Via Machinae

Discovering Stellar Streams and Modeling the Galaxy with Normalizing Flows

David Shih

March 17, 2021

Meeting on Deep Generative Models for Fundamental Physics, BIDS

Thanks to my collaborators

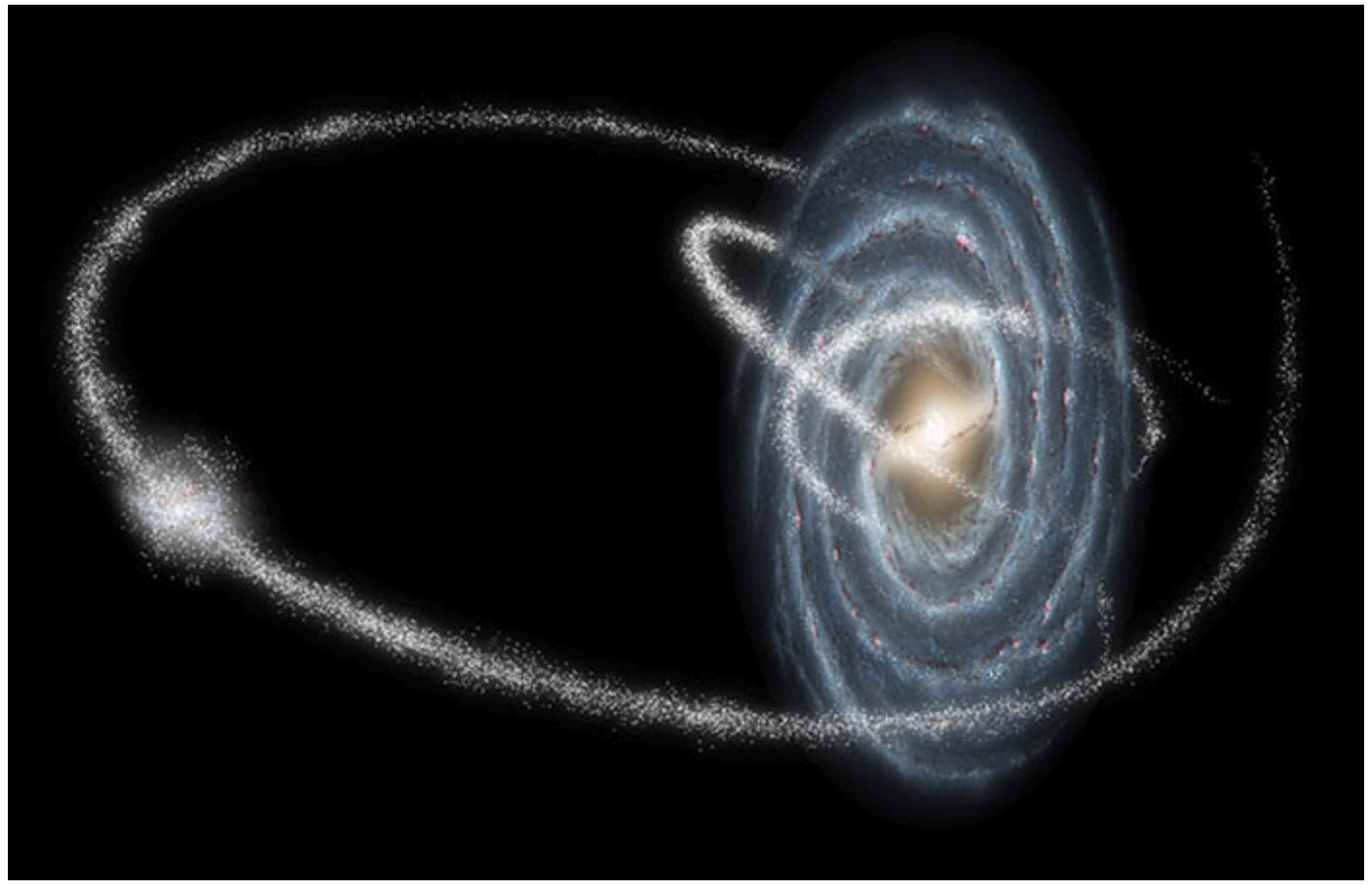
Matt Buckley

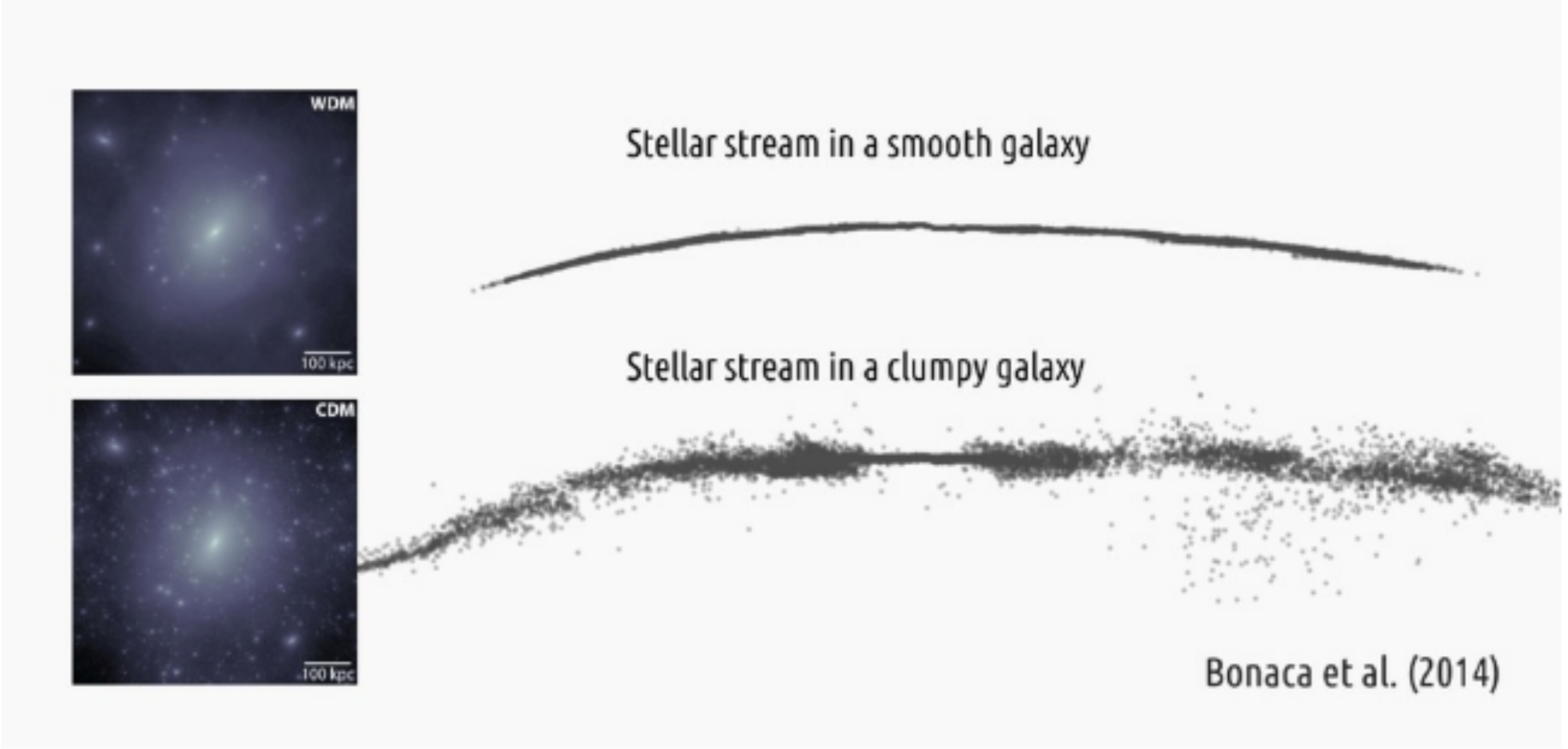
Lina Necib

John Tamanas

Stellar Streams

Stellar streams are cold, tidally-stripped remnants of globular clusters and dwarf galaxies, falling into and orbiting our galaxy.

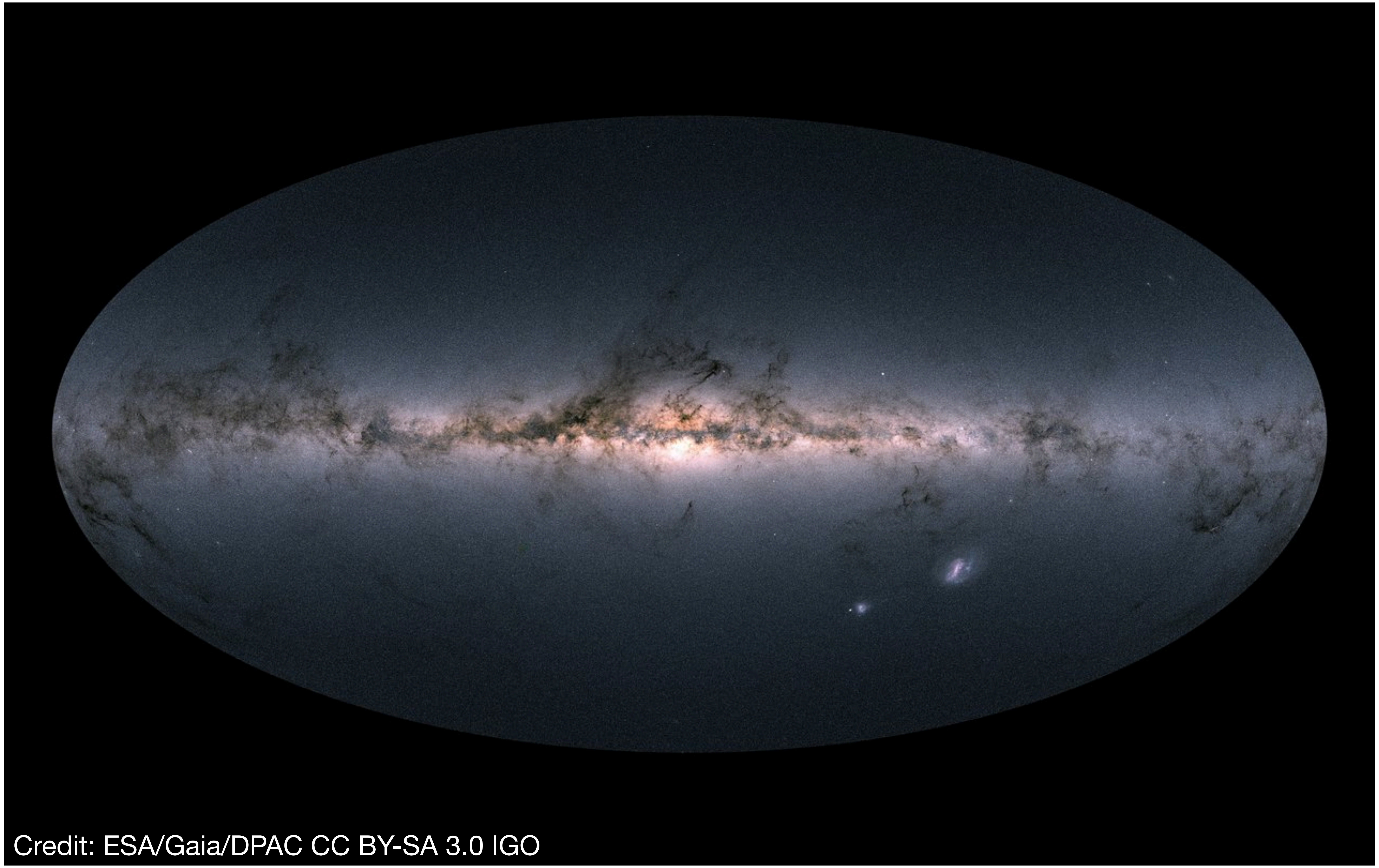




They are very interesting objects of study for astrophysicists and particle physicists.

In particular, they could be unique probes into dark matter substructure.

The Gaia satellite

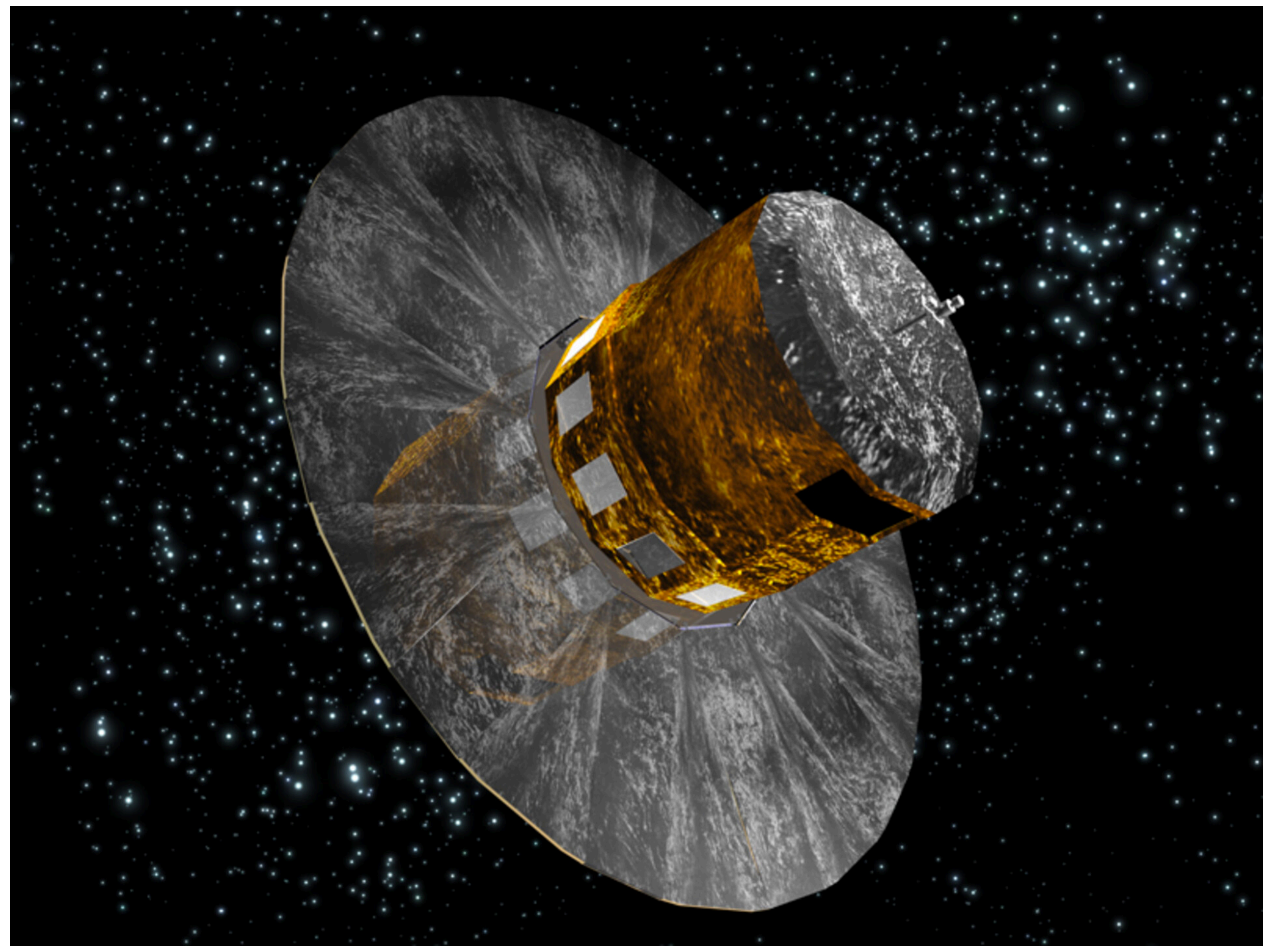


Gaia's image of the Milky Way

The Gaia satellite

The Gaia satellite is providing an unprecedented window into the stellar population of our Galaxy:

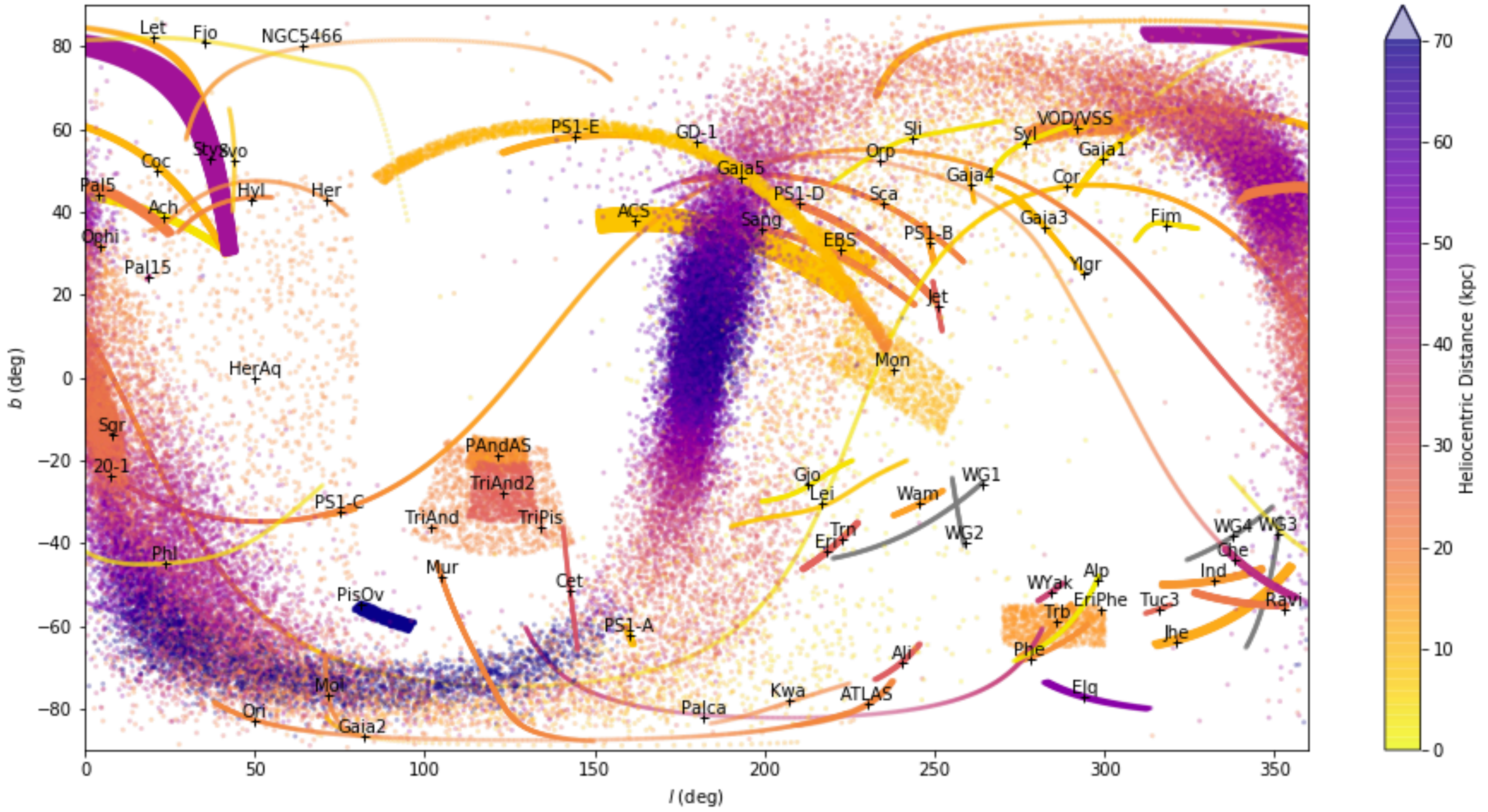
- Launched in 2013; extended to 2025
- Mission: map out the full 6d phase space + photometry of the stars in our galaxy
- **Angular positions, velocities, color, magnitude** of over 1 billion stars in our galaxy
- radial positions and velocities for a smaller subset of nearby stars (not used in this work)



A potential gold mine for stream finding!

