Title

Weakly supervised anomaly detection in the Milky Way

Permalink

https://escholarship.org/uc/item/6zj335f9

Journal

Monthly Notices of the Royal Astronomical Society, 527(3)

ISSN

0035-8711

Authors

Pettee, Mariel
Thanvantri, Sowmya
Nachman, Benjamin
et al.

Publication Date

2023-11-27

DOI

10.1093/mnras/stad3663

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **527,** 8459–8474 (2024) https://doi.org/10.1093/mnras/stad3663
Advance Access publication 2023 November 27

Weakly supervised anomaly detection in the Milky Way

Mariel Pettee,[1★] Sowmya Thanvantri,[2] Benjamin Nachman,[1] David Shih,[3] Matthew R. Buckley[3] and Jack H. Collins[4,5]

[1] *Lawrence Berkeley National Laboratory, Physics Division, Berkeley, CA, 94720, USA*
[2] *University of California, Berkeley, Dept. of Electrical Engineering and Computer Sciences, Berkeley, CA, 94720, USA*
[3] *Rutgers University, Dept. of Physics and Astronomy, New Brunswick, NJ, 08854, USA*
[4] *SLAC National Accelerator Laboratory, Menlo Park, CA, 94025, USA*
[5] *Bosch Research North America, Sunnyvale, CA, 94085, USA*

Accepted 2023 November 20. Received 2023 November 8; in original form 2023 May 17

ABSTRACT

Large-scale astrophysics data sets present an opportunity for new machine learning techniques to identify regions of interest that might otherwise be overlooked by traditional searches. To this end, we demonstrate how Classification Without Labels (CWOLA), a weakly supervised anomaly detection method, can help identify cold stellar streams within the more than one billion Milky Way stars observed by the *Gaia* satellite. CWOLA operates without the use of labelled streams or knowledge of astrophysical principles. Instead, it uses a classifier to distinguish between mixed samples for which the proportions of signal and background samples are unknown. As a proof of concept, we demonstrate that this computationally lightweight strategy is able to detect both simulated streams and the known stream GD-1 in data. Originally designed for high-energy collider physics, this technique may have broad applicability within astrophysics as well as other domains interested in identifying localized anomalies.

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

1 INTRODUCTION

1.1 Motivation

The history of our home Galaxy, the Milky Way, has been marked by the ongoing aggregation of stars, gas, and dark matter from various sources throughout the Universe.These accumulation events include mergers with other galaxies as well as smaller scale gravitationally bound groupings of stars such as globular clusters. Though many such collisions occurred in the distant past, lingering remnants from more recent collisions contain crucial information revealing the Milky Way's merger history (Johnston 1998; Helmi & White 1999; Belokurov et al. 2006, 2018; Helmi et al. 2018; Malhan et al. 2021), underlying gravitational potential (Dehnen et al. 2004; Eyre & Binney 2009; Law & Majewski 2010; Kamdar, Conroy & Ting 2021; Reino et al. 2021; Nibauer et al. 2022), and dark matter content (Carlberg, Grillmair & Hetherington 2012; Purcell, Zentner & Wang 2012; Erkal et al. 2016; Sanders, Bovy & Erkal 2016; Banik & Bovy 2019; Bonaca et al. 2019, 2020; Necib et al. 2019).

Since 1971, astronomers have observed collections of stars called *stellar streams*: thin, ribbon-like arcs orbiting the Milky Way's Galactic centre (Eggen 1971). These dynamically cold associations of stars are thought to be the result of gravitational tidal forces from the Milky Way disrupting and warping nearby low-mass progenitors – dwarf galaxies or globular clusters – until the stars are no longer gravitationally self-bound. Due to their shared origin, the stars tend to share many characteristics ranging from proper velocity to age.

To date, around 100 stellar streams have been identified in the Milky Way (Mateu 2023). They are challenging to discover and study due to their sparse densities and wide angular extents. The number of streams has notably grown in recent years, however, thanks to large data releases from the *Gaia* mission (Prusti et al. 2016). The *Gaia* mission is poised to release even more substantial data sets surveying stars in the Milky Way in the coming years – a catalogue of 66 months of data around 2026 and the full archive of all mission data around 2031. Lightweight computational methods designed to efficiently identify stellar streams will be essential for analysing these upcoming data releases.

Gaining a more detailed understanding of the structure of known streams, as well as uncovering additional streams, will be critical for deepening our understanding of the Milky Way. An extensive catalogue of high-precision stream measurements would greatly improve our estimation of the particularities of Galactic large-scale and small-scale structures, including contributions from cold dark matter.

1.2 Related work

Traditional methods of identifying stellar streams look for groupings of stars that are similar along various metrics: colour and magnitude (Rockosi et al. 2002; Balbinot et al. 2011), velocity (Duffau et al. 2005; Arifyanto & Fuchs 2006; Williams et al. 2011), or position along great circle paths across the sky (Johnston, Hernquist & Bolte 1996; Mateu, Read & Kawata 2017). More recently, an automated technique by Malhan & Ibata (2018) called STREAMFINDER leverages

[★] E-mail: mpettee@lbl.gov

both position and kinematic information to construct volumes called ‘hypertubes’ in a multidimensional phase space around a stream candidate, and then iterates to optimize a statistic similar to a log-likelihood to determine the best parameters for a given hypertube. While this algorithm has led to the discovery of multiple low-density stellar streams, it makes several assumptions based on astrophysical principles. For instance, it assigns the distance to a stream based on a particular choice of isochrone and assumes a specific Milky Way potential to calculate a stream’s orbit. Data mining and clustering techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) have also been applied to *Gaia* data for stellar stream searches (Borsato, Martell & Simpson 2019).

Machine learning techniques have also been deployed in search of stellar streams. Recently, an unsupervised machine learning technique called VIA MACHINAE (Shih et al. 2021; Shih, Buckley & Necib 2023), based on the ANODE (ANOmaly detection with Density Estimation) method (Nachman & Shih 2020) originally designed for high-energy particle physics searches, was applied towards the automated discovery of stellar streams within the *Gaia* Data Release 2 (DR2) catalogue. VIA MACHINAE combines a normalizing flow density estimation technique (Rezende & Mohamed 2015) for anomaly detection with a line-finding algorithm to identify stellar stream candidates without the use of detailed assumptions about isochrones or stream orbits.

1.3 CWoLa: classification without labels

Machine learning has been widely and successfully applied in the context of fundamental physics to the classification and description of various physical phenomena ranging from subatomic to cosmological scales. These techniques excel at identifying complex patterns in a data set without imposing any prior assumptions about their distributions. When dealing with real data with partially inaccurate or incomplete labels, weakly supervised machine learning methods can be particularly helpful.

The fields of high-energy particle physics and astrophysics share a common interest in identifying localized features, meaning overdensities of data concentrated in contained regions of phase space, within vast and high-dimensional data sets. Model-independent forms of anomaly detection – the process of identifying these localized features that deviate from a data set’s typical characteristics – can efficiently filter these large data sets in an unbiased manner and aid in potential discoveries.

In this paper, we demonstrate the first astrophysical application of *Classification Without Labels* (CWoLa[1]; Metodiev, Nachman & Thaler 2017), a weakly supervised machine learning technique based on a simple, lightweight neural network (NN) classifier. CWoLa was originally designed for identifying particles within high-energy particle physics data sets, and it has been applied as a promising model-agnostic anomaly detection method for searching for localized features such as the potential signatures of new fundamental particles at the Large Hadron Collider (Collins, Howe & Nachman 2018, 2019; Aad et al. 2020). Until now, it has not been used on astrophysics data sets.

We apply CWoLa to the search for stellar streams by looking for localized anomalies in proper motion. Stellar streams are kinematically cold, meaning their constituent stars tend to have similar velocities. Compared with background stars, stream stars will therefore have narrower distributions of proper motion.

We scan for these overdensities in regions defined by one proper motion coordinate and train NNs to assign an anomaly score to stars using five input variables: two angular position coordinates, one proper motion coordinate (the one not used for the scan), magnitude, and colour. Once the anomaly scores are assigned, we look at the subsets of stars from each scanning window with the highest anomaly scores, cluster them, and apply fiducial selections to further refine them. As a proof-of-concept result to motivate a future full-sky scan, we apply these techniques on a known stream called GD-1 as well as simulated stellar streams as benchmarks.

This analysis uses the same data sets as in Shih et al. (2021) and follows the same general analysis structure. We have implemented a few key methodological differences, however, with the aim of achieving similar performance in anomaly detection with a much more computationally lightweight framework. A detailed comparison of these analyses is presented in Section 5.

1.4 Outline

This paper is organized as follows. First, in Section 2, we describe the *Gaia* data set and how it is processed for use in our anomaly detection studies. Then, in Section 3, we explain the methodology of applying CWoLa for anomaly detection and our particular implementation of CWoLa on *Gaia* data, including how we define the signal and sideband regions as well as the NN model architecture and training procedure. The results of applying CWoLa to the known stellar stream GD-1 are listed in Section 4. Finally, in Section 5, we conclude with a discussion of CWoLa’s potential usefulness in aiding future stellar stream discoveries and some further steps for this work that can bring us closer to that goal.

2 *Gaia* DATA SET

The *Gaia* catalogues are extensive astronomical data releases mapping the stars populating the Milky Way (Prusti et al. 2016). The *Gaia* satellite itself was launched in 2013, and its data catalogue is being released in discrete stages throughout its operational lifespan through 2025. The data analysed for this work come from *Gaia* DR2, the collection of *Gaia*’s observations from 2014 July 25 to 2016 May 23 (Brown et al. 2018). *Gaia* DR2 contains position, proper motion, and photometric information for approximately 1.3 billion stars, representing around 1 per cent of the total star population of the Milky Way. While this analysis was already in process, *Gaia* released two additional data releases: Early DR3 (eDR3) and DR3. Were this analysis to be extended using eDR3 or DR3 data, we would likely see some further improvements due to reduced measurement errors. Other changes in eDR3 and DR3 include improved distance and radial velocity measurements for a small subset of stars, but these variables are not considered in this analysis.

While *Gaia* DR2 also contains information on parallax and mean radial velocity, these variables are not considered as input variables to CWoLa for this analysis. We exclude parallax because it is not as reliable a feature as our other observables for stars as distant as the stream members in which we are most interested. However, we do use parallax to apply a cut on the *Gaia* data to restrict stars to a maximum parallax of 1, meaning stars at a distance of at least 1 kpc. This aligns well with the catalogue in Mateu (2023), in which all reported streams have distances >1 kpc, with a mean distance of about 15 kpc. Mean radial velocity, or motion along the axis between the Earth and each star, is also measured for some stars in the *Gaia* data set, but *Gaia* DR2 contains radial velocities for only about 7 million stars, representing less than a per cent of the overall star

[1] Note: CWoLa is pronounced ‘koala’.

catalogue. We therefore omit radial velocity in order to maximize available training statistics.

As in Shih et al. (2021), we use the stellar stream GD-1, discovered in 2006 (Grillmair & Dionatos 2006), as the main candidate for evaluating the performance of CWoLa as a stream-finding technique. Most likely the remains of a tidally disrupted globular cluster, GD-1 consists of primarily metal-poor stars totalling approximately $2 \times 10^4\,\mathrm{M}_{\odot}$ (Koposov, Rix & Hogg 2010). It lies at a distance of approximately 10 kpc from the Sun and 15 kpc from the Galactic Centre. GD-1 is especially narrow (Grillmair & Dionatos 2006), dynamically cold (de Boer, Erkal & Gieles 2020), and bright (Erkal et al. 2016) compared to other stellar streams in the Milky Way. It also contains various known physical peculiarities including gaps and wiggles (de Boer et al. 2018), offshoots ('spurs'), and overdensities ('blobs'; Bonaca et al. 2019), and even a surrounding 'cocoon' of stars (Malhan et al. 2019). While CWoLa is designed primarily for stream discovery, not comprehensive stream population labelling, we expect that the most stream-like stars identified by the model will cover a wide range of angular positions along the stream due to the breadth of stars with similar proper motions. The reconstruction of some of the density perturbations along the angular extents of GD-1 in the CWoLa outputs can therefore serve as an additional metric indicating the physical validity of the identified stream stars.

