time t we have used a simple MLP with 128 neurons with a SiLU activation function. Then we passed $\tilde{d}$ through 3 stacked SAB with skip connection (dashed line) to encode the correlations between the particles, modulating the output of each block on t using FiLM. Then we used a CAB, with FiLM modulation, to focus the Attention mechanism on finding the relevant feature in $\tilde{d}$ to regress the parameters θ . The vector field v_ϕ is obtained by compressing the output of the CAB through an MLP with output dimension equal to the dimensionality of θ , in our case 13 dimensions.

2.4.1. Flow Matching Posterior Estimation

In recent years, many SBI applications, performed using Normalizing Flow architecture as Neural Posterior Estimator, have ((Nguyen et al. 2023), (Sante et al. 2025), (Viterbo & Buck 2024), (Ho et al. 2024)). The core idea is to leverage the flexibility of a Neural Network to learn a series of invertible transformations, conditioned on the observations, that project the parameters in a latent space where it is easy to sample from. In this way, instead of relying on the expensive gold standard MCMC, the transformations, via the change of variable formula, take care of keeping track of how the parameter space has been changed through the flow to sample and evaluate the posterior distribution. The expressivity of this technique is limited by the necessity of using an invertible transformation (usually spline functions as described in Durkan et al. 2019). For this reason, Neural Posterior Score Estimation (NPSE) ((Geffner et al. 2022)) and Flow Matching Posterior Neural Estimation (FMPE) (Dax et al. (2023)) tackle the problem of approximating a conditional distribution with a different approach. As described in Dax et al. (2023), inspired by the promising results in generative tasks for which they were developed, these models transform noise into samples via trajectories parametrized by a continuous time variable t . In particular, for FMPE, the goal is to regress the vector field v_ϕ , parametrized by a Neural Network with weights ϕ , that describes these trajectories by solving an ordinary differential equation (ODE). The key advantage over Normalizing Flows is

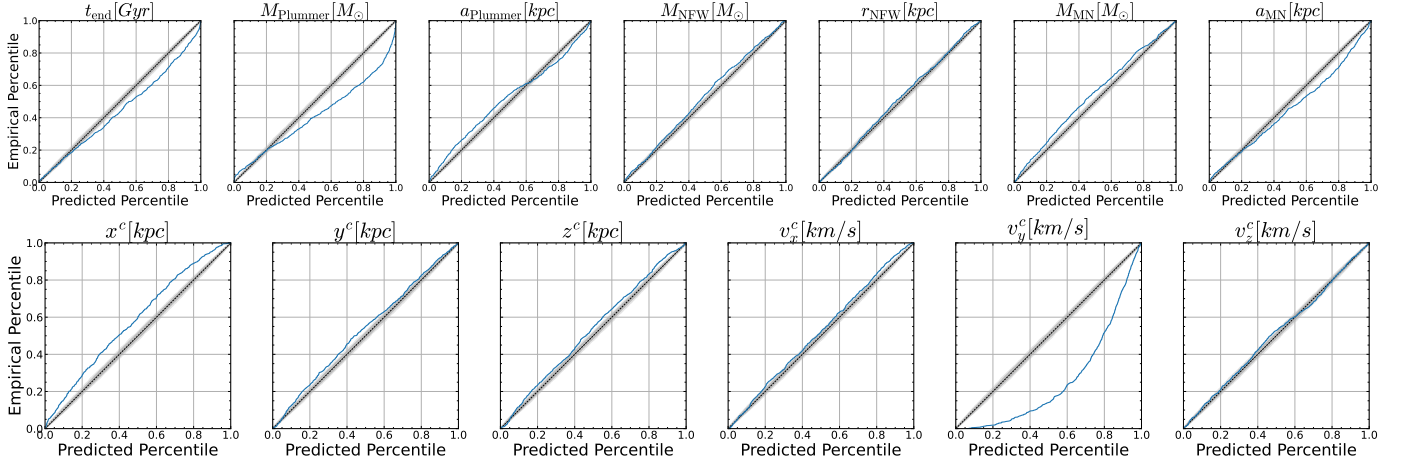


Fig. 6. Percentile-Percentile plot for the marginal posterior distributions over the test set.