Stream finding: previous approaches

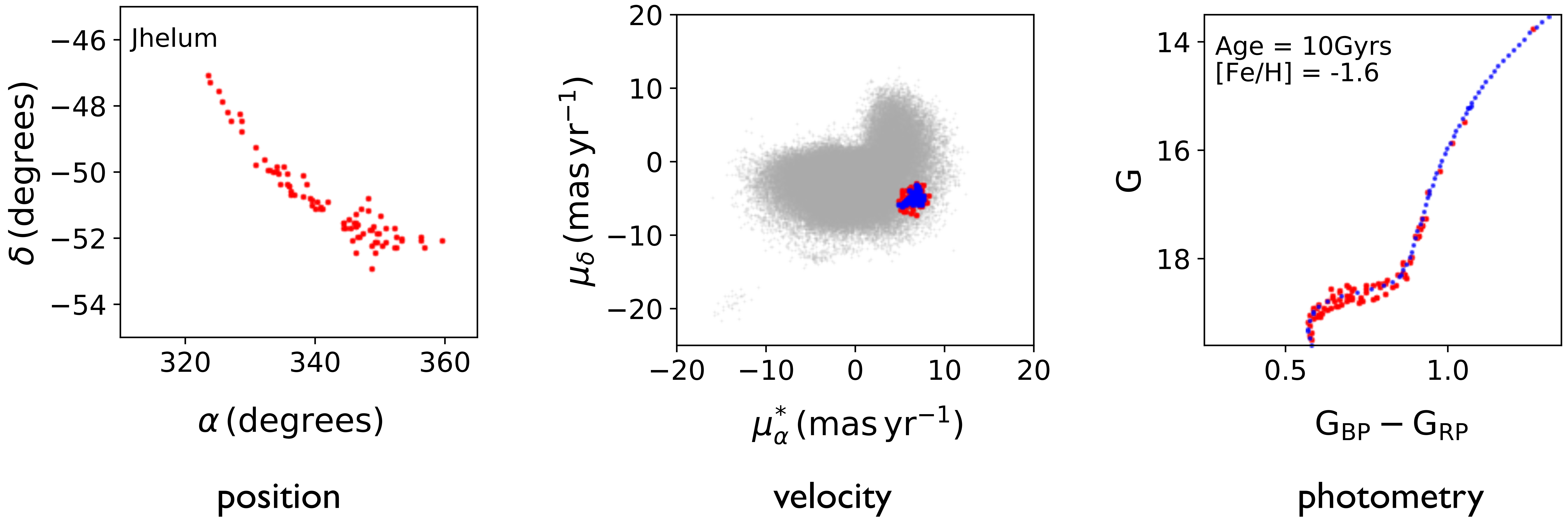
https://github.com/cmateu/galstreams



Stream finding: previous approaches

- Some streams (eg Sagittarius) are large and bright enough to even see by eye
- Many streams were previously discovered in other deep surveys (eg DES, SDSS) and were reconfirmed in Gaia data, often using special tracer stars like RR Lyrae
- Several automated algorithms for stream finding in the bulk of the Gaia data exist; the most successful so far is STREAMFINDER (Malhan & Ibata 2018). These algorithms have found many new streams in the Gaia data, but they all make a number of model-dependent assumptions (form of the galactic potential, orbits, isochrones, galactic merger history…).
- **Our goal:** an automated stream-finding algorithm that
  - Uses only bulk *Gaia* data
  - Does not assume a Galactic potential or orbit
  - Does not assume stream stars lie on a particular isochrone

Stream finding



Streams are local overdensities in position, velocity and photometric space.

- Since they are cold, the stars in the stream are clustered in velocity
- The stars in a (globular cluster) stream are all born at approximately the same time — they should lie on an isochrone in color-magnitude space

Anomaly Detection for Streams

- The problem: we have data, drawn from some probability distribution $p(\vec{x})$

“features”
(position, velocity, color, etc)

- The signal and background probability distributions are different:

$$p(\vec{x}) = \alpha\ p_{\rm sig}(\vec{x}) + (1-\alpha)\ p_{\rm bg}(\vec{x}) \qquad \alpha \ll 1$$

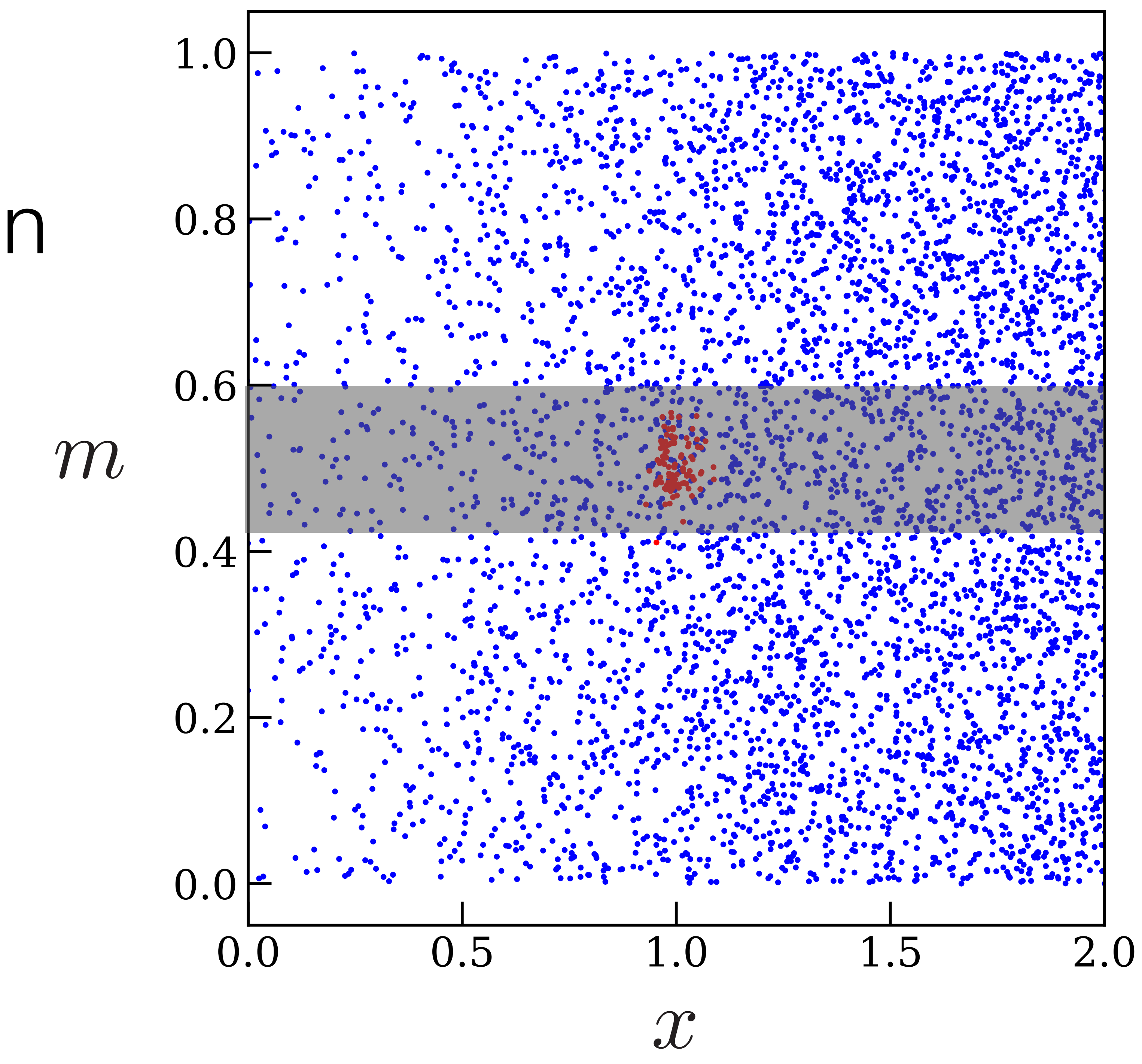
- The optimal statistic for distinguishing signal from background is the ratio

$$R(\vec{x}) = \frac{p(\vec{x})}{p_{\rm bg}(\vec{x})}$$

  - Signal dominates wherever $R(\vec{x}) \gg 1$.
- The problem: How do we determine both $p(\vec{x})$ and $p_{\rm bg}(\vec{x})$? Especially in something as complicated as the Galaxy.

ANODE (Nachman and Shih, 2020)

- Pick one feature m where signal is known to be localized. Define a "search region" (SR) by $m \in [m_0 \pm \frac{\Delta m}{2}]$
- Learn conditional densities $p(x|m{\in}SR)$ and $p(x|m{\notin}SR)=p_{bg}(x|m{\notin}SR)$
    - Made possible in high dimensional data using recent progress in density estimation (esp normalizing flows; see also GIS Dai & Seljak 2020)
    - Via Machinae uses *Masked Autoregressive Flows (MAF)* (Papamakarios et al 1705.07057)
- Interpolate $p_{bg}(x|m{\notin}SR)$ in m to obtain $p_{bg}(x|m{\in}SR)$

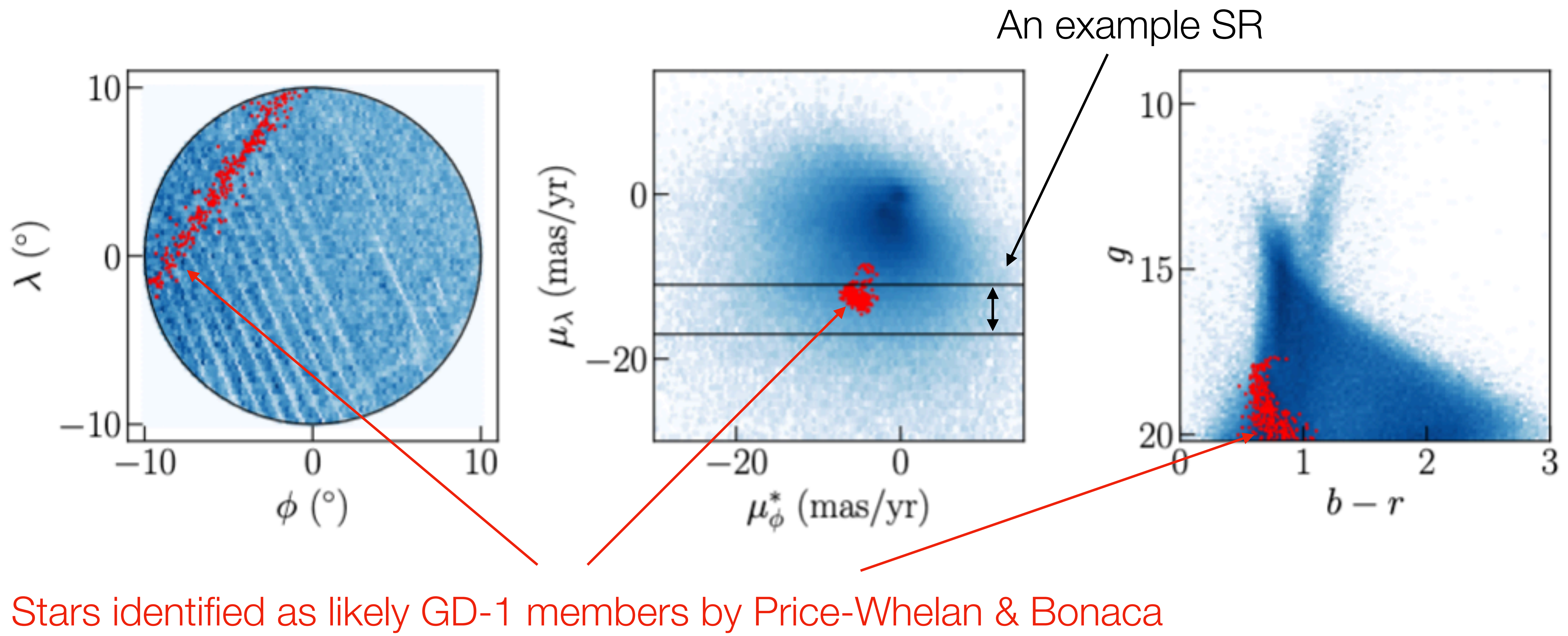


$\Longrightarrow$ Directly construct optimal discriminant in the SR:

$$R(x|m) = \frac{p(x|m)}{p_{\mathrm{bg}}(x|m)}$$

GD-1 Example

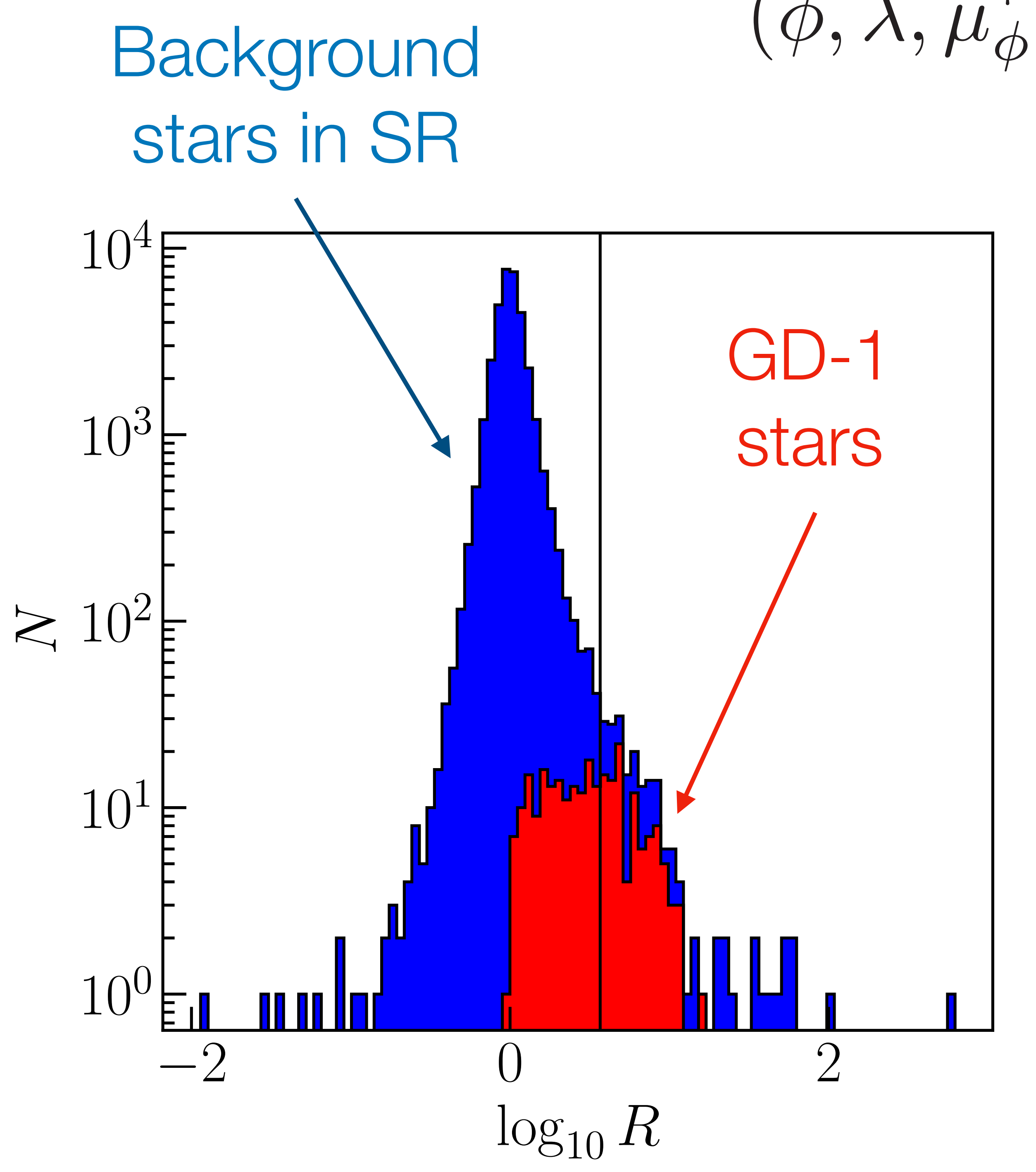
- GD-1 is a bright stream with stellar catalogues of stream membership (Price-Whelan and Bonaca, 2018)
  - Provides a good worked example for Via Machinae (Shih, Buckley, Necib, Tamanas *in prep*)
- Streams are concentrated in both μ_λ and μ_ϕ^*, with a width of a few mas/yr.
  - We will pick μ_λ as the feature m to define our overlapping *search regions* (SRs)
  - Width 6 mas/yr for each SR, neighboring SRs separated by 1 mas/yr

