Via Machinae

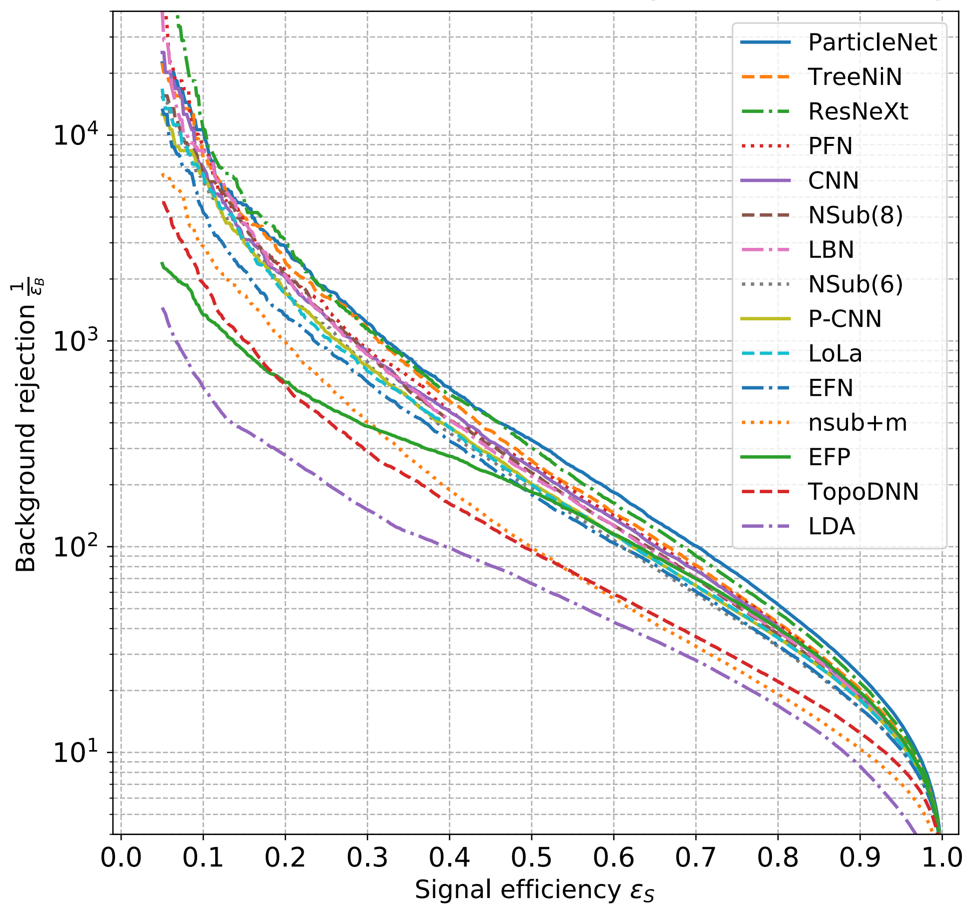
Discovering Stellar Streams using Machine Learning

Matthew Buckley, on behalf of David Shih, Lina Necib, John Tamanas
arXiv:2104.12789

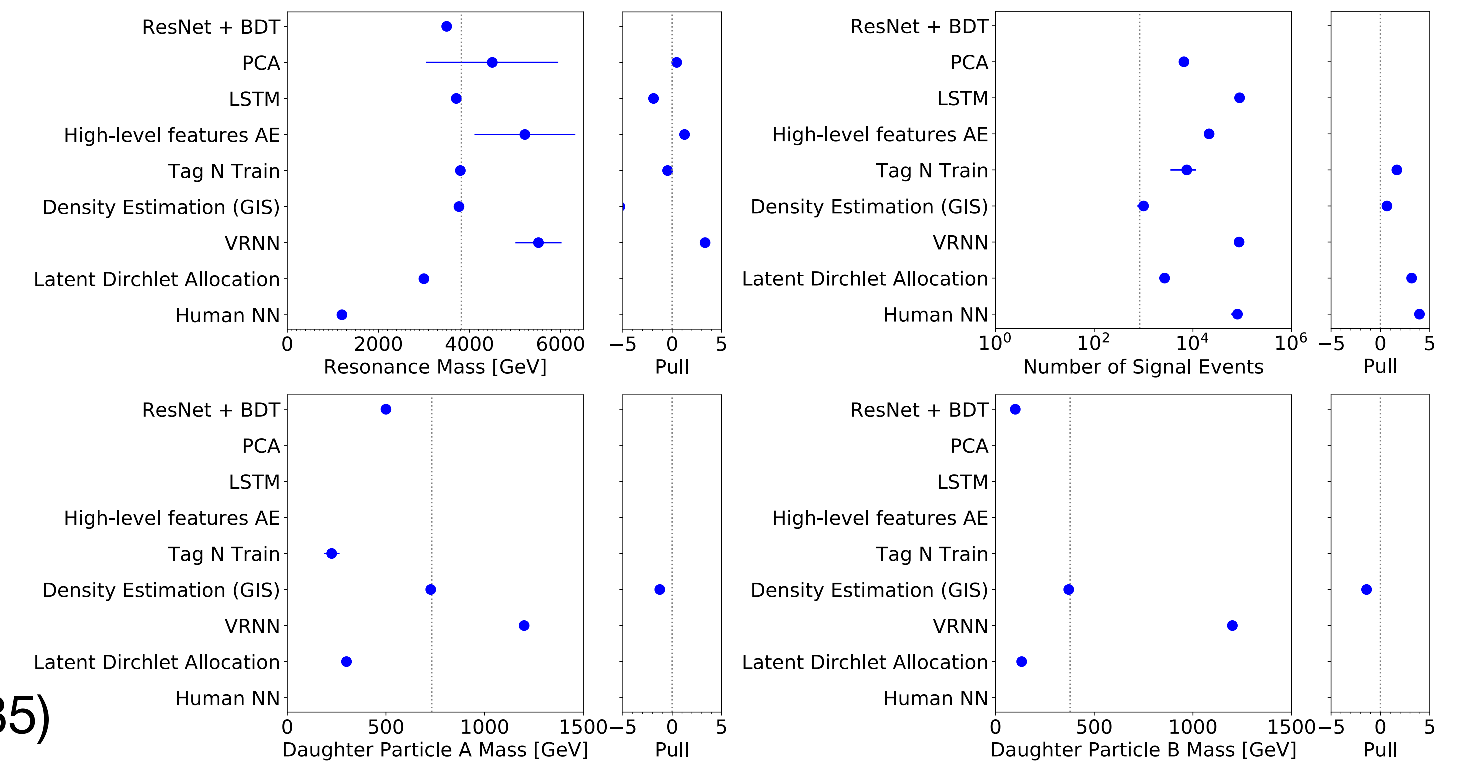
Machine Learning out of the Colliders

- Phenomenology/high-energy experiment has been an early adopter of machine learning.
  - We have vast, complicated datasets, within which is buried some small signal.
  - Forced us to innovate to find new ways to classify events, identify anomalies, and generate simulated data.
- Astrophysics is also in an era of Big Data
  - Facing similar issues as pheno, but with some new twists.
  - What techniques can we transfer over to this related field?
  - What can we learn from solving astrophysical problems?

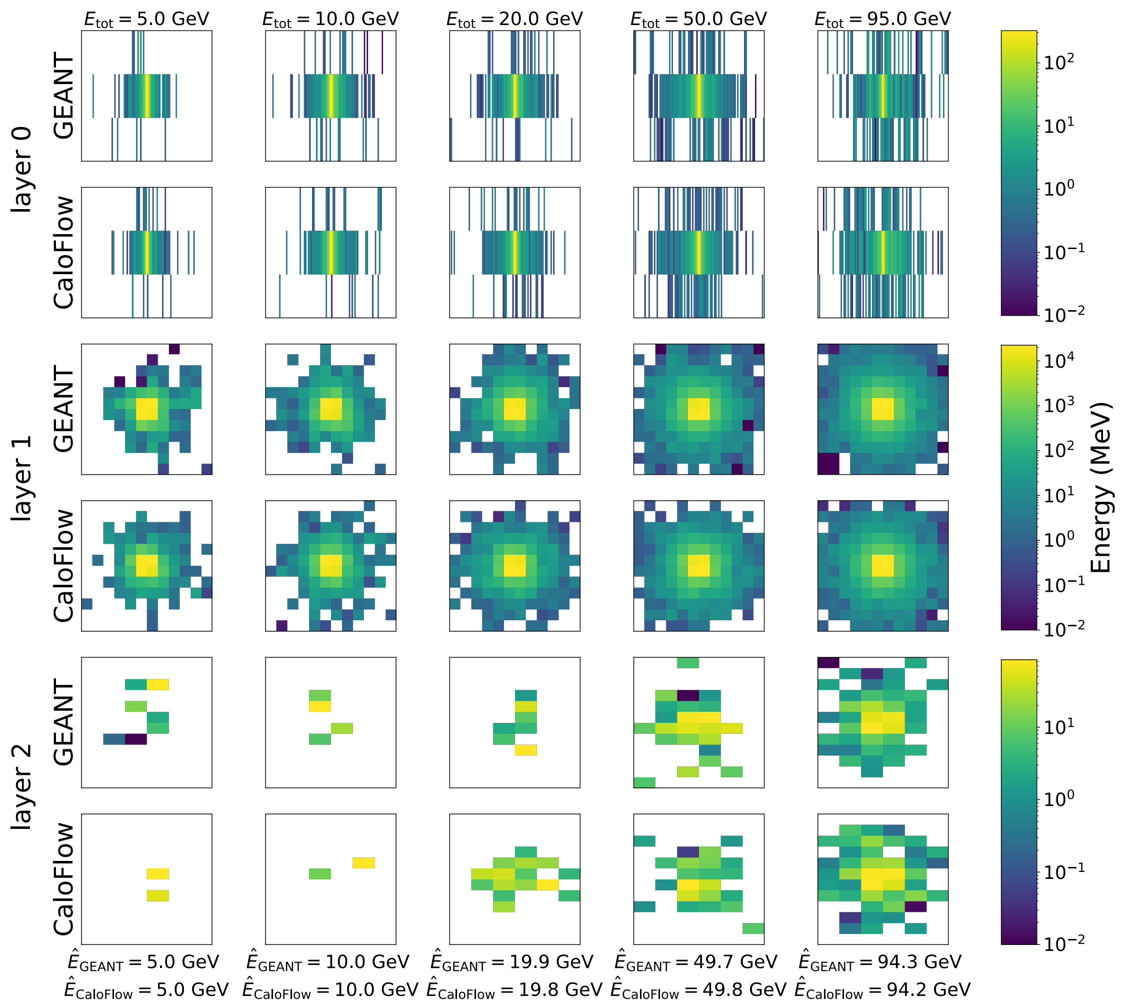
Classification (1902.09914)