2.1 Data pre-processing

We use the same data set of *Gaia* stars as in Shih et al. (2021). Following the same data processing methodology, we train our model on a series of 21 overlapping circular 'patches' of the *Gaia* data set with radius 15°. While the natural angular position coordinates in DR2 are right ascension (α) and declination (δ), we use rotated and centred coordinates ϕ and λ (as well as rotated proper motion coordinates μ_ϕ and μ_λ) such that each patch has a Euclidean distance metric and is centred at $(\alpha_0,\, \delta_0) = (0°,\, 0°)$. This transformation is performed using ASTROPY (Astropy Collaboration 2013, 2018, 2022). Each patch's centre location is also documented in Shih et al. (2021).

Beyond these two rotated and centred angular positions, we consider four additional features associated with each star in the data set: two angular proper motions [$\mu_{\phi*}$, where $\mu_{\phi*} \equiv \mu_\phi \cos(\lambda)$, and μ_λ], colour ($b - r$, where b represents the brightness of the blue photometer and r represents the brightness of the red photometer), and magnitude (g). Distributions of the six relevant variables used in the analysis (ϕ, λ, $\mu_{\phi*}$, μ_λ, $b - r$, and g) are shown for one such patch in Fig. 1. These patches cover an irregularly shaped region stretching between approximately $\alpha \in [-250°, -100°]$ and $\delta \in [-10°, 80°]$.

The *Gaia* data set has inherent measurement uncertainties for each observable. At the median magnitude in our data set of $g \approx 17$, the two proper motion coordinates have median uncertainties of 0.158 mas yr^{-1} ($\mu_\alpha*$) and 0.137 mas yr^{-1} (μ_δ) (Lindegren et al. 2018) – about 3 per cent of their respective median values in our data set. Since CWoLa searches for group anomalies in proper motion, individual effects on star measurements are not of primary concern. Collectively, though, this uncertainty could slightly widen the apparent dispersion of the proper motion distribution of a stream, weakening CWoLa's discrimination power, particularly in sparser regions of the stream. The auxiliary variables also have inherent uncertainties. Both position coordinates have median uncertainties of about 0.1 mas. The uncertainty in g varies from 1 mmag for the brightest stars to a mere 20 mmag for the faintest stars. On the other hand, the uncertainty in $b - r$ varies from 1 mmag for the brightest stars to 200 mmag for the faintest stars, where this uncertainty could result in a smeared distribution of stream stars located in the range of $b - r \in (0.5, 1)$.

Following these selections, 1957 of the approximately 8 million total stars considered for this analysis are labelled as likely belonging to the GD-1 stream using the catalogues developed by Price-Whelan & Bonaca (2018a, b). This choice of labelling is based on selections in position, proper motion, and along an isochrone in colour and magnitude. While these labels cannot be considered fully accurate or complete, they serve as a helpful reference for evaluating our model's efficacy.

3 METHODS

3.1 Classification without labels

CWoLa is a weakly supervised machine learning technique designed to find anomalous features in a data set that are localized along at least one dimension. It was originally designed for applications within high-energy particle physics, where mixtures of particle classes with unknown proportions of signal and background are common. It detects anomalies by scanning along a localized dimension and learning to distinguish between mixtures of data classes where the precise class proportions within each mixture need not be known. In the original particle physics context, the localized dimension could be a property of a particle in the final state of a collision such as the invariant mass, while in this result, the localized dimension is the proper motion of a star. Simply by learning to differentiate regions with higher versus lower proportions of signal, i.e. 'signal' versus 'sideband' regions, CWoLa can be a powerful indicator of patterns of anomalous events.

Consider a signal-enriched mixture M_1 and a signal-depleted mixture M_2, as shown in Fig. 2(a). 'Signal' refers to an object class of interest – here, a member of a localized anomalous feature, such as a stellar stream, that one would like to detect – while 'background' refers to objects not part of the anomaly. Both mixtures contain signal and background events, but the signal-enriched mixture has significantly more signal events relative to the signal-depleted mixture (i.e. $f_1 > f_2$, where f_i indicates the fraction of signal events in each mixture). We exploit the fact (see proof of theorem 1 in Metodiev et al. 2017) that an optimal classifier trained to distinguish events between M_1 and M_2 is the same as an optimal classifier trained in a fully supervised manner to distinguish signal from background events. Importantly, the exact proportions of signal in each mixture (f_1 and f_2) need not be known for this to hold. This theorem relies on the Neyman–Pearson lemma (Neyman & Pearson 1933) that states that an optimal classifier $h(\boldsymbol{x})$ is any function monotonic to the likelihood ratio constructed from the probability distributions of signal and background $p_{\mathrm{S}}(\boldsymbol{x})$ and $p_{\mathrm{B}}(\boldsymbol{x})$ for input variables $\boldsymbol{x}$.

We can apply CWoLa as a model-agnostic, data-driven anomaly detection technique (Collins et al. 2018, 2019) by identifying a certain feature of our data set that might contain a localized anomaly, as illustrated in Fig. 2(b). We then train a fully supervised classifier to distinguish between events from a 'signal region' and a surrounding 'sideband region', as defined by ranges of this feature. The inputs to this classifier are auxiliary variables that should be decorrelated from the characteristic used to define the signal and sideband regions if no anomaly is present. If an anomaly is present and contained primarily in the signal region, then we expect the anomalous events to be ranked more highly by the classifier. We can then repeat this process by sliding the signal and sideband windows across a range of values. For each choice of signal and sideband, we apply a threshold

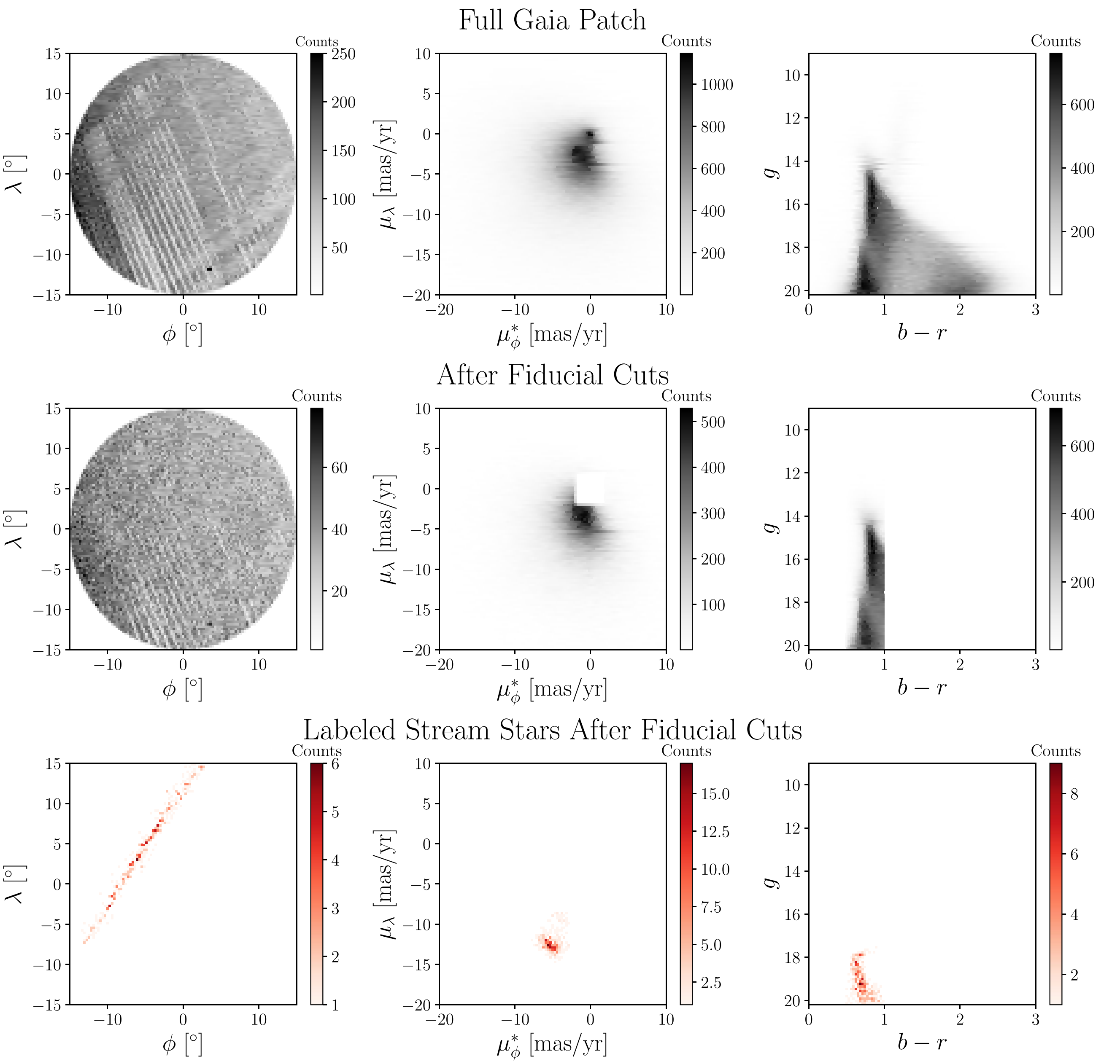


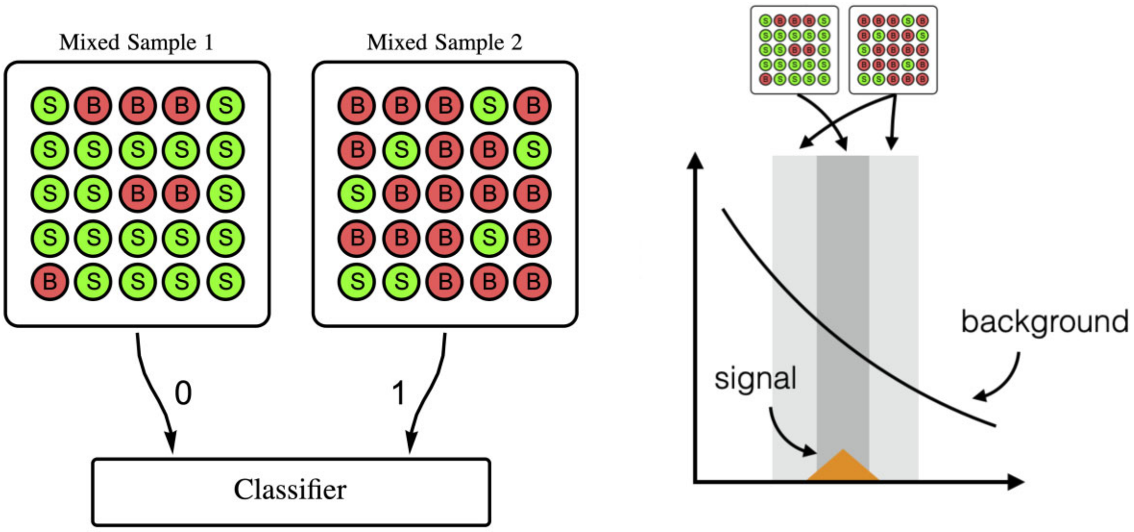
Figure 1. Two-dimensional histograms of the six features used in this analysis are illustrated for a single patch in the sky containing some GD-1 stars. This patch is centred at Galactic longitude $l = 207.0$ and latitude $b = 50.2$. The top row shows the full patch with no selections applied. The second row shows the patch with fiducial selections applied: $g < 20.2$ to reduce streaking; $|\mu_\lambda| > 2$ mas yr^{-1} or $|\mu^*_\phi| > 2$ mas yr^{-1} to remove too-distant stars; and $0.5 \leq b - r \leq 1$ to focus on identifying cold stellar streams. The third row indicates the six features for the GD-1 stream following the fiducial selections.

on the classifier output score (i.e. the top N events or a top percentile of the test set) such that only the highest score events remain. When an anomaly is present, these highest score events will contain an enhanced signal-to-noise ratio of events.

The CWoLa anomaly search makes two key assumptions during its procedure. First, it requires that the anomaly is localized in the dimension over which we search. Because stellar streams are kinematically cold, they are relatively localized in both coordinates of proper motion. We select μ_λ as the primary coordinate used to define the signal and sideband regions for each patch of DR2. A histogram demonstrating the highly localized nature of μ_λ within an example patch of GD-1 is shown in Fig. 3. Secondly, it expects that background and signal events are indistinguishable between the signal and sideband regions. This is a reasonable assumption, as shown for an example patch of data in Fig. 4.