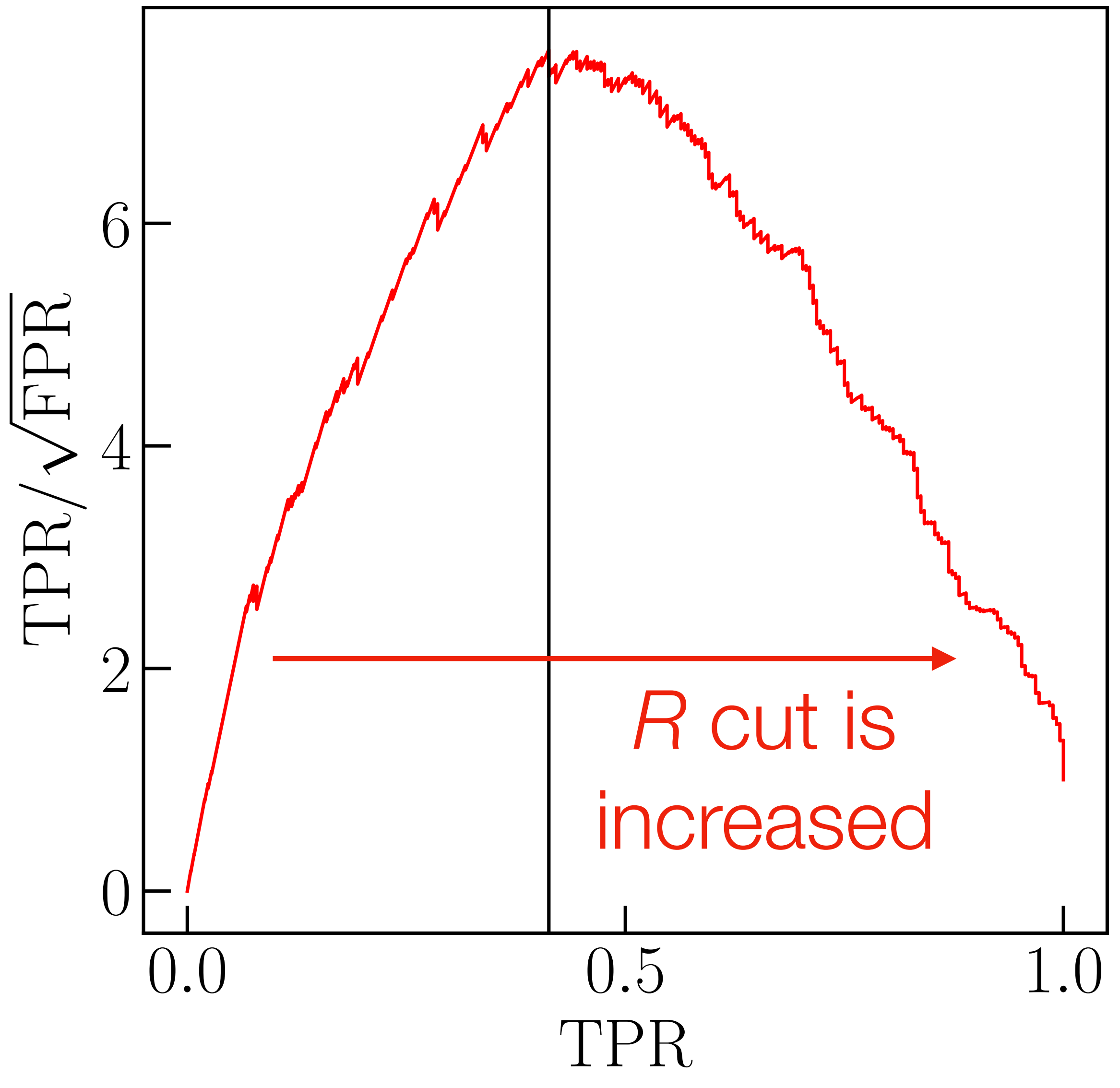


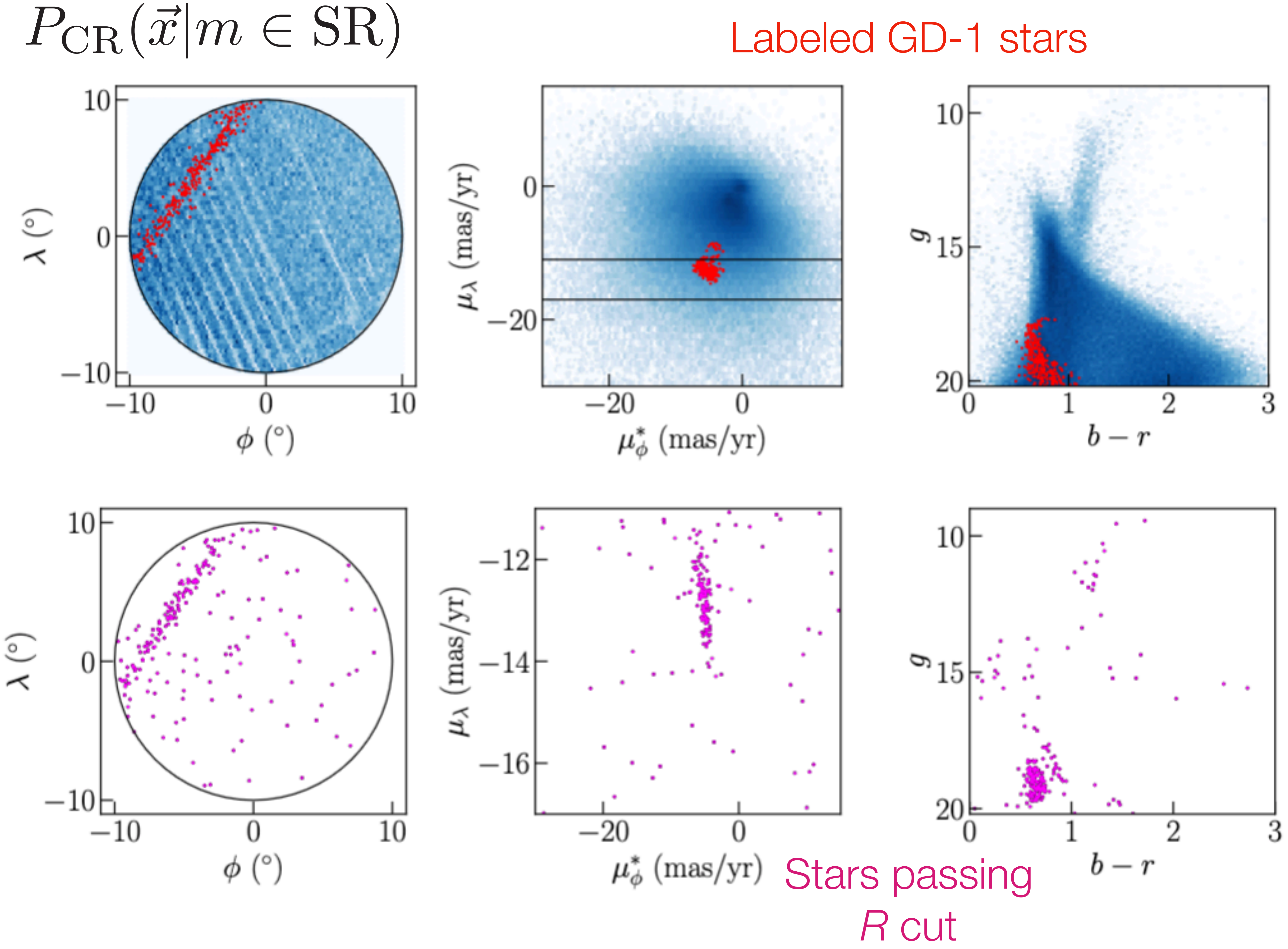
GD-1 Example

- For each SR within each patch, we train ANODE on the stars in the SR, using the complement of the SR as the control region.
  - For each star, we now have $R(\vec{x}|m \in \mathrm{SR}) = \dfrac{P(\vec{x}|m \in \mathrm{SR})}{P_{\mathrm{CR}}(\vec{x}|m \in \mathrm{SR})}$

$(\phi, \lambda, \mu_\phi^*, g, b-r)$ μ_λ

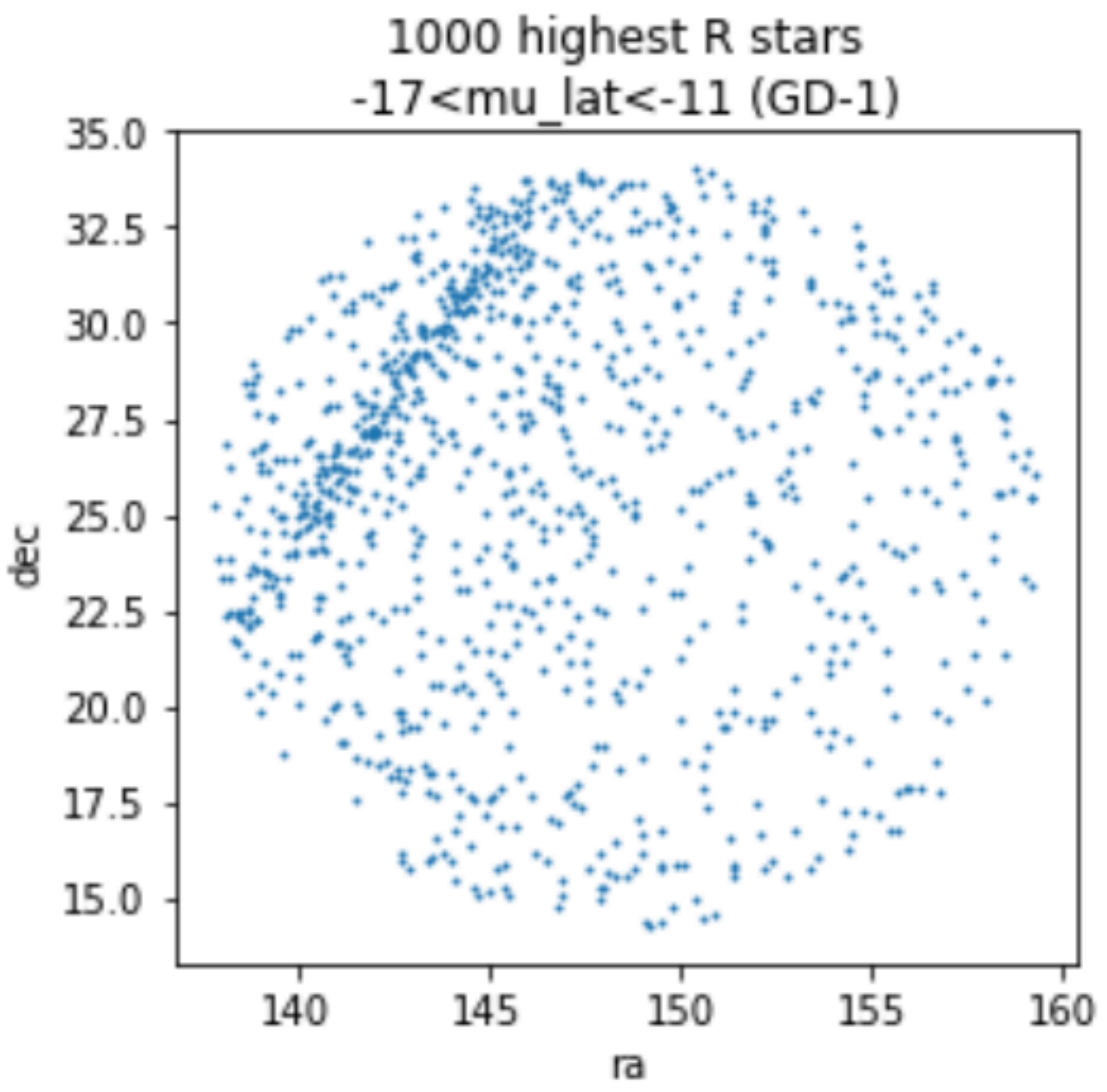






Beyond GD-1

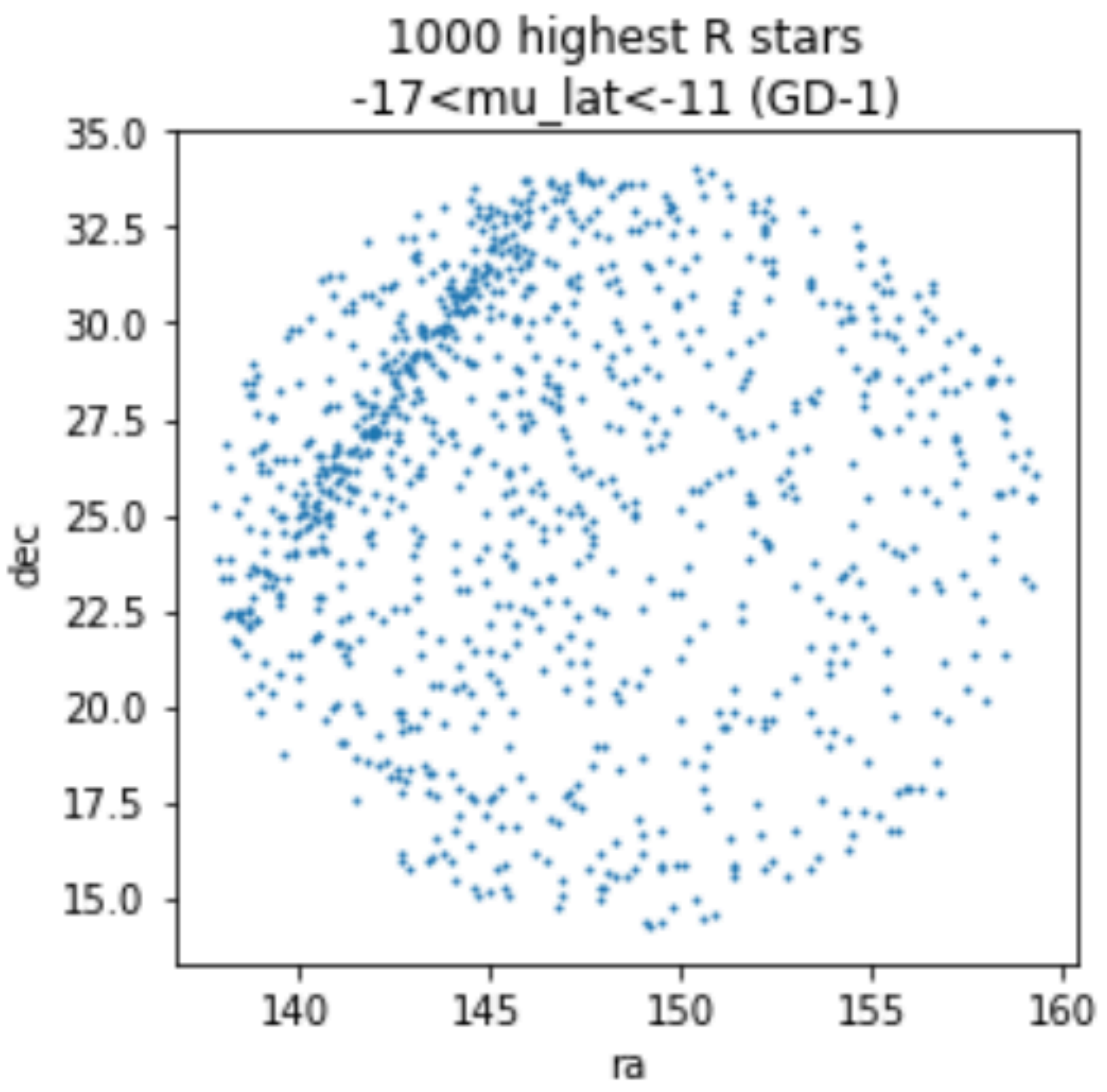
For GD-1 it is enough to cut on R(x) and inspect the stars passing the cut by eye.