Anomaly Detection (2101.08320)



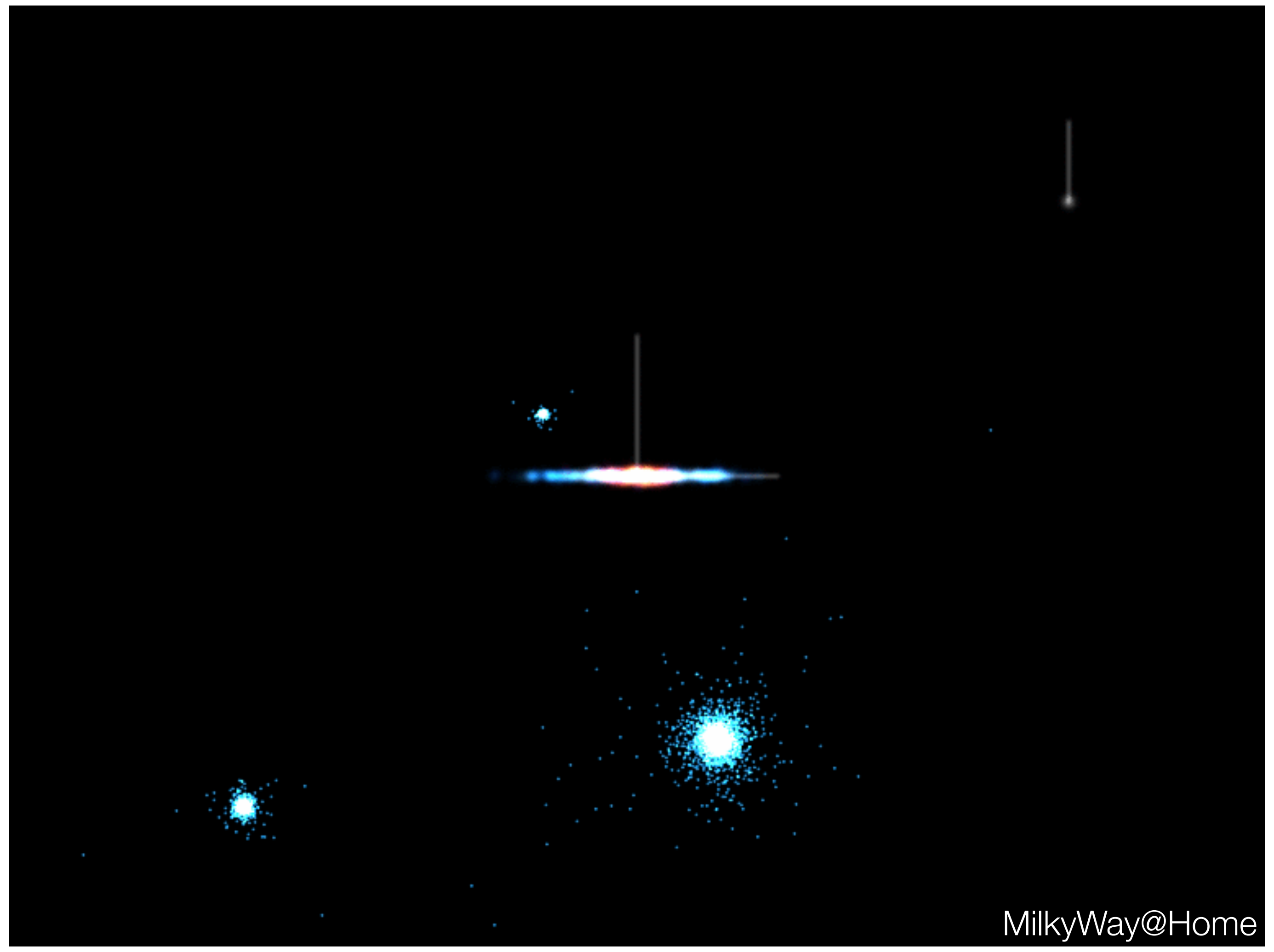
Simulation (2106.05285)