3.2 Defining signal and sideband regions

For each of the 21 patches of DR2 considered in this study, we construct a signal region to contain the bulk of the stream stars available and neighbouring sideband regions such that the background stars in both signal and sideband regions are as close to indistinguishable as possible. Ideally, the stars in both regions should have similar characteristics: background stars in the signal



(a) Signal-enriched (M_1) vs. Signal-depleted (M_2)

(b) A localized anomaly

Figure 2. Signal-enriched and signal-depleted groups are pictured above. The data points labelled 'S' represent signal events, while the data points labelled 'B' represent background events. The signal and sideband regions are chosen such that more signal events (shown as a triangular peak) are located in the central signal region than the surrounding sideband region.

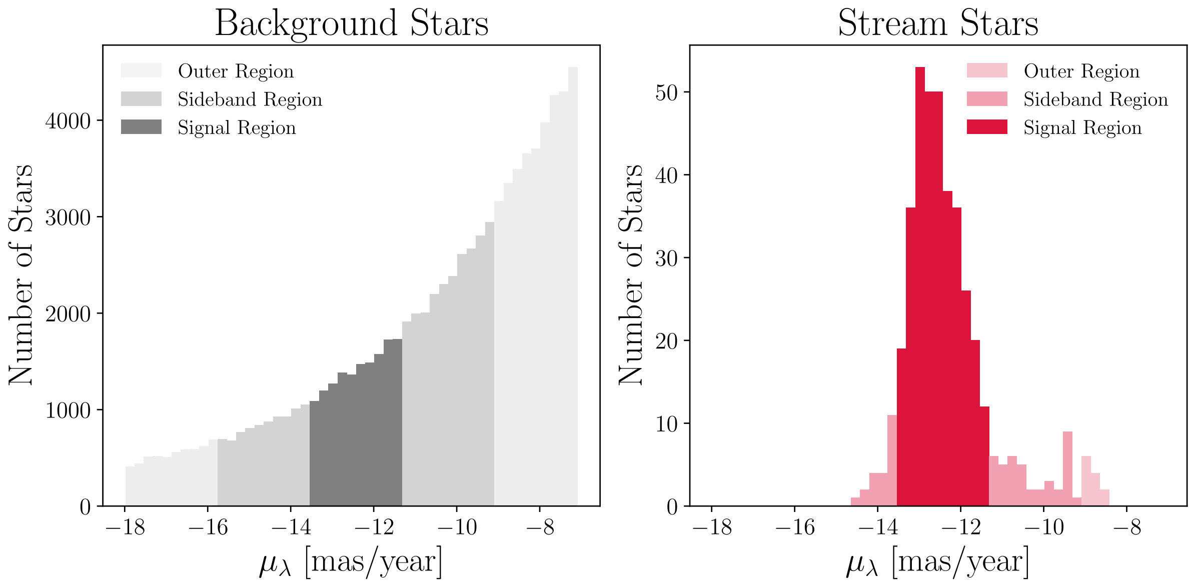


Figure 3. Stars associated with the stellar stream GD-1 are highly localized in μ_λ space in comparison with background stars for the same patch of *Gaia* data seen in Fig. 1. The signal region, shown in the darkest regions in each plot, is defined by taking $\pm 1\sigma$ from the median μ_λ value for the stream stars, which in this case is $[-13.6, -11.4]$. The sideband region is defined by taking $\pm 3\sigma$ from the stream's median μ_λ value, excluding the signal region: $[-15.8, -13.6)$ and $(-11.4, -9.3]$.

region should closely resemble background stars in the sideband region, and same logic applies for the stream stars. Stars belonging to a stellar stream make up a small fraction of total stars within each patch, so in our case, each signal region will still be dominated by background stars not labelled as belonging to the stream. However, each signal region should have a higher signal-to-background ratio than the sideband region.

Signal regions are ideally defined by a range of μ_λ values that encompass the bulk of the stream stars. There are many valid ways to define these regions in general, and in some cases, the best definitions may be model-dependent.

For this proof-of-concept result, we opted for idealized signal and sideband limits based on where we know the stream to be concentrated in proper motion. However, it is crucial to note that for a full-scale anomaly search, one could scan across a range of μ_λ values, meaning that one should still be sensitive to streams even if different signal and sideband regions were selected.

In this case, we define the signal region in each patch as the region within one standard deviation of the median μ_λ of the GD-1 stars in the patch. The sideband region within each patch is then defined as the stars falling within $[-3\sigma, -\sigma]$ or $[\sigma, 3\sigma]$ of the median. Given that the signal region encompasses a bulk of the stream, the sideband

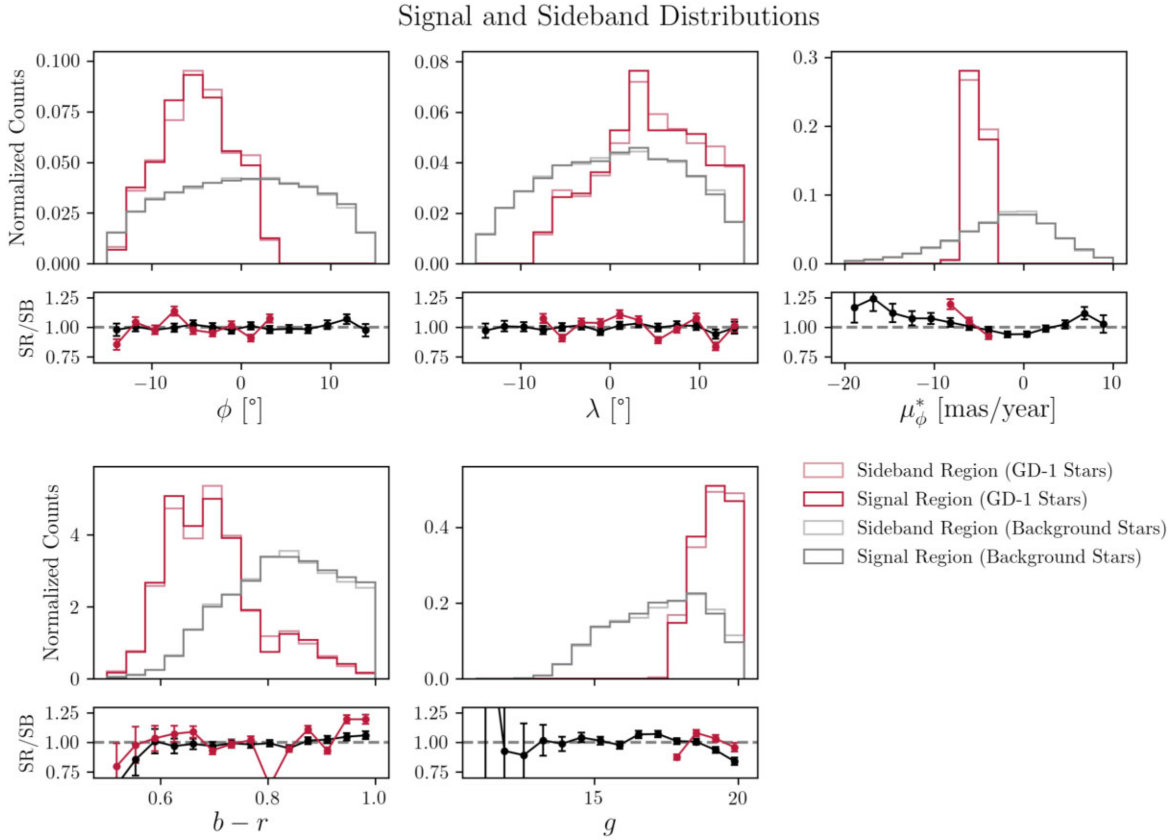


Figure 4. Distributions for the five NN inputs are compared for both GD-1 stars (in red) and background stars (in grey) across signal and sideband regions. The patch shown here is the same example patch from Fig. 1. For both stream and background stars, the distributions for these five variables across the signal and sideband regions are approximately indistinguishable.

regions will have significantly fewer stream stars and will be signal-depleted, as desired. In practice, the average width along μ_λ across the 21 patches was 2.34 ± 0.36 mas yr^{-1} for the signal regions and 7.02 ± 1.09 mas yr^{-1} for the sideband regions.

3.3 Neural network architecture and training procedure

We implement CWoLa with an NN built in Keras (Chollet et al. 2015) with a TensorFlow backend (Abadi et al. 2015). The model consists of 3 hidden fully connected layers, each with a layer size of 256 nodes and a Rectified Linear Unit (ReLU) activation (Nair & Hinton 2010). Each fully connected layer is followed by a dropout operation with a dropout rate of 20 per cent (Srivastava et al. 2014). These layers are followed by a final output layer of a single node with a sigmoid activation. Hyperparameter values for layer size, batch size, and number of k-folds (described below) were chosen via an optimization using Optuna (Akiba et al. 2019).

For each of the 21 *Gaia* patches considered in our search for GD-1, we train a series of classifiers to separate stars labelled as part of the signal region from stars labelled as part of the sideband region. This quality defines CWoLa as being 'weakly supervised': it operates with a little more information than a fully unsupervised network, as we expect an optimal signal region to contain a higher fraction of GD-1 stars than in the sideband region, but it does not have access to the actual GD-1 labels. The training procedure, which closely aligns with other CWoLa searches (Collins et al. 2018, 2019), unfolds as follows:

(i) ***k*-folding:** We implement stratified k-folding ($k = 5$) to randomly divide all the stars in a given patch into five sections, or 'folds'. Each fold is chosen such that the overall percentage of stars labelled as 'signal' versus 'sideband' is also maintained within each fold. The first fold (20 per cent of all stars) is reserved as a test set. The second fold (another 20 per cent of all stars) is used as a validation set. The remaining 60 per cent of stars are used for training.

(ii) **Train:** Next, a classifier with the architecture specified above is trained on the training set with a batch size of 10 000 for up to 100 epochs, though early stopping with a patience of 30 typically halts the training process well before this limit. The large batch size is necessary due to the low number of labelled stream stars in the overall data set – for example, one patch on the tail end of GD-1 has a stream star population of just 0.15 per cent. Large batch sizes therefore help ensure that more than a handful of stream stars will be seen at a time by the network during training. We use the binary cross-entropy loss function and Adam optimizer (Kingma & Ba 2014). The validation set is used to monitor the validation loss for early stopping.

(iii) **Repeat:** The classifier training is repeated twice more, each time with a random initialization of trainable parameters. Of the three distinct trainings, the weights are stored for the model with the lowest validation loss.

(iv) **Cycle through validation sets:** This process is repeated using each of the remaining folds as a validation set with the exception of the test set, which remains unchanged. For each configuration, the remaining three folds besides those used for the test and validation sets are used for training.

(v) **Evaluate on test set:** Each of the best models trained using the four k-fold options for the validation set is evaluated on the test set. The final CWoLa score for each star in the test set is defined as the average across the four scores.

(vi) **Combine test sets:** This entire process is repeated, cycling through each of the five possible k-folds as a test set. These test sets are then concatenated into a single data set such that every star in the patch ends up in the test set exactly once.

3.4 Model evaluation

After training the NN classifiers inherent in the CWoLa methodology, a series of fiducial selections is applied to each patch to further refine the results and optimize for the highest possible signal-to-noise ratio. The fiducial selections used are almost identical to their counterparts in Shih et al. (2021, 2023):

(i) $g < 20.2$, to ensure uniform acceptance by the *Gaia* satellite.

(ii) $|\mu_\lambda| > 2\,\mathrm{mas\,yr^{-1}}$ or $|\mu_\phi^*| > 2\,\mathrm{mas\,yr^{-1}}$, to remove very distant stars that are predominantly concentrated near zero proper motion and therefore not equally distributed throughout the patch.

(iii) $0.5 \leq b - r \leq 1$, to isolate old and low-metallicity stellar streams in colour space.

Unlike in Shih et al. (2021, 2023), however, we do not need to apply a cut restricting the patch radius from 15° to 10° after training. Unlike CWoLa, ANODE is a density estimation technique, so this cut removes areas of the phase space near the boundaries where it could be more challenging to model the distribution of the data.

VIA MACHINAE employs a sophisticated line-finding strategy using modified Hough transforms (Duda & Hart 1972) to search for line-like structures in the identified anomalous stars and then combines these line segments into an overall stream candidate. We achieve similar results with a relatively lighter computational load using k-means clustering (Lloyd 1982) ($k = 2$) in proper motion space. Following the grouping of stars into two clusters, we select the cluster with the largest population of stars and discard the stars in the other cluster. This is motivated by our expectation that the stellar stream should be kinematically cold, therefore the velocities of its constituent stars should be densely clustered in velocity space. This clustering strategy is likely best used as a post-discovery tool and may not perform well in contexts with high numbers of contaminant stars not belonging to a stream. In these cases, opting for a larger $k > 2$ or a line-finding technique could instead be a better choice.

