Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **509,** 5992–6007 (2022) https://doi.org/10.1093/mnras/stab3372
Advance Access publication 2021 November 24

VIA MACHINAE: Searching for stellar streams using unsupervised machine learning

David Shih [1]★ Matthew R. Buckley,[1] Lina Necib [2,3,4] and John Tamanas[5]

[1]*NHETC, Department of Physics and Astronomy, Rutgers, Piscataway, NJ 08854, USA*
[2]*Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, CA 91125, USA*
[3]*Center for Cosmology, Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA*
[4]*Observatories of the Carnegie Institution for Science, 813 Santa Barbara St., Pasadena, CA 91101, USA*
[5]*Department of Physics, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA*

Accepted 2021 November 17. Received 2021 November 15; in original form 2021 July 12

ABSTRACT

We develop a new machine learning algorithm, VIA MACHINAE, to identify cold stellar streams in data from the *Gaia* telescope. VIA MACHINAE is based on ANODE, a general method that uses conditional density estimation and sideband interpolation to detect local overdensities in the data in a model agnostic way. By applying ANODE to the positions, proper motions, and photometry of stars observed by *Gaia*, VIA MACHINAE obtains a collection of those stars deemed most likely to belong to a stellar stream. We further apply an automated line-finding method based on the Hough transform to search for line-like features in patches of the sky. In this paper, we describe the VIA MACHINAE algorithm in detail and demonstrate our approach on the prominent stream GD-1. Though some parts of the algorithm are tuned to increase sensitivity to cold streams, the VIA MACHINAE technique itself does not rely on astrophysical assumptions, such as the potential of the Milky Way or stellar isochrones. This flexibility suggests that it may have further applications in identifying other anomalous structures within the *Gaia* data set, for example debris flow and globular clusters.

Key words: stars: kinematics and dynamics – galaxy: stellar content – galaxy: structure.

1 INTRODUCTION

Stellar streams, the tidally stripped remnants of dwarf galaxies and globular clusters, provide a unique window into the properties of the Milky Way and its formation history. Streams trace the historical record of the mergers that built the Milky Way (Johnston 1998; Helmi & White 1999; Belokurov et al. 2006; Belokurov et al. 2018; Helmi et al. 2018; Malhan et al. 2021a). Their orbits allow measurements of the underlying gravitational potential of the Milky Way (Johnston et al. 1999; Ibata et al. 2001; Koposov, Rix & Hogg 2010; Newberg 2010; Varghese, Ibata & Lewis 2011; Sanders & Binney 2013; Küpper et al. 2015; Malhan & Ibata 2019; Reino et al. 2021). The presence of gaps and density perturbations within streams can inform the population of dark matter substructure, and subsequently the properties of dark matter (Carlberg, Grillmair & Hetherington 2012; Sanders, Bovy & Erkal 2016; Erkal, Koposov & Belokurov 2017; Banik & Bovy 2019; Bonaca et al. 2019; Bonaca et al. 2020). They can also be used to empirically track the underlying distribution of dark matter (Purcell, Zentner & Wang 2012; Necib et al. 2019b).

Starting with the Sloan Digital Sky Survey (SDSS) (York et al. 2000), numerous surveys have increased the number of catalogued stellar streams (Odenkirchen et al. 2001; Newberg et al. 2002; Belokurov et al. 2006; Grillmair 2006; Shipp et al. 2018). Most recently, the *Gaia* Space Telescope (Gaia Collaboration 2018; Lindegren et al. 2018) has opened a new frontier of Galactic kinematics and thus new opportunities for the discovery and study of stellar streams.

Numerous successful stream-finding techniques have been applied to the *Gaia* data (Malhan & Ibata 2018; Malhan et al. 2018a; Yuan et al. 2018; Meingast & Alves 2019; Meingast, Alves & Fürnkranz 2019; Borsato, Martell & Simpson 2020; Ibata et al. 2021). In some cases cross-referencing *Gaia* with other spectroscopic catalogues can provide additional kinematic or spectroscopic information, although statistically limiting the sample size (see e.g. STARGO (Yuan et al. 2018), which identifies streams in the cross match of *Gaia* DR2 with LAMOST DR5 (Luo et al. 2015)). Of the methods relying exclusively on *Gaia*, the STREAMFINDER algorithm (Malhan & Ibata 2018; Malhan et al. 2018a) leverages the fact that stars within a stellar stream would have similar orbits through the Galaxy. By searching for stars occupying the same 'hypertubes' through six-dimensional position/velocity space, STREAMFINDER has discovered a number of new stellar streams (Malhan, Ibata & Martin 2018b; Ibata, Malhan & Martin 2019; Malhan et al. 2019; Ibata et al. 2021). In order to construct these orbits, STREAMFINDER must assume a form for the Galactic potential, and search for stars on an isochrone as part of a kinematically cold stream.

In this paper, we present VIA MACHINAE, a new algorithm for automated stellar stream searches with *Gaia* data (Gaia Collaboration 2018; Lindegren et al. 2018). Based on unsupervised machine learning techniques, we identify streams as local overdensities in the angular position, proper motion, and photometric space of stars in *Gaia* DR2. Importantly, we do not assume the stars in question lie on a particular orbit or stellar isochrone. In fact, the initial (and most computationally intensive) machine learning training steps of

★ E-mail: dshih@physics.rutgers.edu

VIA MACHINAE are designed to find all anomalous structures first, in an agnostic manner. Only then do we implement selections based on prior knowledge of the properties of known stream candidates (particularly that the stars are distributed in an approximately linear structure over small angles on the sky). Such choices can be modified to target structures with other distributions in stellar photometry and proper motion, for example globular clusters or debris flow (Kuhlen, Lisanti & Spergel 2012; Lisanti & Spergel 2012).[1] This flexibility may allow our technique to be sensitive to a wider variety of stellar streams than previous methods, and can be generalized to other anomalous features within the *Gaia* data set (or other astrophysical surveys).

VIA MACHINAE has two main components: an anomaly finding algorithm, and a line finding algorithm. The first component is the ANODE (ANOmaly detection with Density Estimation) algorithm (Nachman & Shih (2020) hereafter referred to as NS20). Originally developed to search for new physics at the Large Hadron Collider, ANODE is a general machine learning algorithm for finding localized overdensities in any data set. To accomplish this, ANODE leverages recent advances in density estimation using neural networks, specifically the idea of normalizing flows (for a recent review and original references, see e.g. Papamakarios et al. 2021). In this paper, following the original ANODE work (NS20), we use masked autoregressive flows (MAF) (Papamakarios, Pavlakou & Murray 2018) to estimate the probability densities of stars in the *Gaia* data set.

The ANODE algorithm begins by slicing up the data set into search regions and their complements, the control regions. As kinematically cold streams are expected to be fully localized in both proper motions, we choose to split the data set into search regions consisting of slices in one of the proper motion coordinates. Then we use the MAF to estimate the probability distribution in position/proper motion/colour/magnitude space of the stars in each search region in two different ways: (1) directly with the stars in the search region; and (2) indirectly with the stars in the control region, followed by interpolation into the search region. The interpolation step is a 'free' byproduct of the density estimation, because we actually learn a conditional probability density conditioned on the proper motion used to define the search region. If the search region contains a stream while the control region does not, then (2) can be thought of as a data-driven estimate of the probability density of the 'background' (i.e. non-stream) halo stars in the search region. Taking the ratio of these two density estimates forms a discriminant R, which is sensitive to anomalous overdensities (or underdensities) in the search region. By selecting the stars with the largest likelihood ratios, we can preferentially enhance the presence of stream stars versus background stars in any given search region.

After performing such a selection in each search region, we are left with a much reduced set of stars spread across the sky. Only some of these stars will correspond to stellar streams. The rest may be other interesting structures (e.g. globular clusters or debris flow) or spurious false positive fluctuations of the ANODE algorithm. This leads to the second major component of the VIA MACHINAE algorithm: an automated method to search for linear features in a collection of stars in an angular patch of the sky. Simply fitting the stars to a line using (for example) least-squares regression yields extremely unsatisfactory results, owing to the presence of noise and outliers (i.e. in a collection of stars, only a small fraction might belong to the stream). Instead, we have developed a method based on the Hough transform. This is an age-old machine learning technique that was originally developed for finding lines and edges in photographs (Hough 1959; Duda & Hart 1972), but which we adapt here to accomplish the same purpose in scatter plots.[2] The idea of the Hough transform is to convert the problem of line finding to counting intersections of curves in an auxiliary parameter space (the Hough space). In this way, one can also give a (rough) figure-of-merit to the best-fitting line detection, based on the contrast between regions of high and low curve density in Hough space.

The major steps and key terms of VIA MACHINAE are summarized in Fig. 1. Moving from left to right in this figure:

(i) We divide the sky into overlapping patches of stars, each a circular region of radius 15°.

(ii) These patches are then divided into overlapping search regions based on one proper motion coordinate. The estimated probability ratio R for each star in each search region is obtained by ANODE training. We then limit ourselves to the inner 10° of the patch to avoid edge effects (among other fiducial cuts).

(iii) Each search region is then subdivided into regions of interest using the orthogonal proper motion coordinate which was not used to define the search region. In order to further purify signal to noise, a cut on colour is imposed to focus on old, metal-poor stars that comprise the majority of known streams.

(iv) The 100 stars with the highest R values in each region of interest are mapped to Hough space and the most line-like feature is assigned a significance σ_L.

(v) In overlapping regions of interest, we combine coincident lines and σ_L values to obtain a protocluster for the patch, with an accompanying total significance $\sigma_L^{\rm tot}$.

(vi) Protoclusters in neighbouring patches are combined into a stream candidate.

In this paper, we will use the GD-1 stream to illustrate the steps of the VIA MACHINAE algorithm. GD-1 (Grillmair & Dionatos 2006), is an exceptionally long and dense stellar stream located at ∼10 kpc, most likely originating as a globular cluster of mass $\sim 2 \times 10^4 M_\odot$ (Koposov et al. 2010). When first detected using SDSS, GD-1 was thought to span ∼60° in the sky. Using the second data release of *Gaia* (*Gaia* DR2), it has been extended by as much as 20° (Price-Whelan & Bonaca 2018b) (hereafter PWB18), and was found to include gaps that could be evidence for dark matter substructure (Price-Whelan & Bonaca 2018b; Banik et al. 2019, 2021; Bonaca et al. 2019; Malhan, Valluri & Freese 2021b). Though most stellar streams are not nearly as long, dense, narrow, or well-defined as GD-1, it nevertheless provides an excellent testbed for VIA MACHINAE, as stellar membership of the stream has been extensively studied (see e.g. Price-Whelan & Bonaca 2018b; Bonaca et al. 2019, 2020), and its distinctiveness allows for clear demonstrations of the utility of the algorithm.

This paper is organized as follows: In Section 2, we introduce the *Gaia* data and its processing into inputs that will be used for anomaly detection. We then present the algorithm in Section 3, with each step illustrated by its action on a segment of the GD-1 stream. In Section 4, we apply VIA MACHINAE to the entire length of the GD-1 stream. Finally, in Section 5 we conclude with a summary and a list of interesting future directions motivated by this work. In a subsequent

[1]Debris flow refers to structure localized in velocity space, but incoherent in physical space (Helmi & White 1999; Kuhlen et al. 2012; Lisanti & Spergel 2012). This is usually the case for older mergers, e.g. the *Gaia* Sausage/Enceladus (Necib, Lisanti & Belokurov 2019a).

[2]The Hough transform has also been proposed for stellar stream identification in (Pearson et al. 2019; Pearson et al. 2021) in the context of M31.