that by regressing on the vector field v_t , we are free in the choice of the network architecture, at the cost of multiple network passages for sampling, since we need to solve the ODE for $t \in [0, 1]$. Moreover, contrary to NPSE, which needs to solve a Stochastic Differential Equation (SDE), we are also able to track the posterior density directly. As described in Sec. 2.4, we are going to train on tuples (d, θ) , respectively observations and parameters. Our objective is to learn a continuous transformation for θ (or equivalently a probability path q_t) with constraints $\theta \sim q_0$ and $\theta \sim q_1 = p(\theta | d)$, where q_0 is a sampling distribution, in our case a normal distribution with zero mean and identity covariance matrix. The ODE that controls this continuous process can then be expressed as:

$$d\theta_t = v_\phi(t, d, \theta_t)dt. \quad (9)$$

Flow Matching actually falls in the paradigm of continuous Normalizing Flow, with an alternative objective function. In fact, since continuous Normalizing Flows are trained using negative log likelihood as objective, it is required to track at training time the computationally expensive $p(\theta | x)$ by

$$p(\theta | x) = q_1 = q_0 \exp \left(\int_0^1 \nabla \cdot v_\phi(t, d, \theta) dt \right), \quad (10)$$

making the training of these models not feasible. As described in (Dax et al. 2023), the key of Flow Matching is to directly regress the vector field v_ϕ on a vector field u_t that generates the desired probability path p_t , avoiding the integration of the ODE at training time⁶. The non-trivial solution on how to perform this task, presented in Lipman et al. (2023), is to choose this path depending on the sample θ_1 ⁷ for which we want to model the probability path. In practice, we are going to model the conditioned probability path $q_t(\theta | \theta_1)$, and the corresponding vector field $u_t(\theta | \theta_1)$ with the sample-conditional flow matching loss (CFM)

$$\mathcal{L}_{CFM} = \mathbb{E}_{t \sim U[0,1], d \sim p(d|\theta), \theta_t \sim q_t(\theta|\theta_1)} \|v_\phi(t, d, \theta_t) - u_t(\theta_t | \theta_1)\|^2. \quad (11)$$

One simple choice for the conditioned probability path is the Gaussian path family

$$q_t(\theta | \theta_1) = \mathcal{N}(\theta | t\theta_1, (1 - (1 - \sigma_{min})t)\mathbb{I}) \quad (12)$$

⁶ The integration is still needed at inference time.

⁷ We indicate with θ_1 the samples θ to be consistent with the literature on Flow Matching.

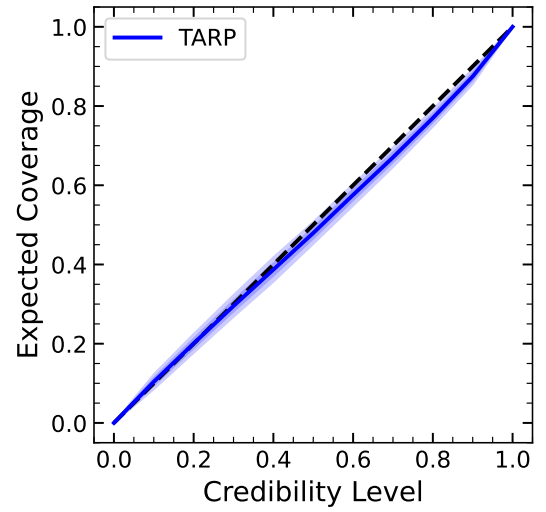


Fig. 7. Tarp plot for the joint posterior distribution over the test set

which generates the velocity field

$$u_t(\theta | \theta_1) = \frac{\theta_1 - (1 - \sigma_{min})\theta}{1 - (1 - \sigma_{min})t}, \quad (13)$$

with $\sigma_{min} > 0$. These choices lead to this problem coinciding with the optimal transport (Lipman et al. 2023) between two Gaussian distributions: the sampling distribution and linear trajectory $t\theta_1$, ending in θ_1 with a smoothing constant given by σ_1 . The steps to be followed to train boil down to sampling $\theta_1 \sim p(\theta)$, and transporting a point $\theta_0 \sim \mathcal{N}(0, \mathbb{I})$ from the sampling distribution to the posterior distribution on the linear trajectory $t\theta_1$ and ending in θ_1 .

In the upper part of Fig. 4, we summarize the SBI task: we first sample from the prior parameters θ , forward model them using `ODISSEO` to get the observation d , and use the couples (d, θ) as our training set. In the lower part of Fig. 4, we show a representation of what the inference task looks like: we model the vector field v_ϕ with a NN that is called for each integration step ($t \in [0, 1]$) of the ODE that transport the sampling distribution $\mathcal{N}(0, \mathbb{I})$ into the posterior distribution $p(\theta | d)$