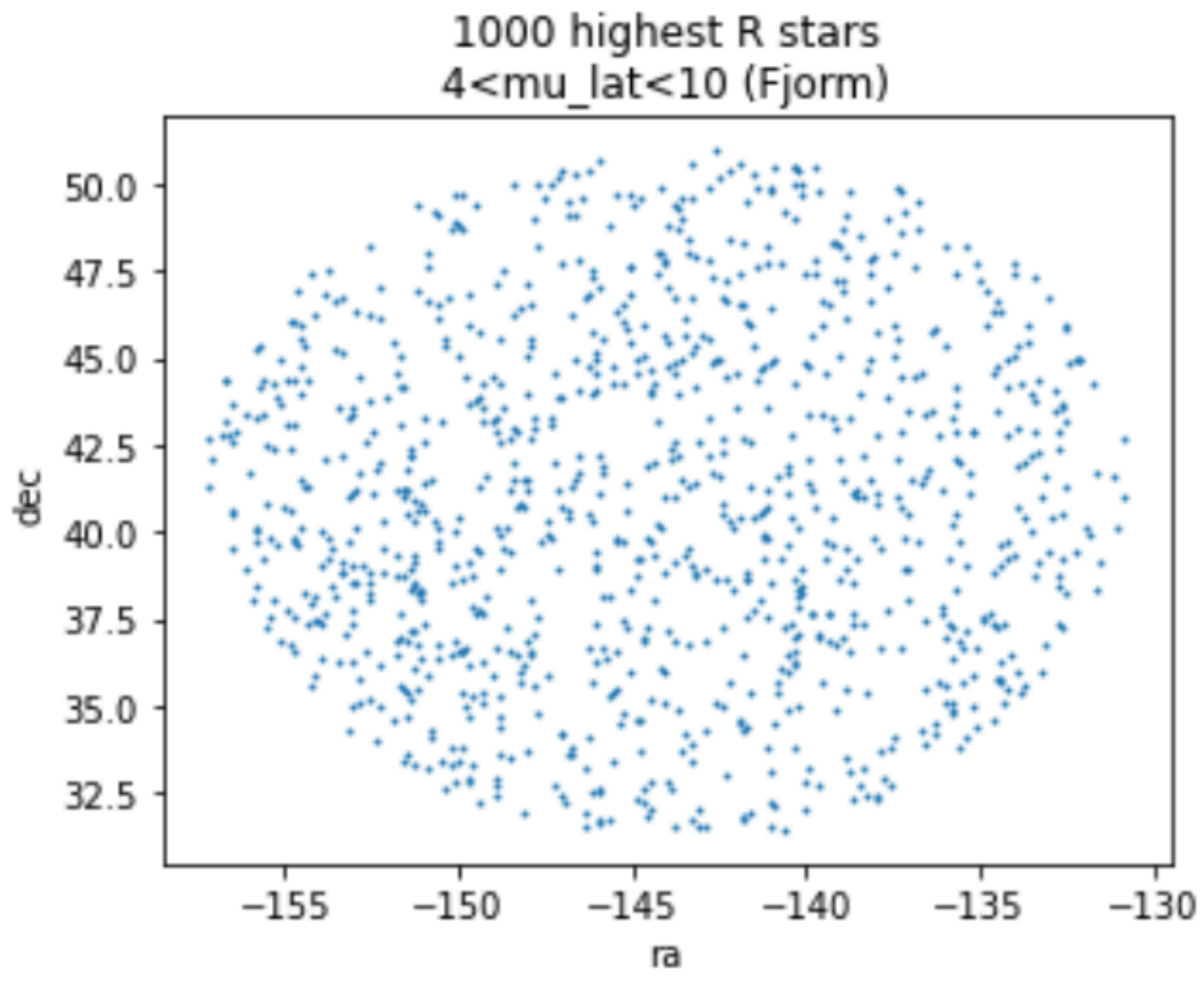


Beyond GD-1

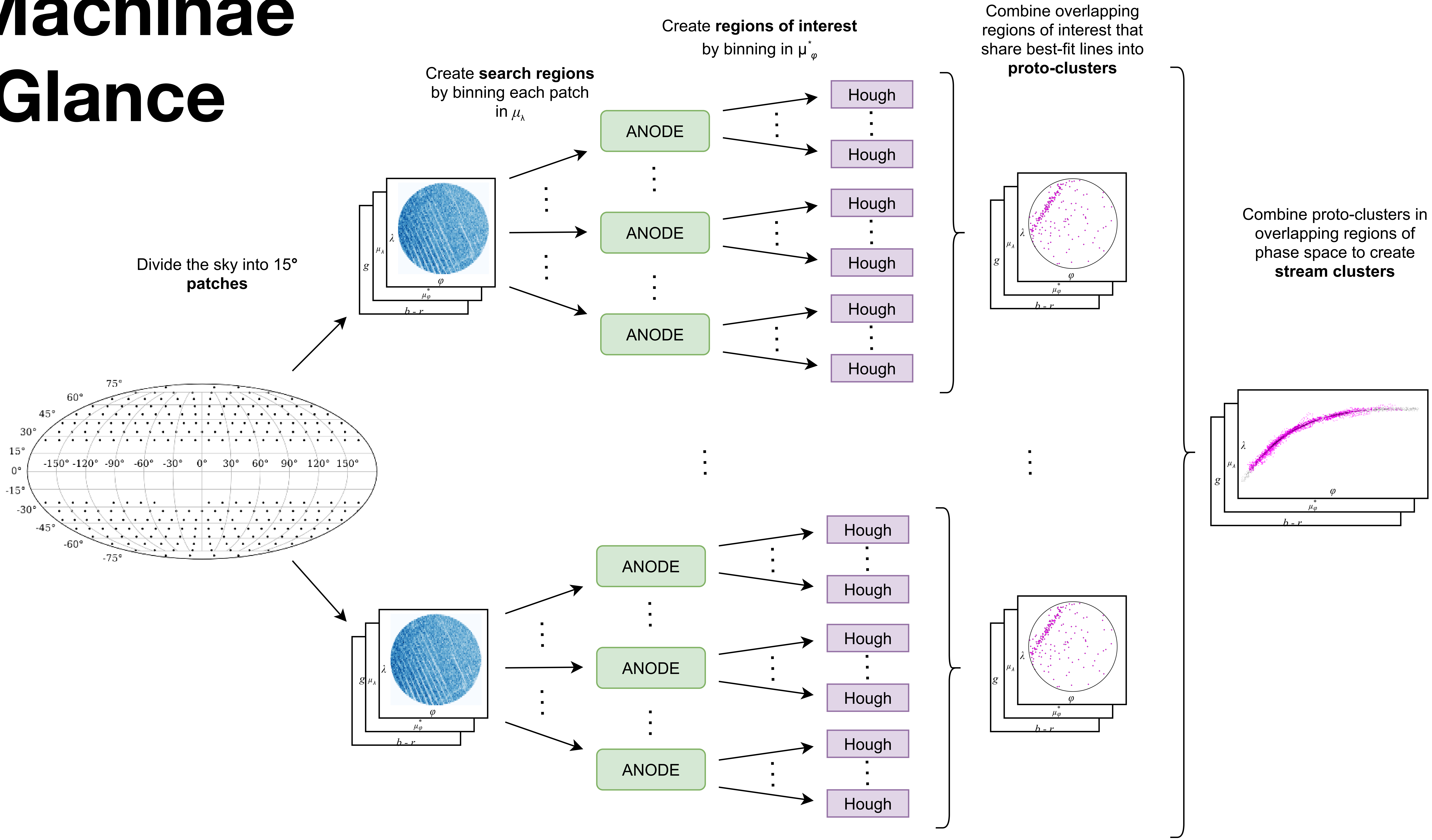
For GD-1 it is enough to cut on R(x) and inspect the stars passing the cut by eye.

For other known streams, we found it is generally not enough.

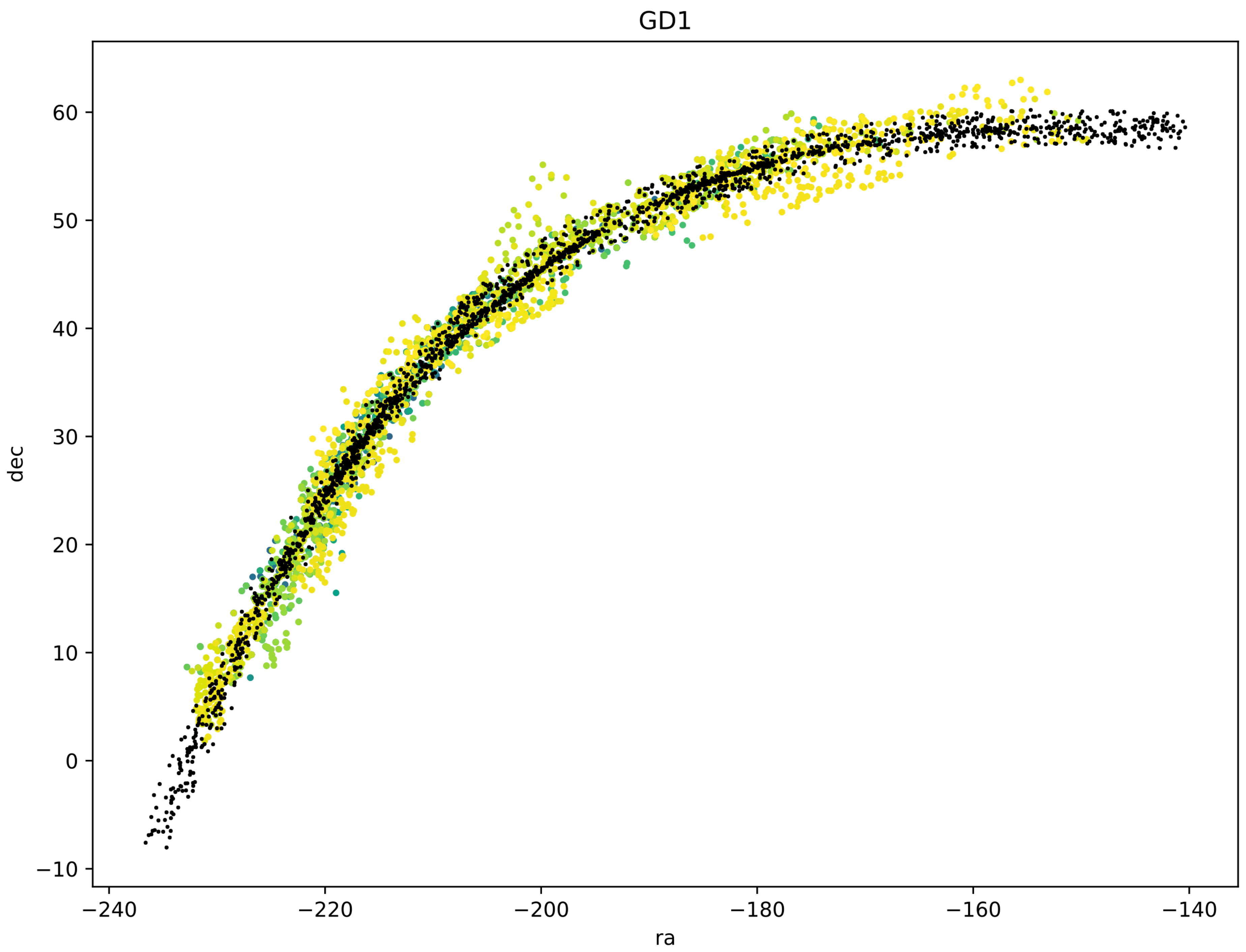


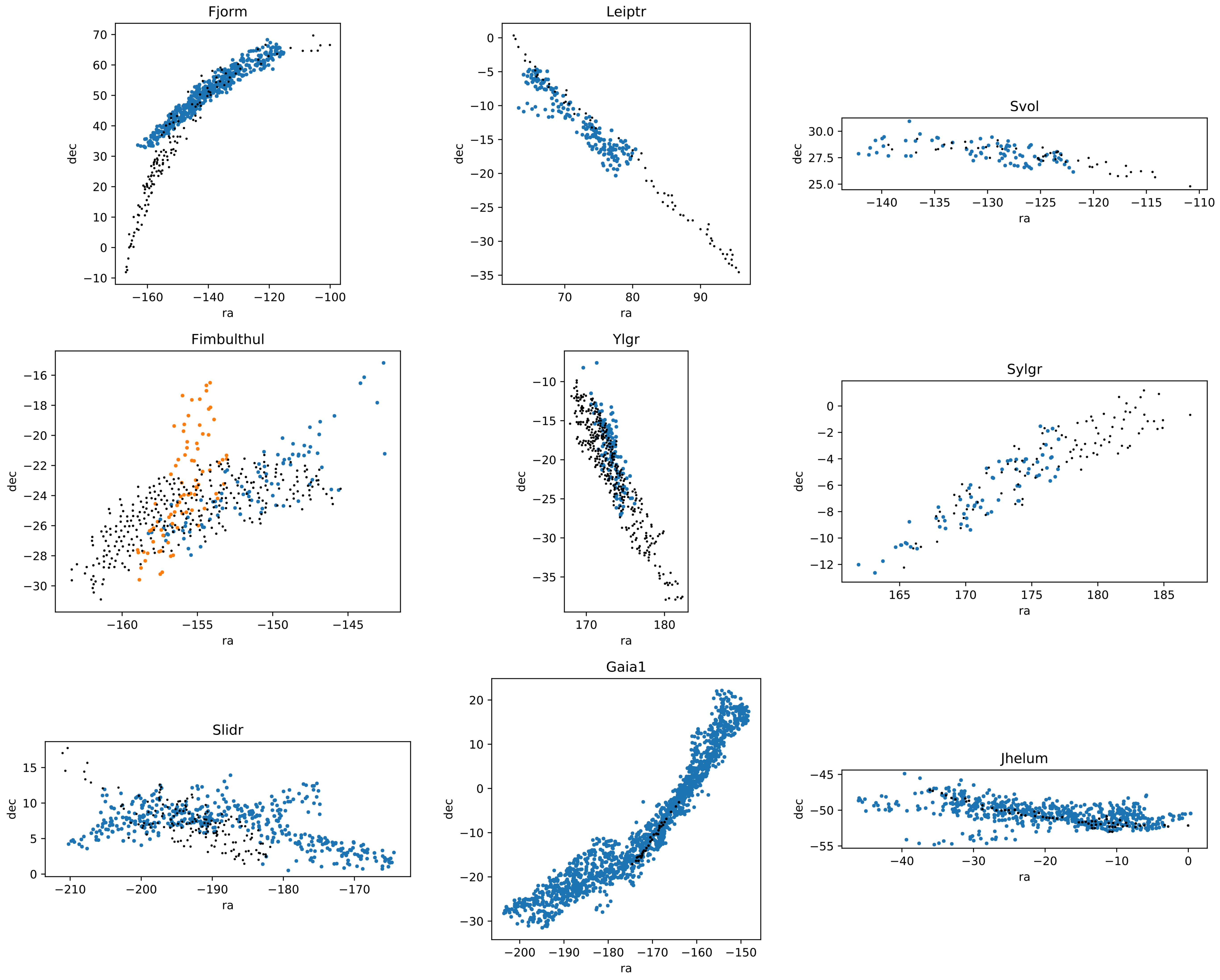


Via Machinae at a Glance



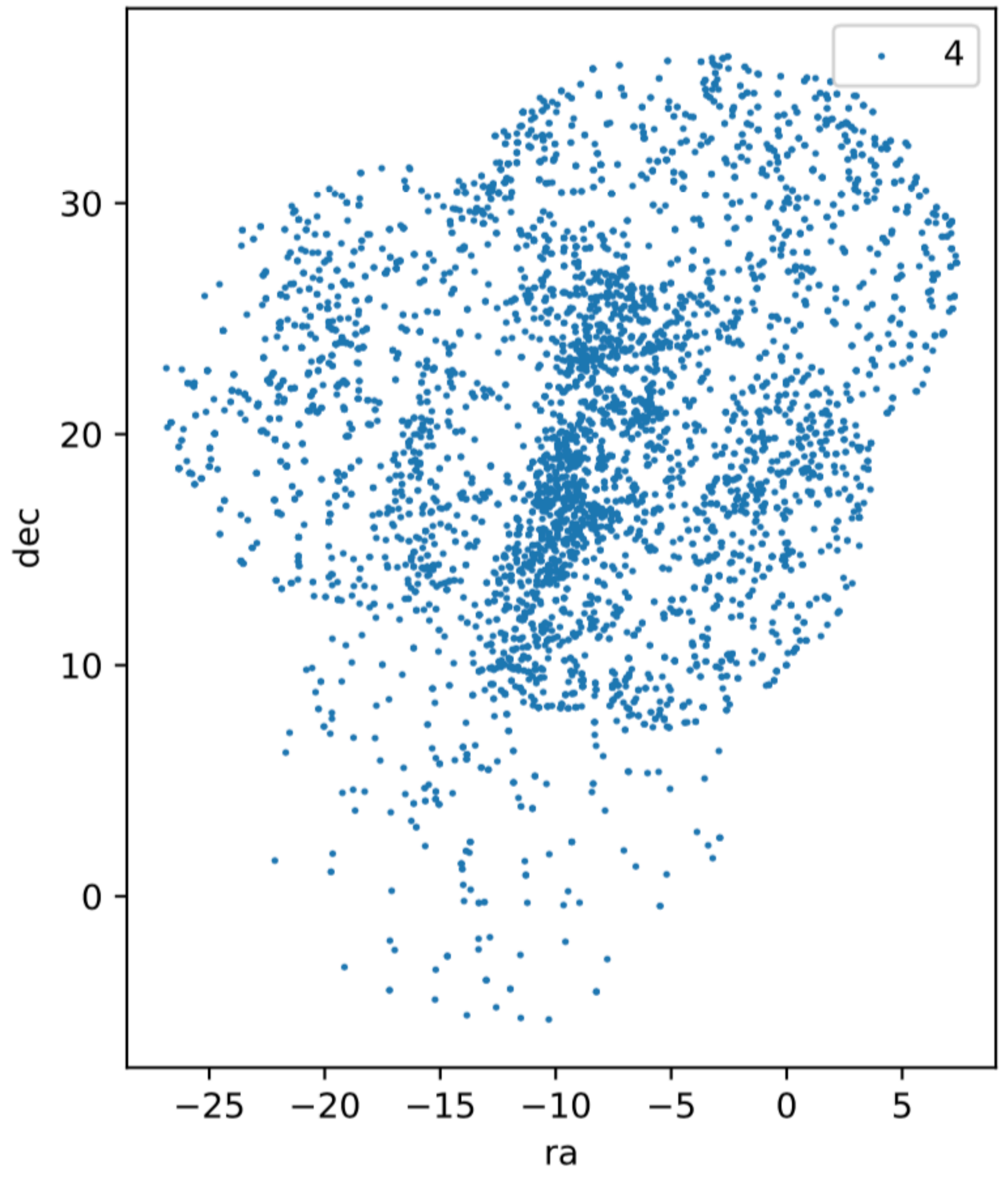
Results: known streams





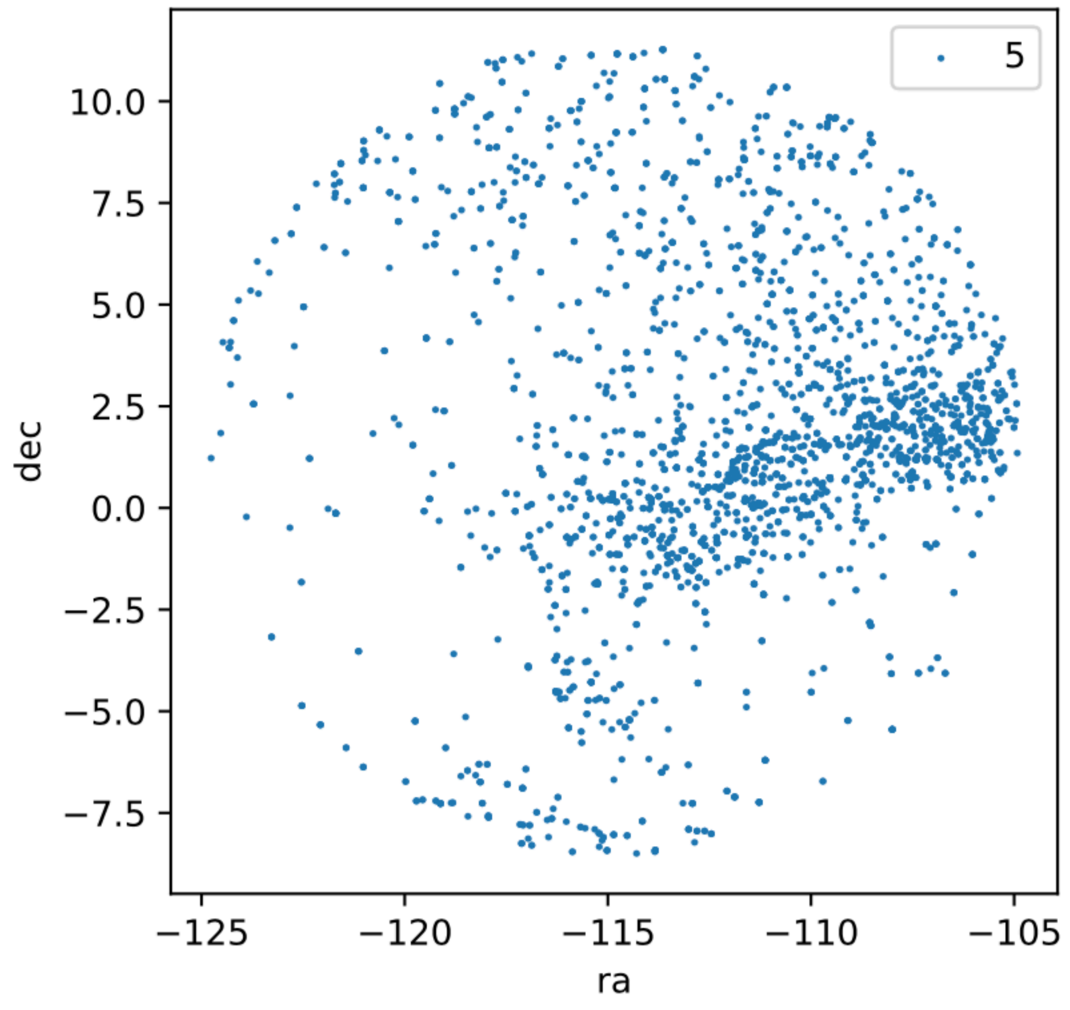
Via Machinae successfully finds known streams!