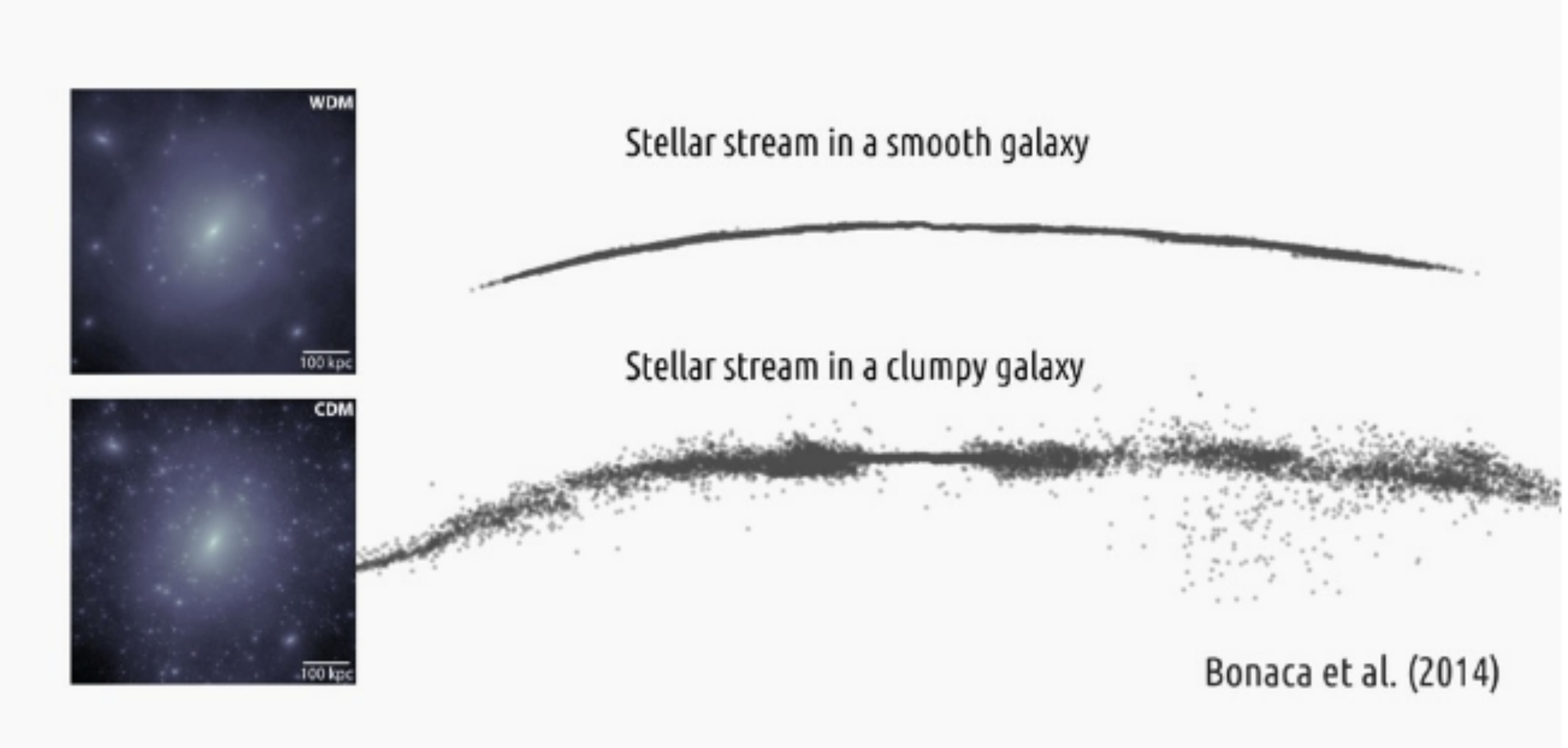


…and many more…

Stellar Streams

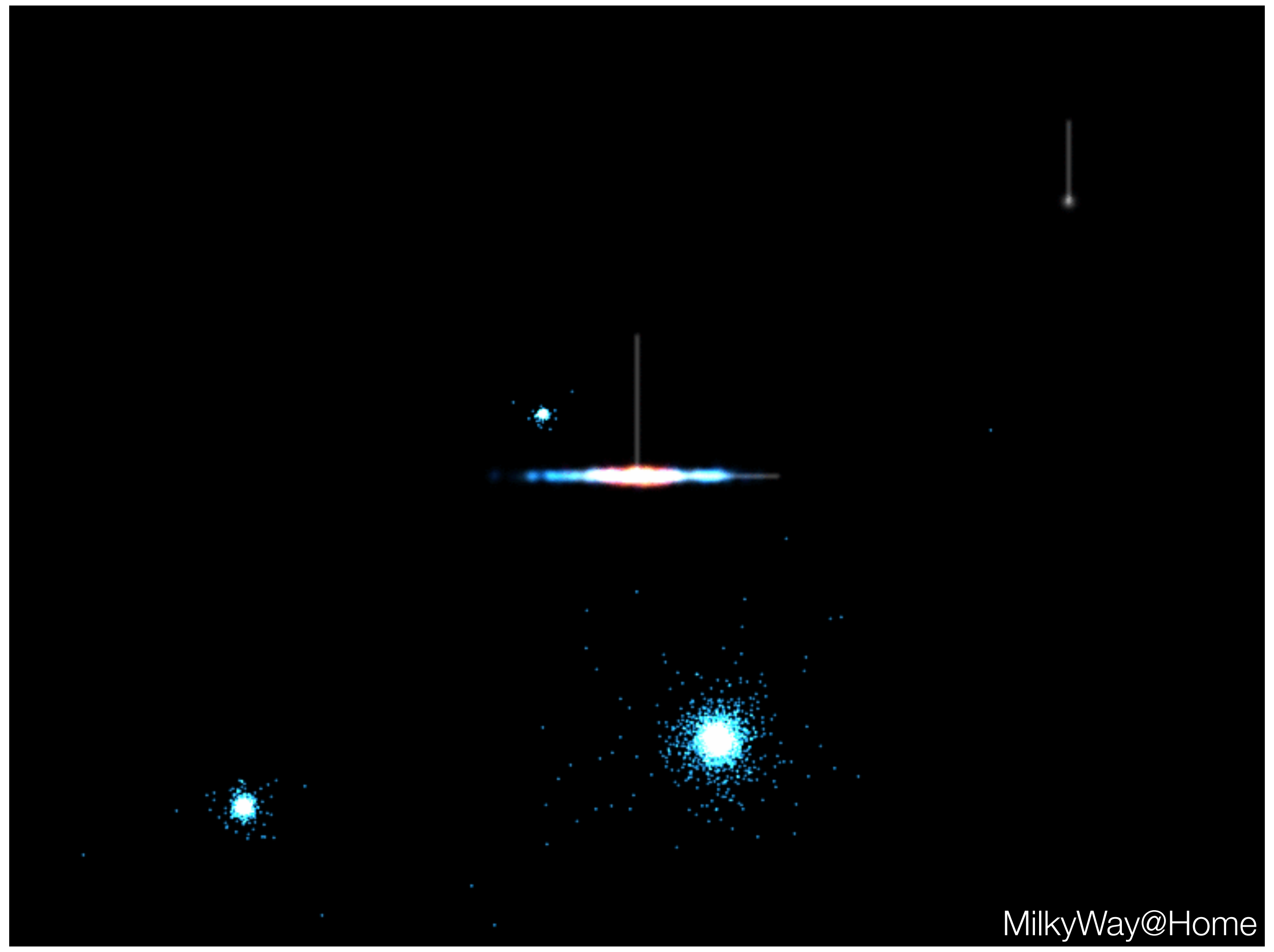
- The Milky Way is built from the merger of smaller objects.
- Compact collections of stars (dwarf galaxies & globular clusters) get tidally stripped during infall and form **stellar streams** before becoming well-mixed with the halo.
  - Provide a probe into the Galactic potential through the stream's orbit.
  - Give a glimpse into the Galaxy's merger history.
  - Can reveal dark matter substructure through gravitational interactions with the stream itself.

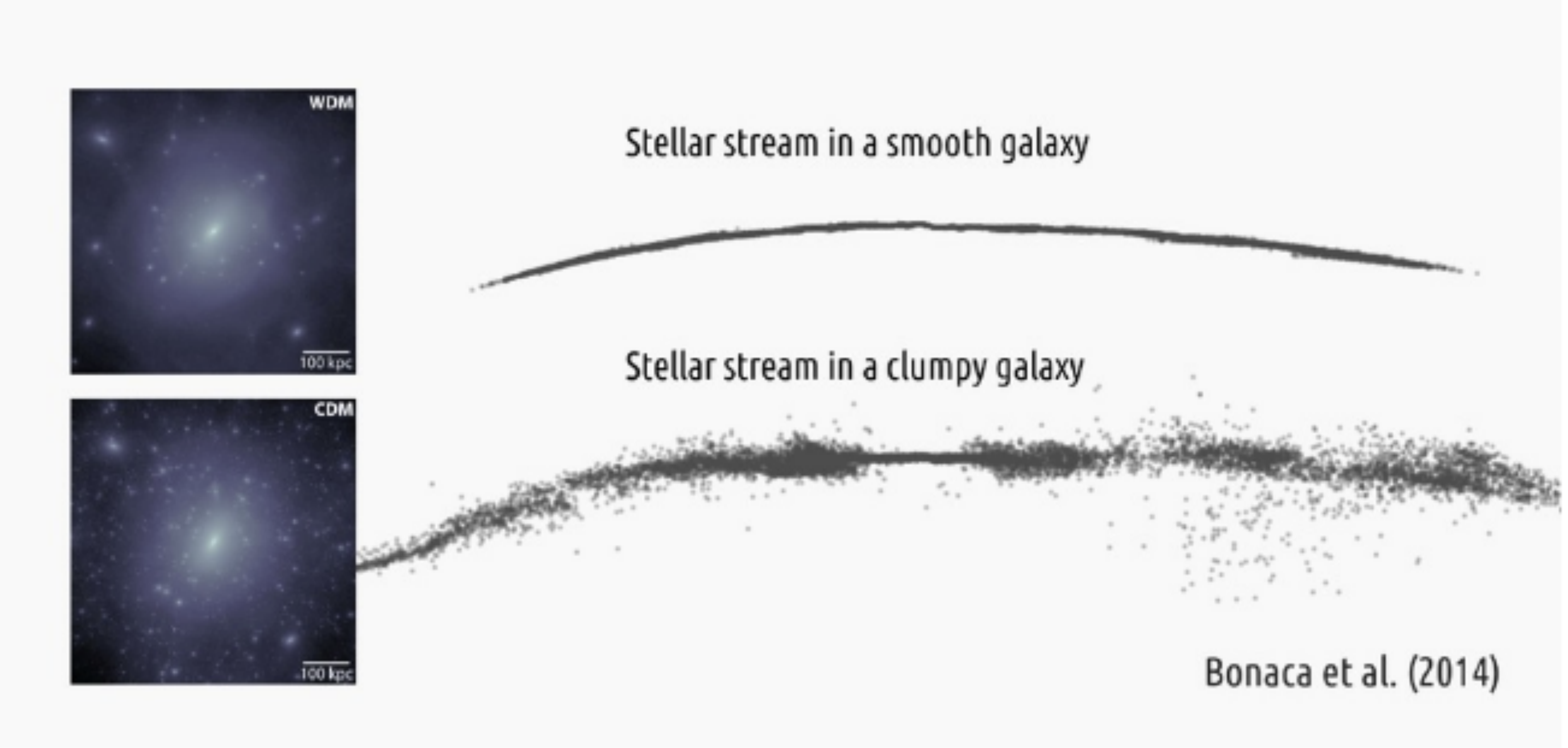




Stellar Streams

- The Milky Way is built from the merger of smaller objects.
- Compact collections of stars (dwarf galaxies & globular clusters) get tidally stripped during infall and form **stellar streams** before becoming well-mixed with the halo.
  - Provide a probe into the Galactic potential through the stream's orbit.
  - Give a glimpse into the Galaxy's merger history.
  - Can reveal dark matter substructure through gravitational interactions with the stream itself.