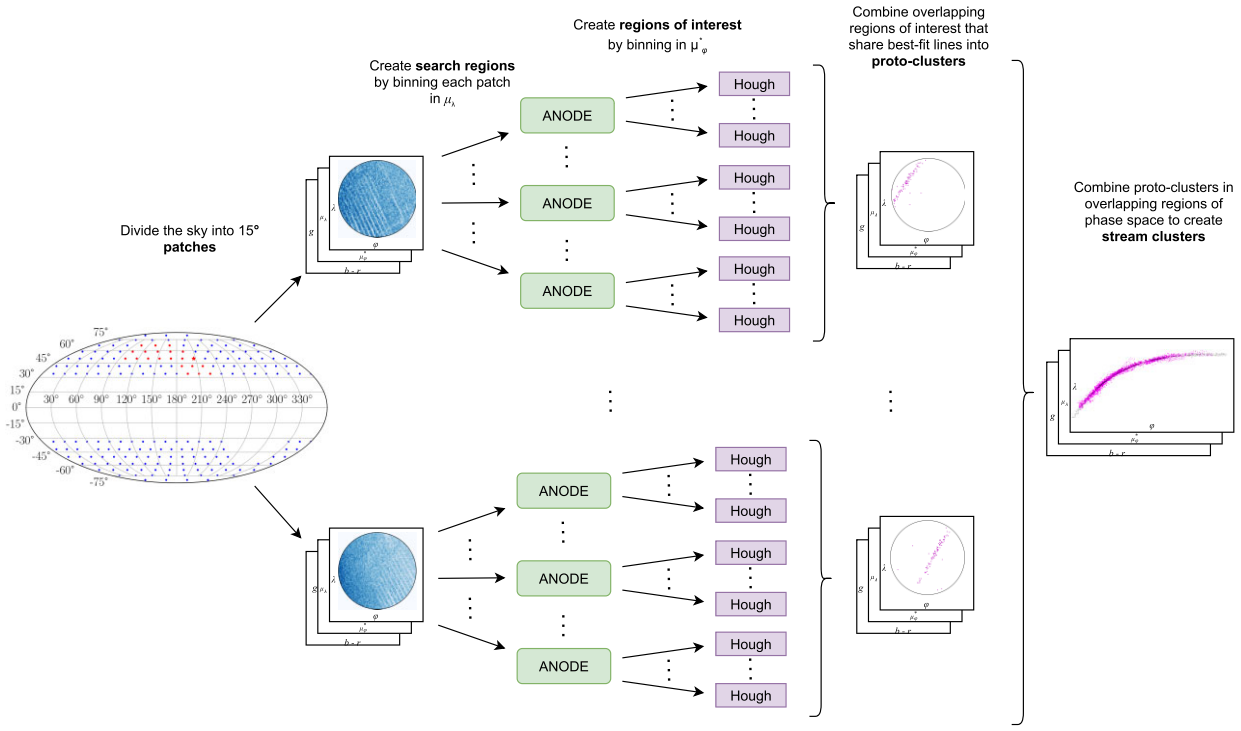

Figure 1. A schematic showing an overview of the VIA MACHINAE algorithm. The bolded and boxed terms are defined in Section 3 (with the exception of patches, which are described in Section 2). First we divide up the sky into evenly tiled 15° patches. Within each patch, we further divide up the stars into search regions defined by a window in μ_λ, one of the proper motion coordinates (the remaining data features for each star are denoted $\vec{x}$). Then we train the ANODE algorithm on the search regions and their complements, to learn a data-driven measure of local overdensities $R(\vec{x})$. To turn this measure into a stream finder, we further divide up the SRs into regions of interest based on the orthogonal proper motion coordinate μ^*_ϕ. We apply an automated line-finding algorithm based on the Hough transform to the 100 highest-R stars in each ROI. Finally, we combine ROIs adjacent in proper motion that have concordant best-fitting line parameters into protoclusters, and cluster these across adjacent patches of the sky into stream candidates.

work (Shih et al., in preparation), we will apply our technique across the full *Gaia* DR2 data set, and demonstrate its ability to detect other known streams, and present new stream candidates.

2 DATA AND INPUT VARIABLES

Before introducing the VIA MACHINAE algorithm, we must first describe the data upon which it will be applied, and the pre-processing required.

Starting with the *Gaia* DR2 data set,[3] we limit ourselves to distant stars with measured parallax less than 1 mas (corresponding to stars beyond 1 kpc). We do not correct for the *Gaia* DR2 zero-point parallax offset; varying the parallax cut by ± 0.05 results in only a ~ 3 per cent change in the number of stars and so is highly unlikely to affect our algorithm. We tile the sky with 15° patches using HEALPY (Górski et al. 2005; Zonca et al. 2019) (with nside $= 5$). This patch size was selected to have a tractable number of stars for the machine learning training step of the algorithm, as will be described in Section 3.2. The patches are also large enough to capture significant portions of most known streams if they should pass through them. As stars in the Galactic disc would overwhelm the training, we limit the analysis to high Galactic latitudes $|b| > 30°$. We also exclude all patches that overlap with the LMC or SMC. The final result is 200 patches in total.

For stars within a patch, our data consists of two position, two kinematic, and two photometric parameters: the angular position on the sky (e.g. right ascension [ra, α] and declination [dec, δ]), the corresponding angular proper motions ($\mu_\alpha \cos\delta$ and μ_δ), the magnitude of the star in the *Gaia* G-band (g), and the difference in the G_{BP} and G_{RP} *Gaia* bands ($b - r$). Throughout this work, we will not correct for dust or extinction; especially since we confine ourselves to high Galactic latitudes, these corrections are generally small ($\lesssim 0.1$ for $b - r$) and do not vary much across a patch. Since we are only interested in local overdensities in each patch and will select a wide range of colour for our final analysis, dust and extinction corrections should not significantly affect our results.

The (α, δ) coordinates do not have a Euclidean distance metric across the sky, and the resulting distortions across the patch, especially at high latitudes, could negatively affect our neural density estimation.[4] Therefore, for each patch (defined by a circle centred on $(\alpha, \delta) = (\alpha_0, \delta_0)$ in angular position), we rotate the positions and proper motions using ASTROPY (Astropy Collaboration 2013, 2018) into a new set of centred longitude and latitude coordinates (ϕ, λ) so that $(\alpha_0, \delta_0) \rightarrow (0°, 0°)$. The unit vectors for the rotated coordinate

[3] As this work was being completed, *Gaia* EDR3 (Gaia Collaboration 2021) was released. While our results likely would have been improved by using this new data set, re-running the ANODE method on *Gaia* EDR3 proved to be too computationally expensive (the full-sky scan of *Gaia* DR2 took $\mathcal{O}(10^5)$ NERSC-hours). We plan to apply our method to *Gaia* EDR3 in a future publication.

[4] Density estimation on spheres and other non-Euclidean manifolds is an active area of research, see e.g. Rezende et al. (2020). We do not use these techniques in this work.

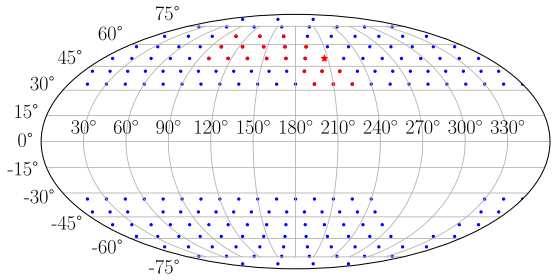

Figure 2. The positions in Galactic ℓ and b coordinates used for the centres for the data sets from the *Gaia* DR2 used in our full-sky analysis. The missing grid centres in the Galactic Southern hemisphere are the patches that overlapped with the Magellanic Clouds. The 21 centres which contain the GD-1 stream are shown in red, and the patch used as the worked example in Section 3 is denoted with a star.

system, $(\hat{\phi}, \hat{\lambda})$, are aligned with those of the previous unit vectors $(\hat{\alpha}, \hat{\delta})$. Within each patch, we will calculate angular distances using a simple Euclidean metric in (ϕ, λ). For notational simplicity, we will define the new proper motion coordinate $\mu_\phi \cos\lambda$ as μ_ϕ^* for the remainder of the work (similarly $\mu_\alpha^* \equiv \mu_\alpha \cos\delta$).

The patches defined above will be used as input for the ANODE method, as will be described in Section 3, using the features $(\phi, \lambda, \mu_\phi^*, \mu_\lambda, b-r, g)$. After training ANODE on each patch, we impose a set of additional fiducial cuts on the data. As we will describe in more detail in Section 3.2, these cuts are driven by the limitations of the MAF density estimator. Specifically, to avoid edge effects in the neural network output, the post-ANODE fiducial region studied in this paper is the inner 10° of each patch with a magnitude cut of $g < 20.2$. Above this magnitude cut, the completeness drops rapidly (Boubert & Everall 2020); this choice also helps reduce (but does not completely eliminate) streaking in the data and other artefacts due to incomplete coverage of the dimmest stars in the *Gaia* DR2 data (Gaia Collaboration 2018).

As described in the Introduction, in this work we focus on demonstrating the VIA MACHINAE algorithm using GD-1 as a worked example. Therefore, we limit ourselves here to patches of the sky that are known to contain portions of the GD-1 stream. We find that 21 patches in our all-sky sample include stars which have been identified by PWB18 as possible members of the GD-1 stream, for a total of 1985 candidate GD-1 stars. Before (after) the ANODE fiducial cuts, the patches containing GD-1 have various numbers of stars, ranging from 8.0×10^5 (2.7×10^5) in the patch with the least number of stars, to 2.1×10^6 (7.0×10^5) stars in the patch with the most number of stars. Fig. 2 shows the locations of all 200 patch centres we use to tile the sky as well as the 21 patches containing GD-1 stars.

We will use the stream membership labels of PWB18 (which can be downloaded at Price-Whelan & Bonaca 2018a) as our point of comparison throughout this work. These were derived through relatively simple means: a visual inspection of the data, combined with polygonal cuts on proper motion, colour and magnitude, and a parallel strip cut (the 'stream track') in angular position. Thus we do not take them as 'absolute truth' labels – indeed, some level of background contamination within this sample is certainly visible by eye. Nevertheless, the GD-1 candidate labels of PWB18 still furnish a very useful and powerful point of comparison.

In Section 3, we will use one of these 21 patches containing GD-1, centred on $(\alpha_0, \delta_0) = (148.6°, 24.2°)$, to provide a worked example of each stage of VIA MACHINAE. Within this patch's fiducial region, there are 334 376 stars, of which 276 have been identified as candidate members of GD-1 by PWB18. The position, proper motion, and photometry of these stars is shown in Fig. 3. In this patch, the candidate GD-1 stars lie in the range $\mu_\lambda \in [-14.6, -8.6]$ mas yr^{-1}.

3 VIA MACHINAE: THE ALGORITHM

3.1 ANODE: Defining the search regions

As described in the Introduction, the first part of VIA MACHINAE is based on the ANODE method (NS20). The starting point of ANODE is the subdivision of the stars within a single patch into search regions (SRs) which are windows in one feature of the data set. The complement of the search region is called the control region (CR). The feature and the width of the window should be chosen so that, if a stream is present, there exists (at least) one SR which fully (or nearly fully) contains the entire stream. As we will explain in the next subsection, this is to enable accurate background estimation from the CR. Defining the SRs by strips of angular position, for example, would not satisfy this requirement, unless the strips coincidentally aligned with the direction of the stream within the patch. However, stellar streams are kinematically cold and so are concentrated in both proper motion coordinates. Thus, we define our SRs using one of the proper motion coordinates. (Selecting SRs based on both proper motion coordinates is possible, but would greatly increase our total training time.) Since streams are localized in both proper motions, in principle it should not matter which one we choose; for this study, we choose μ_λ to be the proper motion coordinate defining the SRs.[5]

Based on the proper motion properties of known streams, we find that a choice of a window in μ_λ of width 6 mas yr^{-1} is optimal. Streams like GD-1, located $\mathcal{O}$(10 kpc) from the Earth, have proper motion dispersions of $\sim$2 mas yr^{-1}. Such streams would be completely contained within our SRs at distances larger than 2–3 kpc (which is more or less commensurate with the parallax cut we placed on our data set). We also note that the stream does not have to be completely enclosed within a given SR for the algorithm to function. Proper functioning of ANODE requires only that the relative distribution of stars within the SR differ significantly from that in the CR; since the CR contains many more stars than the SR, a leakage of stream stars into the CR will not typically invalidate our approach.

Since we do not know a priori which SR contains a stream, we must scan over all regions. In practice, we define a series of SRs by stepping in units of $\mu_\lambda = 1$ mas yr^{-1}, with each SR then defined by the choice of $[\mu_\lambda^{\rm min}, \mu_\lambda^{\rm max}]$:

$$[\mu_\lambda^{\rm min}, \mu_\lambda^{\rm max}] = \ldots, [-10, -4], [-9, -3], \ldots, [3, 9], [4, 10], \ldots \quad (1)$$

in units of mas yr^{-1}. The complement of the proper motion window (i.e. all the stars in the same patch that are not in the SR) defines the control region (CR) for each SR.

Each of these choices of $(\alpha_0, \delta_0, \mu_\lambda^{\rm min})$ furnishes a search region and control region pair for the ANODE training step. Overlapping the SRs in this way allows us to fully capture potential streams in at least one μ_λ window when performing a blind search – if the SRs were not overlapping, then a stream could easily fall at the edge of

[5]The choice of proper motion coordinate can affect the performance of the algorithm through the number of background stars in the SR. For example, if the stream stars have small values of μ_λ but large values of μ_ϕ, then defining the SR in terms of μ_λ would lead to more background stars for the same number of stream stars, and hence a lower *S*/*B*, decreasing the stream detection probability. In Shih et al. (in preparation) we will also incorporate the results of a scan over SRs defined using μ_ϕ and show how this can achieve complementary results.