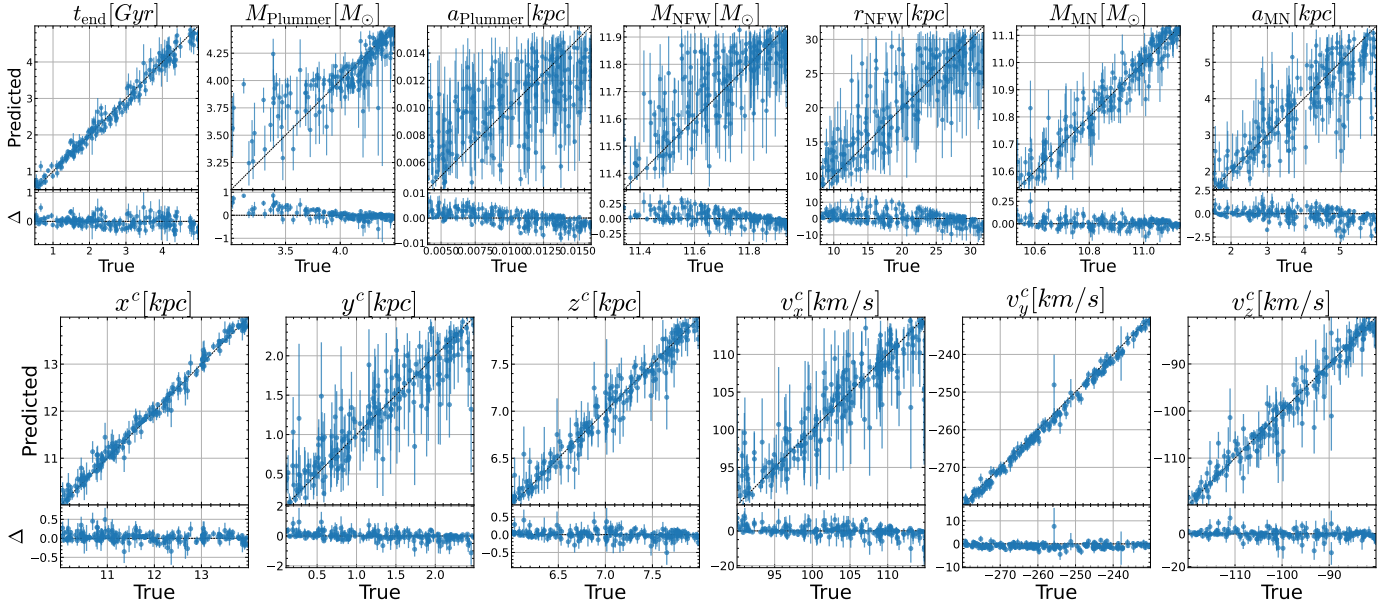


Fig. 8. True-Predicted plots for the test set. We have subsampled the test set to 500 for visual clarity. The circles represent the median, while the error bars report the 16-84 percentiles. We also report the residuals Δ , with error bars being the 16-84 percentiles.

2.5. Architecture: Set Transformer

Given the permutation-invariant nature of our data—the set of N -body particles—we adopt the SetTransformer architecture introduced by (Lee et al. 2019). Our model extends this framework to infer the vector field v_ϕ , conditioned on the flow matching integration time t , the model parameters θ , and the particle observations d .

Each input is first embedded into a latent representation via Multi-Layer Perceptrons (MLPs). The integration time t is further encoded through a sinusoidal time embedding following (Ho et al. 2020). The embedded particle features are processed by a stack of Self-Attention Blocks (SABs) with residual connections, which capture interactions among particles. To incorporate temporal conditioning, the output of each SAB is modulated via Feature-wise Linear Modulation (FiLM; (Perez et al. 2017)) using the embedded time representation. Before producing the vector field, a Cross-Attention layer is applied, where the query Q corresponds to the embedded parameters θ , and the keys K and values V are given by the SAB+FiLM representation of the particle set. In this configuration, the model parameters act as queries that guide the attention toward relevant features within the particle embeddings. The Cross-Attention output is then modulated again through a FiLM layer conditioned on the embedded time. Finally, the predicted vector field v_ϕ is obtained via a last MLP layer. The architecture is described in Fig. 5.

2.6. Training details

We generated 2×10^5 stellar streams, each containing a fixed number of stars ($N = 1000$). Of these, 5×10^4 samples were used for validation and 10^3 for testing. Training employed early stopping with a patience of 30 epochs, resulting in a total of 123 training epochs. We used a batch size of 500 and optimized the model with AdamW, starting from a learning rate of 10^{-3} and applying a reduce-on-plateau schedule with a patience of 5 epochs. The final architecture, illustrated in Fig. 5, was selected after extensive experimentation with different design choices and hyperparameters, including the number of neurons, the num-

ber of SAB and CAB blocks, activation functions, modulation schemes, and related configurations. The training set generation took ~ 7.5 GPU hours on an NVIDIA H200 (140 GB). The training and testing of the FMPE took ~ 8 hours on an NVIDIA A100 (40 GB). All the inference pipeline was carried out using the sbi-sim⁸ package presented in Holzschuh & Thuerey (2024).