Following these fiducial selections, model performance was evaluated by applying the classifier to stars in the combined test set equivalent to the entire patch. The output scores were sorted from highest to lowest, where higher values indicated that the model ranked those stars as more likely to belong to the signal region than the sideband region. The top $N = 250$ stars, ranked by NN output scores, are chosen for evaluation.

The number 250 was chosen following an optimization for both purity (percentage of top-ranked stars overlapping with labelled GD-1 stars) and completeness (percentage of labelled GD-1 stars covered by CWoLa's top-ranked stars). In principle, however, one could isolate a different absolute number or relative percentage of top stars, though it would be advisable to stay under the average of 430 labelled stream stars per patch.

It is worth emphasizing that this method of model evaluation requires ground truth labels. In the absence of reliable stream labels, or in the case of discovering a new stream, we must employ different methods to evaluate model performance, not to mention a modified strategy for the model implementation itself. We discuss this further in Section 5.

4 RESULTS

Before looking at real *Gaia* data, we evaluated the performance of CWoLa when applied to 100 randomly chosen simulated stellar streams. Details of the simulation procedure and the results are shown in Appendix A. With just two passes of CWoLa, 96 per cent of streams are identified with non-zero purity, of which 69 per cent are identified with a purity larger than 50 per cent.

4.1 GD-1 stream identification

The combined results of applying the CWoLa technique to each of the 21 patches of *Gaia* DR2 are shown in Fig. 5. Results for individual patches are detailed in Appendix B. Across the 21 patches, 1498 unique GD-1 stars pass our fiducial selections. 1360 unique stars are identified in the combined top $N = 250$ stars for each CWoLa patch. Of these, 760 are part of the labelled GD-1 star set (Price-Whelan & Bonaca 2018a). Thus, across the entire stream, we achieve a purity of 56 per cent and a completeness of 51 per cent. In our optimization studies, we found that stream purity plateaued at a maximum value of 78 per cent using the top $N = 25$ stars in each patch, but this choice of N only yields a completeness of 13 per cent. Conversely, choosing $N = 300$ yields a reduced purity of just 30 per cent, but a higher completeness of 54 per cent.

The majority of GD-1 is quite narrow, with an average angular width of approximately 0.5° (Malhan et al. 2019), and dense, with approximately 100 stars per 5° bin between $\alpha = -220°$ and $\alpha = -150°$. Within this region, CWoLa can reliably identify the stream stars. The tail ends of GD-1 ($\alpha \leq -220°$ and $\alpha \geq -150°$) are more sparsely populated, with about half the average population per bin of the main body of the stream, and less localized in μ_λ, meaning that stream stars in this region are harder to identify using CWoLa. Some stars in these regions may also have been excluded from the 21 patches due to their proximity to the Galactic disc or the presence of nearby dust. These regions also tend to include stars with small proper motions, meaning that the stream stars are more likely to be overwhelmed by distant background stars.

We can also analyse these results in a rotated set of position coordinates ϕ_1 and ϕ_2 (Koposov et al. 2010) designed to align with the main body of the stream, as shown in Fig. 6. This perspective highlights that CWoLa has identified several of the density perturbations unique to GD-1: two sparsely populated 'gaps' near $\phi_1 = -40°$ and $\phi_1 = -20°$; an offshoot, or 'spur', near $\phi_1 = -35°$; and an overdensity of stars, or 'blob', near $\phi_1 = -15°$. We can more quantitatively demonstrate the identification of these features, as in fig. 5 of Price-Whelan & Bonaca (2018b), by looking at histograms of ϕ_2 in various ranges of ϕ_1 as shown in Fig. 7. This study highlights the 'spur' and 'blob' in particular by fitting a histogram of stars near each feature, along with a third control region, with a three-component Gaussian mixture model assuming a background, the GD-1 stream, and the feature ('blob' or 'spur').

As mentioned above, the underdense regions, or 'gaps', in GD-1 are typically observed near $\phi_1 \approx -40°$ and $\phi_1 \approx -20°$. A third underdense region has also been identified near $\phi_1 \approx -3°$ (de Boer et al. 2020). Our results would not be inconsistent with this third gap, as the density in this area for the CWoLa-identified stars is indeed low, on par with the densities seen at the other two 'gaps', but since this region is so close to the furthest extent of the CWoLa-identified stars, it is not clear whether this underdensity is a feature from the

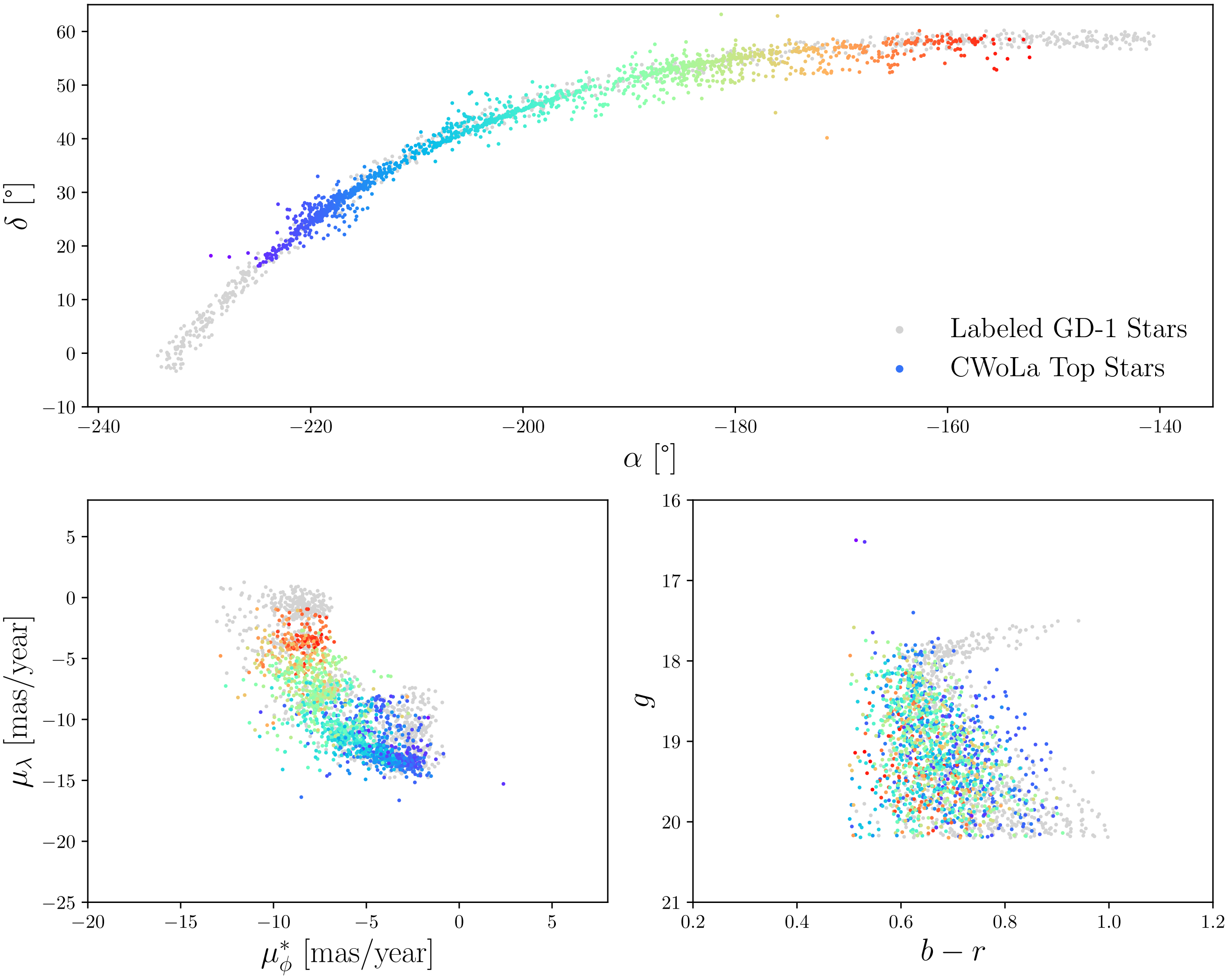


Figure 5. The full scope of stars identified by the CWoLa method in overlapping patches across the angular range corresponding to GD-1. Light grey dots indicate the ground truth labelling of GD-1 stars (Price-Whelan & Bonaca 2018a), while the top 250 stars identified by CWoLa in each patch are indicated in coloured dots. The colours are chosen to correlate with each star's α value.

stream or a reflection of the diminished purity of stars in the patches on the ends of the stream.

As for the overdense regions, de Boer et al. (2020) reported that four overdense regions peaked at $\phi_1 \approx -48°$ (the highest density region of the stream), $\phi_1 \approx -27°$, $\phi_1 \approx -10°$, and $\phi_1 \approx +2°$. While the CWoLa-identified stars do not reliably cover the region above $\phi_1 = 0°$, the remaining three peaks for which $\phi_1 < 0°$ are also seen in the CWoLa-identified stars. By fitting the CWoLa-identified stars with a mixture model of three Gaussian distributions, we can extract approximate overdensity peaks at $\phi_1 \approx -51°$, $\phi_1 \approx -30°$, and $\phi_1 \approx -11°$. It is interesting to note that CWoLa picks up a large $\phi_1 \approx -51°$ peak, in line with the highest density peak reported in de Boer et al. (2020), though this peak is less pronounced in the stars labelled from Price-Whelan & Bonaca (2018a).

Another reported feature of GD-1 is a wider 'cocoon' of stars with a width of around 1° surrounding a much denser core of the stream (Malhan et al. 2019). To probe this feature, we first calculate the median ϕ_2 in broad 5° bins of ϕ_1 to find a smoothed trajectory for the stream. Then, we shift each CWoLa-identified star by its median ϕ_2 location (see Fig. 8a). Once the stream has been centred around this path, we make a 3σ selection, as in Malhan et al. (2019), and then plot the histogram of shifted ϕ_2 (see Fig. 8b). Fitting the distribution to a two-component Gaussian mixture model reveals a narrow peak with a standard deviation of $\sigma \approx 0.3°$ (the core of the stream) and an additional wider peak with $\sigma \approx 1.7°$. This appears to support the observation of such a 'cocoon' of more diffuse stars surrounding the central core of the stream. We estimate potential contamination by running the CWoLa procedure on parts of the sky where GD-1 stars are not expected: signal regions outside of the ones probed in the analysis as well as additional patches lying outside of the considered region. We then run the remaining analysis selections on the top-identified stars. Following this procedure, we find that just 1.7 per cent of stars survive our full selection, suggesting that contamination is minimal.