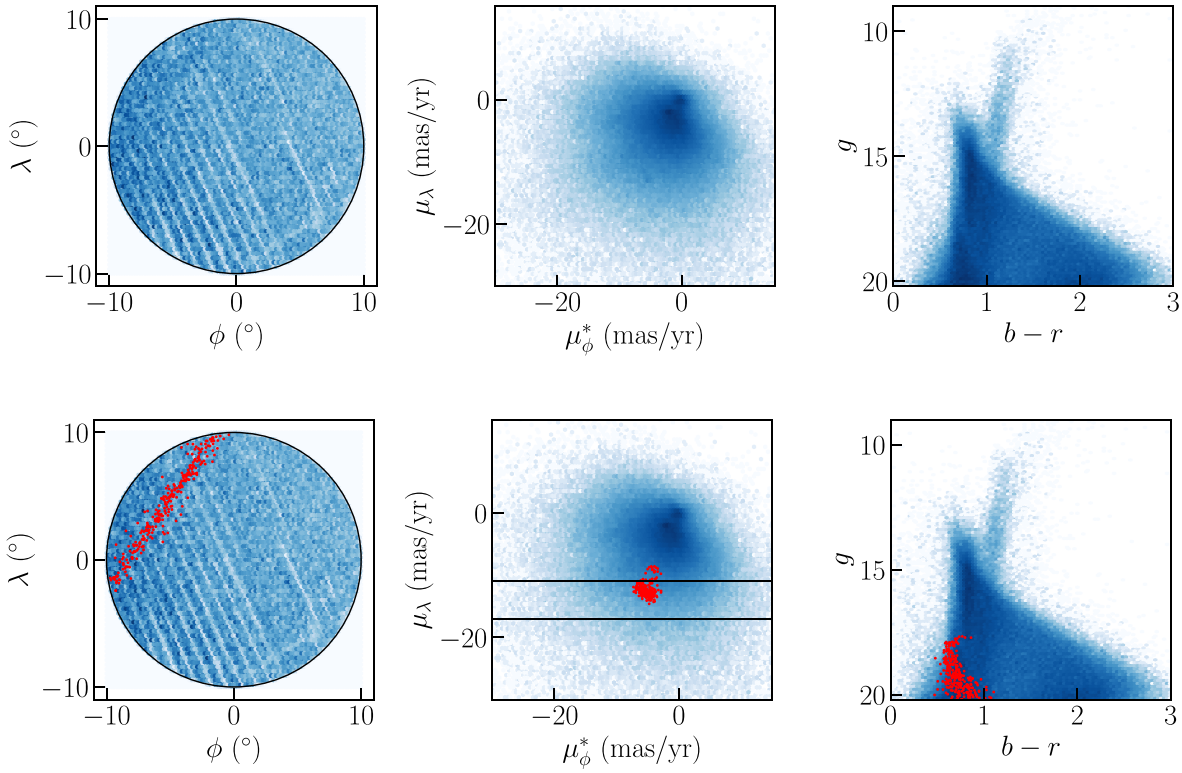

Figure 3. Upper row: Angular position in (ϕ, λ) coordinates (left-hand panel), proper motion in $(\mu_\phi^*, \mu_\lambda)$ coordinates (centre), and photometry (right-hand panel) of all stars in the patch centred on $(\alpha, \delta) = (148.6^\circ, 24.2^\circ)$. (Note the streaking in angular position due to non-uniform coverage in *Gaia* DR2.) Bottom row: As above, with stars identified by PWB18 as likely GD-1 stars shown in red, along with an example search region $\mu_\lambda \in [-17, -11]$ mas yr^{-1} in proper motion.

two SRs, diluting the signal in each. By selecting SRs which are wide enough in proper motion to fully contain a kinematically cold stream and overlapping them by shifts which are smaller than the proper motion width of a typical stream, we minimize the possibility of this dilution.

SRs with fewer than 20 k stars or more than 1 M stars (before the fiducial cuts) are rejected for ANODE training. The former requirement is because too few stars in the SR results in poor density estimation performance, and the latter requirement is to avoid overly long training times. In addition, SRs that contained a GC candidate (identified using a simple algorithm described in Appendix B) were cut from the analysis, as the presence of the GC would completely overwhelm the training (i.e. in an SR containing a GC, the GC would correspond to such a large, delta-function-like overdensity, that ANODE would be unable to identify any other overdensity in the SR, such as one coming from a stream). In the end, we are left with a total of 545 SRs across the 21 patches of the sky containing GD-1.

To provide an example of an SR, we turn to our sample GD-1 patch defined in the previous section, centred on $(\alpha_0, \delta_0) = (148.6^\circ, 24.2^\circ)$. We select the SR defined by $\mu_\lambda \in [-17, -11]$ mas yr^{-1}, which encompasses the majority of the GD-1 stars contained within this patch. This SR is shown in Fig. 3 and contains 34 823 stars in total, of which 252 are tagged by PWB18 as possible GD-1 members.

3.2 ANODE: Density estimation

Having defined the search regions, we turn to the probability density estimation step of the ANODE algorithm. As discussed in Section 2, the stars in our data set are characterized by two position coordinates, two proper motion coordinates, colour, and magnitude. Having set aside one of the proper motion coordinates μ_λ to define the search regions with, the remaining features $(\phi, \lambda, \mu_\phi^*, b-r, g)$ we will refer to collectively as $\vec{x}$.

Suppose the stars in a patch consist of 'signal stars' coming from a cold stellar stream, and 'background stars' coming from the stellar halo. Let the conditional probability density of the background stars be $P_{\rm bg}(\vec{x}|\mu_\lambda)$, and the conditional density for the data (consisting of background stars plus signal stream stars) be $P_{\rm data}(\vec{x}|\mu_\lambda) = (1-\alpha)P_{\rm bg}(\vec{x}|\mu_\lambda) + \alpha P_{\rm sig}(\vec{x}|\mu_\lambda)$, where α is a measure of the signal strength. Then the optimal test statistic for distinguishing data from background is (Neyman & Pearson 1933):[6]

$$R(\vec{x}|\mu_\lambda) = \frac{P_{\rm data}(\vec{x}|\mu_\lambda)}{P_{\rm bg}(\vec{x}|\mu_\lambda)}. \quad (2)$$

If the signal is small ($\alpha \ll 1$) but sufficiently localized in feature space (i.e. a local overdensity), then we expect $R \gg 1$ where the signal is localized and $R \approx 1$ everywhere else. Since R can be computed without knowing α or $P_{\rm sig}$, selecting data points with high R can purify signal to background in a model-agnostic way.

[6]Note that this will in general not be the optimal statistic for distinguishing any particular signal hypothesis from the background, rather it is the optimal test for distinguishing the background-only hypothesis from the data-driven probability distribution. For more discussion of the meaning of optimality in the context of anomaly detection, see the Appendix to NS20.

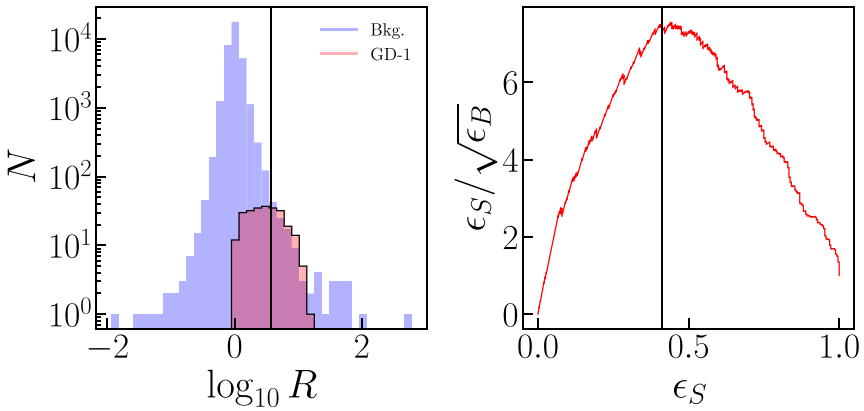

Figure 4. Left-hand panel: R distribution for the SR $\mu_\lambda = [-17, -11]$ mas yr^{-1} in the patch centred at $(\alpha, \delta) = (148.6^\circ, 24.2^\circ)$. Stars identified as likely members of GD-1 by PWB18 are shown in red, while the 'background' stars (those not tagged as likely GD-1 members by PWB18) are in blue. Right-hand panel: Significance improvement characteristic (SIC) curve for the same SR, showing the signal efficiency ϵ_S and the significance improvement (signal efficiency over square root of background efficiency, $\epsilon_S/\sqrt{\epsilon_B}$) as the cut on R is varied. The vertical lines in both plots designate the R value that maximizes the SIC curve.

Probability density estimation of arbitrary distributions is a difficult problem, and so ANODE is only made feasible through recent advances in machine learning. In this paper, as in NS20, we employ the MAF architecture (Papamakarios et al. 2018) for the density estimation task. The MAF uses a specially structured neural network to learn a bijective mapping from the original feature space into a latent space where the data are described by a unit multivariate normal distribution.[7]

Although it is relatively straightforward to train the MAF directly on the stars in the SR to learn $P_{\rm data}(\vec{x}|\mu_\lambda, \mu_\lambda \in {\rm SR})$ (the numerator of the likelihood ratio equation 2), estimating the background density $P_{\rm bg}(\vec{x}|\mu_\lambda)$ takes more consideration. Calculating the denominator $P_{\rm bg}$ from first principles often proves impossible. Instead, one of the key ideas of the ANODE method is to use sideband interpolation from the CR (the complement of the SR) to estimate the background density in the SR. More precisely, we train a second MAF on the CR to learn $P_{\rm data}(\vec{x}|\mu_\lambda, \mu_\lambda \in {\rm CR})$. If there is no stream in the CR, then

$$P_{\rm data}(\vec{x}|\mu_\lambda, \mu_\lambda \in {\rm CR}) = P_{\rm bg}(\vec{x}|\mu_\lambda, \mu_\lambda \in {\rm CR}). \quad (3)$$

If the background distribution in the CR is a smooth and slowly varying function of μ_λ, then the MAF provides an automatic interpolation into the SR and yields an estimate for $P_{\rm bg}(\vec{x}|\mu_\lambda, \mu_\lambda \in {\rm SR})$, the denominator of equation (2).[8]

An important point to note is that the MAF (along with most, if not all unsupervised density estimators) has difficulty matching rapid or discontinuous changes in the probability density as a function of the features $\vec{x}$. This is not a problem for the proper motion and $b - r$ features, which smoothly go to zero. However, in position-space, the selection of stars within a circular patch on the sky results in a sharp cutoff in density at the edge of the patch. Similarly, at high magnitude g, the sensitivity of the *Gaia* satellite drops rapidly. The result is spuriously large R values near the edge of the patch in position space and at large g. To avoid this, we train on a larger data set than the fiducial region in which we perform the subsequent stream-finding steps. As previously discussed in Section 2, after running ANODE, we define a fiducial region of 10° around the centre of the patch in (ϕ, λ) position space and a magnitude cut of $g < 20.2$.

In Fig. 4 (left-hand panel), we show a histogram of the ANODE probability ratio R for the stars in the $\mu_\lambda \in [-17, -11]$ mas yr^{-1} SR within the GD-1 example patch. We see that the likely GD-1 stars identified by PWB18 are disproportionately represented at the high-R tail of the ANODE distribution. By cutting on R, the resulting sample of stars would be enriched with stream stars compared to the full sample. For a given value of R, the signal efficiency ϵ_S is the fraction of candidate stream stars passing the cut on R, and the background efficiency ϵ_B, is the fraction of non-stream-candidate stars passing the threshold. In Fig. 4 (right-hand panel), we show the significance improvement characteristic (SIC) curve, comparing ϵ_S to $\epsilon_S/\sqrt{\epsilon_B}$ as R is varied. We see that cutting on the ANODE output can greatly improve the purity of the sample and enhance the significance of the stream detection. For the sake of illustration, we have indicated in Fig. 4 the optimal $R_{\rm cut}$ value, defined to be the cut on R that maximizes the significance improvement in Fig. 4 (right-hand panel). (In more general settings, without stream-labelled stars, the optimal cut on R would not be known, see the next subsection for further discussion of this.) Starting with 252 stars out of 34 823 identified as candidate GD-1 members by PWB18, the optimal $R_{\rm cut}$ value (corresponding to $\log_{10} R_{\rm cut} = 0.57$ for this SR) selects 206 stars, of which 103 are candidate GD-1 stars (corresponding to $\epsilon_S = 0.41$). This nominally increases the statistical significance of the stream (i.e. $S/\sqrt{B}$) by more than a factor of 7. We emphasize that the R ratio was learned in a completely data-driven, unsupervised manner, and at no point in the training were the stream candidate labels from PWB18 ever used. Here the labels are just used to illustrate the efficacy of the ANODE R-ratio in identifying stream stars.

In Fig. 5 (top), we show all the stars in the $\mu_\lambda \in [-17, -11]$ mas yr^{-1} SR, and (bottom) those stars passing the optimal R cut. GD-1 is an exceptionally dense and distinct stream: unlike other known streams it is visible, albeit barely, before the cut on R. Performing the cut of $R > R_{\rm cut}$, as shown in the lower panel, drastically increases the significance of the stream, as expected.

Finally, we comment on the issue of streaking that can clearly be observed in the position space plots of the stars in many patches (Figs 3 and 5 are prime examples). These streaks are artefacts due to *Gaia*'s scan pattern and incomplete coverage of the sky in DR2. They might seem concerning for the ANODE method, as they

[7]Our selection of hyperparameters is described in Appendix A2.