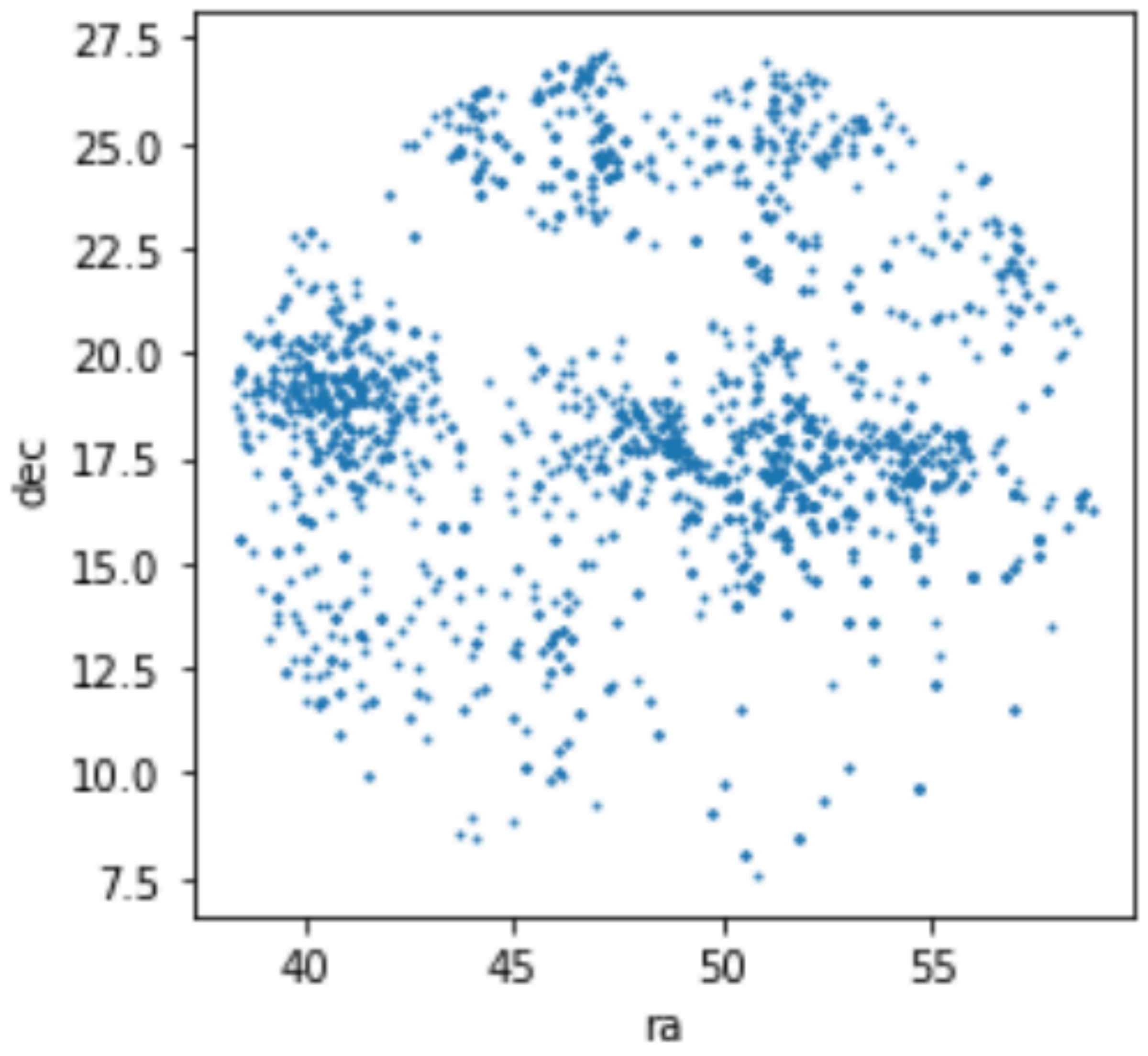
New stream candidates

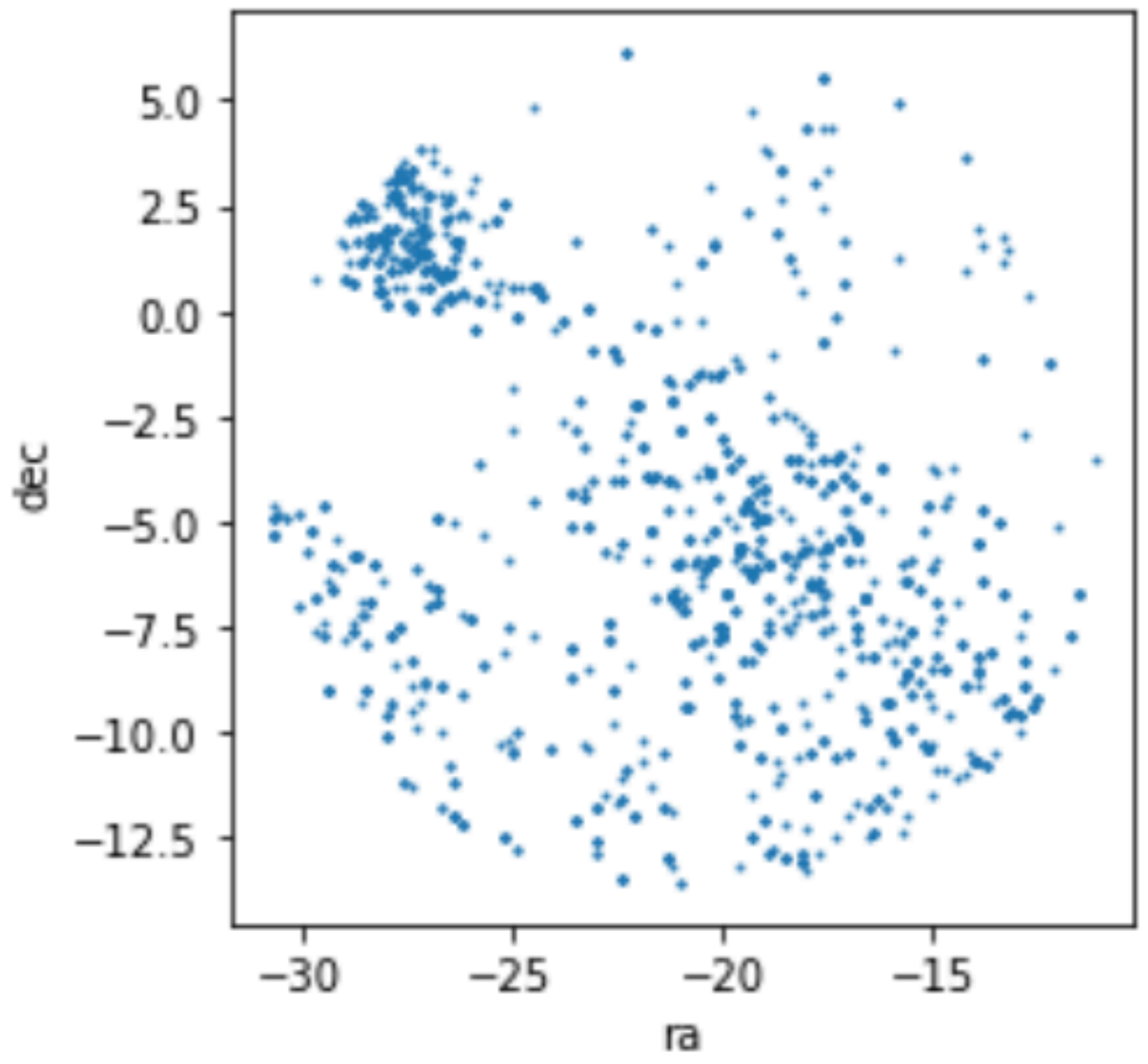


We are currently investigating ~580 new stream candidates.

Some look promising — stay tuned!



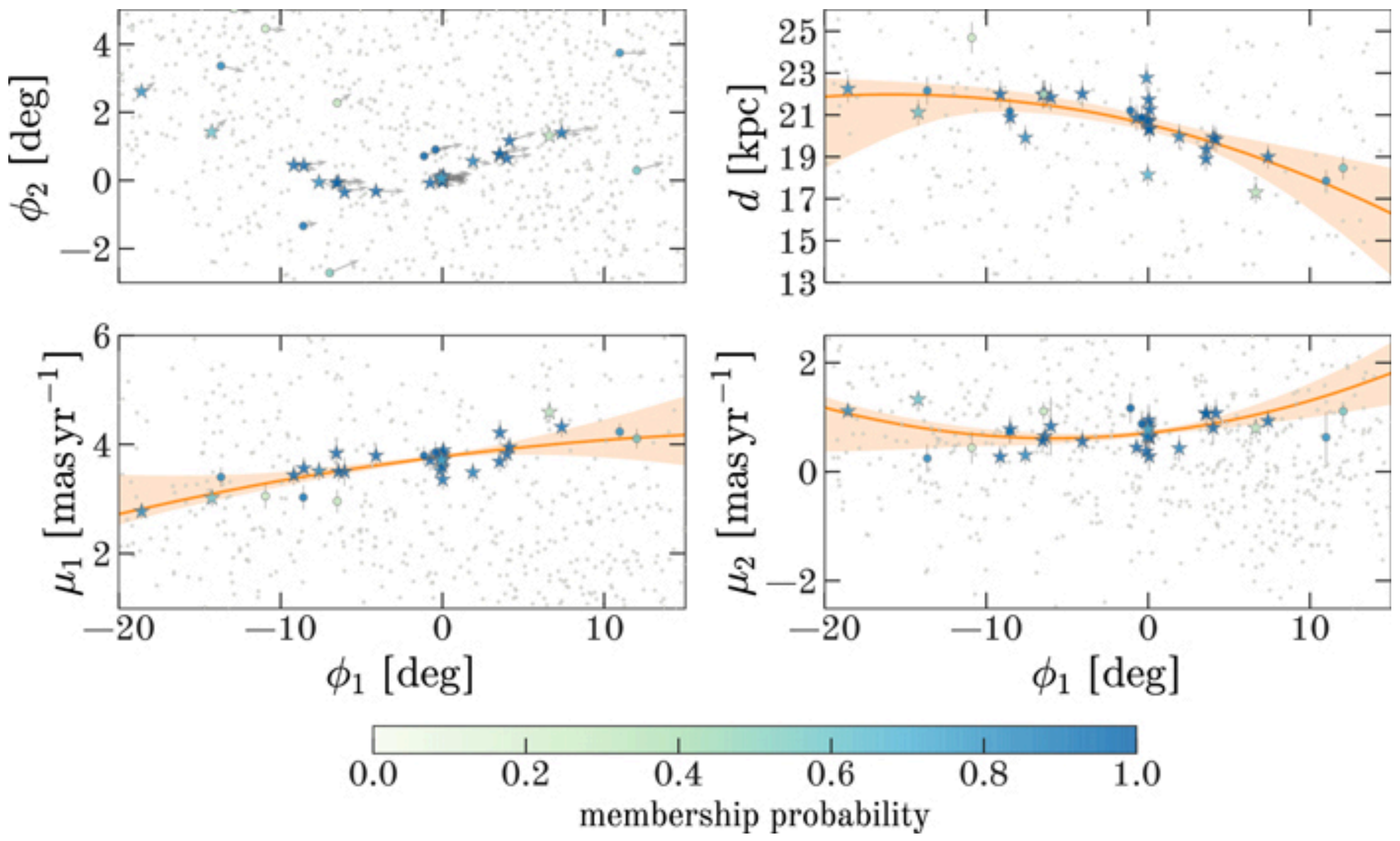




Conclusions

- How to validate the 580 new stream candidates?
  - Cross matching with other catalogues?
  - Follow up observations?
- Improvements to R(x)?
  - More detailed hyperparameter tuning
  - Other even more powerful neural density estimators
  - Alternatives to ANODE method — ***CWoLa in Space?*** Work in progress with Buckley, Collins, Nachman & Thanvantri
- Are we finding other objects besides streams?
  - globular clusters!!
  - debris flow?
- Other uses for density estimation? e.g.
  - stream membership?
  - mock catalogues?

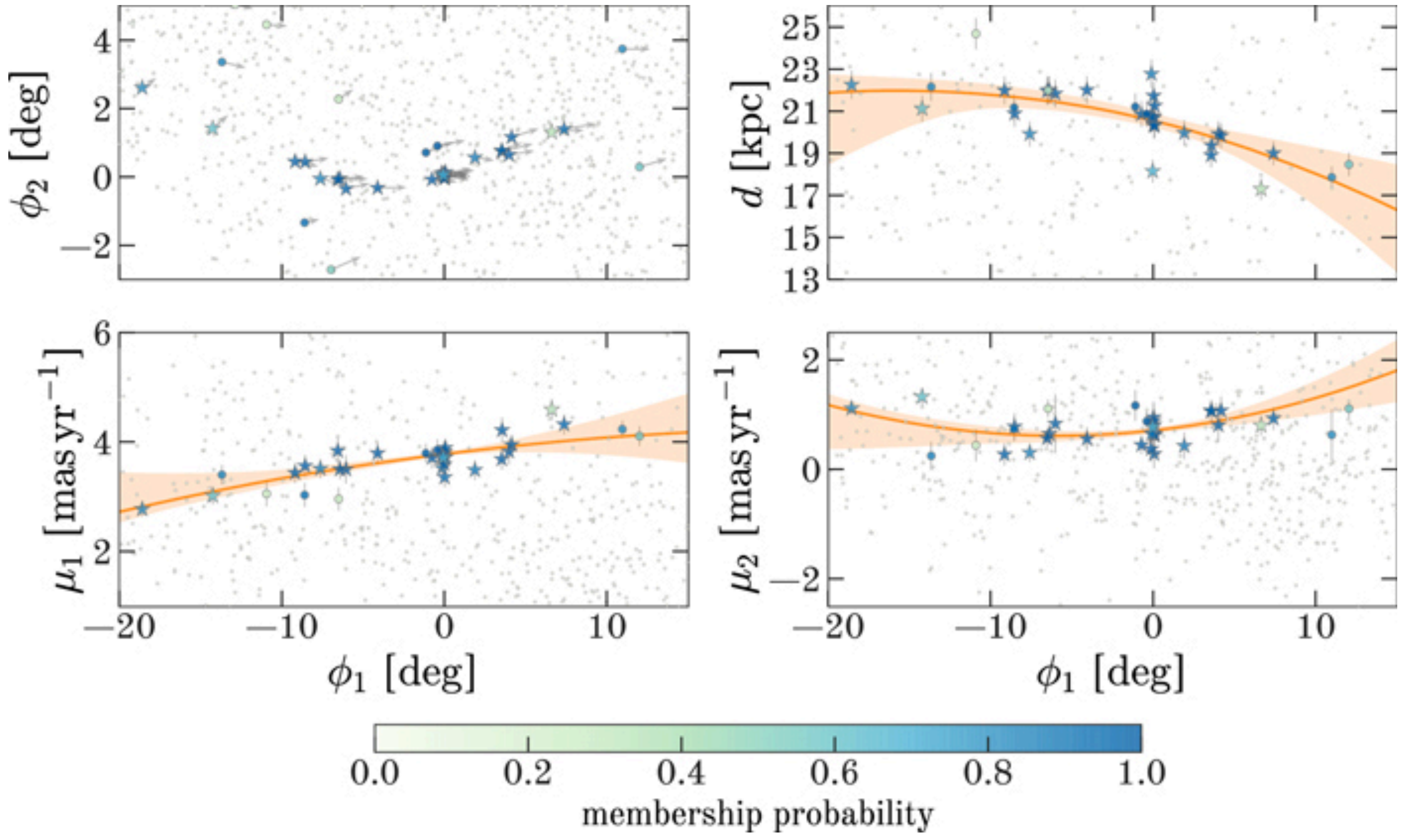
"Kinematics of the Palomar 5 Stellar Stream from RR Lyrae Stars" Price-Whelan et al (2019)