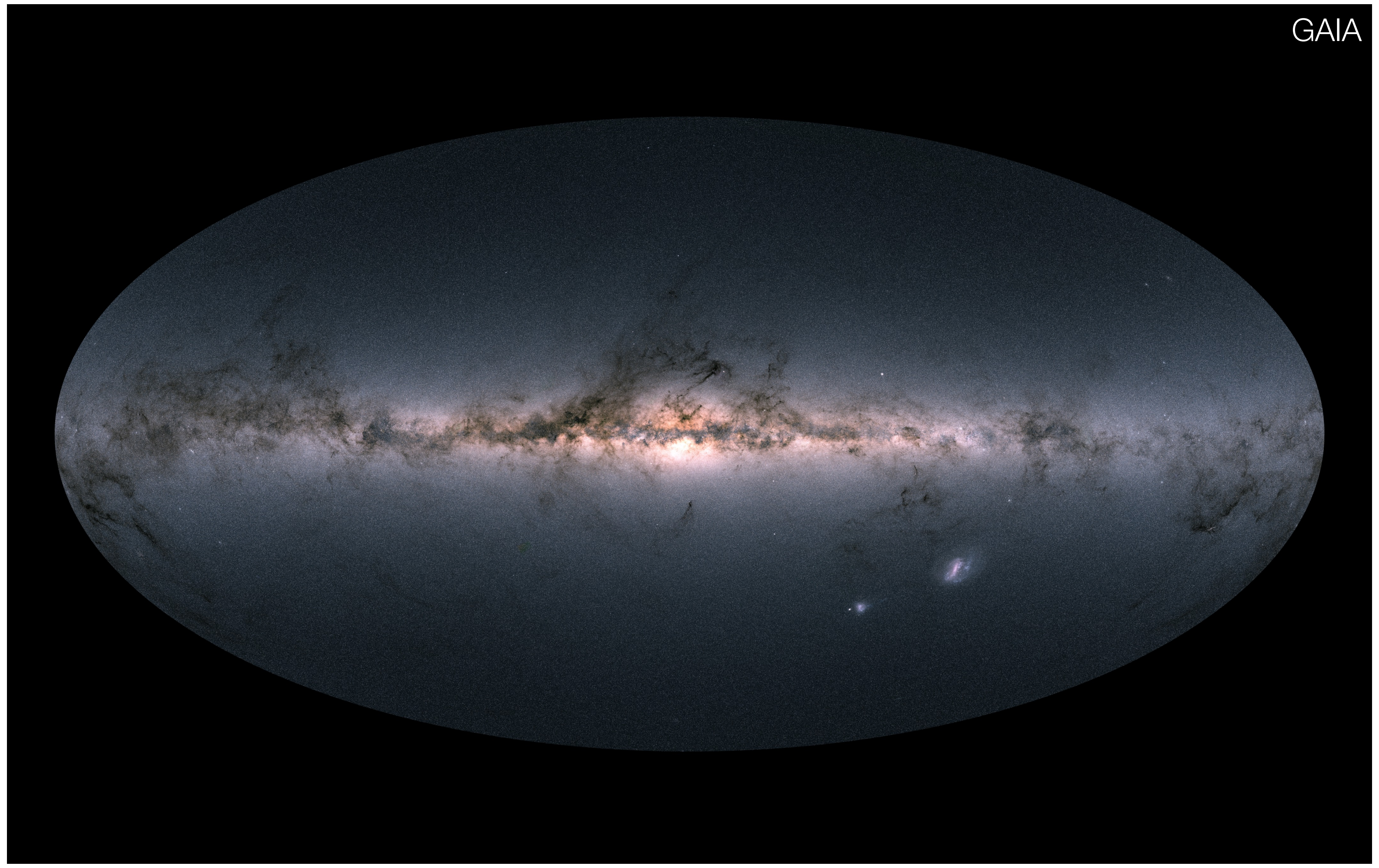




Gaia

- Gaia satellite measures the positions and proper motions of ~billion stars in the Galaxy.
  - Provides *photometry* (color and magnitude) but not *spectroscopy*
  - Accurate parallax distances for ~150 million stars
  - Line-of-sight motion for ~7 million stars

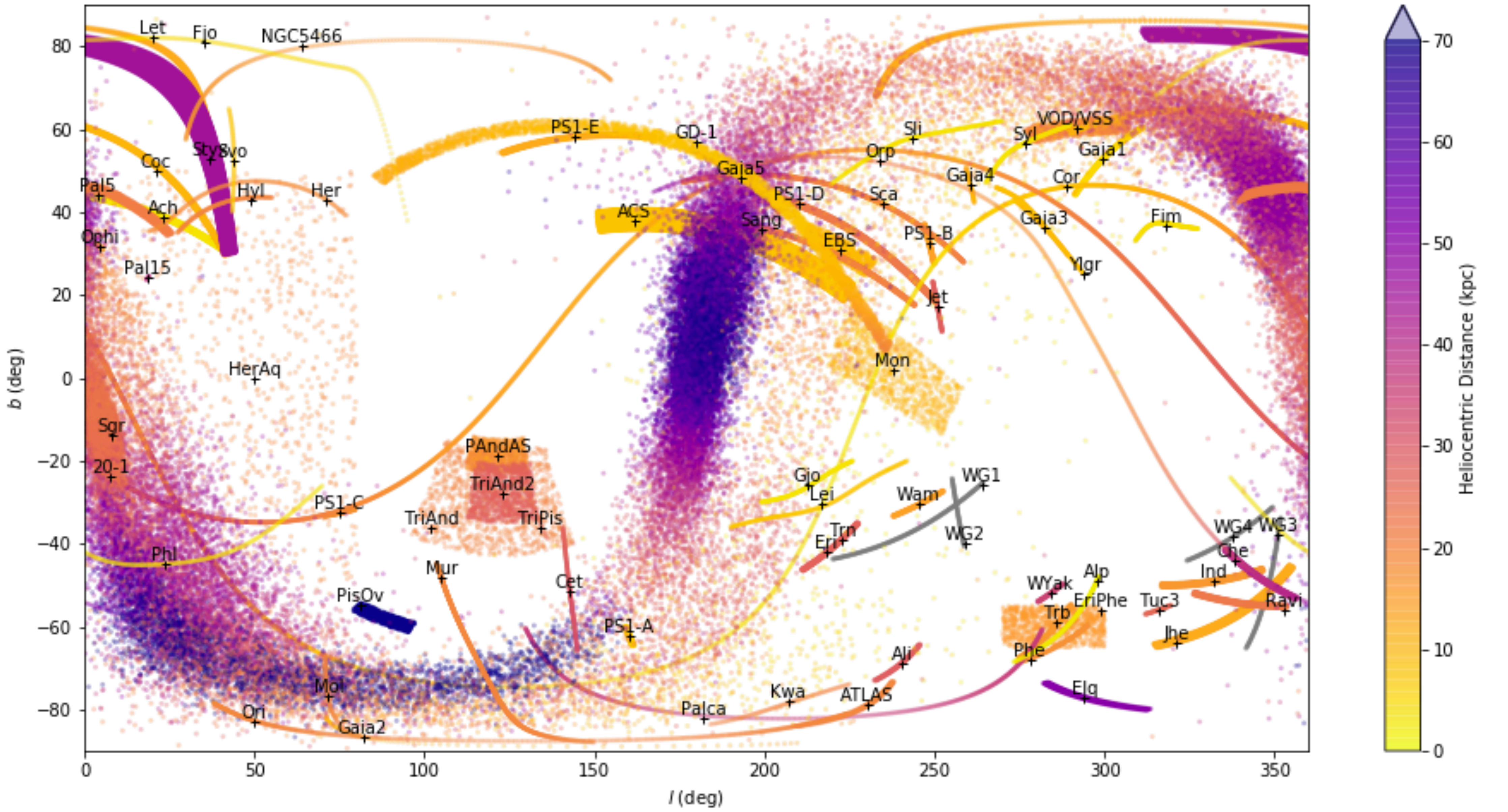
- A huge mine of data for the study of Galactic substructure: including stream-finding.



Previous Approaches

- Some streams can be found by eye, or through other surveys (DES, SDSS) and reconfirmed in Gaia.
- Automated algorithms for Gaia data exist (e.g. STREAMFINDER Malhan et al 2018). Makes assumptions about the composition of stream stars and the Galactic potential.
- Our goal: a stream-finding algorithm that:
  - Uses only Gaia data
  - Does not assume a Galactic potential or orbit
  - Does not assume stream stars lie on a particular isochrone.
  - Use the fact that streams are compact in proper motion space.

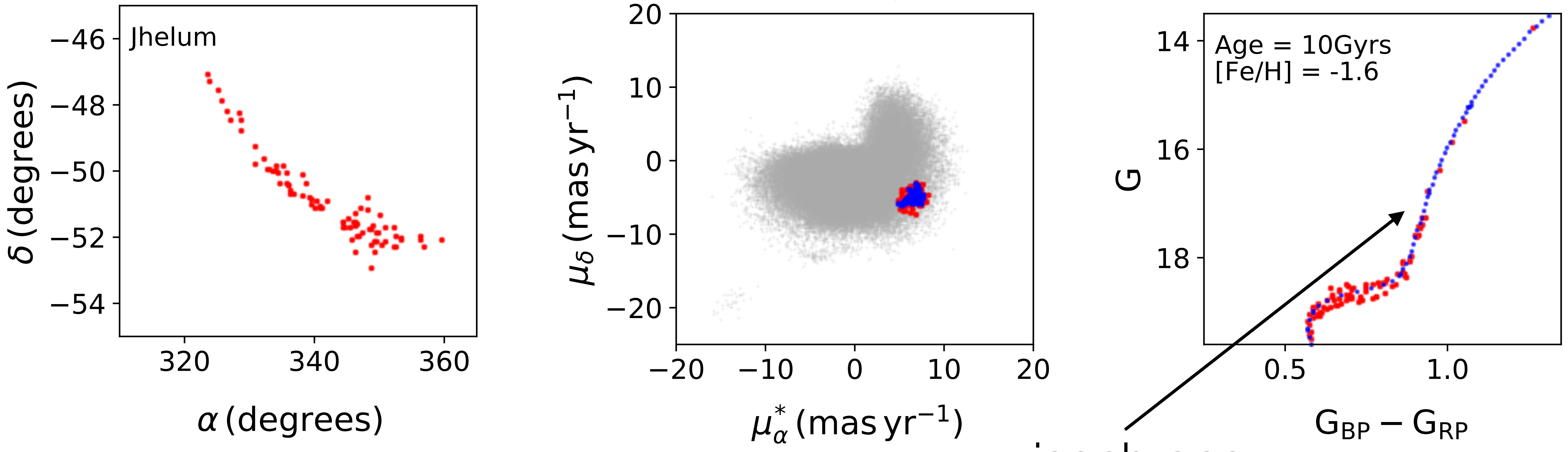
github.com/cmateu/galstreams



Angular position on sky

Angular motion on sky

Stellar brightness and color



isochrone

Malhan et al 2018

Anomaly Detection

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

"features"
(position, velocity, color, etc)

one "feature"
to condition on

- The signal and background probability distributions are different:

$$P(\vec{x}|m) = \alpha P_{\rm sig}(\vec{x}|m) + (1-\alpha) P_{\rm bkg}(\vec{x}|m)$$