[8]If there is signal in the CR, then by assumption it will be a very small perturbation to $P_{\rm data}(\vec{x}|\mu_\lambda, \mu_\lambda \in {\rm CR})$ (i.e. we assume there are many more background stars than signal stars in the CR). Then equation (3) will still be approximately true, and the signal contamination in the CR should not greatly affect the R statistic in the SR.

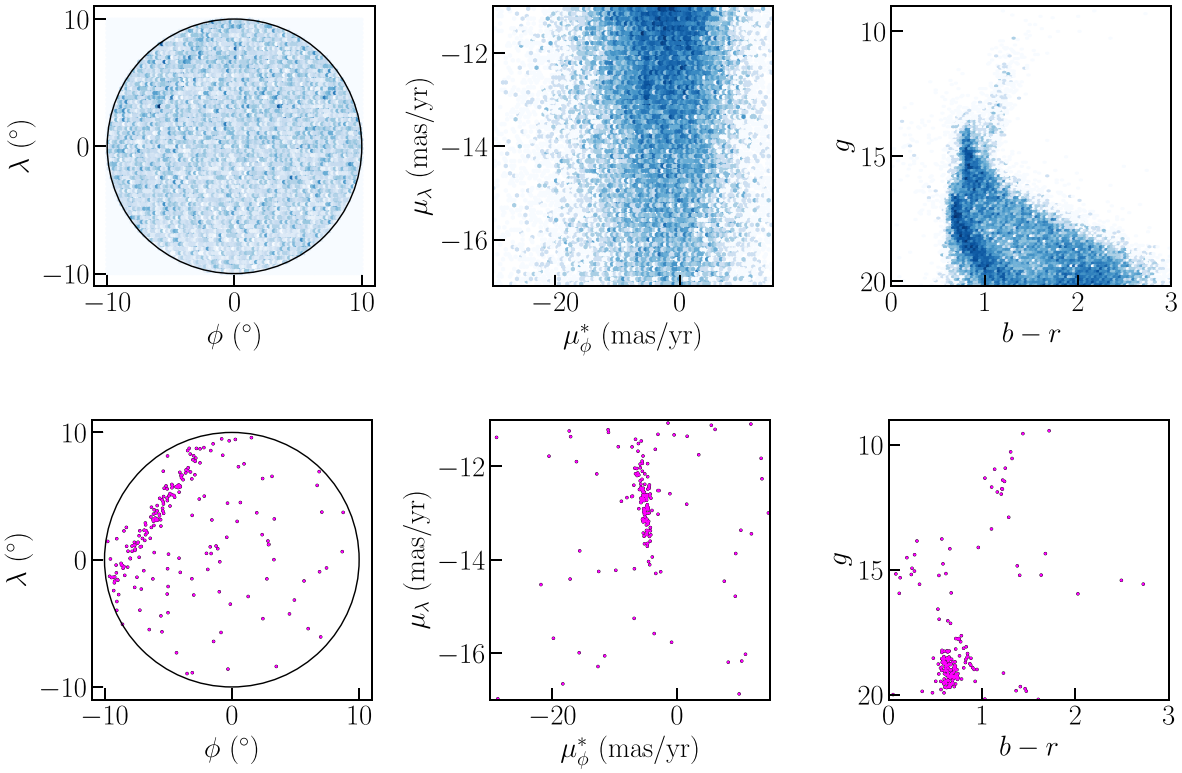

Figure 5. Upper row: Angular position in (ϕ, λ) coordinates (left-hand panel), proper motion in $(\mu_\phi^*, \mu_\lambda)$ coordinates (centre), and photometry (right-hand panel) of all stars (blue) in the $\mu_\lambda \in [-17, -11]$ mas yr^{-1} SR of our example patch centred on $(\alpha, \delta) = (148.6^\circ, 24.2^\circ)$. Bottom row: As the upper row, applying the $R > R_{\rm cut}$ cut on the stars in the SR (purple). The GD-1 stream becomes immediately apparent. See the text for details.

appear as line-like overdensities in the angular coordinates, just like stellar streams would. However, we find no evidence that ANODE is incorrectly selecting for these spurious features. The reason is that ANODE looks for evidence of a local overdensity by comparing the stars in one proper motion slice with the stars outside of it. The streaking patterns are largely uncorrelated with proper motion; therefore, the overdensity they correspond to will actually *cancel* in the construction of the R ratio, and these streaking stars will not be selected for by the ANODE algorithm.

3.3 Regions of interest

Up to this point, our method has been largely agnostic to the astrophysics of stellar streams (beyond the choice to use proper motion as our SR-defining feature). Stars tagged as anomalous by the ANODE training may be streams, globular clusters, debris flow, or some other structure localized in the Milky Way's velocity-space. The steps in this and subsequent subsections are designed specifically to find cold stellar streams similar to the ones identified previously in data; different cuts and/or choices of parameters could be used to focus on other interesting astrophysical structures. The cuts we choose are:

(i) First, we remove all stars within a box around zero proper motion of width 2 mas yr^{-1}. That is, we require

$$|\mu_\lambda| > 2 \text{ mas yr}^{-1} \quad \text{OR} \quad |\mu_\phi^*| > 2 \text{ mas yr}^{-1}. \tag{4}$$

Recall that the ANODE training identifies stars within the SR that are anomalous compared to the interpolation into the SR of the CR density estimate. Stars with proper motion near zero are predominantly distant stars; this population is not well-represented in a CR that does not contain $(\mu_\phi^*, \mu_\lambda) \sim (0, 0)$ mas yr^{-1}. An example can be seen in Fig. 6. If the SR contains this zero-point, the distant stars are (correctly) identified as anomalous relative to the population in the control regions, but their sheer number completely overwhelms

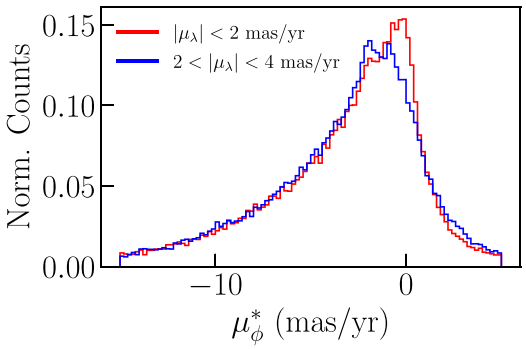

Figure 6. Normalized histogram of μ_ϕ^* values for stars in the 10° patch centred on $(\alpha, \delta) = (148.6^\circ, 24.2^\circ)$, requiring $2 < |\mu_\lambda| < 4$ mas yr^{-1} (blue) and $|\mu_\lambda| < 2$ mas yr^{-1} (red). Note that the high density of stars near $\mu_\phi^* \sim 0$ with $|\mu_\lambda| < 2$ mas yr^{-1} are not represented in the sample which does not overlap $\mu_\lambda \sim 0$. These very distant stars with near-zero total proper motion are absent as a population from search regions which do not include the zero-point of proper motion.

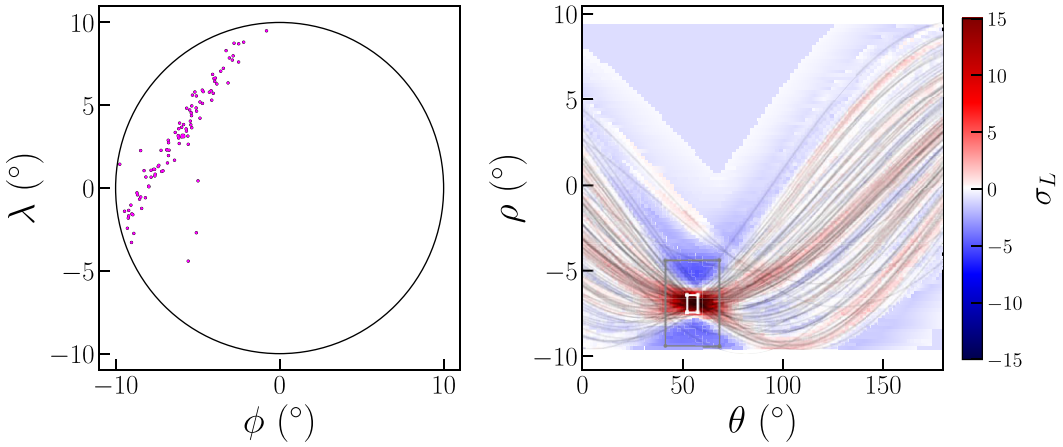

Figure 7. Left-hand panel: Angular position in (ϕ, λ) coordinates for the 100 highest-R stars (purple) in the $\mu_\phi^* \in [-8, -2]$ mas yr^{-1}, $\mu_\lambda \in [-17, -11]$ mas yr^{-1} ROI from our example patch. Right-hand panel: Associated curves in Hough space for these stars (black lines). The significance $\sigma_L(\theta, \rho)$ of a line oriented at each (θ, ρ) value is shown in colour. The region around the point of maximum contrast (as identified by the VIA MACHINAE algorithm) is indicated by the inner white box, with the region defining the background shown as the outer box.

any other signal in the SR, requiring their removal after training is complete.

(ii) Cold stellar streams, produced by tidally stripped globular clusters or dwarf galaxies, are predominantly composed of old, low metallicity stars. Many existing stream-finding algorithms leverage this by fitting stars in the stream candidate to isochrones appropriate to this assumption (see e.g. Malhan & Ibata 2018). Although the ANODE training is agnostic to such assumptions, in this work we are specifically interested in identifying cold streams, and not all anomalous overdensities. To that purpose, we now select stars in a specific colour range in order to further purify signal to background. We require our stream candidates to lie in the broad range of colours $(b - r) \in [0.5, 1]$. This range of colours was chosen so that it will contain (nearly) all of GD-1 and every stream found by STREAMFINDER in Malhan et al. (2018b), Ibata et al. (2019). (STREAMFINDER targeted globular cluster streams composed of stars with ages $\sim$10 Gyr and metallicities -2 dex$\lesssim$ [Fe/H]$\lesssim -1$ dex). But being broader and more general than fitting to specific isochrones, we hope it will also enable the discovery of new streams. There may be interesting anomalous structures outside of this colour range, which will be investigated in a future work.

(iii) To further isolate any potential streams, we subdivide the SRs defined by windows of μ_λ into overlapping windows of μ_ϕ^*, with width 6 mas yr^{-1} and a stride of 1 mas yr^{-1}. We call these windows regions of interest (ROIs) and they are labelled by $(\alpha_0, \delta_0, \mu_\lambda^{\rm min}, \mu_\phi^{*\rm min})$. We exclude any ROI that has fewer than 200 stars as we need larger statistics to determine the presence of a stream.

Applying these cuts and further subdivision of the data to the 21 patches of the sky containing GD-1, we obtain 17 563 ROIs in total.

Within each ROI, we must decide how to apply the cut on the ANODE overdensity function $R(\vec{x}|\mu_\lambda)$. Many different types of cuts are possible, for instance setting a threshold as a percentile cut in each ROI, or a fixed value of R across all ROIs. We have empirically found that selecting the 100 highest R stars in each ROI is effective at finding known streams (more on this in Shih et al., in preparation). An example of this is shown in the left-hand panel of Fig. 7. It is possible that another cut (e.g. the 1000 highest R stars in an ROI) would also be effective or would find other, qualitatively different streams. This would be interesting to explore in future work.

3.4 Line-finding and stream detection

Over large angles on the sky, most streams form arcs in (α, δ) rather than lines (and streams with large line-of-sight velocities may not appear to form lines at all). However, the deviation from a line for the stars in the stream is small across a 10° radius circle on the sky.

Given the large number of ROIs – $\mathcal{O}(10^4)$ for the 21 patches of the sky containing GD-1 alone – we need an automated procedure for line finding. To do so, we adapt a long-standing technique from the field of computer vision based on the Hough transform (Hough 1959; Duda & Hart 1972). A line passing through a point on the plane (ϕ, λ) can be expressed in terms of the distance ρ of closest approach to the origin, and the angle θ between the ϕ axis and the perpendicular from the line to the origin:[9]

$$\rho = \phi \sin\theta - \lambda \cos\theta. \tag{5}$$

Viewing this equation another way leads to the idea of the Hough transform for line finding: for a single point, the collection of lines that pass through it will form a sinusoidal curve in the (θ, ρ) Hough space described by equation (5). If we consider two points in the plane, then their curves in Hough space will intersect for the values of θ and ρ that define a line passing through both points. For a set of points in the plane, a subset of points on a line will manifest itself as overdensity in the (θ, ρ) space as many such curves intersect.

In Fig. 7, we show an example of the Hough transform on position data (left-hand panel) of the 100 highest-R stars in the ROI with $\mu_\phi^* \in [-8, -2]$ mas yr^{-1}, $\mu_\lambda \in [-17, -11]$ mas yr^{-1} from our example patch. As can be seen in the right-hand panel, the Hough curves for the stars on the line all cross at the same point, corresponding to the θ and ρ values of the line on which the stream falls. The Hough transform therefore converts the problem of finding a line among a set of 2D points to the problem of finding the point with the highest density of curves in a 2D plane. Although this overdensity is obvious by eye in the example shown in Fig. 7, this is an extreme case and most overdensities will not be as clear-cut.