4.2 Towards an augmentation of the GD-1 stream labelling

Beyond identifying cold stellar streams without labels, the CWoLa technique may also be useful for improving the labelling systems that indicate as to which stars belong to a particular stream. For instance, CWoLa can identify promising stellar candidates that were not labelled as GD-1 members in Price-Whelan & Bonaca (2018a),

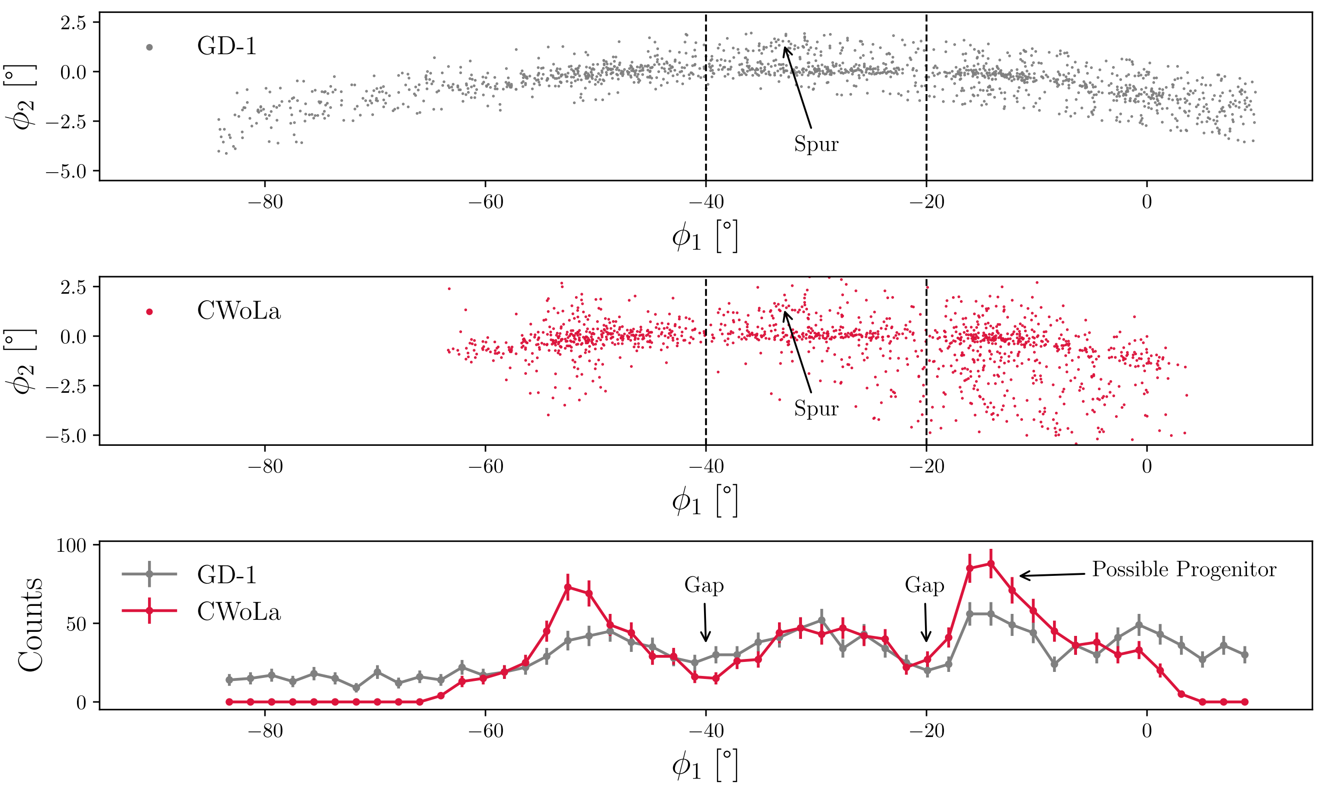


Figure 6. The CWoLa-identified stars across all patches are compared with the labelled GD-1 stars from Price-Whelan & Bonaca (2018a) in stream-aligned coordinates ϕ_1 and ϕ_2. This perspective highlights that CWoLa has identified several of the density perturbations unique to GD-1: two sparsely populated 'gaps' near $\phi_1 = -40°$ and $\phi_1 = -20°$; an offshoot, or 'spur', near $\phi_1 = -35°$; and an overdensity of stars, or 'blob', near $\phi_1 = -15°$. Two additional overdensities are seen near $\phi_1 = -51°$ and $\phi_1 = -30°$.

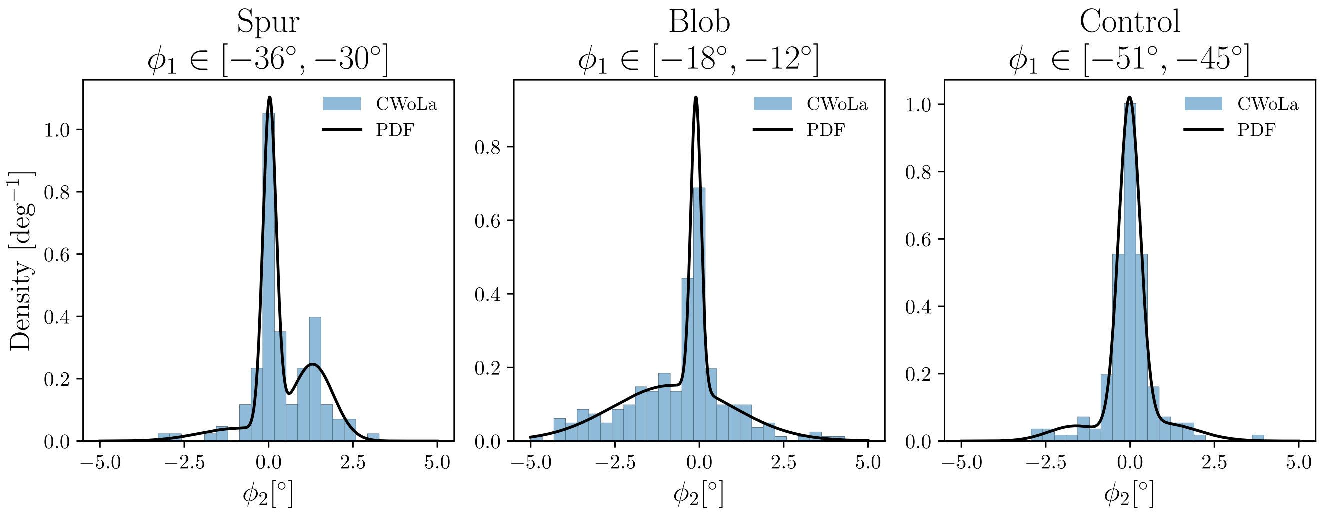
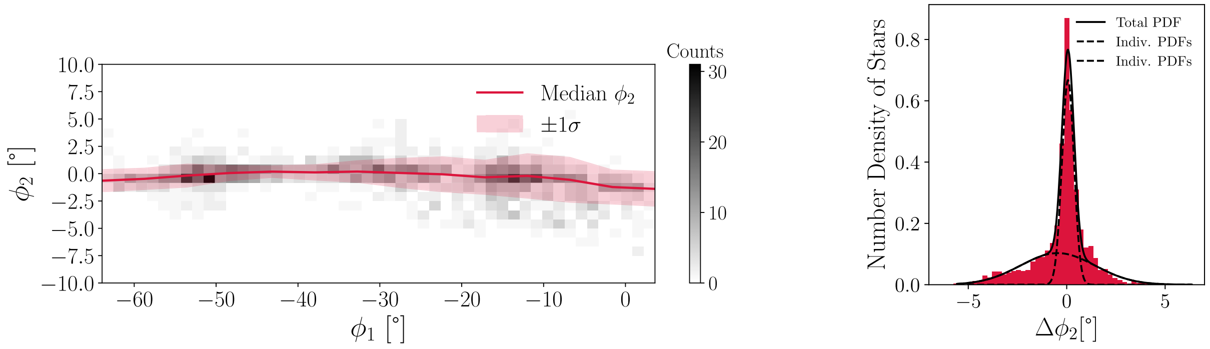


Figure 7. Three subsets of the CWoLa-identified stars (the 'blob', 'spur', and a control region) are selected and fitted with a three-component Gaussian mixture model to highlight the kinematic qualities of the additional feature, if present. In each case, the GD-1 stream corresponds to the primary narrow peak centred near $\phi_2 = 0°$. In the first two plots, we see clear indications of a second peak representing each feature.

but nevertheless have properties closely aligned with labelled stars.

As shown in Section 4, 1360 unique stars are identified by CWoLa across all 21 patches, and of these, 760 (56 per cent) are part of the labelled GD-1 star set. We can further refine the remaining 600 unlabelled stars by identifying the subset of individual stars s that minimize the Euclidean distance d to their respective closest labelled GD-1 star in the test set s' along the 5 dimensions used as CWoLa inputs, individually standardized to have $\mu = 0$ and $\sigma = 1$: [ϕ, λ, $\mu_{\phi *}$, colour ($c \equiv b - r$), and magnitude (g)]:

$$d = \sqrt{(\phi - \phi')^2 + (\lambda - \lambda')^2 + (\mu_{\phi *} - \mu'_{\phi *})^2 + (c - c')^2 + (g - g')^2}. \tag{1}$$



(a) CWoLa-identified stars with the median ϕ_2 in bins of ϕ_1

(b) Width of CWoLa stars

Figure 8. The width of the CWoLa-identified stars is determined by first calculating the median stream position in ϕ_2 for 10 bins of ϕ_1 [the overlaid red line in (a)]. The ϕ_2 coordinates are then shifted by these median values, yielding the histogram in (b). In (b), we use a two-component Gaussian mixture model to show two individual Gaussian components with $\sigma \approx 0.3°$ (the core of GD-1) and $\sigma \approx 1.7°$ (the 'cocoon'). This appears to support the general trend observed in fig. 6 of Malhan et al. (2019).

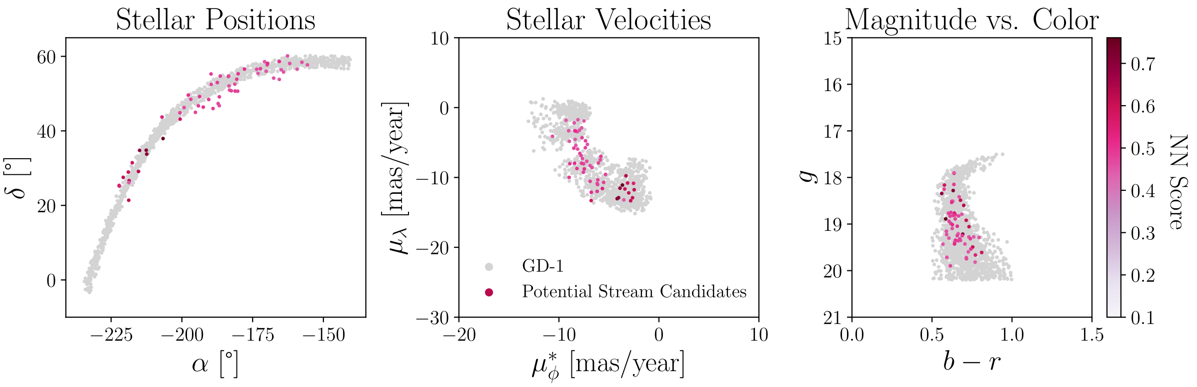


Figure 9. Isolating the subset of the top stars identified by CWoLa with the 10 per cent smallest five-dimensional Euclidean distances d to the nearest labelled star reveals 60 additional stellar candidates for GD-1 membership that may have been omitted from the GD-1 ground truth labelling.

Isolating the stars yielding the smallest 10 per cent of distances d reveals 60 additional stars, shown in Fig. 9 and listed explicitly in Appendix C, that appear to align with the labelled GD-1 stars across these 6 dimensions and would be interesting to investigate as potential GD-1 member candidates. A detailed cross-checking of these candidate GD-1 stream stars with other, more precise GD-1 stream catalogues will be pursued in future work.

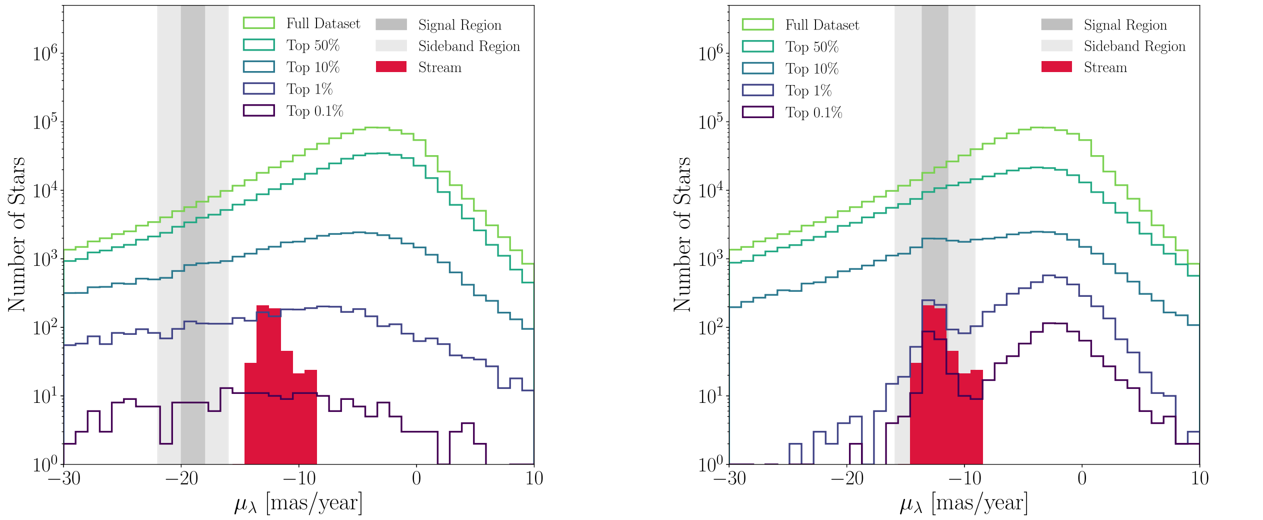
5 DISCUSSION

It is evident that CWoLa successfully identifies significant portions of GD-1, as measured by not only purity and completeness but also the faithful reconstruction of physical characteristics and density perturbation characteristic of this stream. Additionally, CWoLa is highly effective at identifying simulated streams.