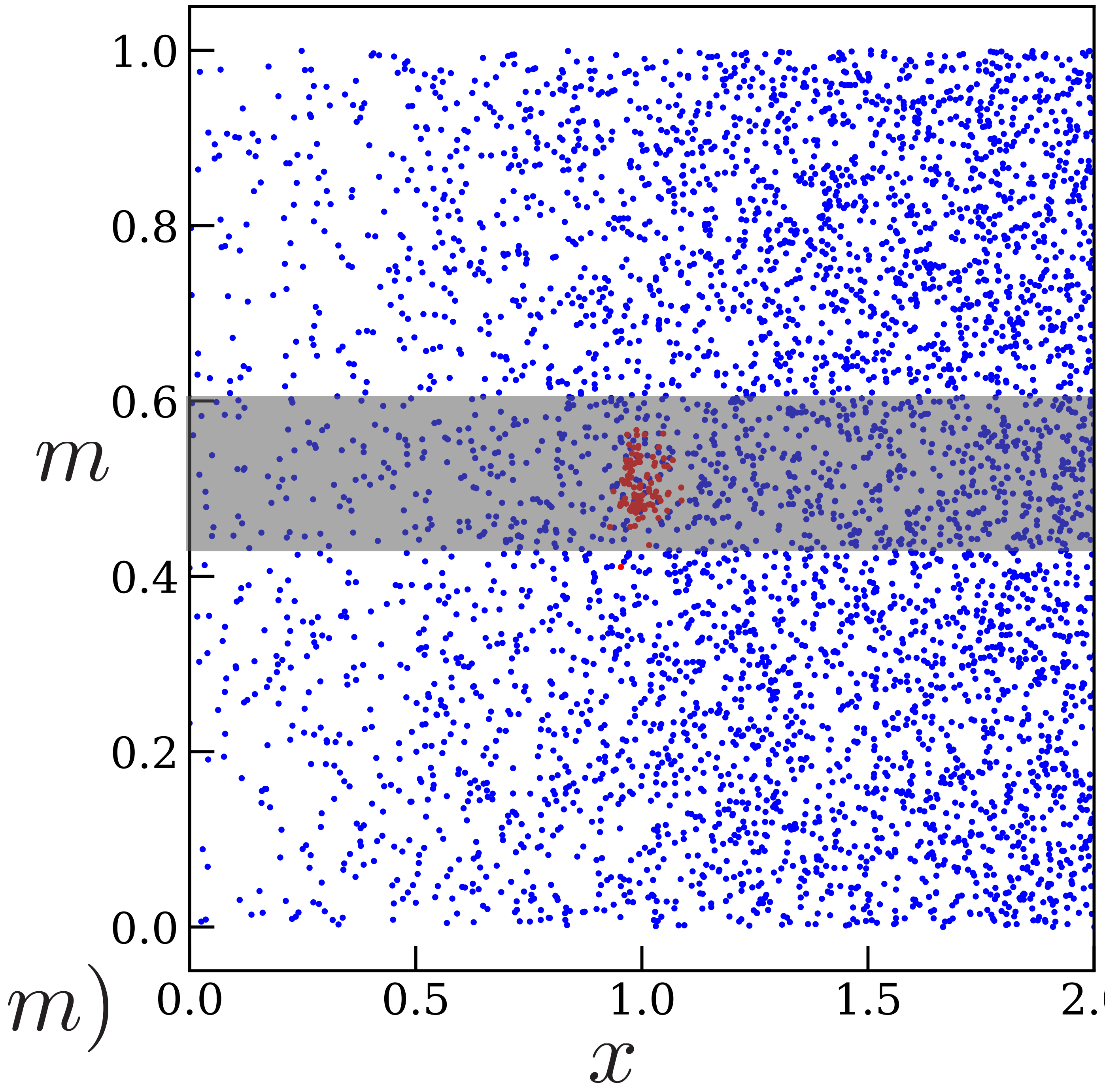
- The optimal parameter for distinguishing signal from background is the ratio

$$R(\vec{x}|m) = \frac{P(\vec{x}|m)}{P_{\rm bkg}(\vec{x}|m)}$$

  - Signal dominates wherever $R(\vec{x}|m) > 1$.
- The problem: How do we determine both $P(\vec{x}|m)$ and $P_{\rm bkg}(\vec{x}|m)$? Especially in something as complicated as the Galaxy.

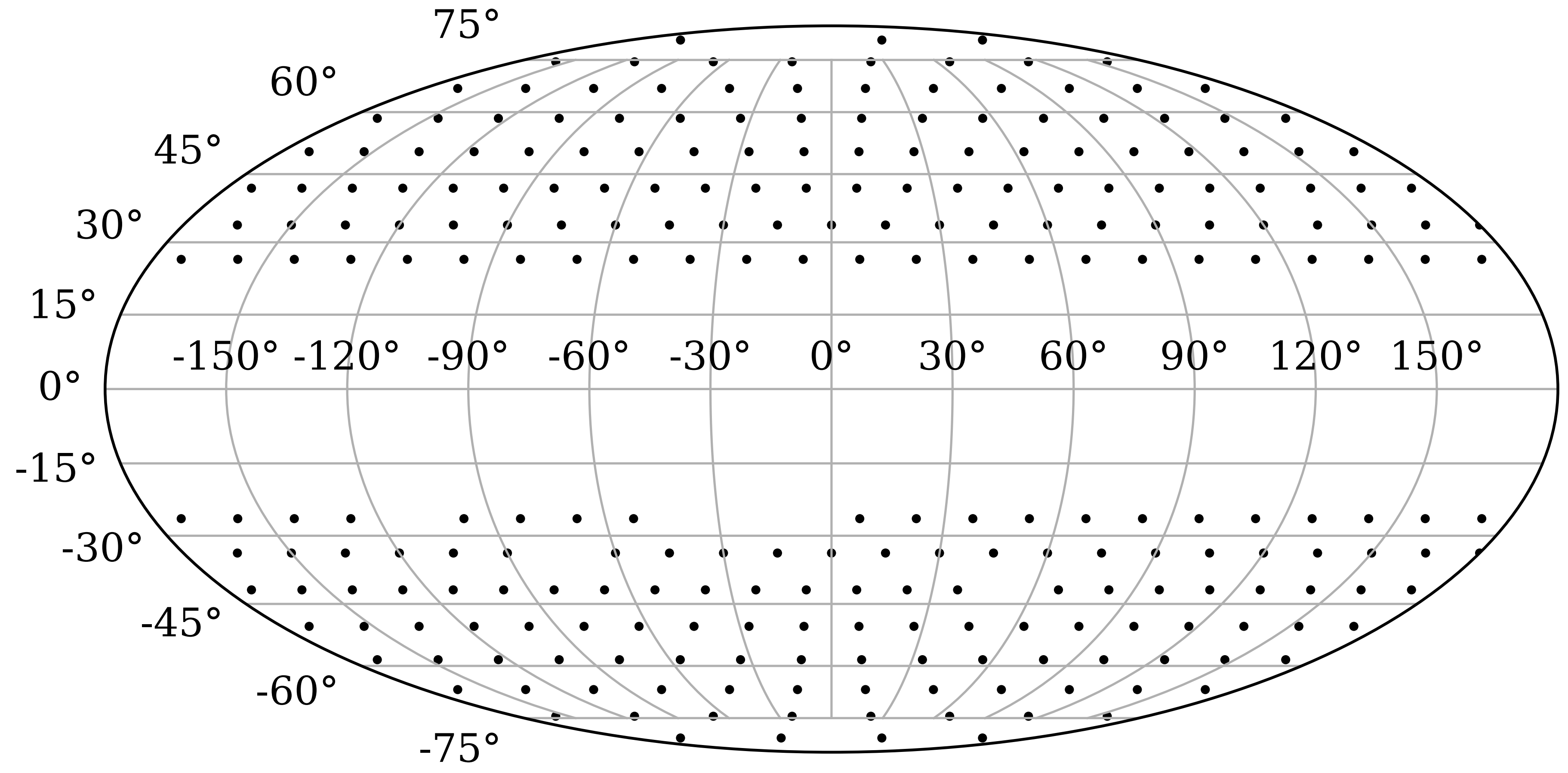
ANODE

- Unsupervised Deep Learning offering new approaches to modeling probability distributions.
  - Normalizing flows: transform from a known distribution (multivariate Gaussians) to the target distribution (the data) through invertible functions.
- ANODE (Nachman and Shih, 2020) uses *Masked Autoregressive Flows* (MAF), which learns a target distribution conditioned on one feature dimension.
  - Learn the probability distribution with $m \in [m_0 \pm \frac{\Delta m}{2}]$ in two ways:
    - 1st by training directly on the data in the region $\approx P(\vec{x}|m)$
    - 2nd by training outside this region, then interpolating in $\approx P_{\rm bkg}(\vec{x}|m)$
  - Allows direct estimation of the ratio *R* inside this region.