3. Results

We validate the results of our inference by reporting a few metrics obtained on the test set. This set of (d, θ) was not used during training. For each of the test set couples, we sample 10^3 samples from the posterior distribution. In order to evaluate accuracy and calibration over the whole test set, we report in Sec 3.1.1 and 3.1.2, respectively, predicted-true plots and percentile-percentile plots. Then, in section Sec 3.2 we report the posterior's samples obtained from inferring on our mock observation of the GD1 stream. For this test, we decided to sample 10^4 times the posterior. Moreover, we perform posterior predictive checks to show that by forward passing these posterior samples, we obtain an observation that resembles the mock observation of GD1.

3.1. Model evaluation

3.1.1. Calibration

Percentile-percentile (P-P) plots are a diagnostic tool used to evaluate the calibration of posterior distributions in probabilistic inference. The key concept is that a well-calibrated model should produce posteriors whose credible intervals contain the true parameter values with the correct empirical frequency. For the marginal distributions of the predicted posterior, P-P plot visualizes this property by comparing the predicted percentiles, under the inferred posterior, against the empirical percentiles (the fraction of true parameters that fall below the corresponding posterior quantile). Ideally, if the posterior is perfectly calibrated, these two quantities coincide, and the curve follows the

⁸ <https://github.com/tum-pbs/sbi-sim/tree/dev>

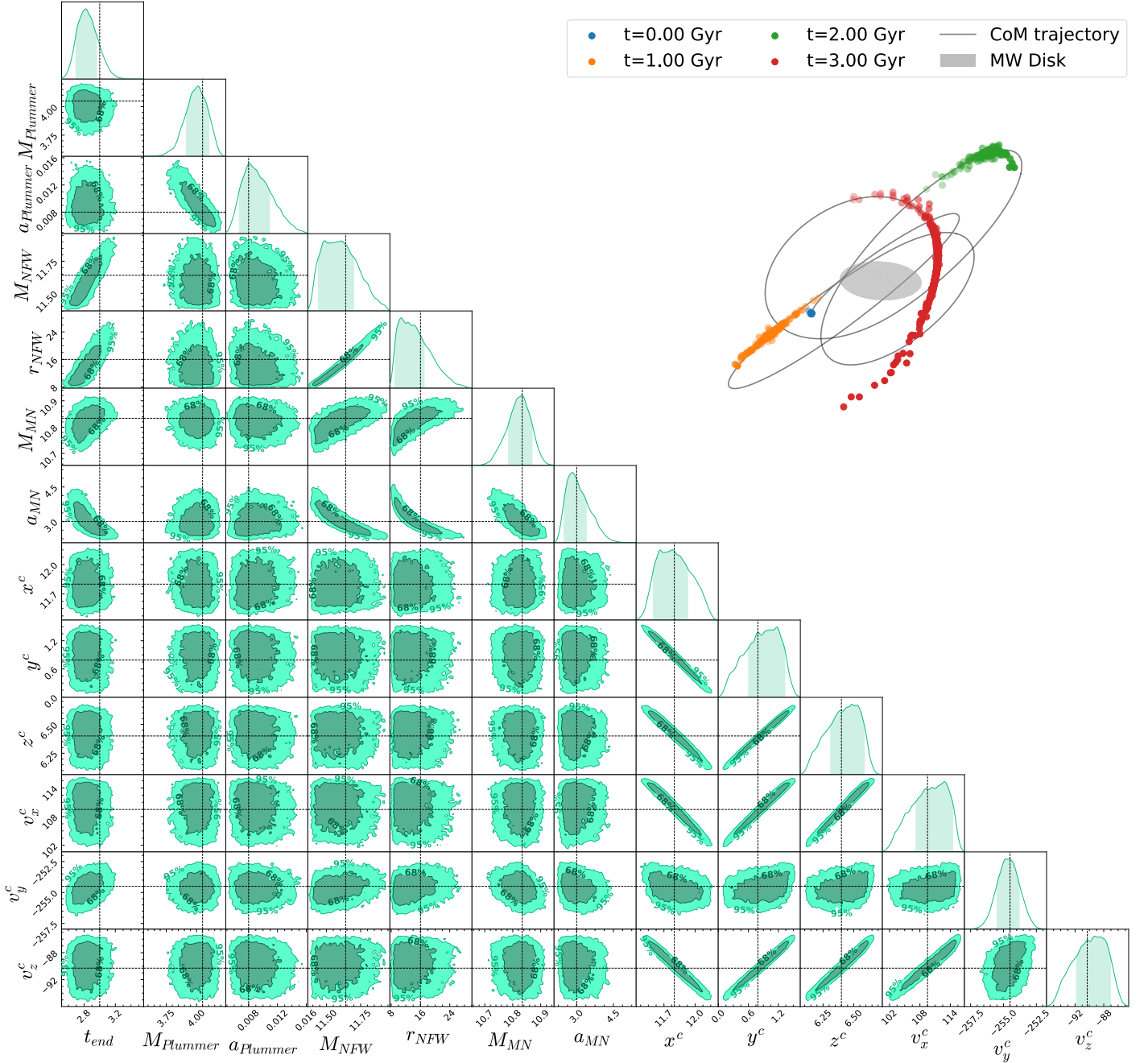


Fig. 9. Lower left: Posterior samples from the fiducial GD1 observation. The dashed line corresponds to the True value column in Tab. 1. Upper right: snapshots of our fiducial GD1 simulation. We reported an 8 kpc Milky Way disk for reference.

diagonal line. By inspecting these plots and their deviations from the diagonal, we can spot deviations from the true posterior distributions, like over/under confidence or positive/negative biases. We report the marginal coverage plots for individual parameters in Fig. 6 and the joint coverage plot in Fig. 7. We can appreciate that for the θ_{host} the P-P plots suggest a good calibration. The two S-shape behaviour in Plummer parameters suggest a possible bias, which for the case of $M_{Plummer}$ is a positive bias as reflected by the over-estimation shown in Fig. 8. Moreover a significant under (over) confidence is shown for x^c (v_y^c).

But marginal posterior coverages do not tell the whole story. We adopt the "Tests of Accuracy with Random Point" (TARP)