Conclusions

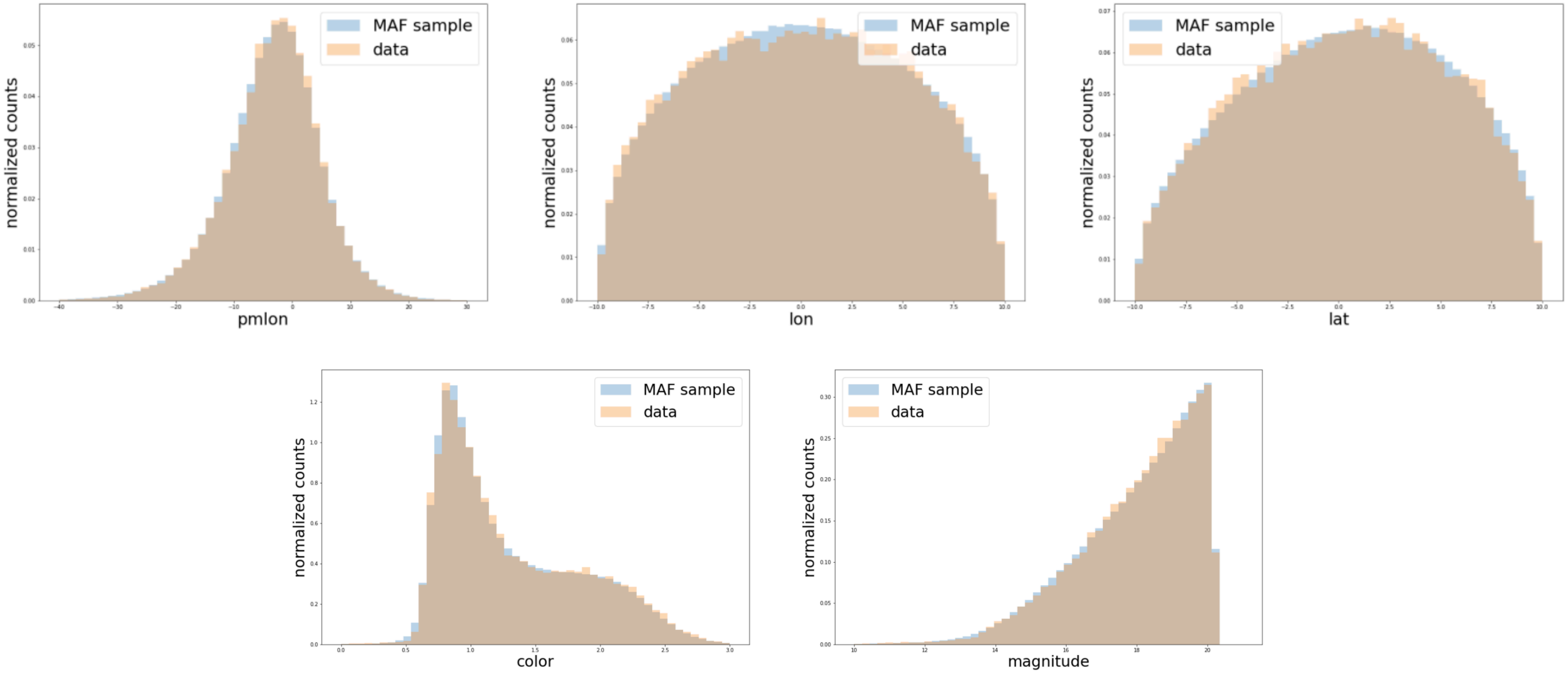
- How to validate the 580 new stream candidates?
  - Cross matching with other catalogues?
  - Follow up observations?
- Improvements to R(x)?
  - More detailed hyperparameter tuning
  - Other even more powerful neural density estimators
  - Alternatives to ANODE method — ***CWoLa in Space?*** Work in progress with Buckley, Collins, Nachman & Thanvantri
- Are we finding other objects besides streams?
  - globular clusters!!
  - debris flow?
- **Other uses for density estimation? e.g.**
  - **stream membership?**
  - **mock catalogues?**

"Kinematics of the Palomar 5 Stellar Stream from RR Lyrae Stars" Price-Whelan et al (2019)



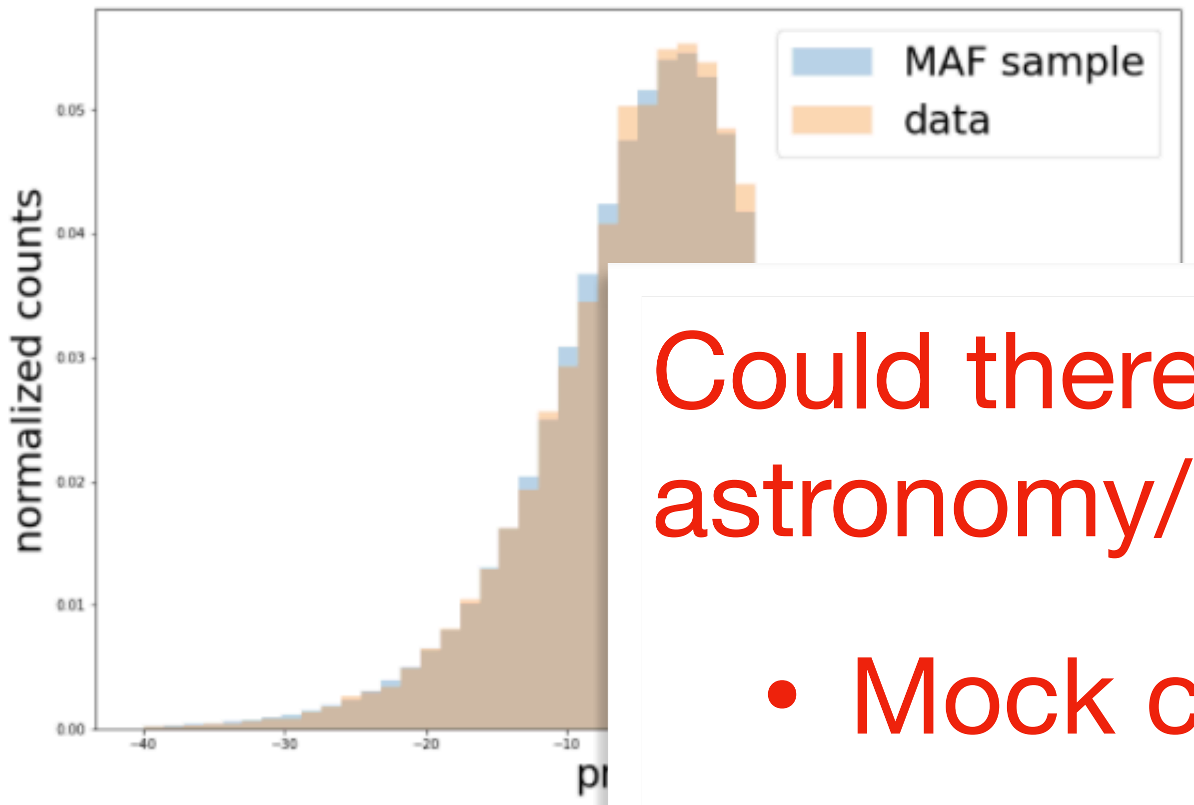
Future directions: modeling the galaxy?

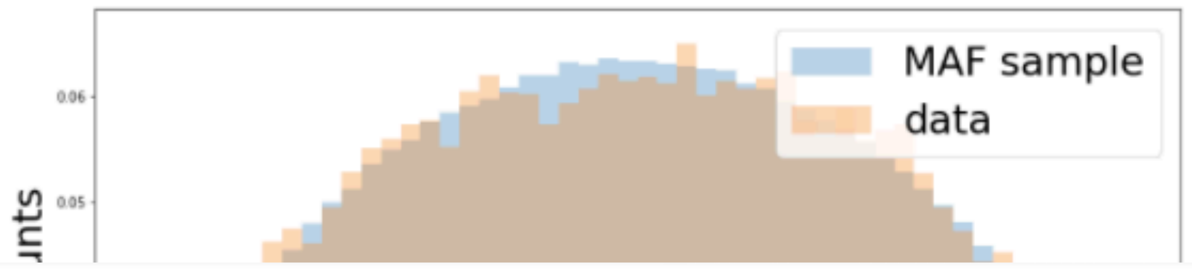
- We've trained normalizing flows on the positions, proper motions, color and magnitude of the stars in the Gaia data.
- These normalizing flows can be sampled from:

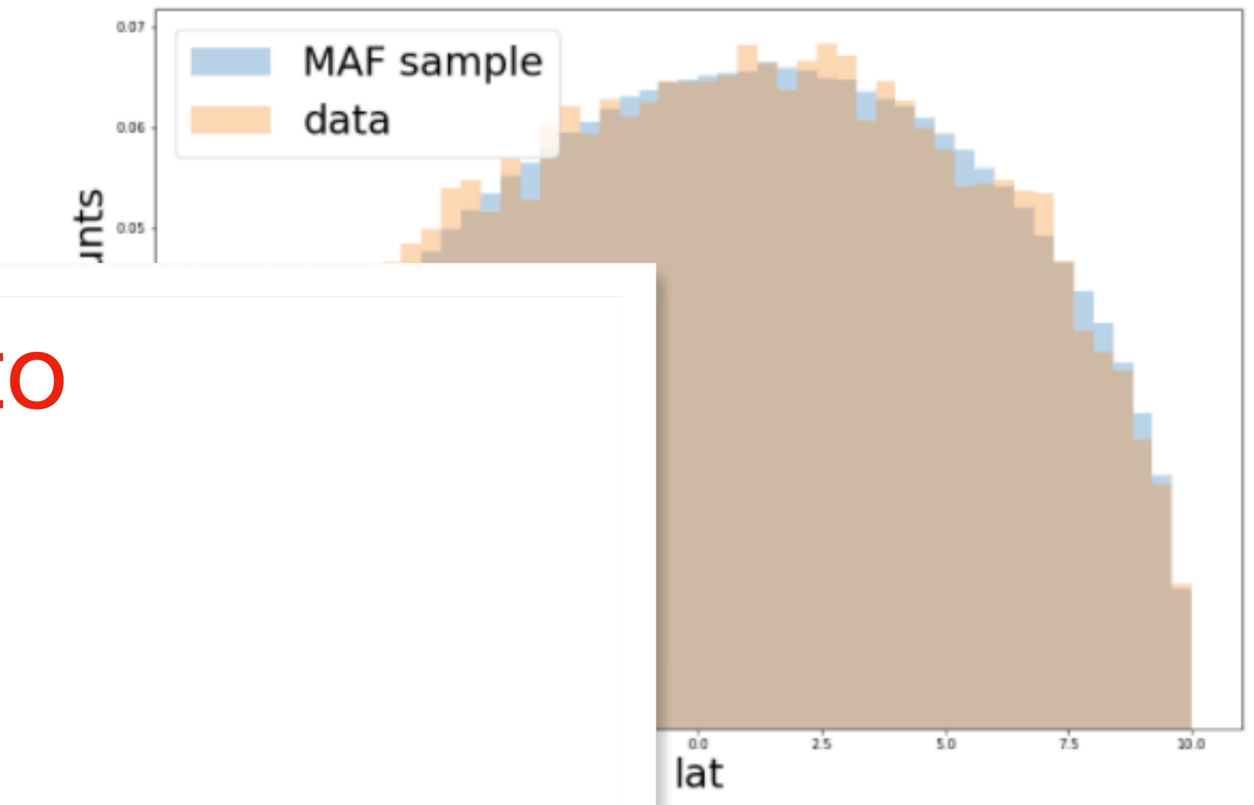


Future directions: modeling the galaxy?

- We've trained normalizing flows on the positions, proper motions, color and magnitude of the stars in the Gaia data.
- These normalizing flows can be sampled from:



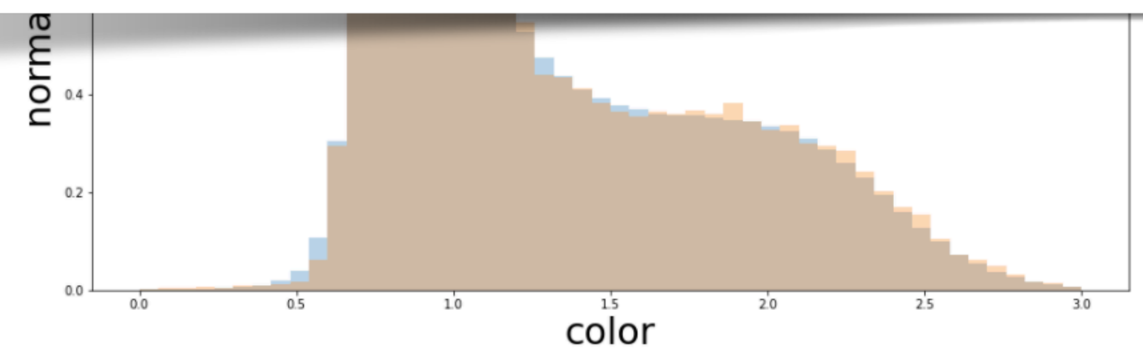


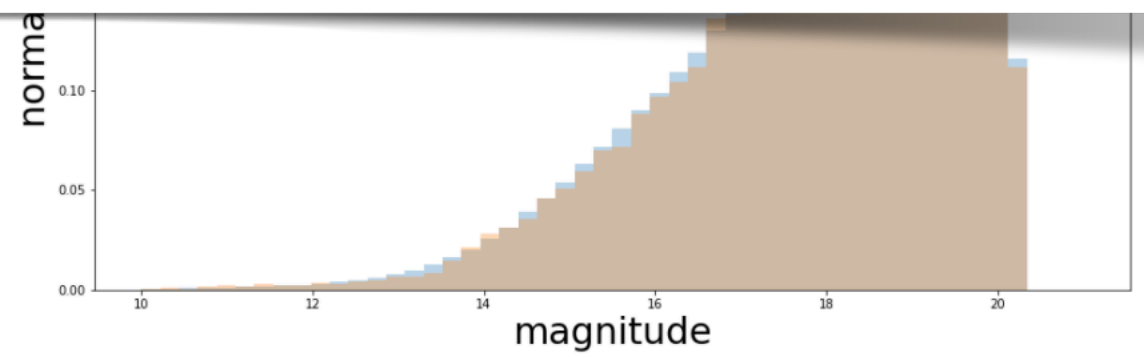


Could there be interesting applications to astronomy/astrophysics?

- Mock catalogs?
- Measuring the potential of the Milky Way?

Would be happy to discuss further!