While the analysis presented here shares many core strategic components and the same data set with VIA MACHINAE, this analysis differs primarily in terms of the mechanisms for how to assign anomaly scores to stars, how to divide the sky into subsections for scanning, and how to cluster stars post-training. CWoLa is implemented via a comparatively simple, lightweight, and easy-to-train NN-based classifier instead of a normalizing flow model to approach the same problem of anomaly detection. When applied to the same example stellar stream, GD-1, CWoLa is able to identify stars with comparable purity with much less computational overhead. We find 760 labelled stars overall, yielding a 56 per cent purity, while VIA MACHINAE's first iteration (Shih et al. 2021) found 738 stars, yielding a 49 per cent purity. VIA MACHINAE's latest iteration (Shih et al. 2023), which includes additional fiducial selections and an augmented scan over both proper motion variables, increases its star yield to 820, or 65 per cent purity.

It is worth emphasizing that our approach does not apply any kind of line-finding or protoclustering algorithms as is done in VIA MACHINAE – the anomalous stars here are simply combined and filtered via k-means clustering. This lightweight clustering strategy is particularly useful in a post-discovery context in which we are interested in refining stream membership catalogues. Another important distinction between these techniques is that CWoLa uses signal and sideband regions of varying widths that are chosen for each patch based on the location of the signal, while VIA MACHINAE searches over regions of interest defined by the orthogonal proper motion coordinate with a fixed width of 6 mas yr^{-1} with centres spaced 1 mas yr^{-1} apart. The fiducial selections also differ slightly between these implementations: VIA MACHINAE restricts each patch to the innermost 10° circle in position space to avoid edge effects, but CWoLa does not exhibit these effects and thus we do not impose this selection.

Training normalizing flows such as those in ANODE can be a time-intensive task, while completing the full CWoLa training paradigm for GD-1 takes just 15 min per patch on an NVIDIA A40 GPU.



(a) When the signal region does not contain the stream, there is no obvious bump anywhere in proper motion space as the cut percentage gets larger.

(b) When the signal region contains the stream, an obvious bump forms in the same area where the stream is localized.

Figure 10. A demonstration of a scan for which the anomaly location is not previously known.

Running over each patch is embarrassingly parallel and can easily be run simultaneously based on GPU access or using multiprocessing on CPUs. Running all 21 patches on a single GPU takes about 5 h in total, making it quite feasible for researchers to optimize their signal and sideband region definitions as well as to combine the results from scans of multiple variables.

This training time can be even further reduced by applying the fiducial selections to the samples *before* training – cutting training time roughly in half and yielding an overall purity of 44 per cent. For a 100 per cent improvement in training time, this technique only reduces the final purity of the identified stars by about 20 per cent, making this a valuable option particularly for coarse-grained scans across wide areas of phase space. CWoLa and VIA MACHINAE can therefore be thought of as complementary tools for stream detection under different circumstances or computational constraints.

If CWoLa can be used to identify known streams, it may also be used to potentially find new, undiscovered streams. Some additional challenges will arise when extending CWoLa to look for new stellar streams within the full *Gaia* data set. Detected anomalies are not necessarily guaranteed to be stellar streams, since CWoLa could identify any localized anomalous features. A lack of ground truth labelling for a stream would also require a re-evaluation of our performance metrics – for instance, streams would need to be evaluated using the standard anomaly detection technique of performing a series of selections (e.g. a range of percentiles of the NN score, or hand-picked thresholds based on the background rate in the sideband region) on a histogram of proper motion. If no anomaly is present, these increasingly harsh selections will reduce the sample statistics without significantly altering the histogram shape, as shown in Fig. 10(a). However, if an anomaly is present and identified by CWoLa, a new shape will emerge with increasingly harsh selections on the distribution in question, as shown in Fig. 10(b). Additionally, multiple passes of CWoLa might be needed with different choices of signal and sideband region widths if the approximate width of the anomaly is not a priori known.

Searching for new stellar streams will require scanning over the full range of proper motion values in the data set, since we will not know where new streams might be localized. By applying CWoLa in a coarse sliding-window fashion across μ_λ, regions of interest may be identified. These regions can be further studied through finer scans until anomalous data points are identified. When combining multiple patches together for an overall result, we might also need to additionally employ a line-finding algorithm for identifying larger scale stream-like results, such as the modified Hough transform used in VIA MACHINAE (Shih et al. 2021).

6 CONCLUSIONS

We have demonstrated a new application of ‘CWoLa hunting’, an anomaly detection technique based on the weakly supervised machine learning classifier CWoLa that is designed to detect localized anomalies in a model-agnostic manner. CWoLa is shown to be easy to train, highly computationally efficient, and, most importantly, effective at identifying anomalies including the stellar stream GD-1 and dozens of simulated streams with high purity. The GD-1 candidate stars identified by CWoLa exhibit the same density perturbations and physical characteristics (the ‘spur’, ‘blob’, ‘cocoon’, gaps, and overdense regions) noted in several independent studies of the stream. The NN output scores also give clues as to which stars might have been accidentally omitted from more formal GD-1 labelling schemes, suggesting several promising candidates. The successful application of CWoLa in this study shows that CWoLa has strong potential to improve the signal-to-noise ratio on the membership of known streams as well as to potentially reveal previously undetected streams throughout the Galactic halo. CWoLa has broad applicability as a weakly supervised anomaly detection technique outside of high-energy physics and could be applied into still more areas of fundamental science.

ACKNOWLEDGEMENTS

MP and BN thank Joshua Bloom for his thoughtful feedback and suggestions. MP also thanks Shirley Ho, Adrian Price-Whelan, David Spergel, and Xiangyang Ju for their helpful comments, and gratefully acknowledges the hospitality of the Flatiron Institute Center for Computational Astrophysics. ST thanks Vanessa Boehm for her help on working with OPTUNA for hyperparameter optimization. This work makes 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. BN, MP, and ST were supported by the Department of Energy, Office of Science under contract number DE-AC02-05CH11231.

DATA AVAILABILITY

A codebase with instructions on how to reproduce each of the plots in this paper is located at https://github.com/hep-lbdl/GaiaCWoLa. The data sets needed to fully reproduce the plots in this paper (with CWOLA already applied) are publicly available (Pettee et al. 2023). The full 21 patches covering GD-1 are also publicly available (Buckley, Shih & Necib 2023).

REFERENCES

Aad G. et al., 2020, Phys. Rev. Lett., 125, 131801
Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Available at https://www.tensorflow.org/
Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Arifyanto M. I., Fuchs B., 2006, A&A, 449, 533
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Astropy Collaboration, 2022, ApJ, 935, 167
Balbinot E., Santiago B. X., da Costa L. N., Makler M., Maia M. A. G., 2011, MNRAS, 416, 393
Banik N., Bovy J., 2019, MNRAS, 484, 2009
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
Bonaca A. et al., 2020, ApJ, 892, L37
Bonaca A., Hogg D. W., Price-Whelan A. M., Conroy C., 2019, ApJ, 880, 38
Borsato N. W., Martell S. L., Simpson J. D., 2019, MNRAS, 492, 1370
Brown A. G. A. et al., 2018, A&A, 616, A1
Brown A., 2013, PyGaia. Available at https://github.com/agabrown/PyGaia
Buckley M. R., Shih D., Necib L., 2023, Gaia DR2 Stellar Stream Test. Available at https://zenodo.org/records/7897936
Carlberg R. G., Grillmair C. J., Hetherington N., 2012, ApJ, 760, 75
Choi J., Dotter A., Conroy C., Cantiello M., Paxton B., Johnson B. D., 2016, ApJ, 823, 102
Chollet F. et al., 2015, Keras. Available at https://keras.io
Collins J. H., Howe K., Nachman B., 2018, Phys. Rev. Lett., 121, 241803
Collins J. H., Howe K., Nachman B., 2019, Phys. Rev. D, 99, 014038
de Boer T. J. L., Belokurov V., Koposov S. E., Ferrarese L., Erkal D., Côté P., Navarro J. F., 2018, MNRAS, 477, 1893
de Boer T. J. L., Erkal D., Gieles M., 2020, MNRAS, 494, 5315
Dehnen W., Odenkirchen M., Grebel E. K., Rix H.-W., 2004, AJ, 127, 2753
Dotter A., 2016, ApJS, 222, 8
Duda R. O., Hart P. E., 1972, Commun. ACM, 15, 11
Duffau S., Zinn R., Vivas A. K., Carraro G., Méndez R. A., Winnick R., Gallart C., 2005, ApJ, 636, L97
Eggen O. J., 1971, PASP, 83, 271
Erkal D., Belokurov V., Bovy J., Sanders J. L., 2016, MNRAS, 463, 102
Eyre A., Binney J., 2009, MNRAS, 400, 548
Grillmair C. J., Dionatos O., 2006, ApJ, 643, L17
Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G. A., 2018, Nature, 563, 85
Helmi A., White S. D. M., 1999, MNRAS, 307, 495
Johnston K. V., 1998, ApJ, 495, 297
Johnston K. V., Hernquist L., Bolte M., 1996, ApJ, 465, 278
Kamdar H., Conroy C., Ting Y.-S., 2021, Stellar Streams in the Galactic Disk: Predicted Lifetimes and Their Utility in Measuring the Galactic Potential , preprint (arXiv:2106.02050)
Kingma D. P., Ba J., 2015, Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)
Koposov S. E., Rix H.-W., Hogg D. W., 2010, ApJ, 712, 260
Law D. R., Majewski S. R., 2010, ApJ, 714, 229
Lindegren L. et al., 2018, A&A, 616, A2
Lloyd S., 1982, IEEE Trans. Inf. Theory, 28, 129
Malhan K., Ibata R. A., 2018, MNRAS, 477, 4063
Malhan K., Ibata R. A., Carlberg R. G., Valluri M., Freese K., 2019, ApJ, 881, 106
Malhan K., Yuan Z., Ibata R. A., Arentsen A., Bellazzini M., Martin N. F., 2021, ApJ, 920, 51
Mateu C., 2023, MNRAS, 520, 5225
Mateu C., Read J. I., Kawata D., 2017, MNRAS, 474, 4112
McMillan P. J., 2016, MNRAS, 465, 76
Metodiev E. M., Nachman B., Thaler J., 2017, J. High Energy Phys., 2017, 174
Nachman B., Shih D., 2020, Phys. Rev. D, 101, 075042
Nair V., Hinton G. E., 2010, Rectified Linear Units Improve Restricted Boltzmann Machines, Proceedings of the 27th International Conference on International Conference on Machine Learning, 807
Necib L., Lisanti M., Garrison-Kimmel S., Wetzel A., Sanderson R., Hopkins P. F., Faucher-Giguère C.-A., Kereš D., 2019, ApJ, 883, 27
Neyman J., Pearson E. S., 1933, Phil. Trans. R. Soc. A, 231, 289
Nibauer J., Belokurov V., Cranmer M., Goodman J., Ho S., 2022, ApJ, 940, 22
Paxton B. et al., 2013, ApJS, 208, 4
Paxton B. et al., 2015, ApJS, 220, 15
Paxton B. et al., 2018, ApJS, 234, 34
Paxton B., Bildsten L., Dotter A., Herwig F., Lesaffre P., Timmes F., 2011, ApJS, 192, 3
Pettee M., Thanvantri S., Nachman B., Shih D., Buckley M. R., Collins J. H., 2023, Finding Stellar Streams in the Milky Way with CWoLa. Available at https://zenodo.org/records/7897840
Price-Whelan A. M., 2017, J. Open Source Softw., 2, 388
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://zenodo.org/records/1295543
Price-Whelan A. M., Bonaca A., 2018b, ApJ, 863, L20
Prusti T. et al., 2016, A&A, 595, A1
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., Mohamed S., 2015, Proceedings of the 32nd International Conference on Machine Learning, in Proceedings of Machine Learning Research, 37, 1530
Rockosi C. M. et al., 2002, AJ, 124, 349
Sanders J. L., Bovy J., Erkal D., 2016, MNRAS, 457, 3817
Shih D., Buckley M. R., Necib L., 2023, Via Machinae 2.0: Full-Sky, Model-Agnostic Search for Stellar Streams in Gaia DR2 , preprint (arXiv:2303.01529)
Shih D., Buckley M. R., Necib L., Tamanas J., 2021, MNRAS, 509, 5992
Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, J. Mach. Learn. Res., 15, 1929
Williams M. E. K. et al., 2011, ApJ, 728, 102

APPENDIX A: SIMULATED STELLAR STREAMS

Our implementation of CWOLA for stellar stream discovery was first tested on simulated stellar streams. These simulated streams are frequently highly localized in the proper motion coordinate μ_λ, meaning that they may tend to be easier to find with our methods than a real stream such as GD-1.