to study the behavior of the joint posterior distribution. As described in Lemos et al. (2023), TARP provides a necessary and sufficient condition for posterior coverage. The method constructs spherical credible regions around randomly chosen points in parameter space and measures the fraction of posterior samples contained within each region. By repeating this process over many random points, TARP estimates the empirical coverage as a function of the nominal credibility level. A perfectly calibrated posterior yields a one-to-one correspondence between nominal and empirical coverage, represented by a diagonal trend in the resulting Tarp in Fig. 7 to show we

have good coverage on the joint distribution evaluated over the test set.

3.1.2. Accuracy

After assessing that our posterior is well calibrated on almost all the marginals and the joint, we report the accuracy that we can expect by using our model in the form of True-Predicted plots. In Fig. 8, the circles represent the median value, and the error bars represent the 16-84 percentile intervals. Moreover, in the lower plots, we report the distribution of residuals to capture any possible residual trends. Our model seems to be extremely capable of capturing the posterior distribution for most of the parameters θ , except for the M_{Plummer} and a_{Plummer} for which a mostly uniform trend seems to emerge when we inspect the residuals distribution. We interpret the mass bias as the fact that under $\sim 10^3 M_\odot$ the feature that the progenitor can imprint of the stars are limited and the stars are mostly dominated by the host potential. The Plummer scale a_{Plummer} seems to play a less important role compare to other quantities and the results of our analysis could be limited by the choice of a not enough broad prior.

3.2. GD1

In the following section, we show the posterior distribution for our mock true observation of the GD1 stream. This simulation is meant to allow for a direct comparison with the results of (Alvey et al. 2024), even though the set of parameters common between the two approaches is limited to $(t_{\text{end}}, M_{\text{Plummer}}, \mathbf{x}^c, \mathbf{v}^c)$. In Fig. 9, we report the corner plot for the posterior. The results suggest that we can strongly constrain these parameters, since the true value always lies inside the 16-84 percentile, and that we can reproduce the results on Alvey et al. (2024). We also capture the correlation between the host potential parameters (e.g. $(M_{\text{NFW}}, r_{\text{NFW}})$, $(M_{\text{MN}}, a_{\text{MN}})$, ...); these correlation are expected since different combination leading to the same enclosed mass at a given radius. We can appreciate also the correlations in the phase space $(\mathbf{x}^c, \mathbf{v}^c)$, reflection of the spherical symmetry of the problem. In agreement with the literature ((Koposov et al. 2010), (Nibauer et al. 2022), (Palau et al. 2025)), these results show that stellar streams can probe the shape and amplitude of the gravitational potential.