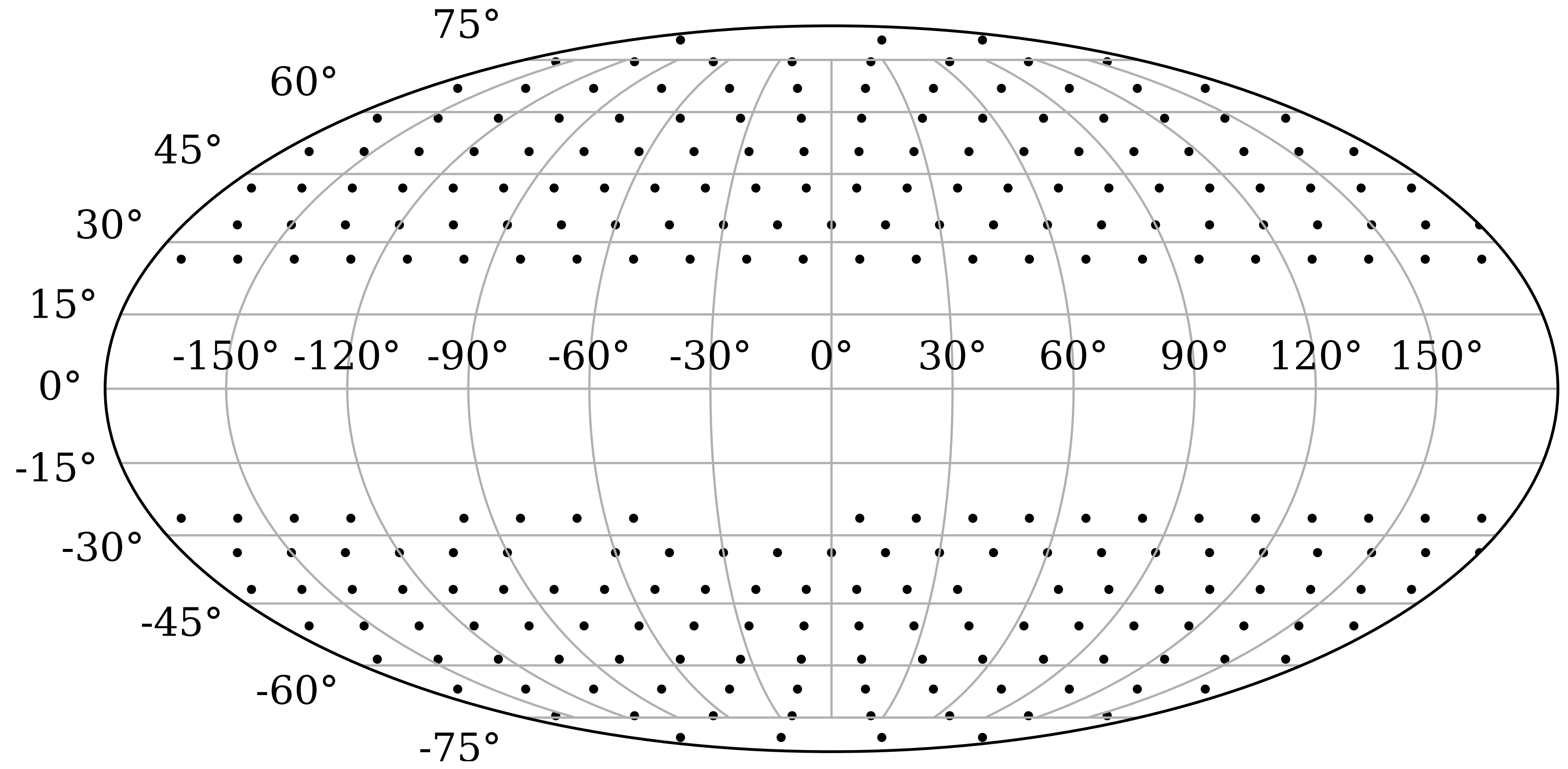


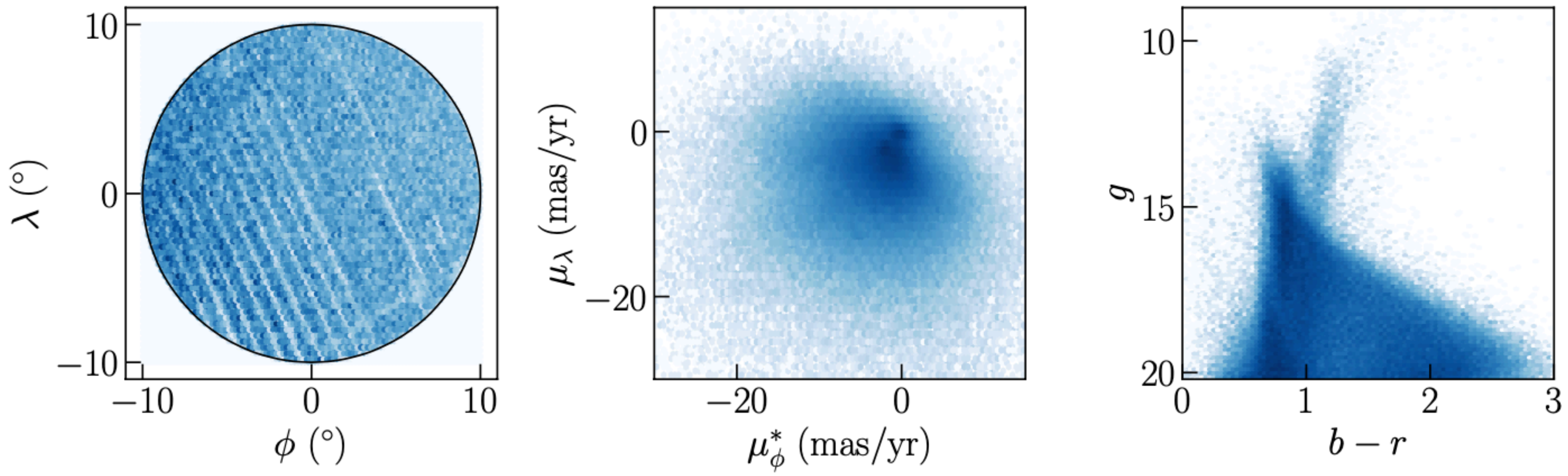
The End

Gaia Data

- We restrict ourselves to distant stars: parallax < 1 mas
- Available features: 2 angular positions, 2 proper motions, magnitude g, color $b - r$
- ANODE training times grow with number of stars, so we select *patches* of stars within 15° of centers that tile the sky, every star within 7° of a center.
  - Discontinuities in probability densities cause errors in the MAF density estimate. We train on the full patch and use fiducial region of inner 10° and $g < 20.2$
- Recenter the angular positions on patch center:

$$(\alpha, \delta, \mu_\alpha^*, \mu_\delta) \to (\phi, \lambda, \mu_\phi^*, \mu_\lambda)$$

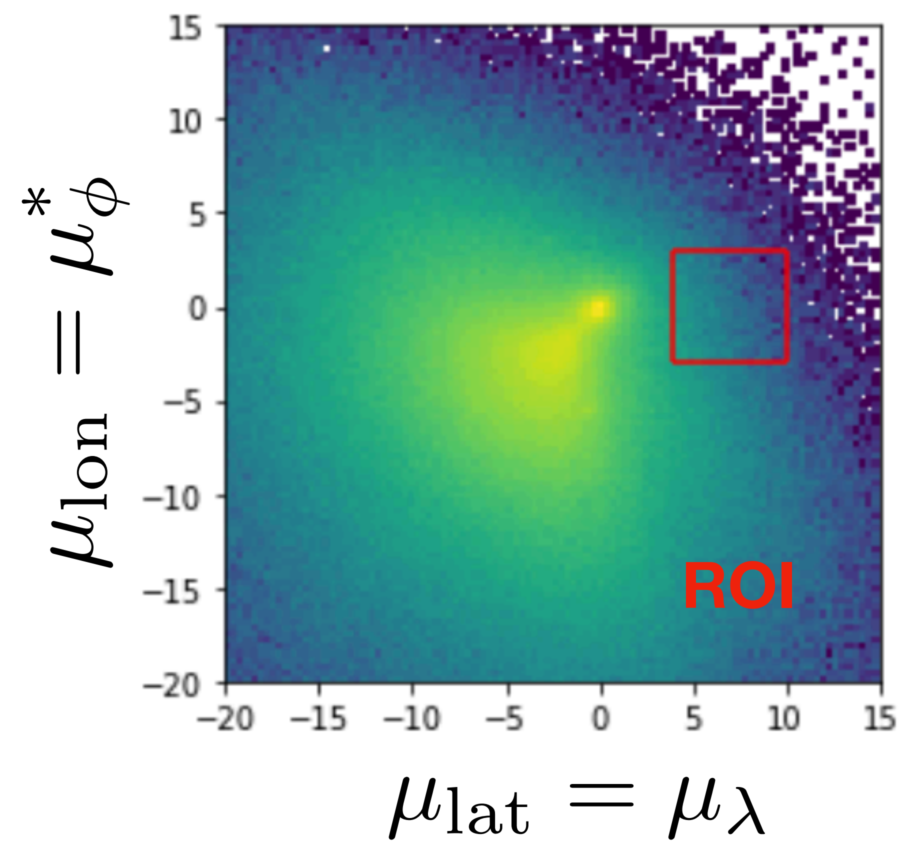




Regions of Interest (ROIs)

Instead, we needed additional cuts, in combination with a cut on R, to improve signal over background.

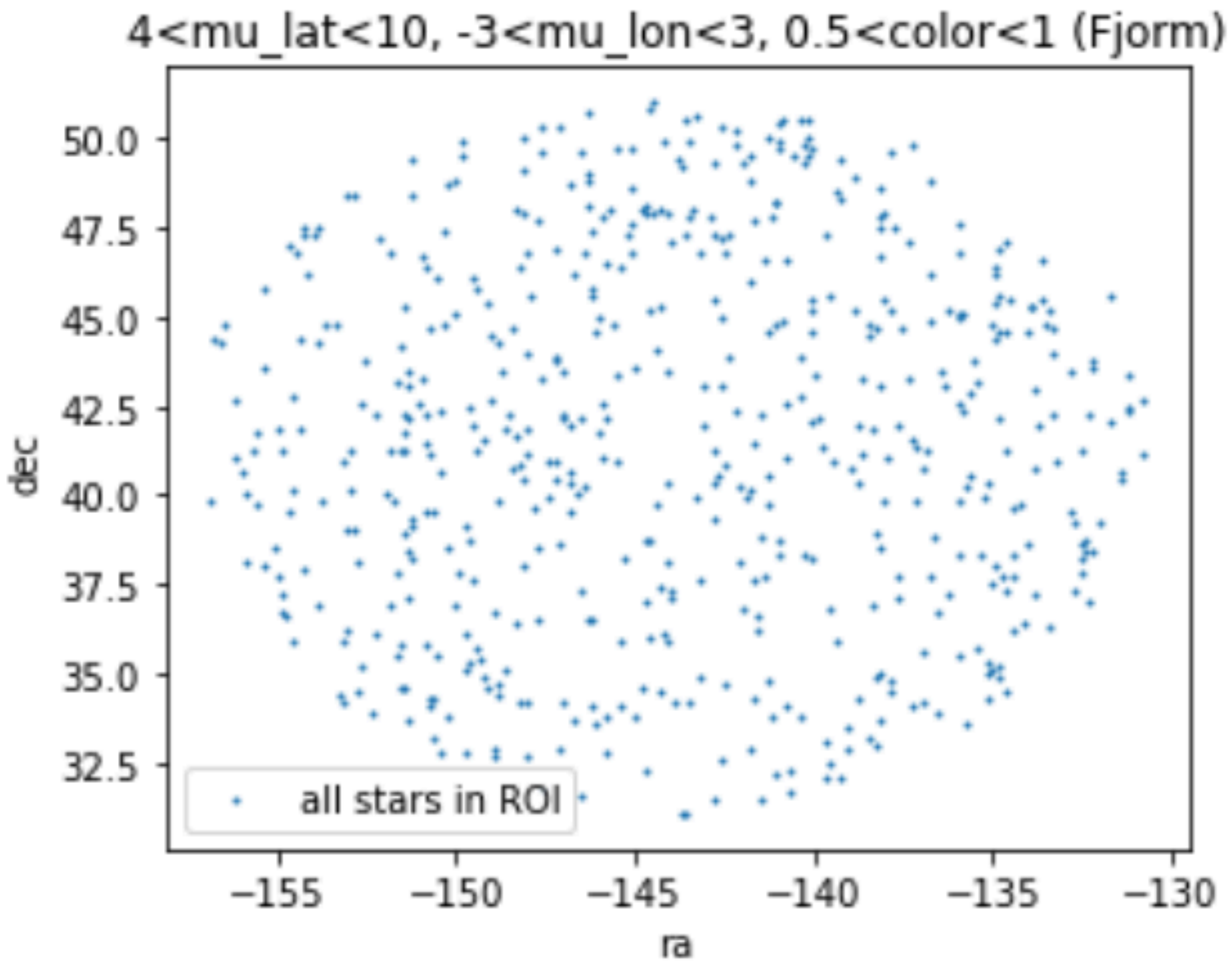
“Region of Interest”: window of width 6 in mu_lon (the other p.m. coord.) + color cut of 0.5<b-r<1

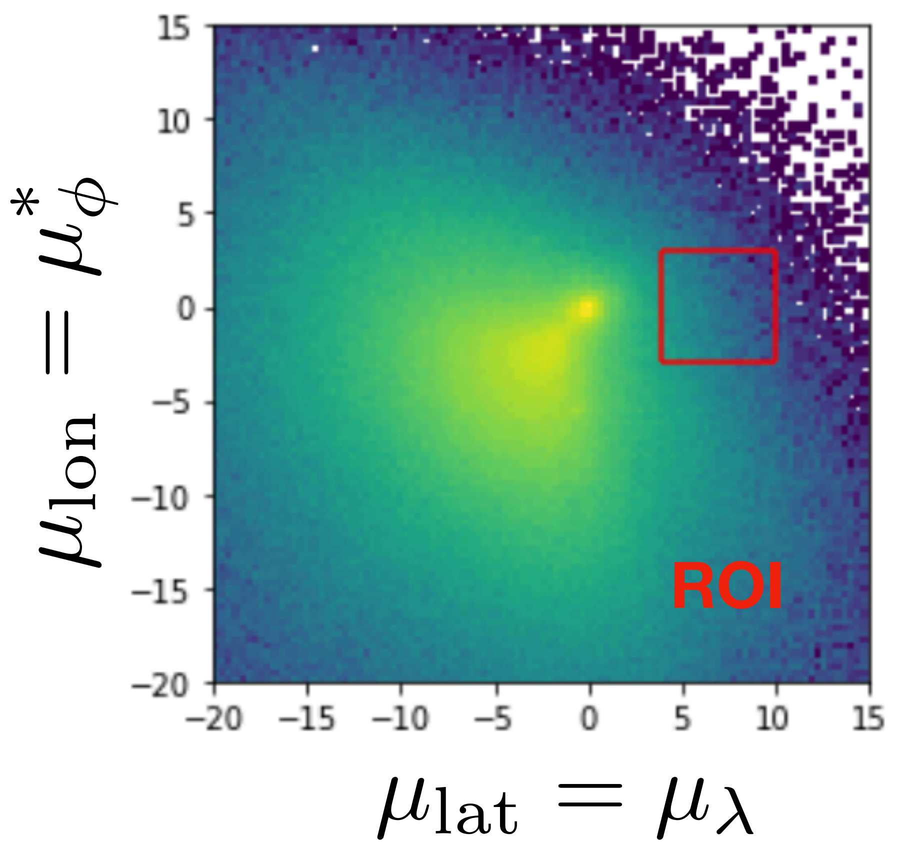


Regions of Interest (ROIs)

Instead, we needed additional cuts, in combination with a cut on R, to improve signal over background.

“Region of Interest”: window of width 6 in mu_lon (the other p.m. coord.) + color cut of 0.5<b-r<1

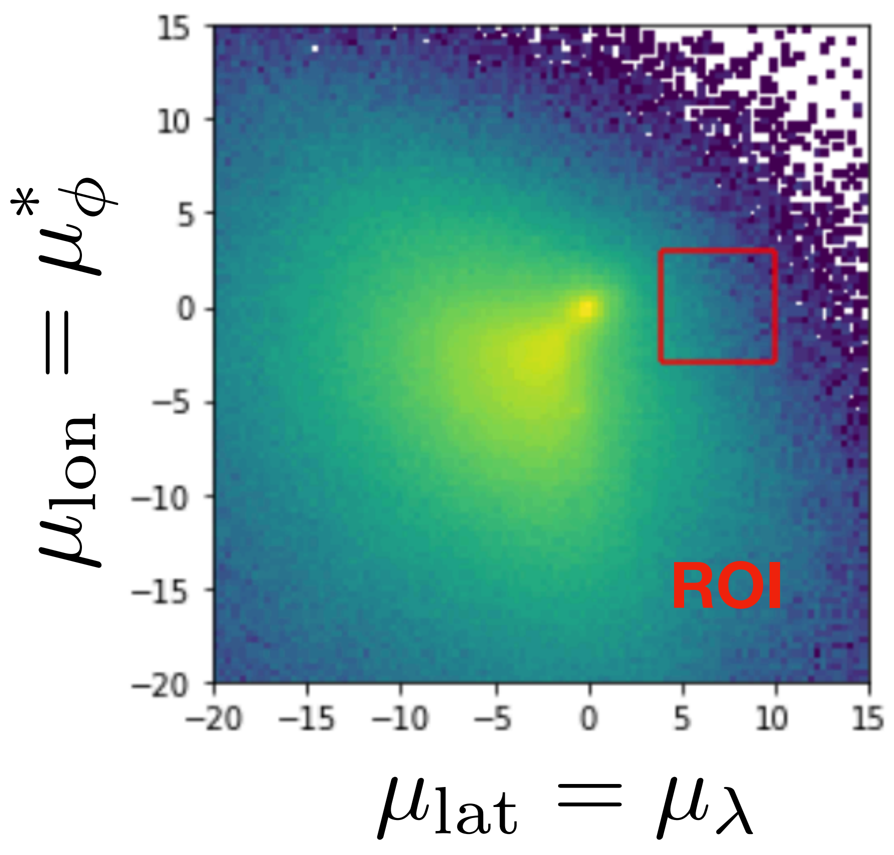


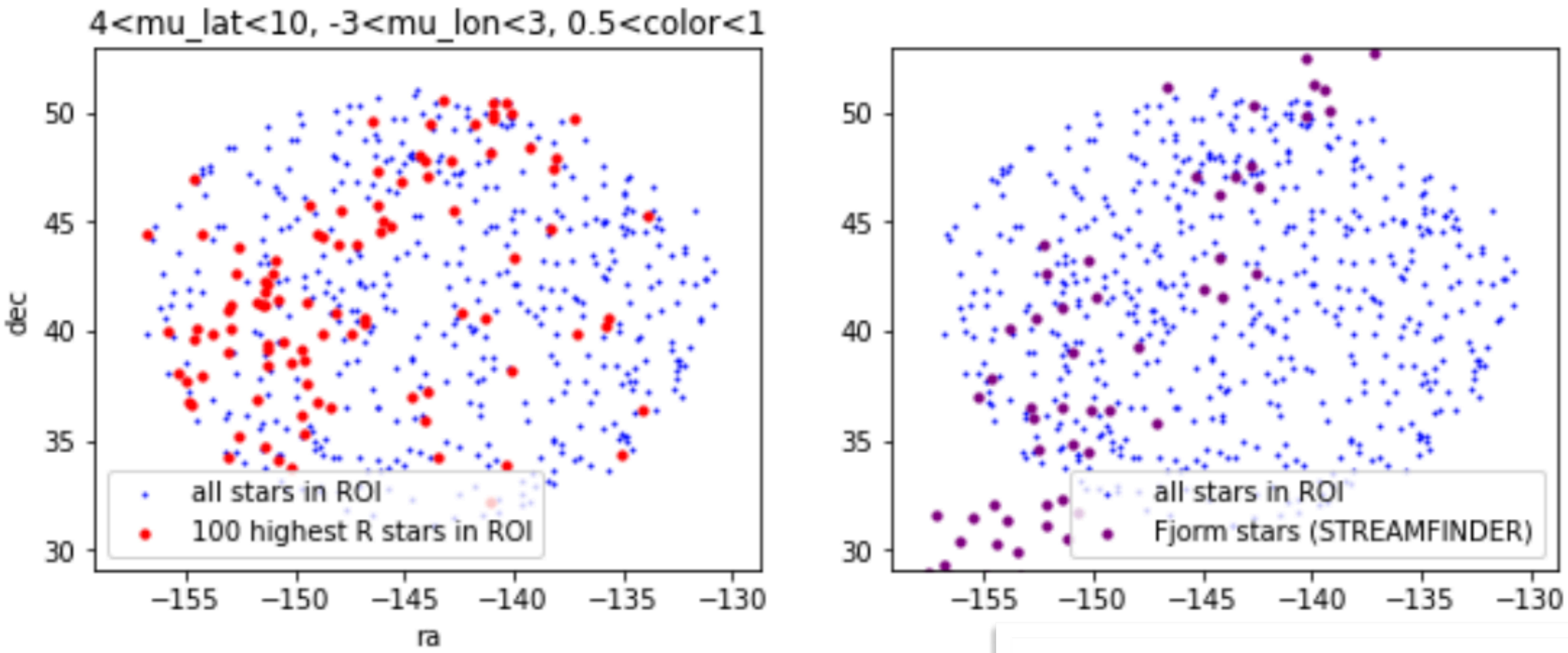


Regions of Interest (ROIs)

Instead, we needed additional cuts, in combination with a cut on R, to improve signal over background.