We automate the line-finding by identifying the region in Hough space with the highest contrast in density compared to the region

[9]Note ρ can take negative values – there is a periodicity in Hough space of the form $(\rho, \theta) \sim (-\rho, \theta \pm \pi)$.

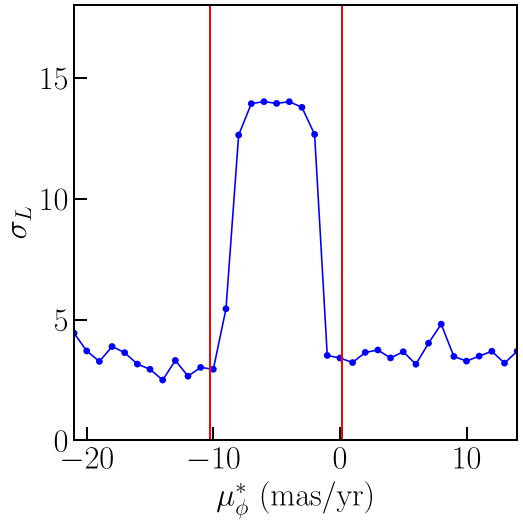

Figure 8. Line significance σ_L versus the central μ_ϕ^* value for each ROI with $\mu_\lambda \in [-17, -11]$ mas yr^{-1} in our example patch. The vertical red lines indicate the minimum and maximum μ_ϕ^* values for the candidate GD-1 stars of PWB18.

surrounding it. We define a filter function which is applied to a box centred on a location (θ, ρ) of width w_θ and height w_ρ. The filter counts the number of stars whose Hough curves pass through the box, allowing us to define a number of curves at each point $n(\theta, \rho)$. We then redo the filtering with a larger box (subtracting the curves which also pass through the initial box) to estimate the 'background' curve count, $\bar{n}(\theta, \rho)$ (being careful to renormalize the counts for the different areas of the patch covered by the two regions in Hough space). Examples of these two filtering regions are shown in Fig. 7. The large and small box dimensions are 'hyperparameters' of the Hough transform line detection method and must be tuned based on known stellar streams to maximize detection efficiency. In this work, we will specialize to $w_\theta = 5.4^\circ$ and $w_\rho = 1^\circ$ for the inner box, and an outer box five times larger. This was found to be optimal for detecting relatively narrow streams such as GD-1. In Shih et al. (in preparation) we will also explore other hyperparameters for the line finder that are sensitive to wider streams.

From the filter function count of Hough curves and background estimate at each point (θ, ρ), we define the line detection significance to be

$$\sigma_L(\theta, \rho) = \frac{n(\theta, \rho) - \bar{n}(\theta, \rho)}{\sqrt{\bar{n}(\theta, \rho)}} \tag{6}$$

We search in Hough space for the parameters that maximize this significance. Concretely, we bin the $(\theta - \rho)$ plane in two dimensions, using a grid of 100 bins for $0 \leq \theta \leq \pi$ and 100 bins for $-10^\circ \leq \rho \leq 10^\circ$. We then select the bin that maximizes σ_L and return this as our line detection in each ROI.[10]

When a stream is present in the SR and within the proper motion range of an ROI, we expect the resulting σ_L value to be much larger than those of ROIs without linear structures. As an example of this, in Fig. 8 we show the σ_L values for every ROI in the SR as a function of the central μ_ϕ^* value defining each ROI, with vertical red lines indicating the maximum and minimum ROIs which contain any GD-1 stars. As can be seen, the high-significance lines fall only in the ROIs containing GD-1 stars. By cutting on σ_L, we are to be able to distinguish ROIs that contain an actual stream in the high-R stars from those without.

3.5 Final merging and clustering

After selecting the 100 highest R stars in each ROI and applying the Hough transform line finder, we obtain the line parameters (θ, ρ) with the highest significance σ_L in each ROI. We wish to use the significances of these lines to select only the most promising stream candidates. However, cutting on the raw σ_L of an individual ROI is not effective in identifying a tractable number of likely stream candidates, because of the large trials factor (the so-called 'look elsewhere effect'). Across only the 21 patches containing GD-1 there are already $\mathcal{O}(10^4)$ ROIs, and random fluctuations could result in spurious line-like features in the background stars. This essentially dilutes the significance of a individual line detection by a correction factor, which may not be entirely trivial to estimate in the presence of correlations between ROIs.

To obtain a meaningful line detection, we use the fact that a stream is likely to be found in multiple ROIs – since the SRs are highly overlapping, each star generally has more than one R value attached to it. Therefore, we aim to cluster the ROIs that have concordant best-fitting line parameters, across proper motions in a given patch, and across patches.

To perform this combination of overlapping ROIs, we have developed a three-step clustering algorithm (see Fig. 1 for a graphical illustration of these steps):

(i) In a given patch, we consider all ROIs with the same value of μ_ϕ^*. We group together ROIs adjacent in μ_λ which have concordant line parameters.[11] In this way, all ROIs in a patch are clustered into seeds which have the same μ_ϕ^* and consecutive values of μ_λ. For each seed, we add the line significances of its ROIs in quadrature to form a combined line significance $\sigma_L^{\rm tot}$.[12]

(ii) Next, we group together seeds at adjacent μ_ϕ^* based on the same criteria for concordance of line parameters. This forms protoclusters, as shown in the second-to-last step of Fig. 1.

(iii) Finally, we merge together protoclusters across adjacent patches using the same criteria for concordance of line parameters. This produces our final stream candidates, as shown in the final step of Fig. 1.

A schematic of steps (i)–(ii) is shown in Fig. 9. There we see 12 hypothetical ROIs that are coloured by seed. Neighbouring seeds are combined into protoclusters which are denoted by dashed boxes. The number of signal regions, $N_{\rm SR}$, in each protocluster is the number of ROIs in its largest seed.

In step (i), the rationale for grouping in μ_λ and not μ_ϕ^* is that in a given patch, ROIs with the same μ_λ but different μ_ϕ^* represent different, highly overlapping slices of the same SR, with each star that appears in multiple ROIs having the same R values from ANODE.

[10] We are implicitly assuming here that each ROI will contain at most one stream. We believe this is a safe assumption, since ROIs are fully localized in both proper motions and angular position.

[11] To be precise, we require the line parameters to be within $\Delta\theta = \pi/10$ and $\Delta\rho = 2^\circ$ of each other.

[12] We are careful not to interpret $\sigma_L^{\rm tot}$ as a meaningful statistical significance in this work; rather we think of it more loosely as a figure of merit or an anomaly score for stream detection. At best, $\sigma_L^{\rm tot}$ would be a local significance (i.e. ignoring an enormous and difficult-to-quantify look-elsewhere-effect), and would be based on the assumption (probably not completely true) that separate ANODE runs in neighbouring SRs return completely uncorrelated, random values of R on background-only stars.

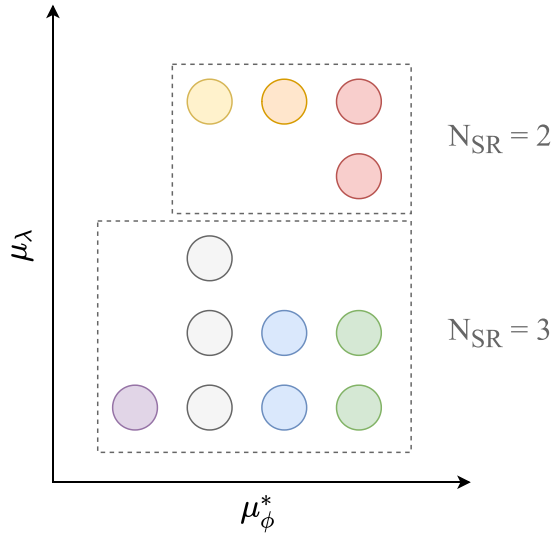

Figure 9. A schematic showing how regions of interest (ROIs) are combined into different protoclusters. The different colours denote different seeds, i.e. clusters of ROIs with adjacent μ_λ and the same μ^*_ϕ values. The boxes show how adjacent seeds are combined into protoclusters with different N_{SR}.

On the other hand, ROIs with the same μ^*_ϕ and different μ_λ represent different SRs, and each SR represents an independent ANODE training. Although the SRs are highly overlapping, the ANODE training is sufficiently stochastic that we take the outcome in different SRs to be quasi-independent. This motivates the adding in quadrature of the line significances of the ROIs in each seed.

In step (ii), for each protocluster, we characterize its significance by the seed with the highest $\sigma_L^{\rm tot}$ that it contains. The size of this seed we will call N_{SR} and is another measure of the significance of the protocluster. Note that we do not add the $\sigma_L^{\rm tot}$ values of different seeds in a protocluster together in quadrature, since these are highly correlated.

Applying the final merging and clustering steps to the 21 patches containing GD-1, we find that the 17 563 ROIs are clustered into 10,955 protoclusters. Of these, 10 267 have $N_{SR} = 1$; 606 have $N_{SR} = 2$; and 82 have $N_{SR} \geq 3$.

All else being equal, we expect real streams to have higher values of $\sigma_L^{\rm tot}$ and N_{SR}. We show in Fig. 10 histograms of the fraction of stars within the best-fitting line of each ROI that have been identified as candidate GD-1 stars by PWB18, for ROIs that belong with protoclusters with different values of N_{SR}. As can be seen, the fraction of candidate GD-1 stars (i.e. the 'purity' of the best-fitting line) is significantly improved when we require $N_{SR} \geq 3$.

In Fig. 11, we show the distributions of $\sigma_L^{\rm tot}$ values across the ROIs (the total line significance is for the protocluster that the ROI has been clustered into), for $N_{SR} \geq 3$. We see that there is clearly a bulk distribution at low $\sigma_L^{\rm tot}$ and then a tail of outliers, with the separation occurring around $\sigma_L^{\rm tot} = 8$. It is reasonable to suppose that the majority of these low-$\sigma_L^{\rm tot}$ corresponds to false positives, while the tail could correspond to real stream detections that should be subjected to more in-depth investigation.

4 DEMONSTRATING THE FULL VIA MACHINAE ALGORITHM WITH GD-1

Having described all the steps of the VIA MACHINAE algorithm, we now demonstrate the full algorithm on the 21 patches of the sky that contain GD-1. For the first step of the algorithm (ANODE), we

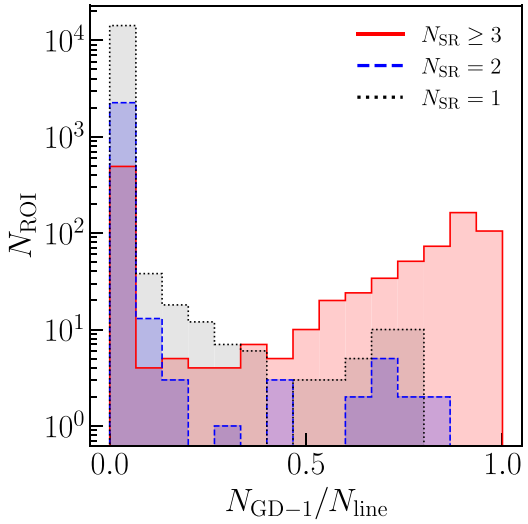

Figure 10. Histograms of the fraction of stars in the best-fitting line of each ROI that were identified as likely GD-1 stars by PWB18, for ROIs which are part of protoclusters with $N_{SR} = 1$ (black, dashed), $N_{SR} = 2$ (blue, dotted) and $N_{SR} \geq 3$ (red,solid). We see that requiring $N_{SR} \geq 3$ greatly increases the fraction of candidate GD-1 stars in the best-fitting line.

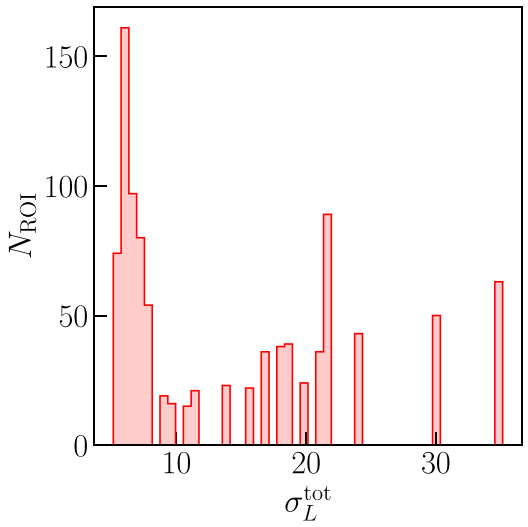

Figure 11. Histogram of the $\sigma_L^{\rm tot}$ values of protoclusters with $N_{SR} \geq 3$, with each protocluster weighted by the number of ROIs it contains.

used the 'Haswell' processors at NERSC, for a total of approximately 10 000 CPU-hours to analyse all 21 patches. For the subsequent steps of VIA MACHINAE (line finding, forming protoclusters, and forming stream candidates), we used the local HEP cluster at Rutgers, for a total of approximately 50 CPU-hours.