The streams were simulated using the GALA (Price-Whelan 2017) PYTHON package to evolve stars in a mock globular cluster along an orbit through the simulated Milky Way potential (McMillan 2016), with the centre of the stream randomly placed on the sky with a distance randomly chosen between 5 and 20 kpc from the Earth. Stellar properties for the stream components were generated from a MESA Isochrones & Stellar Tracks (MIST) (Paxton et al. 2011, 2013, 2015, 2018; Choi et al. 2016; Dotter 2016) isochrone, assuming [Fe/H] = −1.6 and an age of 10 Gyr.Observational errors compatible with the *Gaia* DR2 data set were added to the synthetic stream stars using the PYGAIA (Brown 2013) package.

An example simulated stream, representing just 1161 (0.13 per cent) of the 886 677 stars in the simulated patch, is shown in angular position space in Fig. A1. The simulated streams are presented as standalone patches, so CWoLa is applied to just one simulated patch at a time. No fiducial cuts are applied to the simulated patches.

As a proof of concept, we choose idealized signal and sideband limits with prior knowledge of the location of the stellar stream: the signal is defined as the window of total width $\sigma/4$ surrounding the median μ_λ value of the stream, while the sideband is defined as the additional window of total width $\sigma/2$ surrounding the signal region. These signal and sideband regions for background and stream stars are plotted in Fig. A2.

We train CWoLa to distinguish between events from these signal and sideband regions, and then select the top 250 stars as ranked by CWoLa's classifier output score. As shown in Fig. A3(a), 100 per cent of the top 250 stars selected for this patch are members of the ground truth labelled stream population in this patch.

When this technique is applied across 100 randomly sampled simulated streams, 76 per cent of streams are identified with purity >0 per cent, of which 75 per cent are identified with high purity (defined as purity greater than 50 per cent). However, a large portion of these cases with zero purity are streams with wider distributions along μ_λ, so the results can be further augmented with additional scans choosing different signal and sideband regions. When supplemented with an additional scan with wider signal and sideband region definitions (signal region = $\pm\sigma$ and sideband region = $\pm 3\sigma$ − signal region), 96 per cent of streams are identified with non-zero purity, of which 69 per cent are identified with high purity. Across the 100 streams, the median purity of the CWoLa-identified results is 86 per cent. Fig. A3(b) illustrates that the clear majority of the simulated streams are identified with high purity levels.

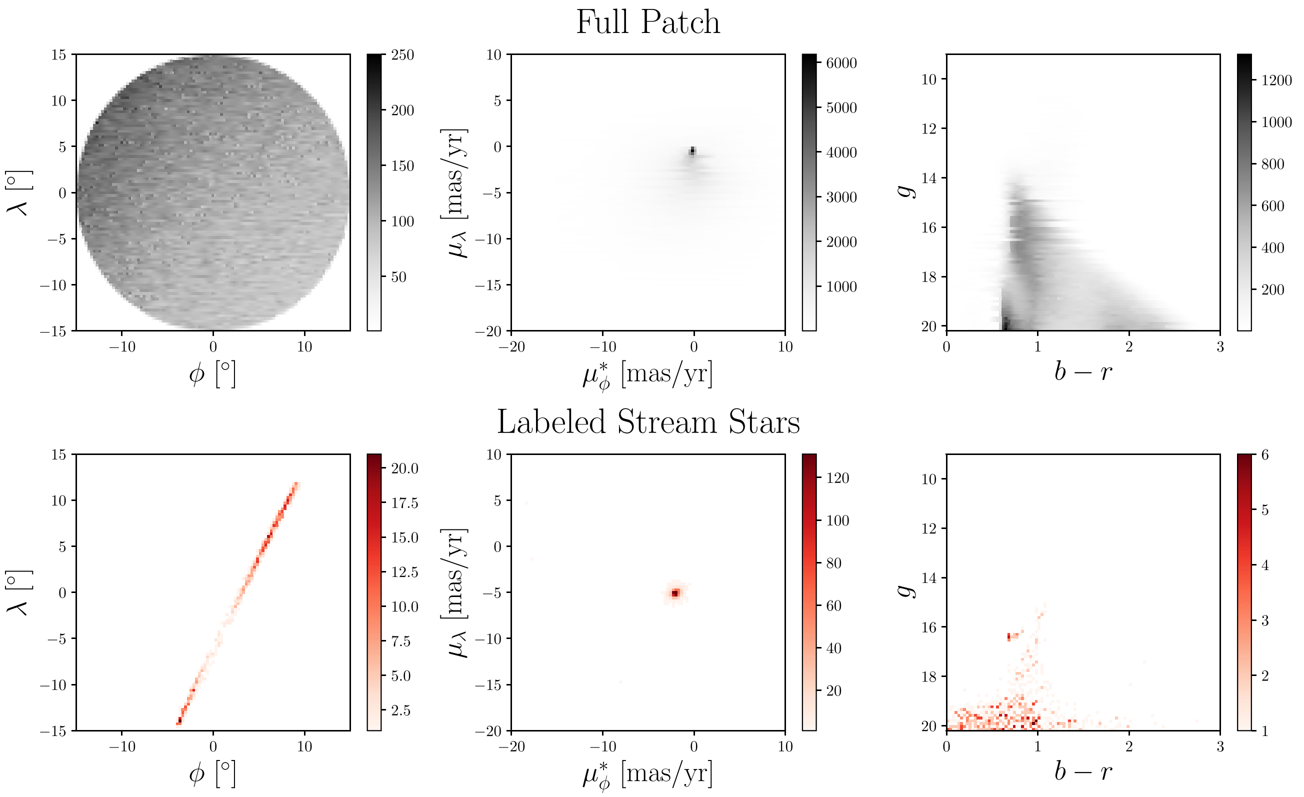


Figure A1. Distributions in position, velocity, and colour space for a simulated patch as well as the simulated stream contained within it. While both background stars and simulated stream stars are highly concentrated in velocity space, the stream stars' peak proper motions are located further from $(\mu_\phi^*, \mu_\lambda) = (0, 0)$ than those of the background stars.

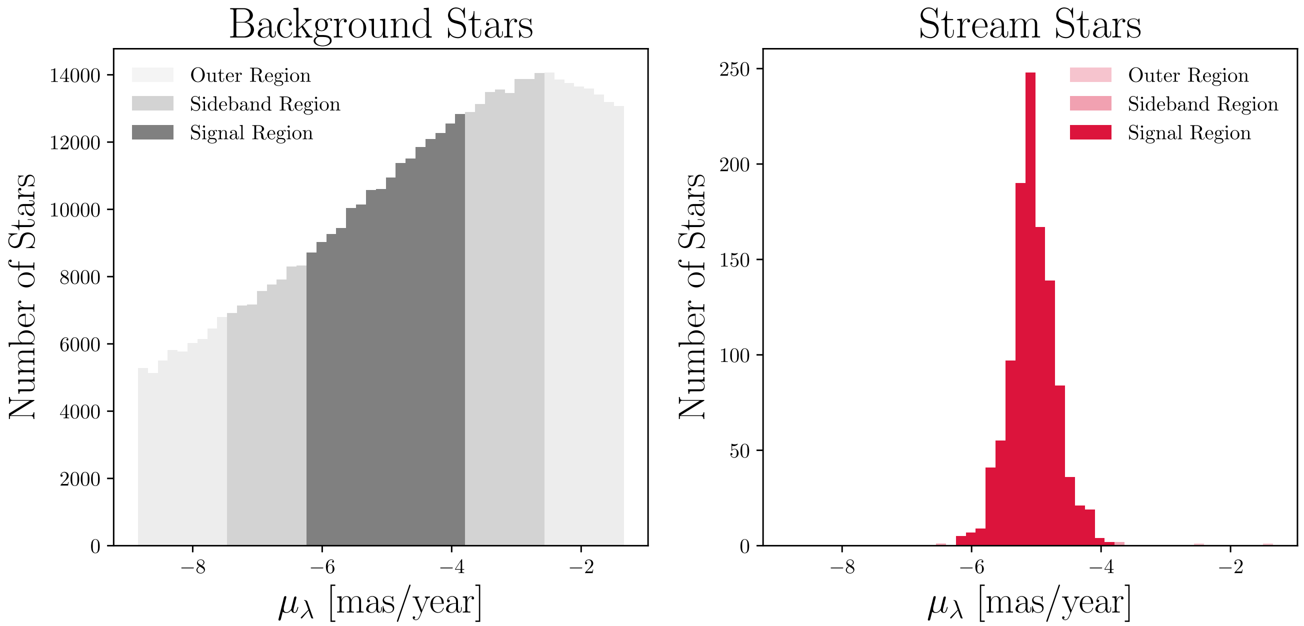
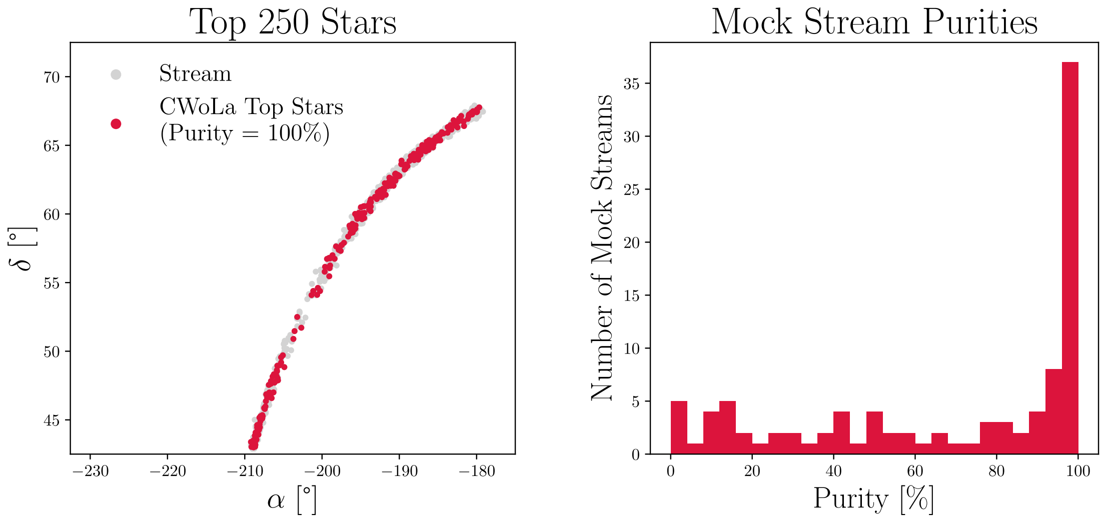


Figure A2. Simulated streams are far more concentrated in angular velocity space than a typical stream in the *Gaia* data set. As a result, the signal and sideband regions are defined within a much narrower band around the median stream μ_λ. The signal region is defined within $\pm\sigma/4$, or $[-6.3, -3.8]$, while the sideband region is defined as $\pm\sigma/2$, excluding the signal region: $[-7.6, -6.3)$ and $(-3.8, 2.6]$. The stream stars are almost exclusively contained within the signal region.



(a) The top 250 stars in one sample simulated stream, as ranked by CWoLa classifier output score, are all members of the ground truth labeled star population.

(b) The vast majority of simulated stream stars are identified with high purity levels after two passes of the CWoLa search method.

Figure A3. CWoLa performance evaluated as a function of purity across multiple simulated streams.

APPENDIX B: PATCH-BY-PATCH PERFORMANCE

Fig. B1 shows the patch-by-patch breakdown of CWoLa applied to GD-1. Each of the 21 patches is considered separately for individual applications of the CWoLa methodology, including fiducial cuts and *k*-means clustering. These results are combined in Fig. 5. CWoLa achieves a high purity across nearly all patches, with the exception of those patches with relatively fewer stream stars located at the leftmost and rightmost edges of the stream (near $\alpha = -230°$ and $\alpha = -150°$).