"Region of Interest": window of width 6 in mu_lon (the other p.m. coord.) + color cut of 0.5<b-r<1





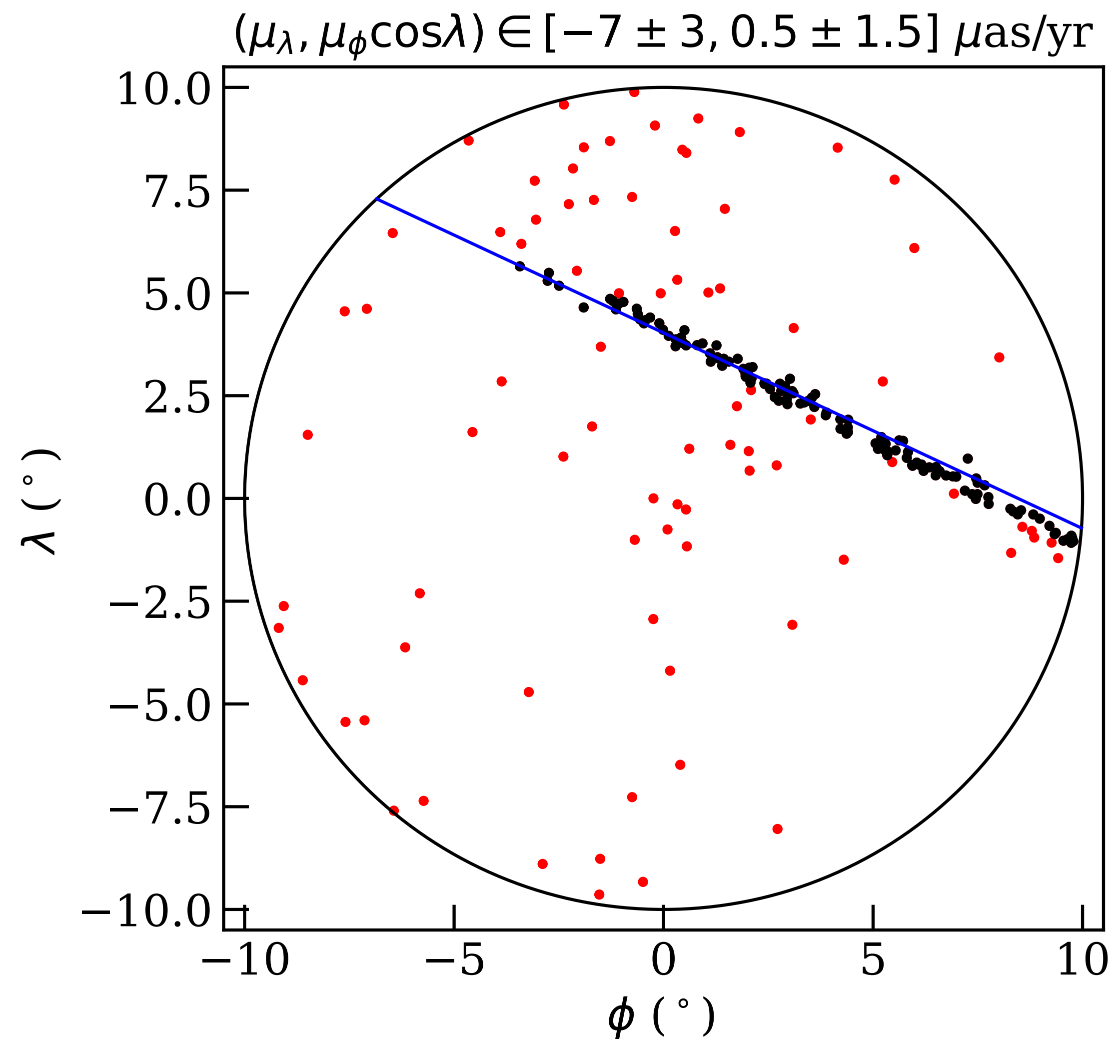
ROI + cut on R => stream can be found!

Hough transform for automated stream detection

There are 140,000 ROIs in the all-sky dataset. Need an automated method for stream detection!

Across a single patch, streams are likely to be line-like (but possibly wide).

Idea: use age-old ML technique (Hough transform, 60s-80s) to automate line detection.



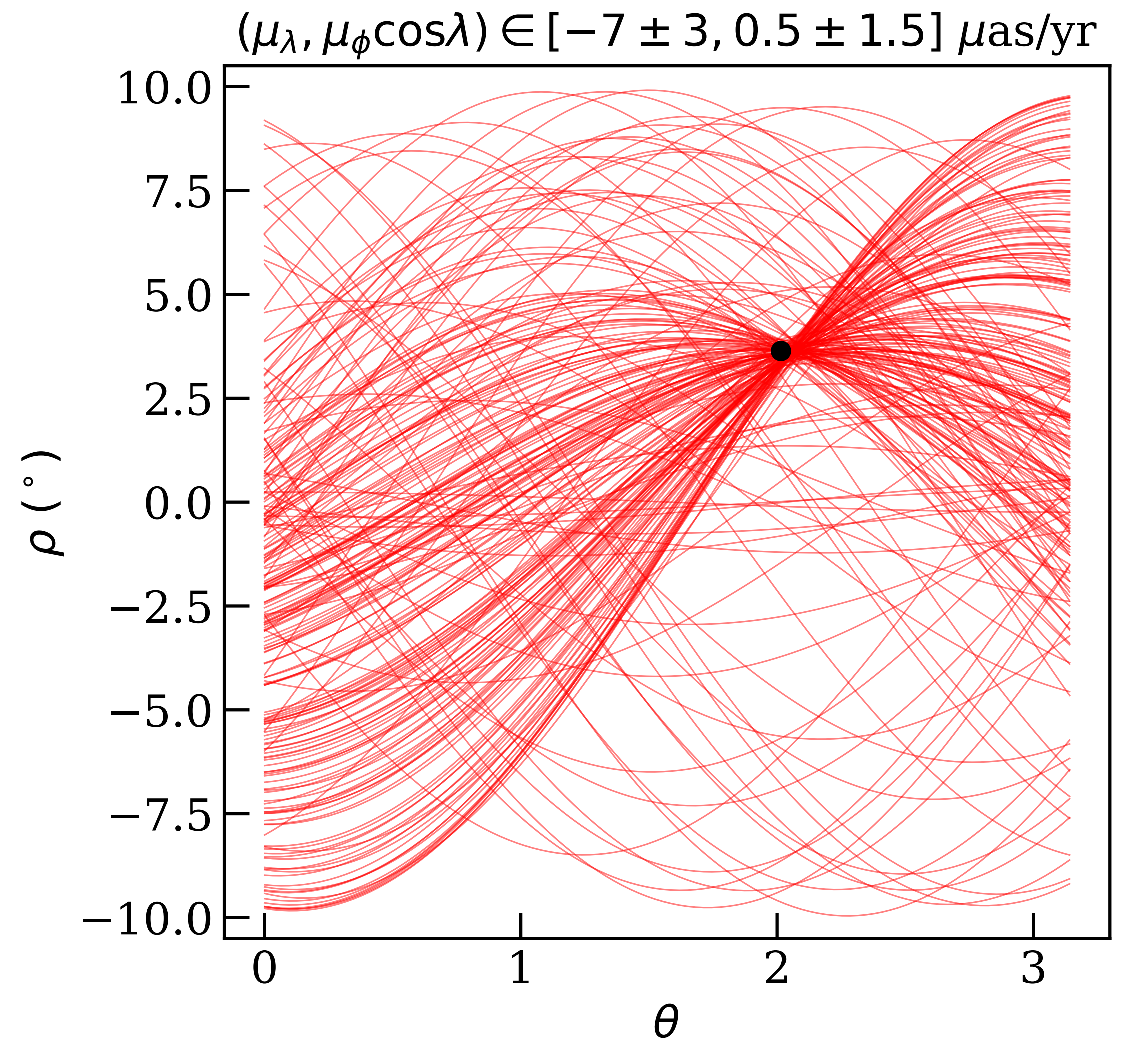
Hough transform:

- Each point in scatter plot seeds a family of lines that pass through it.
- Lines described by parameters (θ,ρ) that lie on a sine curve.

$$\rho = x \sin\theta - y \cos\theta$$

[usual slope/intercept parametrization leads to singularities]

- Significant line detection: many curves intersecting at same point in Hough space

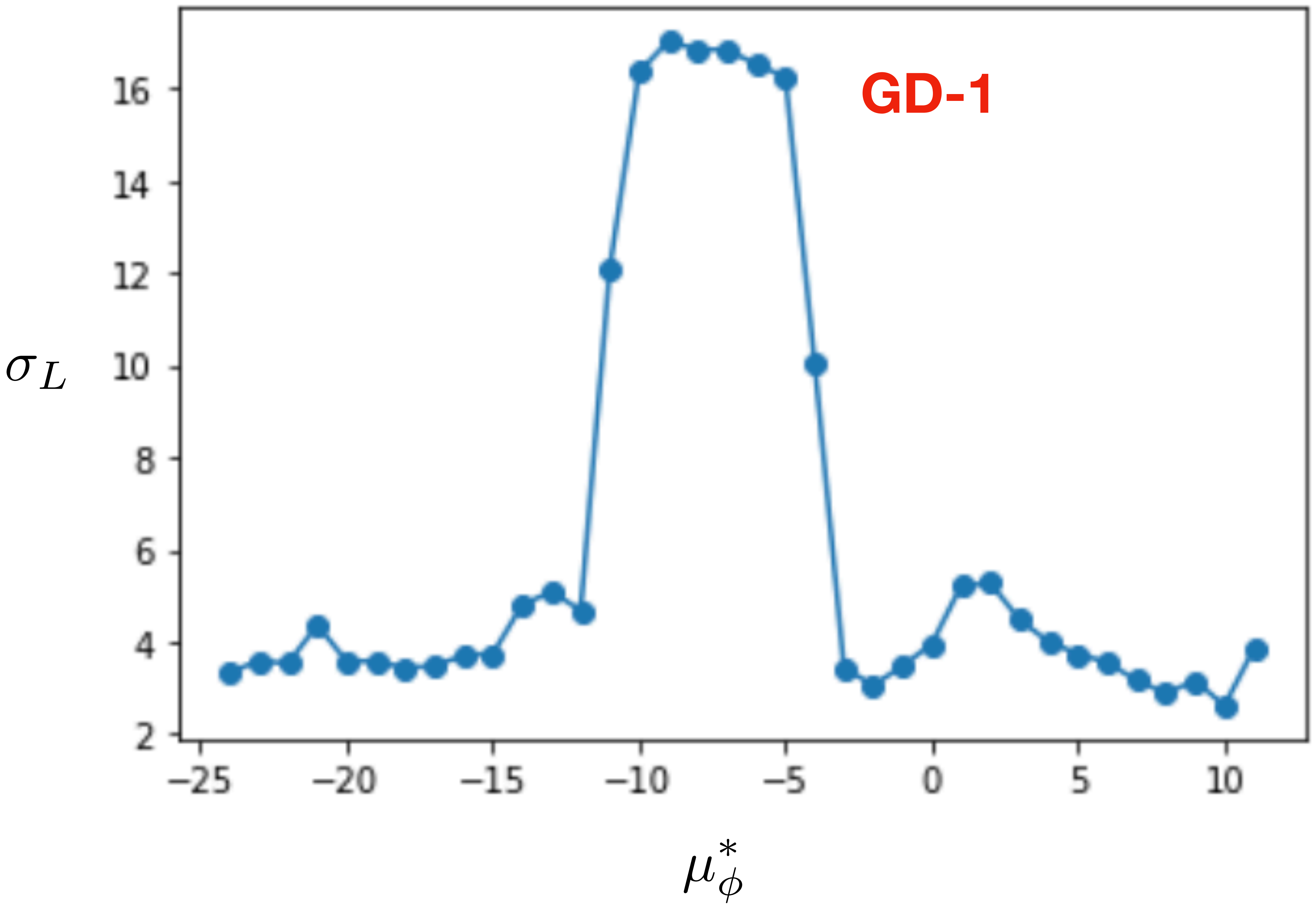


Define line significance using local contrast of curve density

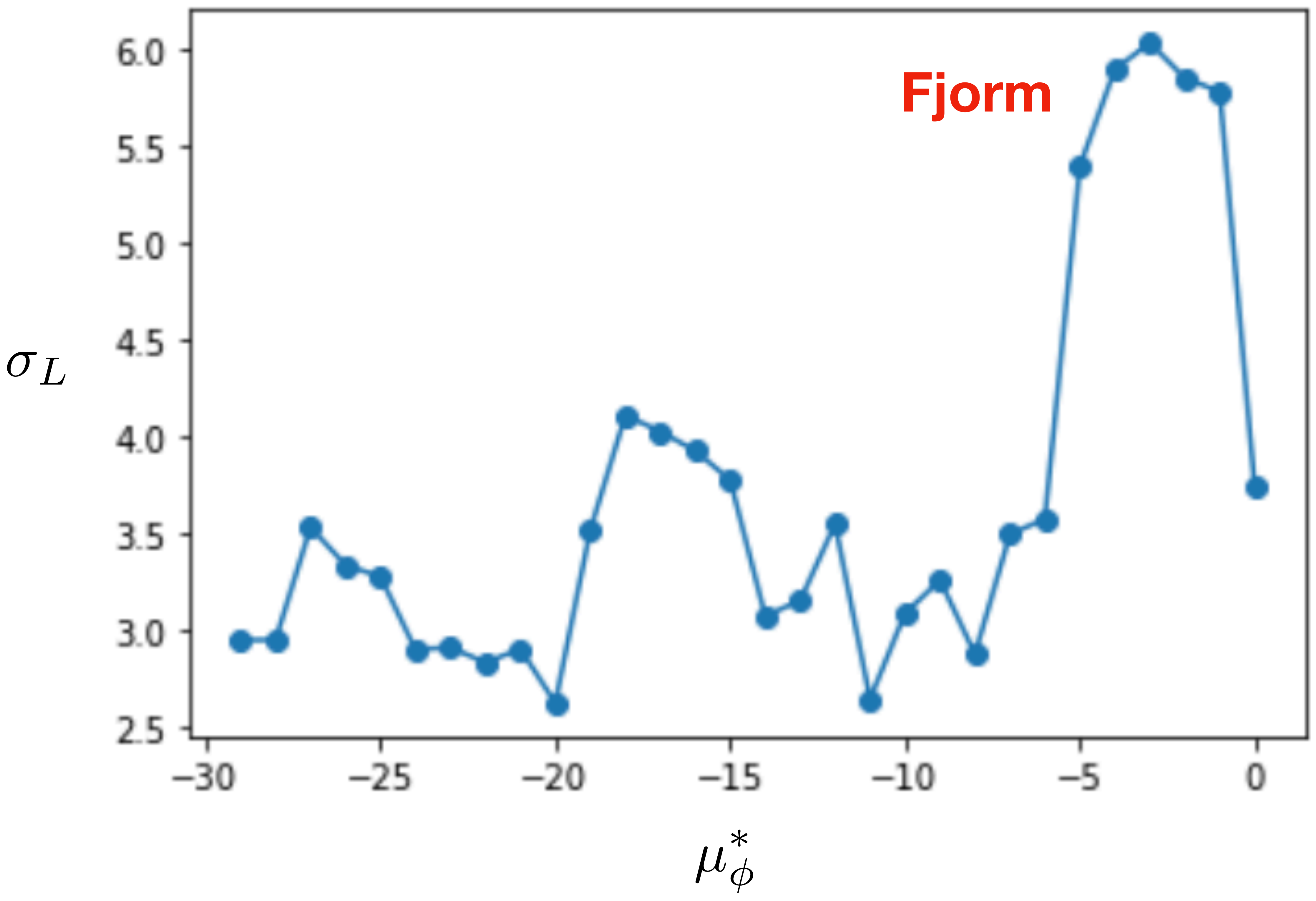
$$\sigma_L(\rho, \theta) = \frac{N(\rho,\theta) - \bar{N}(\rho,\theta)}{\sqrt{\bar{N}(\rho,\theta)}}$$

Scanning over ROIs

$(\alpha, \delta) = (148.6, 24.2), \quad -17 < \mu_\lambda < -11$



$(\alpha, \delta) = (216.0, 41.0), \quad 4 < \mu_\lambda < 10$



Clustering ROIs

- To reduce trials factor and cut down on false positives / strengthen case for stream detection, we require "same" line to be detected by 3 neighboring ROIs (mu_lat=x,x+1,x+2 and same mu_lon).
- Add individual line significances in quadrature to get combined line significance.

"proto-cluster"

Stream name	Significance	ra, dec	pmlat	pmlon
Gaia1	15.52	193.3, -4.5	[-25,-24,-23]	-18
Jhelum	16.83	351.4, -43.0	[-9,-8,-7]	4
Fjorm	10.38	216.0, 41.0	[3,4,5]	-1
Leiptr	11.75	71.1, -12.4	[-16,-15,-14]	8
Svol	6.87	227.6, 23.3	[-9,-8,-7]	1
Fimbulthul	7.20	196.5, -20.9	[-14,-13,-12]	-20
Gaia3/Ylgr	14.61	173.3, -17.2	[-12,-11,-10]	-5
Sylgr	6.73	167.5, -4.2	[-21,-20,-19]	-22
Slidr	8.61	171.4, 3.1	[-10,-9,-8]	-24
GD1	29.1	148.6, 24.2	[-18,-17,-16]	-9

(High success rate — these are nearly all the previously found streams contained in our fiducial dataset!)

$$\sigma_L^{combined} > 6.7$$

This cut on combined line significance would capture all the previously-found streams on this list

Merging protoclusters

- 1755 protoclusters (out of 140,000) after the 6.7sigma cut on combined line significance.
- Merge the protoclusters that “agree” in position and velocity space (neighboring patches of the sky; concordant line parameters and proper motions)
- Final result: **590 clustered stream candidates**