Motivated by the discussion in the previous subsection, we focus on only those protoclusters with $N_{SR} \geq 3$ and $\sigma_L^{\rm tot} \geq 8$. This leaves only 16 protoclusters. Merging these results in only two stream candidates, shown in Fig. 12. One might have expected far more stream candidates, given the enormous trials factors involved (e.g. $\mathcal{O}(10^4)$ ROIs that we started with). This is a sign that the cuts on N_{SR} and $\sigma_L^{\rm tot}$ that we have chosen are indeed effective at reducing the false positive rate.

The less prominent stream candidate, shown in blue, is built from a single protocluster representing 16 ROIs with $\sigma_L^{\rm tot} = 9.5$. It comes from the patch centred at $(\alpha, \delta) = (138.8^\circ, 25.1^\circ)$. The stream

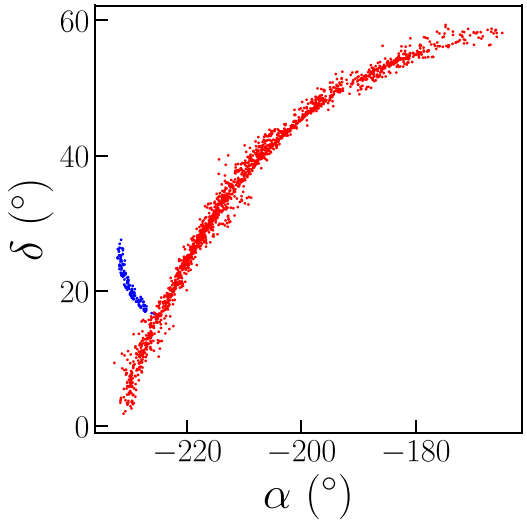

Figure 12. The two stream candidates built out of protoclusters with $N_{\rm SR} \geq 3$ and $\sigma_L^{\rm tot} \geq 7.5$.

candidate does not correspond to any known stream, and a priori it may be a real stream or a spurious detection. Closer inspection reveals that all of the high-*R* stars identified by ANODE are tightly clustered at the edge of the circular patch, almost perfectly aligned with the direction of the Galactic disc (and on the same side of the patch as the disc). Although this patch is $\gtrsim 30^\circ$ off the Galactic plane, we still observe a strong density gradient towards and aligned with the disc. Therefore, we suspect that ANODE has identified disc stars in this case, and not a stellar stream. We discuss this further in Appendix C.

The second, more prominent stream shown in red in Fig. 12, is composed of 15 protoclusters representing 518 ROIs, and is clearly GD-1. In Fig. 13 we show the positions, proper motions, and photometry of the 1688 stars in this stream candidate, overlaid on the locations of the stars tagged as likely GD-1 stream members by PWB18. In Fig. 14, we present another look at the comparison between VIA MACHINAE and PWB18, this time using the coordinate system aligned with the GD-1 stream (Koposov et al. 2010).[13]

Broadly speaking, we see that VIA MACHINAE has done an excellent job finding the GD-1 stars across the 21 patches of the sky considered in this work. Some notable features and caveats which deserve consideration are as follows:

(i) Fig. 14 shows that VIA MACHINAE has successfully reproduced some famous features of GD-1, including both gaps, the possible progenitor, and the 'spur' (PWB18).

(ii) We see that VIA MACHINAE confirms most of the additional 20° of GD-1 discovered in PWB18 (corresponding to $\alpha \lesssim -220^\circ$, or $\phi_1 \lesssim -60^\circ$). The left-most end of GD-1 ($\alpha \lesssim -235^\circ$, $\phi_1 \lesssim -80^\circ$) is missing from our stream candidate; this is because those patches were not included in our analysis as they were deemed too close to the disc ($|b| < 30^\circ$).

(iii) On the right-hand side of GD-1, we see that we are also missing stars compared to PWB18. Closer inspection of this missing region reveals that this segment of the stream was captured by only a single patch, centred on $(\alpha, \delta) = (212.7^\circ, 55.2^\circ)$, and the proper motion of GD-1 on this end of the stream is closer to $\mu_\lambda = 0$, increasing the number of background stars in the relevant SRs. We will return to this point and elaborate on it further below.

(iv) The feature protruding from the stream at $\alpha \sim -215^\circ$ and $\delta \sim 40^\circ$ (see Fig. 13) is most likely an artefact of our line finding procedure.

(v) Of the 1985 stars identified as likely members of GD-1 from PWB18, 1519 are in our fiducial (colour and magnitude) region, and 738 (49 per cent) overlap with the membership of our stream candidate.

(vi) The remaining 950 stars in our stream candidate were not tagged by PWB18. Some of these may very well be additional members of GD-1. However, the proper motion and colour–magnitude plots of Fig. 13 make clear that our method also picks up a significant number of non-stream stars, see e.g. the group of bright stars that are clearly not associated with the main GD-1 isochrone.

Regarding the stars missing from the right-most end ($\alpha \gtrsim -160^\circ$ or $\phi_1 \gtrsim -10^\circ$) of GD-1, it is notable that this side of the stream has values of μ_λ closest to zero (this is apparent in Fig. 13 after taking into account that $\mu_\lambda \approx \mu_\delta$ for these patches). Hence, this segment of GD-1 falls primarily in SRs with an increased number of background stars compared to the rest of the stream. The fact that we do not recover this part of GD-1 strongly suggests that successful stream-finding with VIA MACHINAE may require a minimum admixture of stream stars in the SR. In Fig. 15, we use the stream candidates from PWB18 as a proxy for the GD-1 stars, and plot the fraction of PWB18-tagged stars which are also identified as stream candidates by VIA MACHINAE (in each SR) versus the fraction of stars in each SR which are tagged by PWB18. As can be seen, the fraction of PWB18 stars also identified as stream members by VIA MACHINAE is strongly correlated with the fraction of stream stars in the SR. In particular, the overlap fraction drops precipitously when the stream makes up $\lesssim 0.1$ per cent of the total stars in the SR. All of the SRs through which the missing right-hand side of the GD-1 stream pass have a low fraction of stream stars. We believe this goes a long way toward explaining why VIA MACHINAE missed these members of GD-1. Further work is needed (including a study of other streams beyond GD-1) to determine if this threshold is a more general requirement of ANODE and VIA MACHINAE for stream detection.

Apart from the apparent required minimum *S*/*B* detection thresholds for ANODE and VIA MACHINAE, the fact remains that our stream candidate does not include all of the likely GD-1 stars tagged by PWB18 (completeness), and appears to include a substantial number of non-GD-1 stars (purity). However, this is not necessarily a drawback of the method. Rather, it reflects the emphasis placed by VIA MACHINAE on stream discovery rather than stream membership. When designing our algorithm, our choices were motivated to identify stream candidates at a sufficiently high statistical significance to overcome the random background. Decisions such as the number of high-*R* stars to include in each ROI and the line width in the Hough transform were made with this in mind, rather than maximizing accurate stream membership of the candidate. Loosening these criteria would likely recover more of the tagged stream stars than the 49 per cent identified here – this would have to be weighed against increasing the number of false-positive stream candidates identified across the full sky. Thus, the resulting stream candidates should be taken as signs for discovery, rather than an accurate membership study of particular stars and whether they belong to a stream. After stream discovery, the candidate must be considered individually, loosening or eliminating some of the algorithmic choices that are part of VIA MACHINAE. The density estimates from ANODE may continue to aid in this a posteriori analysis, but this is beyond the scope of this paper.

[13]To allow for direct comparison of our results with PWB18 in this section, we show the stars of the latter without correction for extinction, and apply the same cuts on their (uncorrected) magnitudes and colours of $g < 20.2$ and $0.5 < b - r < 1$ as we do for our fiducial sample.

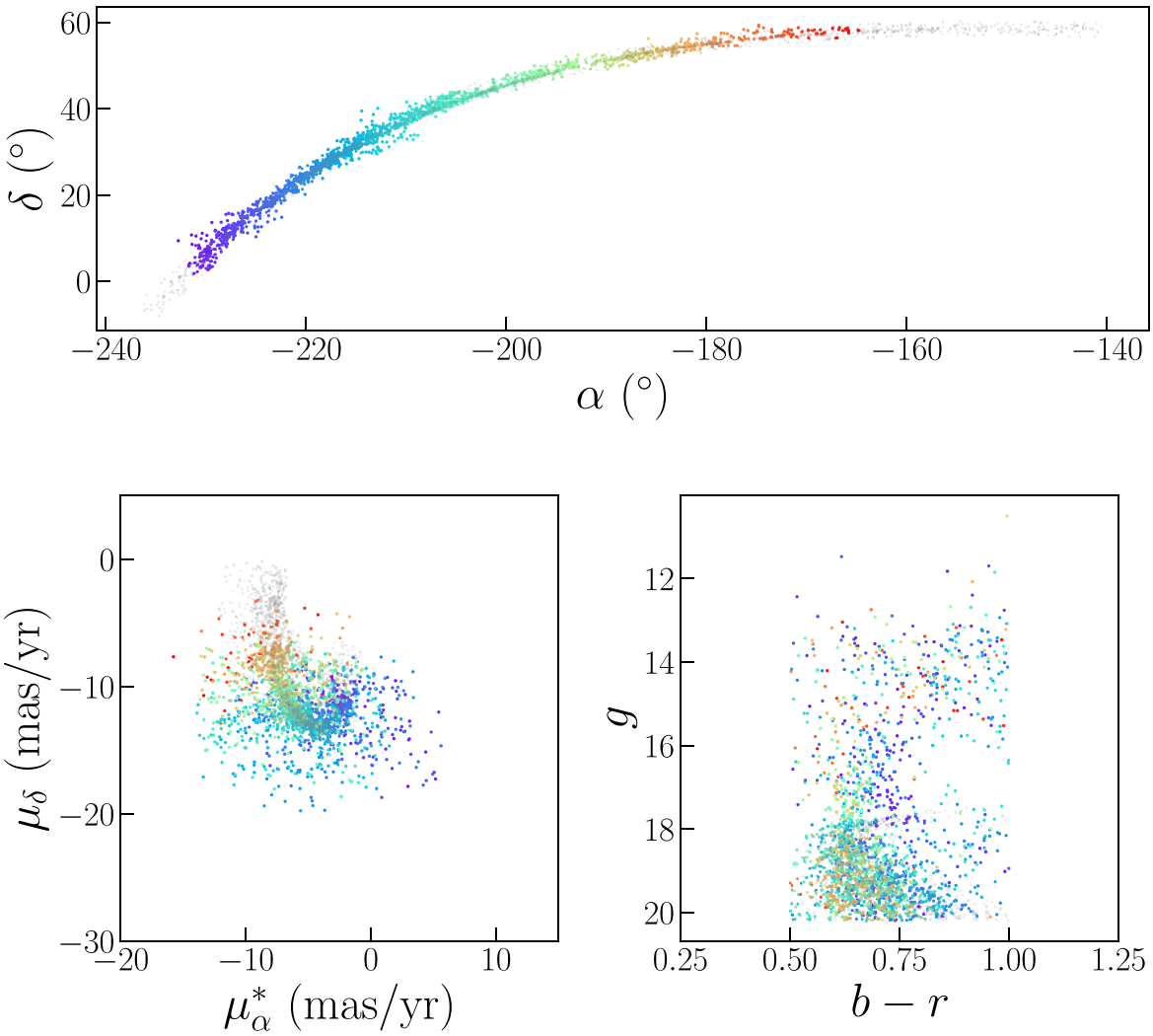

Figure 13. Scatter plots of the angular positions, proper motions, and colour/magnitudes of the 1688 stars in the more prominent of the two stream candidates identified by VIA MACHINAE, overlaid on the likely GD-1 stars tagged by PWB18 (grey) in the same region of g and $b - r$ space. The VIA MACHINAE stars are colour-coded by position in α, to facilitate cross referencing between the three individual scatter plots.

5 CONCLUSIONS

In this work, we described a new machine learning-based algorithm called VIA MACHINAE for stellar stream detection using *Gaia* DR2 data, and applied this technique to identify the GD-1 stream. As a particularly distinct stream with readily available membership catalogues to use for detailed comparisons, GD-1 is an excellent testbed for our algorithm.

The core of our technique is ANODE, a data-driven, unsupervised machine-learning algorithm that uses conditional probability density estimation to identify anomalous data points in a search region without having to explicitly model the background distribution. This approach is made possible by advances in deep learning that approximate the probability densities in an unsupervised way. We take as input for the ANODE training the angular position, proper motion, and photometry of the stars in *Gaia* DR2. No astrophysical knowledge is embedded into ANODE, other than in our choice to condition the probability estimation on one of the proper motion coordinates μ_λ – which is also used to define the search regions. This allows us to identify potential anomalies while remaining agnostic to the Galactic potential, orbits, or stellar composition of the streams.