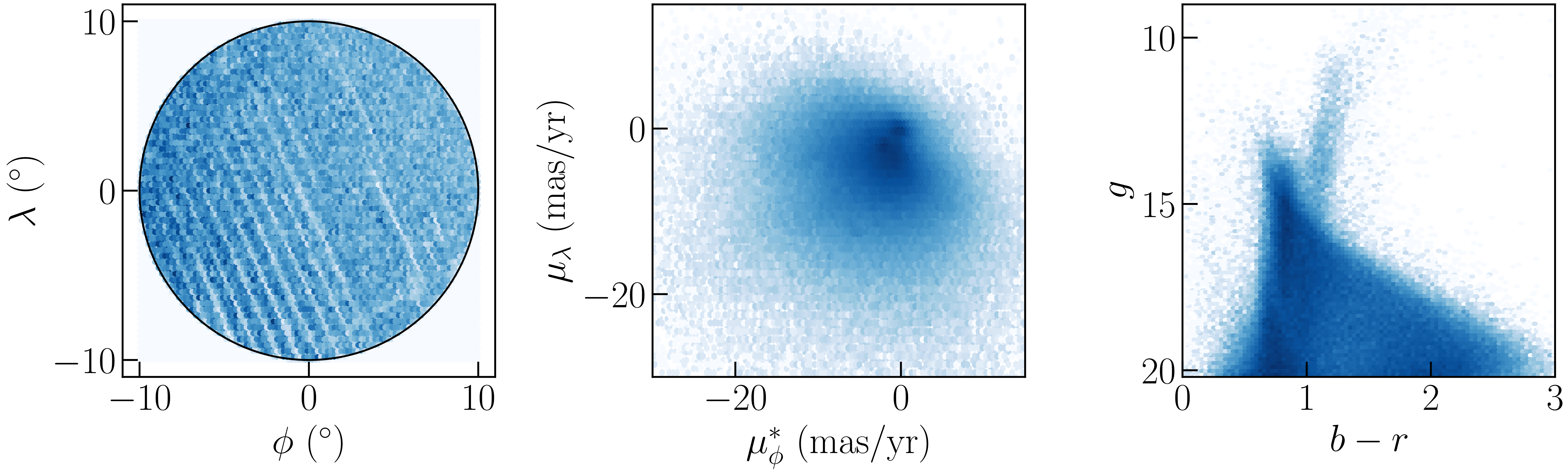
Gaia Data

- We restrict ourselves to distant stars: $\varpi < 1\ \mathrm{mas}$
- 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 15º of centers that tile the sky, every star within 7º 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 10º and $g < 20.2$



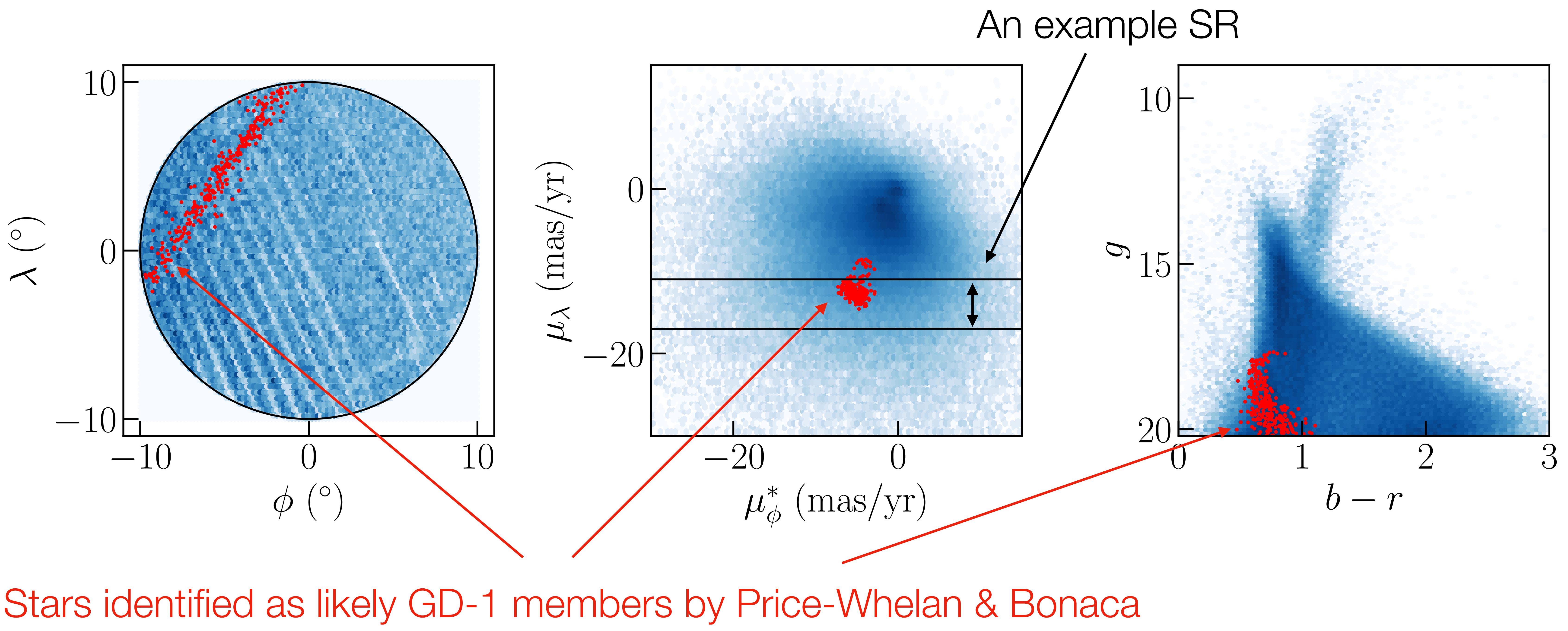
- Recenter the angular positions on patch center:

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



GD-1 Example

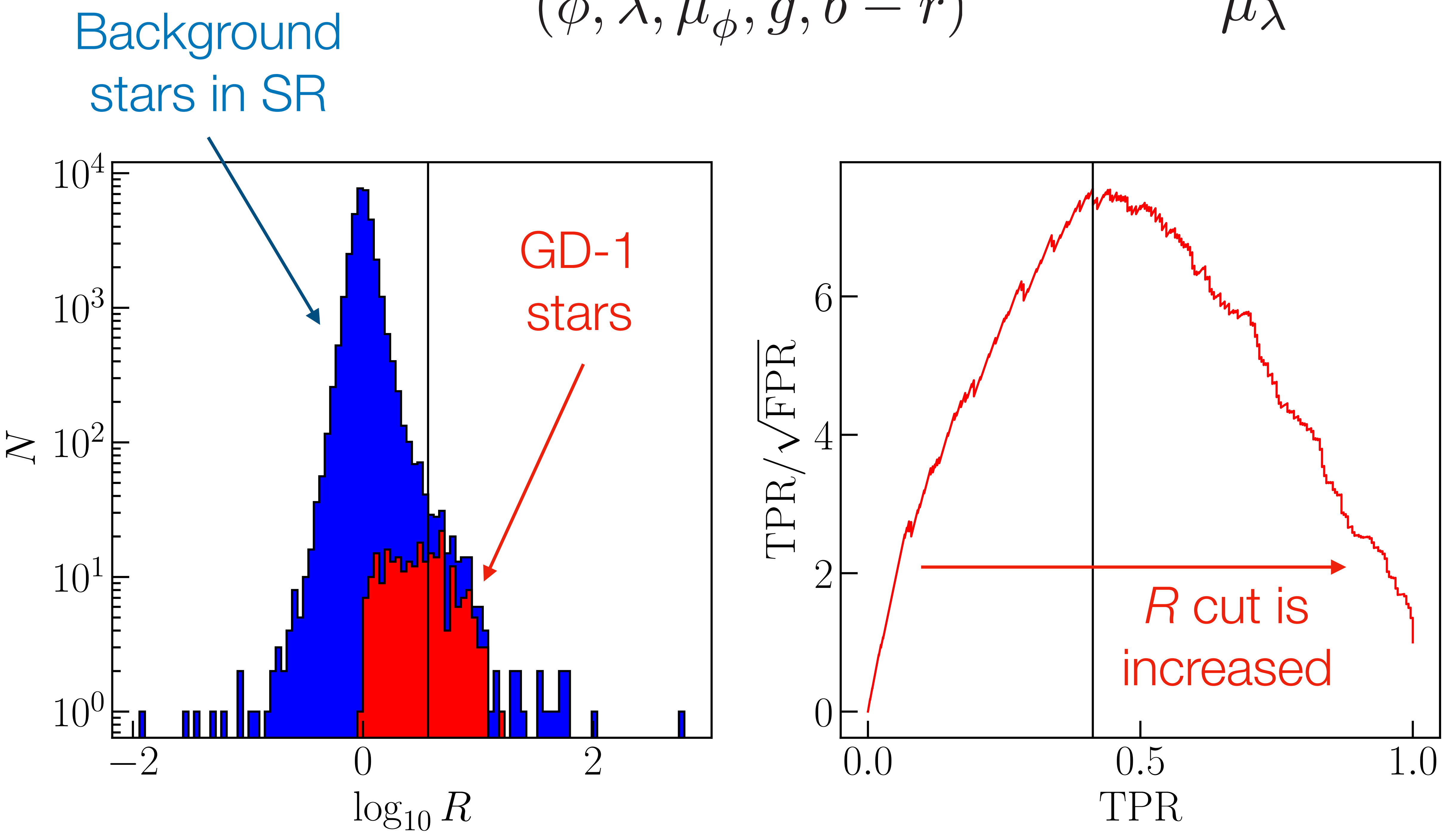
- GD-1 is a bright stream with stellar catalogues of stream membership (Price-Whelan and Bonaca, 2018)
  - Provides a good worked example for our technique.
- Streams are concentrated in both μ_λ and μ_ϕ^*, with a width of a few mas/yr.
  - We will pick μ_λ as the feature m to define our overlapping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