3.2.1. Posterior Predictive Check

As a complementary check, we forward pass the posterior samples from Fig. 9 to obtain $d \sim P(d | \theta_{\text{Posterior}})$. We aim to capture biases introduced by the inference, which would be reflected as deviation of $d \sim P(d | \theta_{\text{Posterior}})$ from d_{True} . For each of the samples in Fig. 9, we generate a stellar stream with 1000 stars and simulate the tidal disruption as described in Sec. 2.3.1. We then combine all the streams and check if the distribution in observable space d obtained by this procedure deviates from the stars in our mock observation. We report the results in Fig. 10. In this plot, the "ground truth" is not a single point but a distribution of particles, so the black contours are given by displaying 1000 test set simulation, each with 1000 star particles. We can appreciate that all major stream features are captured nicely by our model. Moreover, in Fig. 11 we report the observation obtained by forward modelling our estimate of the parameters, the posterior mean, obtained by the $N = 1000$ samples in our posterior as $\langle \theta_{\text{Posterior}} \rangle = 1/N \sum_{i=0}^N \theta_i$. Also in this test, we can see strong agreement with the mock observation.

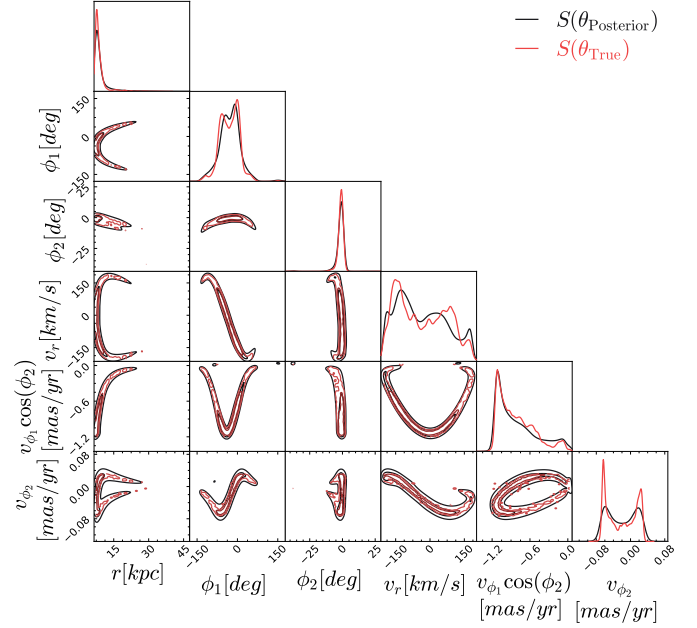


Fig. 10. Posterior Predictive Check: We report in red our mock observation of the GD1 stream obtained using the True values in Tab. 1, and in black we report the joint results of forward modelling the 1000 samples in the posterior distribution.

4. Conclusion

With this work, we have presented the first use of Simulation Based Inference applied to the GD1 stream to jointly constrain parameters of the progenitor and the host potential. With the use of the **ODISSEO**, we ran a large set of mock GD1 stream analogues to reproduce the tidal stripping of a Globular cluster, which we have used to train our neural density estimator. We adopted a **SetTransformer** architecture to automatically extract summary statistics from the observations x , enabling a robust inference implemented via a Flow Matching strategy. Our results are summarized below.

- We produce a large set of publicly available mock GD1 streams using the **ODISSEO** simulator.
- We have extended the work of (Alvey et al. 2024) to also study the gravitational potential of the host galaxy, achieving an amortized inference that does not require sequential training.
- We provide the validity of our model with extensive testing in the form of P-P plots for coverage, True-Predicted plots for accuracy and Posterior Predictive Checks for self-consistency.
- We recover the fiducial parameters over our fiducial GD1 observation in Fig 9 for this controlled experiment; however, performance under more realistic survey conditions remains to be tested. We capture strong correlations between masses and scale lengths parameters $(M_{\text{NFW}} - r_{\text{NFW}}; M_{\text{MN}} - a_{\text{MN}}; r_{\text{NFW}} - a_{\text{MN}})$, and surprisingly a weak positive correlation between the masses $(M_{\text{NFW}} - M_{\text{MN}})$. We also capture a strong correlation between the present-day phase space of the progenitor of GD1 $(\mathbf{x}^c, \mathbf{v}^c)$, a result of the axisymmetry of the problem. A notable exception is v_y^c , for which our model is also not well calibrated, as we can observe in Fig. 6.
- We leverage the freedom given by the Flow Matching technique to adopt a robust Transformer architecture, capable of handling this high dimensional inference task.