The output of ANODE is a likelihood ratio R, with large R values corresponding to stars whose phase space density in the search region is larger than expected based on interpolation from the control regions. To turn these anomalies into stream detection, VIA MACHINAE engages in a number of additional steps. Some of these steps – concentrating on old, metal-poor stars (identified with a cut on $b - r$, without requiring the stars to lie on an isochrone), further slicing the data into regions of interest based on the other proper motion coordinate μ^*_ϕ – are designed to improve signal-to-noise of stream detection, without sacrificing too much of the model-independence. Other steps – automated line finding using the Hough transform, merging concordant best-fitting lines in adjacent ROIs and patches of the sky – are intended to build a stronger case for a stream detection (as opposed to some other anomalous structure or spurious false positive). The upshot is that VIA MACHINAE produces a list of stream candidates that have been found in multiple overlapping search regions with high significance.

Using this method, we recover the GD-1 stream across the 21 patches in our full-sky scan that include it. Although GD-1 is an atypically dense, cold and narrow stream, it is still non-trivial that our method is able to recover it in an unsupervised and fully automated

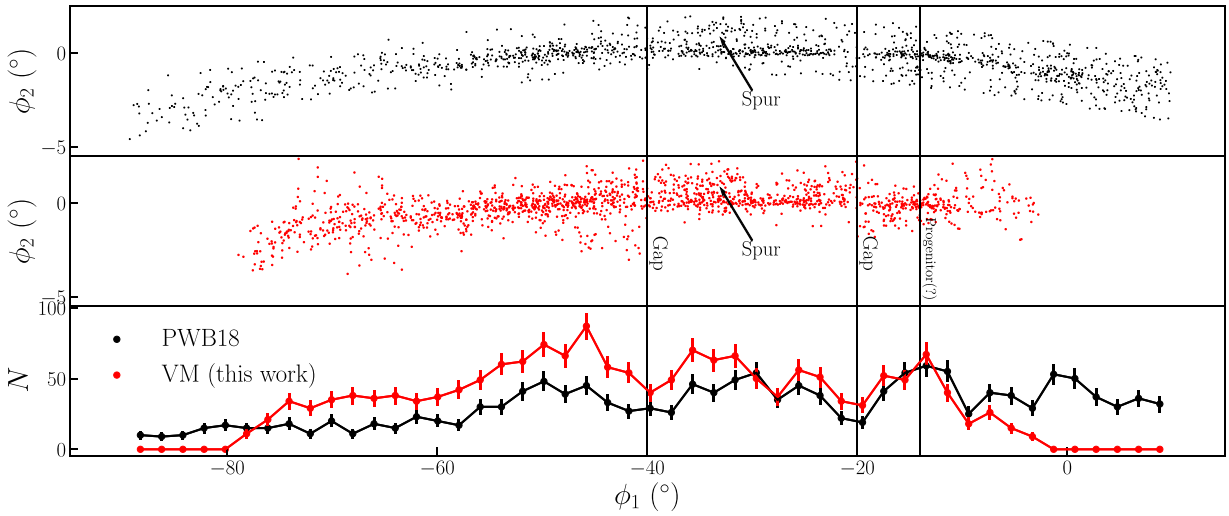

Figure 14. Comparison of the likely GD-1 stars from PWB18 (top, black) and the stream candidate stars identified by VIA MACHINAE (middle, red), in the GD-1 stream-aligned coordinate system (ϕ_1, ϕ_2) (Koposov et al. 2010). The location of previously identified features of GD-1 (two gaps, the possible progenitor location, and the spur) are indicated. Bottom row shows the number of candidate stream stars identified by PWB18 (black) and VIA MACHINAE (red) in ϕ_1 bins of width 2°; the error bars are purely statistical (Poissonian). In top and bottom panels, a cut on $g < 20.2$ and $0.5 < b - r < 1$ has been applied to the stars from PWB18 so that a direct comparison can be made with the stars in this analysis.

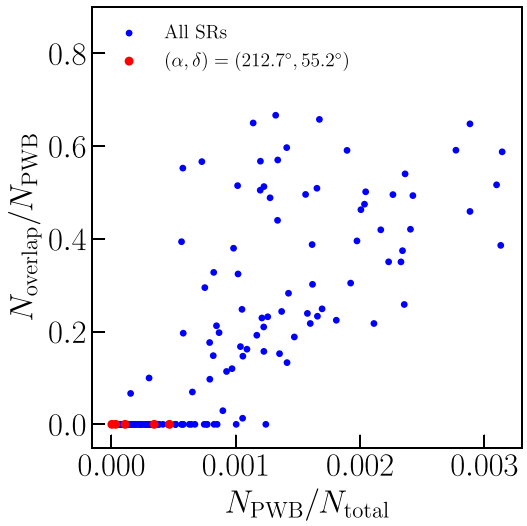

Figure 15. For each SR, we plot the fraction of total stars $N_{\rm total}$ in an SR which are identified as likely members of GD-1 by PWB18 ($N_{\rm PWB}$), compared to the fraction of $N_{\rm PWB}$ which are also identified by VIA MACHINAE as likely members of GD-1 ($N_{\rm overlap}$). The SRs which lie in the patch centred on (α, δ) = (212.7°, 55.2°) are shown in red. This patch contains the majority of the right-hand side of GD-1 which is not identified in our analysis.

way. Moreover, we chose to focus on GD-1 in this paper as it provides a clear, step-by-step introduction to our algorithm. The application of VIA MACHINAE to other known streams and to the full-sky data set will be discussed in a forthcoming paper Shih et al. (in preparation).

This initial application of unsupervised density estimators for stellar stream discovery, suggests other potentially interesting directions which recent advances in deep learning have made possible. Most obviously, our method can in principle be adapted to look for other interesting cold objects in the Milky Way, such as debris flow, tidal tails, and other stellar substructure (Johnston, Hernquist & Bolte 1996; Johnston 1998; Robertson et al. 2005; Font et al. 2006, 2011; Helmi 2020). Other methods of density estimation beside the MAF may also prove to be useful: we used the MAF because it is reasonably fast and easy to train and was demonstrated to perform well in the ANODE anomaly detection task in NS20. However, neural autoregressive flows (Huang et al. 2018), neural spline flows (Durkan et al. 2019), and mixture density networks (Bishop 1994) may possibly have improved performance in some or all contexts.

Having data-driven measures of 'signal' and 'background' densities may prove to be useful for problems beyond discovery. Sampling from these density estimators is possible, and might be a way to construct mock catalogues. The density estimates themselves might be useful for answer questions of stream membership. This could help in going beyond the VIA MACHINAE discovery steps outlined in this paper, and further establish the validity and accuracy of the proposed stream candidates.

ACKNOWLEDGEMENTS

We would like to thank A. Bonaca, D. Hogg, S. Pearson, A. Price-Whelan for helpful conversations; and Ting Li, Ben Nachman, and Bryan Ostdiek for comments on the manuscript. MB and DS are supported by the DOE under Award Number DOE-SC0010008. LN is supported by the DOE under Award Number DESC0011632, the Sherman Fairchild fellowship, the University of California Presidential fellowship, and the fellowship of theoretical astrophysics at Carnegie Observatories. LN is grateful for the generous support and hospitality of the Rutgers NHETC Visitor Program, where this work was initiated.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the

DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

DATA AVAILABILITY

This paper made use of the publicly available *Gaia* DR2 data. For the GD-1 stars identified through our analysis, please email the corresponding author.

REFERENCES

Astropy Collaboration 2013, A&A, 558, A33
Astropy Collaboration 2018, AJ, 156, 123
Banik N., Bovy J., 2019, MNRAS, 484, 2009
Banik N., Bovy J., Bertone G., Erkal D., de Boer T. J. L., 2021, J. Cosmol. Astropart. Phys., 10, 043
Banik N., Bovy J., Bertone G., Erkal D., de Boer T. J. L., 2021, MNRAS, 502, 2364
Belokurov V. et al., 2006, ApJ, 642, L137
Belokurov V., Erkal D., Evans N. W., Koposov S. E., Deason A. J., 2018, MNRAS, 478, 611
Bishop C. M., 1994, Mixture Density Networks. Technical Report NCRG/94/004. Aston University, Birmingham, UK
Bonaca A., Hogg D. W., Price-Whelan A. M., Conroy C., 2019, ApJ, 880, 38
Bonaca A. et al., 2020, ApJ, 892, L37
Borsato N. W., Martell S. L., Simpson J. D., 2020, MNRAS, 492, 1370
Boubert D., Everall A., 2020, MNRAS, 497, 4246
Carlberg R. G., Grillmair C. J., Hetherington N., 2012, ApJ, 760, 75
Duda R. O., Hart P. E., 1972, Commun. ACM, 15, 11
Durkan C., Bekasov A., Murray I., Papamakarios G., 2019, Neural Spline Flows,preprint (arXiv:1906.04032)
Erkal D., Koposov S. E., Belokurov V., 2017, MNRAS, 470, 60
Font A. S., Johnston K. V., Bullock J. S., Robertson B. E., 2006, ApJ, 646, 886
Font A. S., McCarthy I. G., Crain R. A., Theuns T., Schaye J., Wiersma R. P. C., Dalla Vecchia C., 2011, MNRAS, 416, 2802
Gaia Collaboration 2018, A&A, 616, A1
Gaia Collaboration 2021, A&A, 649, A1
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Grillmair C. J., 2006, ApJ, 645, L37
Grillmair C. J., Dionatos O., 2006, ApJ, 643, L17
Helmi A., 2020, ARA&A, 58, 205
Helmi A., White S. D. M., 1999, MNRAS, 307, 495
Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G. A., 2018, Nature, 563, 85
Hough P. V. C., 1959, Conf. Proc. C, 590914, 554
Huang C.-W., Krueger D., Lacoste A., Courville A., 2018, Neural Autoregressive Flows. preprint (arXiv:1804.00779)
Ibata R., Lewis G. F., Irwin M., Totten E., Quinn T., 2001, ApJ, 551, 294
Ibata R. A., Malhan K., Martin N. F., 2019, ApJ, 872, 152
Ibata R. et al., 2021, ApJ, 914, 123
Johnston K. V., 1998, ApJ, 495, 297
Johnston K. V., Hernquist L., Bolte M., 1996, ApJ, 465, 278
Johnston K. V., Zhao H., Spergel D. N., Hernquist L., 1999, ApJ, 512, L109
Kingma D. P., Ba J., 2017, Adam: A Method for Stochastic Optimization,preprint (arXiv:1412.6980)
Koposov S. E., Rix H.-W., Hogg D. W., 2010, ApJ, 712, 260
Kuhlen M., Lisanti M., Spergel D. N., 2012, Phys. Rev. D, 86, 063505
Küpper A. H. W., Balbinot E., Bonaca A., Johnston K. V., Hogg D. W., Kroupa P., Santiago B. X., 2015, ApJ, 803, 80
Lindegren L. et al., 2018, A&A, 616, A2
Lisanti M., Spergel D. N., 2012, Phys. Dark Univ., 1, 155
Luo A. L. et al., 2015, Res. Astron. Astrophys., 15, 1095
Malhan K., Ibata R. A., 2018, MNRAS, 477, 4063
Malhan K., Ibata R. A., 2019, MNRAS, 486, 2995
Malhan K., Ibata R. A., Goldman B., Martin N. F., Magnier E., Chambers K., 2018a, MNRAS, 478, 3862
Malhan K., Ibata R. A., Martin N. F., 2018b, MNRAS, 481, 3442
Malhan K., Ibata R. A., Carlberg R. G., Bellazzini M., Famaey B., Martin N. F., 2019, ApJ, 886, L7
Malhan K., Yuan Z., Ibata R., Arentsen A., Bellazzini M., Martin N. F., 2021a, ApJ, 920, 51
Malhan K., Valluri M., Freese K., 2021b, MNRAS, 501, 179
Meingast S., Alves J., 2019, A&A, 621, L3
Meingast S., Alves J., Fürnkranz V., 2019, A&A, 622, L13
Nachman B., Shih D., 2020, Phys. Rev. D, 101, 075042 (NS20)
Necib L., Lisanti M., Belokurov V., 2019a, ApJ, 874, 3
Necib L., Lisanti M., Garrison-Kimmel S., Wetzel A., Sanderson R., Hopkins P. F., Faucher-Giguère C.-A., Kereš D., 2019b, ApJ, 883, 27
Newberg H. J., 2010, AAS/Division of Dynamical Astronomy Meeting #41, 5.01
Newberg H. J. et al., 2002, ApJ, 569, 245
Neyman J., Pearson E. S., 1933, Phil. Trans. R. Soc. A, 231, 289
Odenkirchen M. et al., 2001, ApJ, 548, L165
Papamakarios G., Pavlakou T., Murray I., 2018, Masked Autoregressive Flow for Density Estimation. preprint (arXiv:1705.07057)
Papamakarios G., Nalisnick E., Rezende D. J., Mohamed S., Lakshminarayanan B., 2021, J. Mach. Learn. Res., 22, 1
Pearson S., Starkenburg T. K., Johnston K. V., Williams B. F., Ibata R. A., Khan R., 2019, ApJ, 883, 87
Pearson S., Clark S. E., Demirjian A. J., Johnston K. V., Ness M. K., Starkenburg T. K., Williams B. F., Ibata R. A., 2021, The Hough Stream Spotter: A New Method for Detecting Linear Structure in Resolved Stars and Application to the Stellar Halo of M31. preprint (arXiv:2107.00017)
Price-Whelan A. M., Bonaca A., 2018a, Gaia data, Pan-STARRS Photometry, and Stream Selection Masks for the Region around the GD-1 Stream, Available at: https://doi.org/10.5281/zenodo.1295543
Price-Whelan A. M., Bonaca A., 2018b, ApJ, 863, L20 (PWB18)
Purcell C. W., Zentner A. R., Wang M.-Y., 2012, J. Cosmol. Astropart. Phys., 2012, 027
Reino S., Rossi E. M., Sanderson R. E., Sellentin E., Helmi A., Koppelman H. H., Sharma S., 2021, MNRAS, 502, 4170
Rezende D. J., Papamakarios G., Racaniere S., Albergo M. S., Kanwar G., Shanahan P. E., Cranmer K., 2020, Normalizing Flows on Tori and Spheres. Accepted to the International Conference on Machine Learning (ICML),preprint (arXiv:2002.02428)
Robertson B., Bullock J. S., Font A. S., Johnston K. V., Hernquist L., 2005, ApJ, 632, 872
Sanders J. L., Binney J., 2013, MNRAS, 433, 1813
Sanders J. L., Bovy J., Erkal D., 2016, MNRAS, 457, 3817
Shipp N. et al., 2018, ApJ, 862, 114
Varghese A., Ibata R., Lewis G. F., 2011, MNRAS, 417, 198
York D. G. et al., 2000, AJ, 120, 1579
Yuan Z., Chang J., Banerjee P., Han J., Kang X., Smith M. C., 2018, ApJ, 863, 26
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Softw., 4, 1298