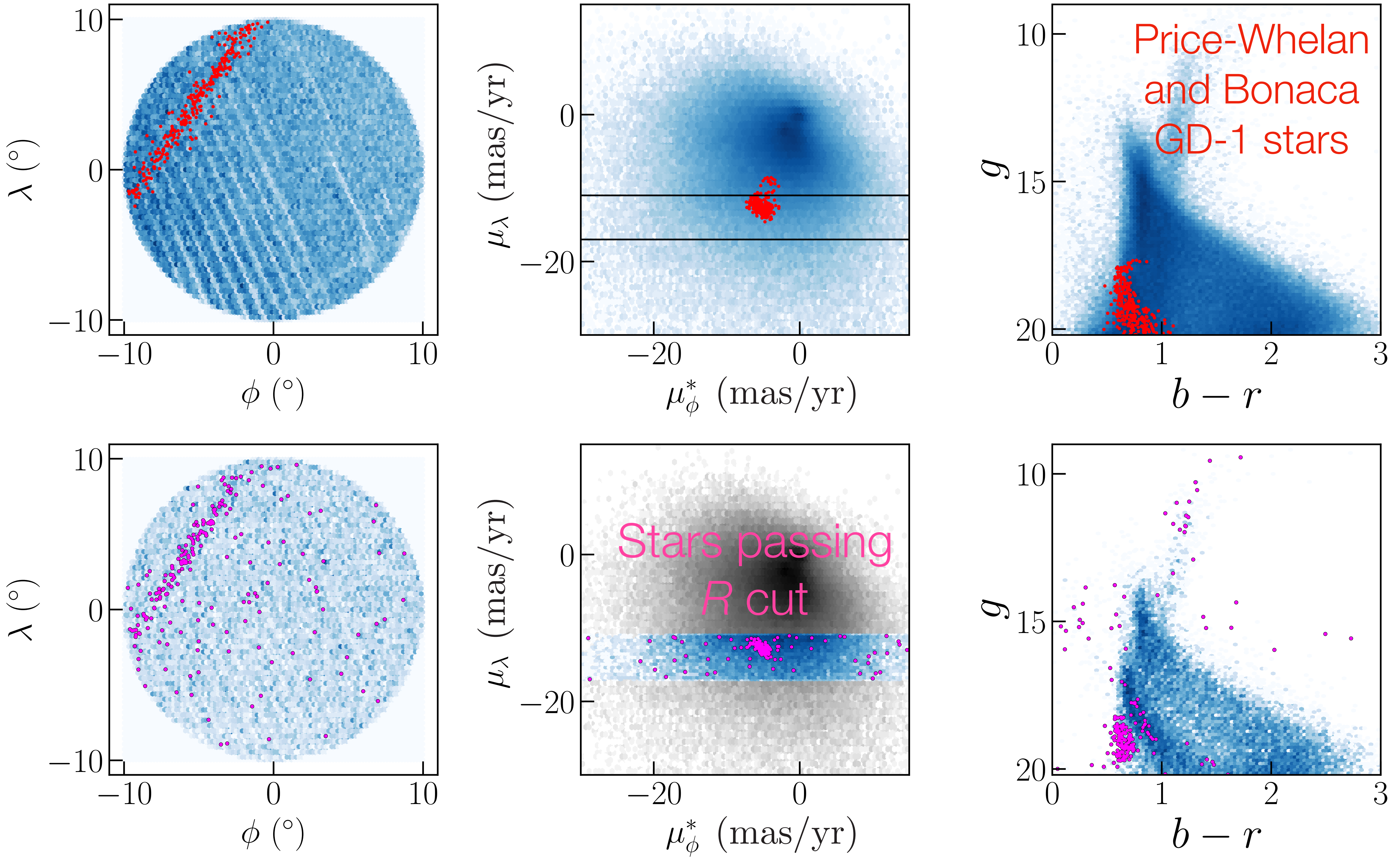


GD-1 Example

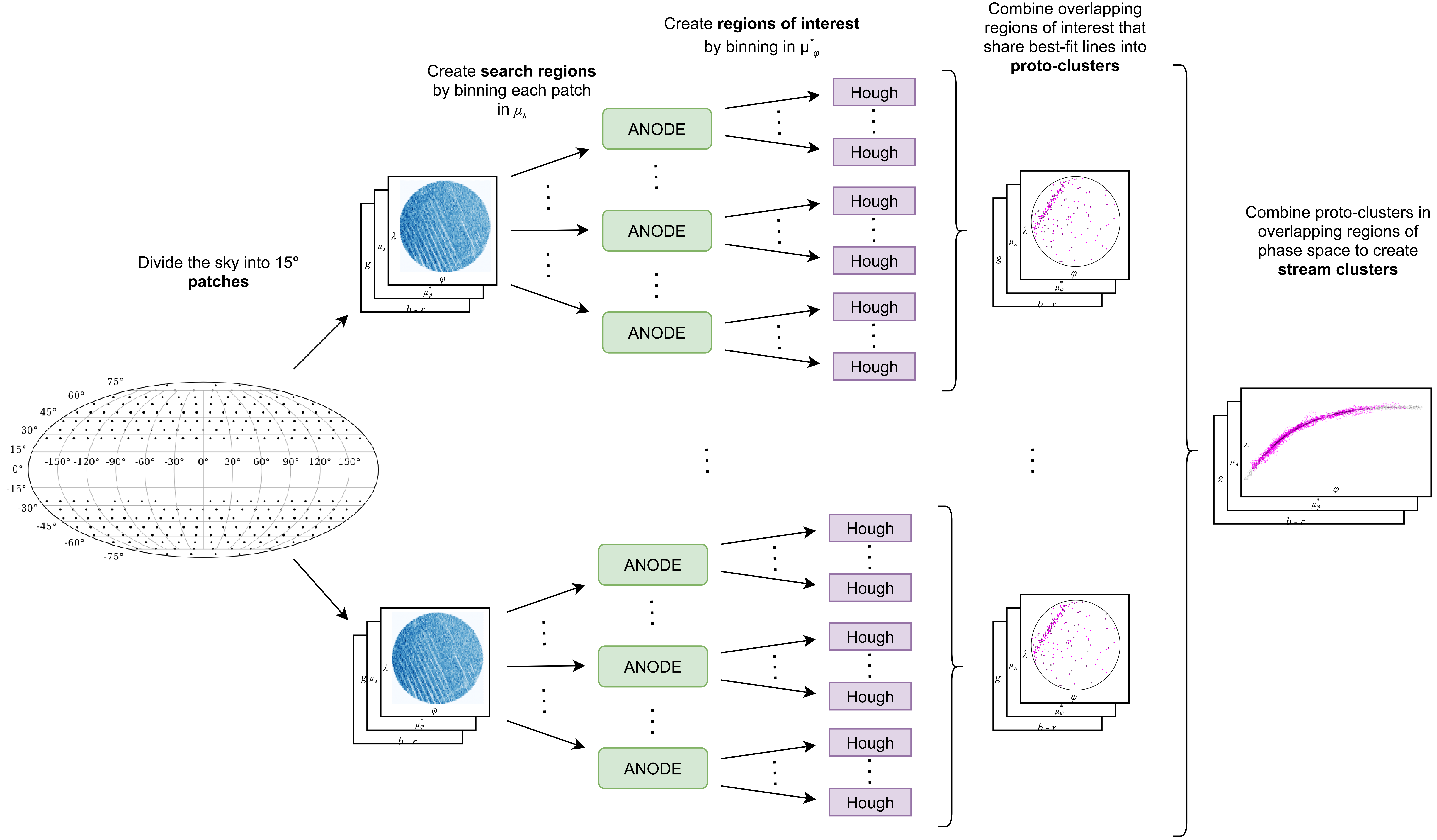
- For each SR within each patch, we train ANODE on the stars in the SR, using the complement of the SR as the control region.
  - For each star, we now have $R(\vec{x}|m \in \mathrm{SR}) = \dfrac{P(\vec{x}|m \in \mathrm{SR})}{P_{\mathrm{CR}}(\vec{x}|m \in \mathrm{SR})}$