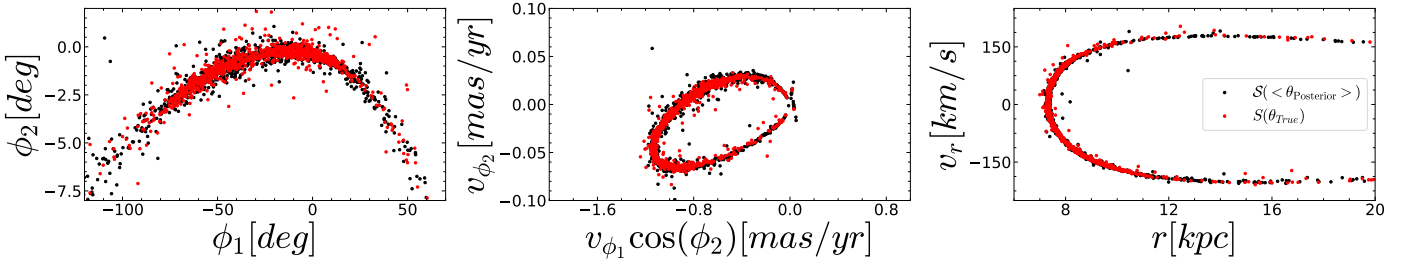


Fig. 11. Posterior Predictive Check: we report in red our mock observation of the GD1 stream just like in Fig. 10, and in black the forward model of the mean of the posterior samples, which is our estimate of the parameters θ .

Our aim is to quantify, in a controlled simulation setting, the information that a single cold stellar stream (GD-1 analogues) carries about both the progenitor properties and the host galaxy potential. We do this by combining self-consistent N-body forward modelling (ODISSEO) with amortized, likelihood-free inference (Flow Matching), and by validating calibration and accuracy on held-out synthetic data. In future work, we will extend this method with a more robust, realistic and survey-dependent handling of the observational errors, which we have simplified in this work, magnitude-dependent selection functions, and background contamination, which were both ignored. Also, we intend to extend the pipeline to leverage multi-stream observation, inspired by the results obtained in (Bovy et al. 2016). Lastly, since ODISSEO is capable of calculating the gradient of the simulation with respect to the input parameters, we aim to incorporate this as additional information to guide the inference pipeline, as shown in (Holzschuh & Thuerey 2024).

5. Code availability

We publicly release our code to reproduce all the figures via Github: https://github.com/vepe99/sbi-sim/tree/odisseo_branch.

Acknowledgements. This project was made possible by funding from the Carl-Zeiss-Stiftung. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). We acknowledge the usage of the AI-clusters Tom and Jerry funded by the Field of Focus 2 of Heidelberg University. We are grateful to Paola Ziero for the design of the project logo and their creative support.

References

Aarseth, S. J., Henon, M., & Wielen, R. 1974, *Astronomy and Astrophysics*, 37, 183, publisher: EDP ADS Bibcode: 1974A&A....37..183A
 Alvey, J., Gerdes, M., & Weniger, C. 2024, *Albatross: A scalable simulation-based inference pipeline for analysing stellar streams in the Milky Way*, arXiv:2304.02032 [astro-ph]
 Binney, J. & Tremaine, S. 2008, *Galactic Dynamics: Second Edition* (Princeton University Press)
 Bovy, J. 2014, *The Astrophysical Journal*, 795, 95
 Bovy, J., Bahmanyar, A., Fritz, T. K., & Kallivayalil, N. 2016, *The Astrophysical Journal*, 833, 31
 Bowden, A., Belokurov, V., & Evans, N. W. 2015, *Monthly Notices of the Royal Astronomical Society*, 449, 1391, arXiv:1502.00484 [astro-ph]
 Bradbury, J., Frostig, R., Hawkins, P., et al. 2018, *JAX: composable transformations of Python+NumPy programs*
 Buck, T., Günes, B., Viterbo, G., Oliver, W. H., & Buder, S. 2025, *A&A*, 702, A184
 Buder, S., Buck, T., Skúladóttir, Á., et al. 2025a, arXiv e-prints, arXiv:2510.11284
 Buder, S., Buck, T., Skúladóttir, Á., et al. 2025b, arXiv e-prints, arXiv:2510.20233
 Chen, Y., Valluri, M., Gnedin, O. Y., & Ash, N. 2025, *ApJS*, 276, 32