APPENDIX A: DETAILS OF THE MAF

A1 Training and model selection

For each SR – defined by a patch of the sky and a slice in μ_λ – we train two separate MAFs, one on the stars $\mu_\lambda \in [\mu_\lambda^{\rm min}, \mu_\lambda^{\rm max}]$ in the SR and one on the stars in its complement (the CR), $\mu_\lambda \notin [\mu_\lambda^{\rm min}, \mu_\lambda^{\rm max}]$. Before training, the data are standardized by shifting the mean in each feature to zero and normalizing the standard deviation to unity.

We opt not to divide the data up into training and validation sets, as doing so would dilute the significance of any stream detection. Based on direct inspection, we do not find any evidence for overfitting. For the density estimation, overfitting would typically correspond to $p(x)$

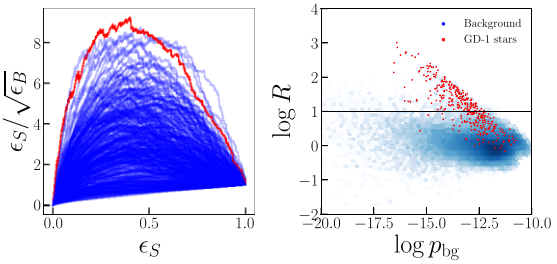

Figure A1. Left-hand panel: SIC curve of signal efficiency ϵ_S to $\epsilon_S/\sqrt{\epsilon_B}$ (for a background efficiency ϵ_B) as a cut is placed on log R, for all hyperparameters tested on the GD-1 example data set. Right-hand panel: Density plot of log $p_{\rm bg}$ versus log R for stars in the signal region of the GD-1 data set used for hyperparameter optimization, trained using the neural network parameters that maximize the true-positive over root false-positive rate.

degenerating into a set of delta functions centred on each point in the training data. This is generally not a concern for the MAF, and in fact it generally has the opposite problem (not being able to fit extremely sharp distributions). Also, the lower bound on data set size (SRs must have at least 20 000 stars, otherwise they are rejected) should be sufficient to mitigate overfitting for the dimensionality of the feature space.

For each MAF, we train for 150 epochs using the Adam optimizer (Kingma & Ba 2017). The learning rate is a hyperparameter that will be included in the scan to be described in Section A2. This number of epochs seemed to be sufficient for convergence, and training for significantly longer is computationally prohibitive. To smooth out fluctuations in the MAF from epoch to epoch arising from stochastic gradient descent, we calculate a running average for each star's probability density over the output of 20 consecutive training epochs.

To select the best model for each MAF, we employ the following approach. On general grounds, we expect the log R distribution to be roughly symmetric around 0 in the absence of any signal; any deviation from $R = 1$ is due to random fluctuations in the MAFs estimating the numerator or the denominator of the likelihood ratio. The better the performance of the density estimation, the more sharply peaked the R distribution should be around $R = 1$. Furthermore, we expect astrophysical signals (such as streams) to typically correspond to overdensities, not underdensities. Putting all this together, we select the 'best' epoch by considering the log R distribution for log $R < 0$, reflecting this across 0, and choosing the epoch with the smallest standard deviation in this symmetrized distribution. Strictly speaking, we only perform this for the MAF trained on the CR; for the MAF trained on the SR, we take the last 20 epochs, as we found this led to the best performance on tests with the labelled GD-1 stars.

A2 Hyperparameter optimization

Here we describe the hyperparameter optimization for the MAF neural network used for density estimation in this work. For the MAF architecture, these hyperparameters include the number of blocks in the neural network that make up the affine transformations, the number of hidden layers in the network, and the number of nodes in each hidden layer. Then there are the usual hyperparameters involved in training (mini-batch size, learning rate, etc.). The optimal values for these hyperparameters are not derivable from first principles; instead, we must perform a scan over the hyperparameters and select the configuration that maximizes the performance of the neural network. To measure performance, we will use the GD-1 labelled stars from PWB18 and quantify the signal/background discrimination power of the ANODE method as in Section 3.2.

To optimize the hyperparameters, we used a 15° patch of the sky centred on $(\alpha, \delta) = (140°, 30°)$, which contains a segment of the GD-1 stream (this patch was hand-selected, and is not one of the 200 centres described in Section 2). The patch contains 1.2×10^6 stars, of which 574 were tagged as stream stars by PWB18. The inner 10° fiducial region has 4.3×10^5 stars, 374 of which are stream-tagged.

Using these tagged stars, we hand-pick an SR defined by $\mu_\alpha^* \in [-8.75, -15]$ mas yr^{-1}, which contains all the stream stars and 1.7×10^5 total stars (6.4×10^4 in the fiducial region).

We varied the hyperparameters over batch size, number of epochs, learning rates, number of blocks, and number of hidden layers with:

$$\begin{aligned} \text{batch size} &= [256, 512, 1024] \\ \text{num.blocks} &= [16, 18, 20, 25] \\ \text{num.hidden} &= [16, 32, 64] \\ \text{num.epochs} &= [125, 150, 175] \\ \text{learning rate} &= [5 \times 10^{-5}, 7 \times 10^{-5}, 2 \times 10^{-5}, 4 \times 10^{-5}]. \end{aligned} \tag{A1}$$

We train the MAF for each combination of these five parameters. For each hyperparameter set in the scan, we calculate a log R value for each star in the fiducial (inner 10°) region. After training, we use the last epoch to construct the significance improvement characteristic (SIC) curve by varying a cut on log R. The SIC curves for each hyperparameter configuration in the scan are plotted in the left-hand panel of Fig. A1, with the optimal choice that maximizes $\epsilon_S/\sqrt{\epsilon_B}$ highlighted in red. On the right-hand panel of Fig. A1, we show the distribution of log $p_{\rm bg}$ versus log R for stars in the SR for the optimal set of hyperparameters. The optimal hyperparameters – 150 epochs, a batch size of 512, 20 blocks, 64 hidden blocks, and a learning rate of 7×10^{-5} – are then used for all MAF trainings in this work.

APPENDIX B: GLOBULAR CLUSTER DETECTION

Here we describe the simple algorithm we use to remove SRs that contain a suspected globular cluster. The presence of such overdensities in an SR is enough to distort the density estimation; the MAF cannot fit the delta function that is a globular cluster while simultaneously accurately describing the rest of the patch.

Based upon inspecting many patches pre- and post-ANODE, we find that the GCs that spoil the MAF are usually visible as a single bright pixel in a simple 2D density plot of the stars' positions in an SR. Given that there are thousands of SRs to sift through, we make a 2D histogram of the latitude and longitude of all the stars in an SR (recall, the patch size is 15°). With some tuning, we find that a good resolution is 120 × 120 bins across the 15° × 15° region. We then compute the mean number of counts $\bar{N}$, the max number of counts $N_{\rm max}$, and the standard deviation of the number of counts (as measured by the interquartile range) σ. We declare the SR to contain a likely GC if

$$\frac{N_{\rm max} - \bar{N}}{\sigma} > 4 \qquad \text{and} \qquad N_{\rm max} > 25. \tag{B1}$$

In other words, the bin with the maximum number of stars had to have at least 25 stars, and had to be at least '4σ' significant over the background stellar distribution.

Using these simple criteria, we find 1381 (out of 6117) SRs in the full-sky data set contain a GC candidate. We have visually inspected

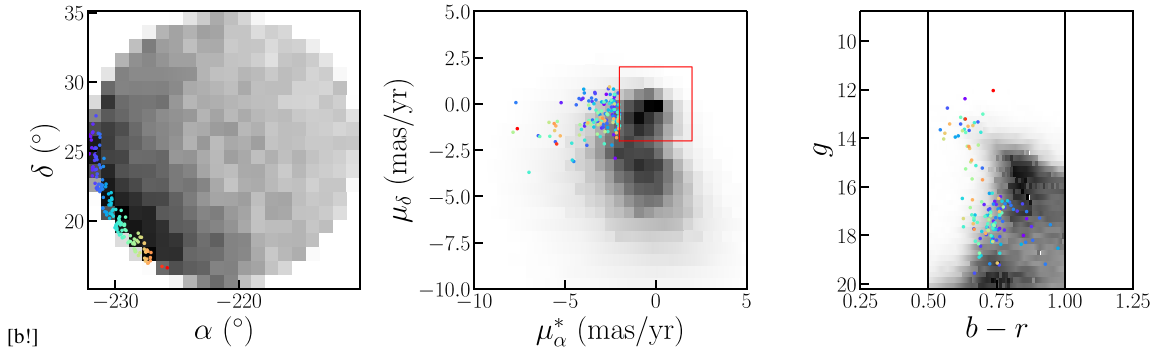

Figure C1. Scatter plots of the angular positions, proper motions, and colour/magnitudes of the stars in the second, less prominent stream candidate identified by VIA MACHINAE, overlaid on 2D histograms of all the stars in the circular patch that contains this stream candidate (darker pixels indicate higher density of stars). As in Fig. 13, the VIA MACHINAE stars are colour-coded by position in α, to facilitate cross referencing between the three individual scatter plots.

all of the SRs containing GC candidates and confirmed that the selections appear to be reasonable.

APPENDIX C: COMMENTS ON STREAM 2 AND DISC STARS

Here we elaborate further on the second, less prominent stream candidate tagged by the full VIA MACHINAE algorithm in the 21 patches containing GD-1. As described in Section 4, this stream candidate is contained completely in a single patch (centred on $(\alpha, \delta) = (138.8^\circ, 25.1^\circ)$), and all of the high-$R$ stars follow tightly the edge of the circular patch, aligned and on the same side as the Galactic disc. We illustrate this further in Fig. C1, which show the stars of the stream candidates in angular position, proper motion, and colour/magnitude space, overlaid on top of density plots of all the stars in the patch containing the stream candidate. This shows clearly how the stream candidate is aligned with the density gradient in the patch (which in turn is aligned with the Galactic disc, which one can check by transforming to Galactic coordinates (ℓ, b)). We also see that the stream candidate is clustered in proper motion space close to $\mu^*_\alpha,\ \mu_\delta \sim 0$, which as we have noted in Section 3.3 is a significant source of false positives for the ANODE method. Finally, we note that (unlike for GD-1 and other known streams), there is no noticeable correlation between the position along the stream and the proper motion. Taken together, we view this as strong evidence that this second stream candidate is likely to be a false positive.

More generally, we observe a strong gradient in stellar density towards the Galactic disc in many patches and SRs. There is also likely a strong correlation between disc stars and proper motion within a patch.[14] Therefore, it is potentially concerning that VIA MACHINAE could systematically misidentify disc stars as stream stars.

A conservative approach to avoid this misidentification is to reject all ROIs where the line-finder returned best-fitting parameters that are at the edge of the patch closest to the Galactic disc and parallel to it. Specifically, we propose to cut out all ROIs whose best-fitting line radius has $|\rho| > 9.5^\circ$, slope less than 0.2 radians in Galactic ℓ, b coordinates (that is, aligned with the disc), and are localized on the side of the patch nearest to the disc. This requirement removes only 91 ROIs (out of $\approx 17\,000$) from our sample. Such cut would eliminate the second stream that we find in Section 4, but it would not affect the GD-1 stream candidate at all.

[14] Understanding this correlation requires modelling stellar orbits in the Milky Way, and a detailed understanding of projection and line-of-sight effects. This is beyond the scope of the present work.

This paper has been typeset from a TEX/LATEX file prepared by the author.