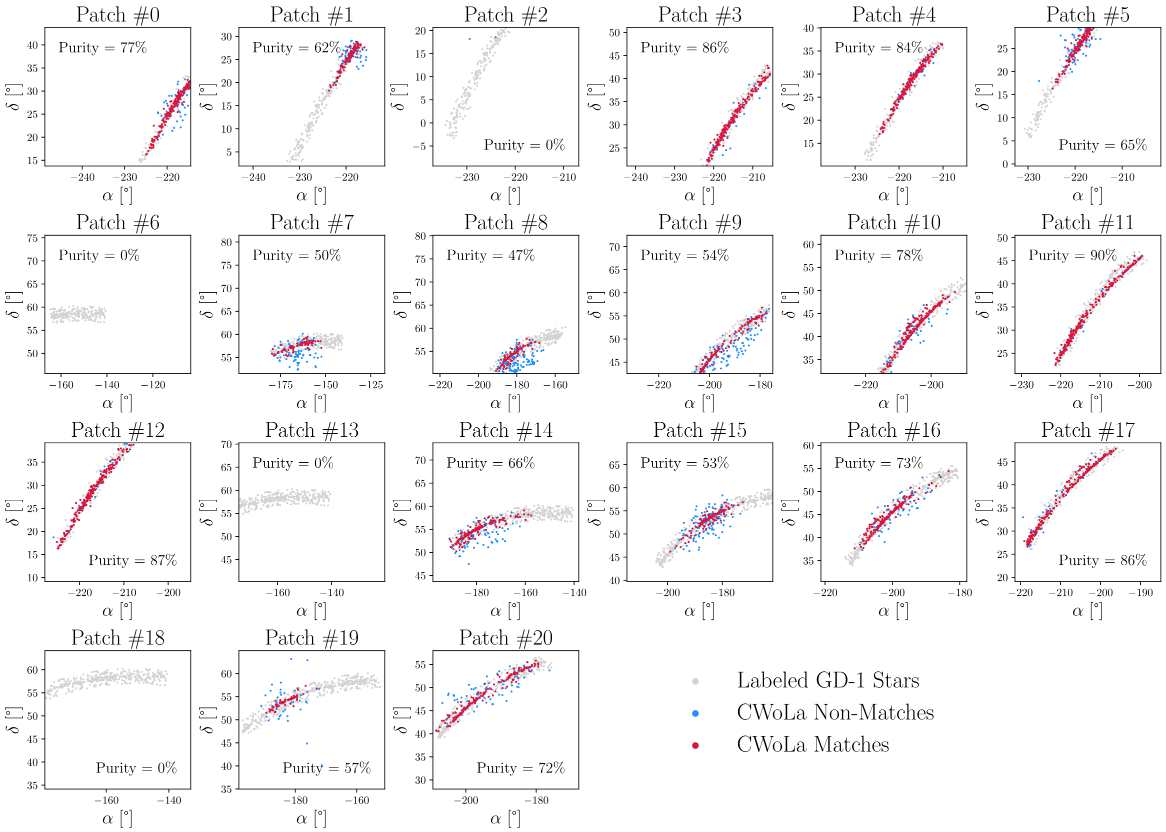


Figure B1. The top 250 identified stars across each of the patches of GD-1 from the *Gaia* data set show that CWOLA is able to effectively identify GD-1 stars with high purity levels across all patches, with the exception of patches on the very furthest tails of the stream.

APPENDIX C: POTENTIAL GD-1 CANDIDATE MEMBERS

Table C1 details the subset of unlabelled stars identified by CWOLA that fall within the smallest 10 per cent of five-dimensional distances d (see equation 1) to stars in the labelled GD-1 set, ranked in descending order by CWOLA's NN classifier score.

Table C1. Stars not part of the labelling from Price-Whelan & Bonaca (2018a) that also belong to the highest ranked subset of stars identified by the CWoLa scan of GD-1. The stars falling in the smallest 10 per cent of five-dimensional distances d (see equation 1) to stars in the labelled GD-1 set are shown here, ranked in descending order by CWoLa's NN classifier score.

Index	Patch	α	δ	μ^*_ϕ	μ_λ	$b-r$	g	d	NN score
0	3	153.343 750	37.944 386	− 3.648 079	− 11.087 369	0.585 669	18.889 666	0.199 219	0.760 671
1	3	147.407 532	34.807 056	− 4.174 263	− 13.030 541	0.632 448	18.280 497	0.161 793	0.710 143
2	3	147.581 451	33.771 442	− 4.031 364	− 12.893 708	0.692 181	19.217 056	0.200 405	0.703 576
3	3	145.067 444	34.774 712	− 3.889 829	− 11.528 839	0.560 968	18.343 340	0.208 592	0.687 666
4	0	144.668 640	29.123 871	− 3.294 825	− 9.761 044	0.698 603	18.599 613	0.157 535	0.628 063
5	0	139.284 058	27.534 752	− 2.474 001	− 11.759 004	0.678 160	18.488 266	0.202 521	0.608 509
6	0	141.217 377	21.388 325	− 3.016 653	− 12.211 497	0.731 615	19.056 505	0.207 587	0.607 377
7	0	141.366 760	26.641 132	− 2.638 692	− 12.904 817	0.754 404	19.493 120	0.169 499	0.603 854
8	0	137.850 464	25.221 642	− 2.823 553	− 13.327 346	0.713 997	18.919 308	0.181 071	0.600 161
9	0	141.072 662	28.944 424	− 3.398 272	− 11.644 784	0.810 871	19.610 243	0.077 491	0.590 473
10	10	159.375 610	43.155 155	− 6.771 805	− 13.304 443	0.577 917	18.163 719	0.193 080	0.582 036
11	0	142.465 027	31.453 848	− 3.129 723	− 10.795 928	0.638 067	17.908 943	0.162 005	0.582 026
12	0	137.821 594	25.365 973	− 2.537 415	− 10.836 835	0.771 456	19.665 091	0.186 863	0.547 747
13	10	152.977 966	43.714 725	− 5.651 521	− 11.580 858	0.640 087	18.764 666	0.173 924	0.541 810
14	14	190.253 159	58.102 585	− 8.347 640	− 3.359 662	0.640 982	19.427 656	0.206 074	0.513 133
15	14	194.429 871	58.634 842	− 8.420 991	− 7.967 748	0.623 426	18.534 977	0.183 633	0.506 830
16	15	182.109 512	56.585 564	− 6.765 204	− 12.116 959	0.627 577	19.014 071	0.191 147	0.505 615
17	16	162.275 238	49.602 287	− 7.460 149	− 8.779 883	0.658 072	19.046 940	0.198 527	0.505 006
18	16	166.262 054	49.207 397	− 8.196 286	− 8.721 247	0.743 465	19.535 479	0.206 786	0.498 103
19	9	167.548 462	46.377 983	− 6.176 190	− 11.992 228	0.646 498	19.345 844	0.088 779	0.497 648
20	14	179.886 322	55.335 400	− 8.907 069	− 8.033 201	0.637 199	18.149 529	0.158 184	0.492 758
21	15	172.536 072	52.980 537	− 7.740 779	− 7.965 540	0.740 067	19.315 273	0.179 650	0.487 783
22	15	170.305 725	55.276 653	− 8.972 501	− 9.987 629	0.618 397	18.751 921	0.121 189	0.484 890
23	8	178.793 396	50.679 008	− 7.688 651	− 4.308 014	0.688 902	19.250 017	0.209 937	0.484 879
24	19	179.387 024	53.861 988	− 7.553 021	− 9.258 564	0.633 465	18.817 064	0.177 797	0.484 814
25	14	173.568 787	54.718 918	− 8.506 504	− 5.805 184	0.657 850	19.249 598	0.127 450	0.484 371
26	9	169.494 263	48.418 388	− 6.111 476	− 10.342 608	0.673 502	19.317 274	0.128 495	0.484 024
27	8	187.757 996	56.693 752	− 7.339 619	− 4.749 725	0.576 080	18.252 359	0.151 545	0.484 008
28	9	165.833 618	46.725 655	− 5.527 166	− 10.702 893	0.613 531	18.830 116	0.150 776	0.483 593
29	9	171.009 125	45.997 330	− 5.812 016	− 6.994 314	0.617 441	18.986 839	0.169 053	0.483 200
30	9	173.235 535	49.141 151	− 5.340 712	− 9.560 485	0.606 213	18.940 699	0.100 418	0.479 875
31	8	179.718 201	50.653 923	− 7.778 558	− 6.846 837	0.664 310	19.358 950	0.182 553	0.477 425
32	9	159.310 638	44.860 550	− 6.879 270	− 10.555 117	0.646 944	18.803 308	0.209 683	0.477 115
33	9	164.777 161	44.976 212	− 6.129 719	− 11.014 638	0.603 947	18.769 825	0.195 550	0.476 689
34	9	176.500 488	54.726 501	− 7.533 120	− 8.043 803	0.614 882	19.896 984	0.196 763	0.475 622
35	19	177.458 374	50.769 260	− 7.356 768	− 5.745 450	0.645 157	18.890 451	0.145 692	0.475 277
36	9	172.817 383	47.330 818	− 6.053 680	− 7.922 386	0.642 633	19.386 862	0.202 310	0.474 907
37	9	163.159 119	48.486 298	− 6.168 635	− 10.865 490	0.666 342	18.826 414	0.182 255	0.472 844
38	15	170.374 420	52.475 494	− 7.904 741	− 6.552 203	0.715 446	19.745 758	0.196 269	0.472 336
39	9	160.771 454	46.261 520	− 5.855 455	− 7.474 462	0.591 482	19.225 233	0.171 725	0.464 586
40	7	186.439 041	56.541 378	− 10.644 468	− 4.102 360	0.592 939	19.715 214	0.175 948	0.463 402
41	8	178.696 075	52.531 986	− 7.652 631	− 7.286 445	0.669 975	18.414 928	0.178 466	0.462 668
42	7	190.129 303	57.698 757	− 8.130 330	− 5.936 471	0.590 197	19.154 623	0.140 599	0.462 299
43	8	176.834 106	50.947 758	− 7.213 428	− 7.684 932	0.678 570	19.351 543	0.182 765	0.461 110
44	9	173.211 731	46.760 998	− 6.690 597	− 8.577 130	0.730 625	19.276 901	0.171 330	0.460 547
45	7	202.413 101	58.419 315	− 8.423 603	− 3.303 266	0.626 270	19.095 215	0.075 321	0.460 429
46	7	198.092 728	58.096 100	− 7.937 617	− 3.228 269	0.629 757	18.509 893	0.170 395	0.459 476
47	8	176.415 344	54.566 105	− 7.152 437	− 5.265 400	0.794 621	19.286 375	0.115 260	0.458 131
48	7	192.434 601	54.163 914	− 9.025 725	− 3.314 333	0.714 384	19.720 222	0.199 648	0.457 166
49	7	197.343 521	60.127 289	− 8.470 557	− 2.185 798	0.705 921	19.428 307	0.190 847	0.456 135
50	8	176.661 682	52.064 625	− 6.882 434	− 8.300 988	0.686 596	18.789 845	0.134 871	0.454 971
51	7	194.420 959	53.819 744	− 7.492 790	− 1.907 870	0.605 728	19.301 640	0.188 744	0.452 361
52	7	195.111 771	56.396 278	− 8.144 449	− 3.451 988	0.727 108	19.597 523	0.174 442	0.452 323
53	7	204.361 725	57.704 208	− 8.309 439	− 5.412 242	0.568 638	19.430 788	0.205 758	0.452 242
54	7	195.939 850	55.778 358	− 7.729 141	− 2.890 564	0.624 128	18.510 710	0.162 695	0.451 727
55	7	189.208 954	58.567 673	− 8.094 695	− 1.741 354	0.663 921	19.207 617	0.192 408	0.449 935
56	7	186.898 468	55.441 170	− 8.953 770	− 4.824 331	0.760 971	19.275 148	0.133 778	0.448 510
57	7	189.780 121	56.009 251	− 9.288 628	− 1.816 544	0.769 598	19.816 544	0.209 066	0.447 054
58	7	200.713 043	56.548 923	− 8.371 297	− 4.555 718	0.603 445	18.767 347	0.129 104	0.446 953
59	1	141.269 653	26.119 368	− 3.506 178	− 13.209 450	0.637 341	17.898 008	0.194 991	0.444 487

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