Cranmer, K., Brehmer, J., & Louppe, G. 2020, *Proceedings of the National Academy of Science*, 117, 30055
 Dax, M., Green, S. R., Gair, J., et al. 2025, *Nature*, 639, 49
 Dax, M., Wildberger, J., Buchholz, S., et al. 2023, *Flow Matching for Scalable Simulation-Based Inference*, arXiv:2305.17161 [cs]
 Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. 2019, *Neural Spline Flows*, arXiv:1906.04032 [stat]
 Fardal, M. A., Huang, S., & Weinberg, M. D. 2015, *MNRAS*, 452, 301
 Geffner, T., Papamakarios, G., & Mnih, A. 2022, arXiv e-prints, arXiv:2209.14249
 Günes, B., Buder, S., & Buck, T. 2025, arXiv e-prints, arXiv:2507.05060
 Ho, J., Jain, A., & Abbeel, P. 2020, *Denosing Diffusion Probabilistic Models*, arXiv:2006.11239 [cs]
 Ho, M., Bartlett, D. J., Chartier, N., et al. 2024, *The Open Journal of Astrophysics*, 7, 54
 Holzschuh, B. & Thuerey, N. 2024, *Flow Matching for Posterior Inference with Simulator Feedback*, arXiv:2410.22573 [cs]
 Kidger, P. 2022, *On Neural Differential Equations*, arXiv:2202.02435 [cs]
 Koposov, S. E., Rix, H.-W., & Hogg, D. W. 2010, *The Astrophysical Journal*, 712, 260
 Küpper, A. H. W., Balbinot, E., Bonaca, A., et al. 2015, *The Astrophysical Journal*, 803, 80
 Lee, J., Lee, Y., Kim, J., et al. 2019, *Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks*, arXiv:1810.00825 [cs]
 Lemos, P., Coogan, A., Hezaveh, Y., & Perreault-Levasseur, L. 2023, *Sampling-Based Accuracy Testing of Posterior Estimators for General Inference*, arXiv:2302.03026 [stat]
 Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. 2023, *Flow Matching for Generative Modeling*, arXiv:2210.02747 [cs]
 Malhan, K. & Ibata, R. A. 2019, *Monthly Notices of the Royal Astronomical Society*, 486, 2995
 Mestre, M. F., Argüelles, C. R., Carpintero, D. D., Crespi, V., & Krut, A. 2024, *Astronomy & Astrophysics*, 689, A194, arXiv:2404.19102 [astro-ph]
 Nguyen, T., Mishra-Sharma, S., Williams, R., & Necib, L. 2023, *Phys. Rev. D*, 107, 043015
 Nibauer, J., Belokurov, V., Cranmer, M., Goodman, J., & Ho, S. 2022, *The Astrophysical Journal*, 940, 22, arXiv:2205.11767 [astro-ph]
 Nibauer, J. & Bonaca, A. 2025, *The Astrophysical Journal Letters*, 985, L22
 Nibauer, J., Bonaca, A., Spergel, D. N., et al. 2024, *StreamSculptor: Hamiltonian Perturbation Theory for Stellar Streams in Flexible Potentials with Differentiable Simulations*, arXiv:2410.21174 [astro-ph]
 Palau, C. G., Wang, W., & Han, J. 2025, *Constraining the potential of the Milky Way using stellar streams and the Inverse Time Integration method*, arXiv:2412.15091 [astro-ph]
 Perez, E., Strub, F., Vries, H. d., Dumoulin, V., & Courville, A. 2017, *FiLM: Visual Reasoning with a General Conditioning Layer*, arXiv:1709.07871 [cs]
 Price-Whelan, A. M., Hogg, D. W., Johnston, K. V., & Hendel, D. 2014, *The Astrophysical Journal*, 794, 4
 Reino, S., Rossi, E. M., Sanderson, R. E., et al. 2021, *Monthly Notices of the Royal Astronomical Society*, 502, 4170, arXiv:2007.00356 [astro-ph]
 Sante, A., Kawata, D., Font, A. S., & Grand, R. J. J. 2025, *MNRAS*, 542, 1776
 Saoulis, A. A., Piras, D., Jeffrey, N., et al. 2025, *MNRAS*, arXiv:2505.21215
 Skúladóttir, Á., Ermandes, H., Feuillet, D. K., et al. 2025, *ApJ*, 986, L21
 Starkman, N., Price-Whelan, A., & Nibauer, J. 2024, *GalacticDynamics/galax: v0.0.2*
 Viterbo, G. & Buck, T. 2024, arXiv e-prints, arXiv:2411.17269
 Viterbo, G. & Buck, T. 2025, arXiv e-prints, arXiv:2511.22468
 Walder, M., Erkal, D., Collins, M., & Martinez-Delgado, D. 2024, *Probing the dark matter haloes of external galaxies with stellar streams*, arXiv:2402.13314 [astro-ph] version: 1