$\vec{x} = (\phi, \lambda, \mu_\phi^*, g, b-r)$, $m = \mu_\lambda$



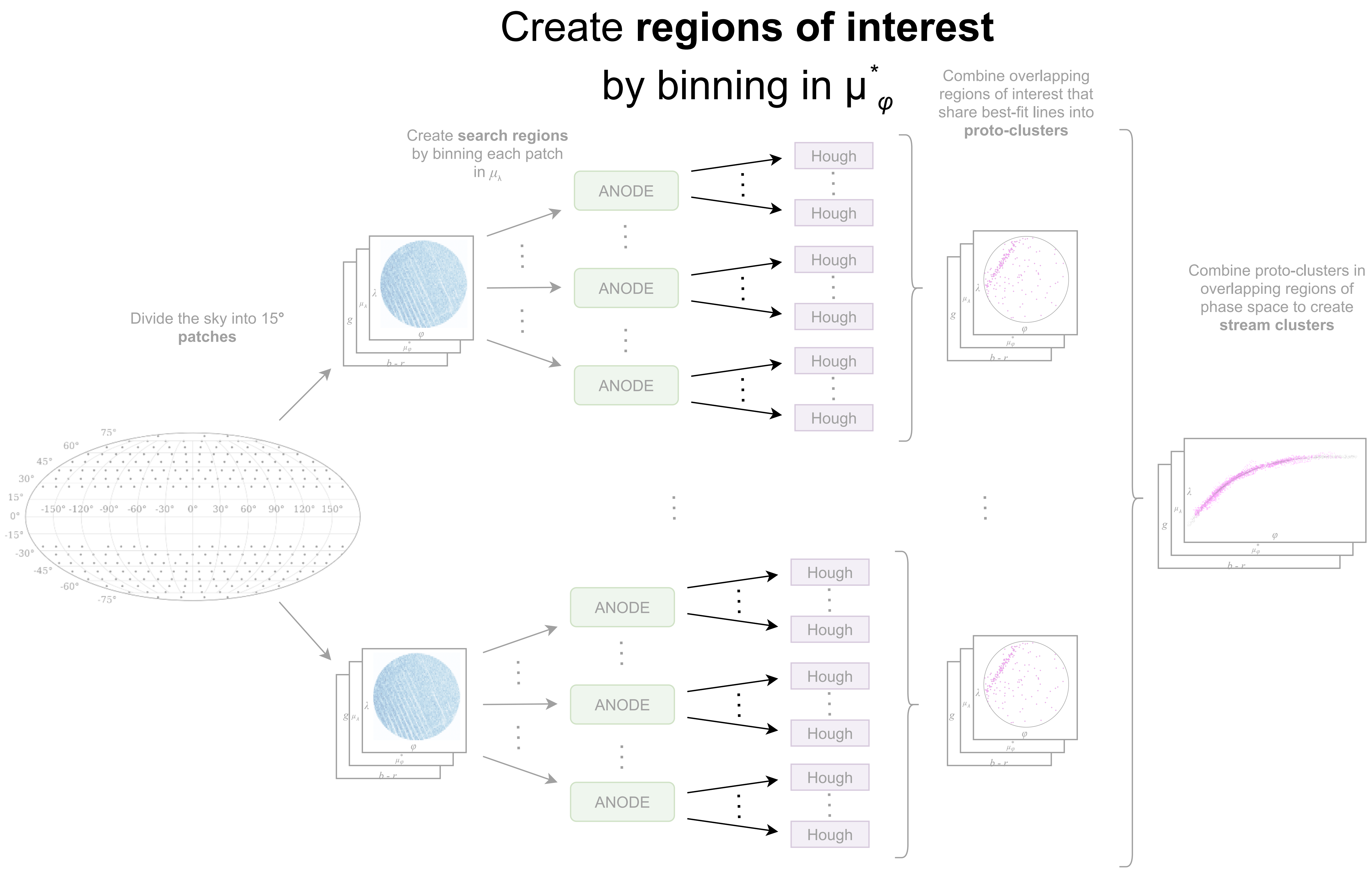


Via Machinae: An Overview

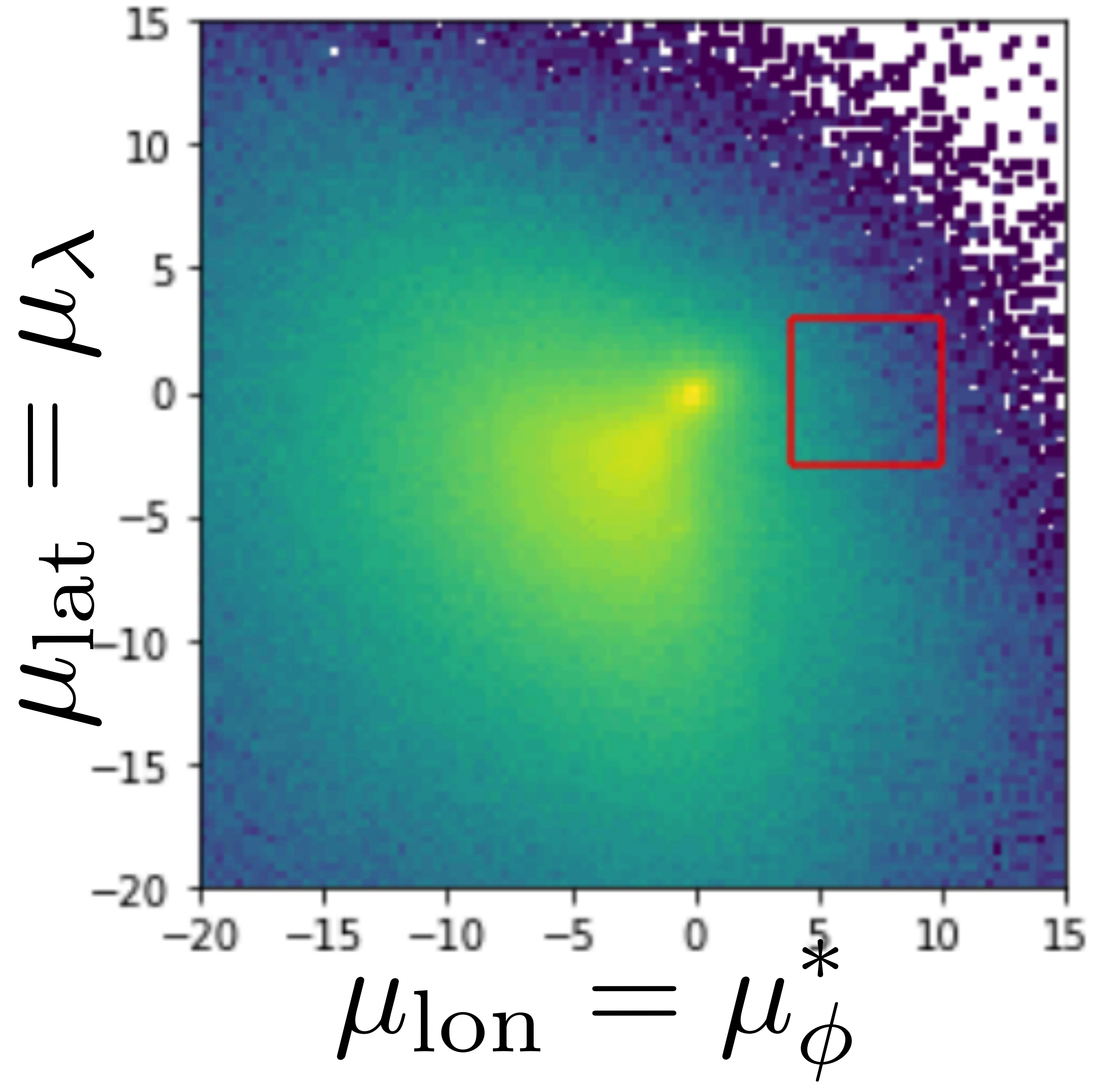


2104.12789

Via Machinae: Regions of Interest

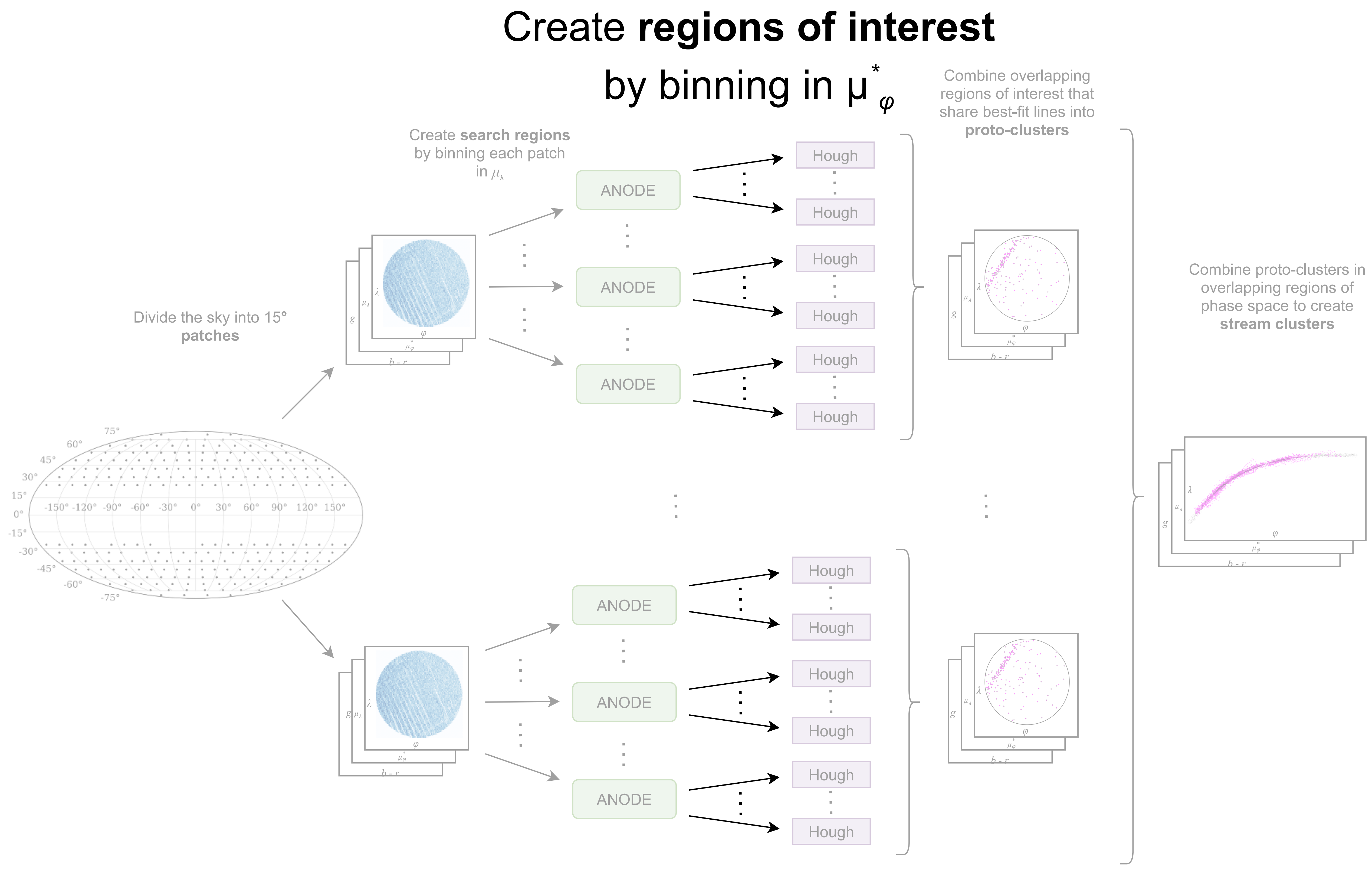


- For known streams less distinct than GD-1, we need further subdivide the search regions (which were defined using μ_λ).
- Each SR divided into overlapping *Regions of Interest* using μ^*_ϕ

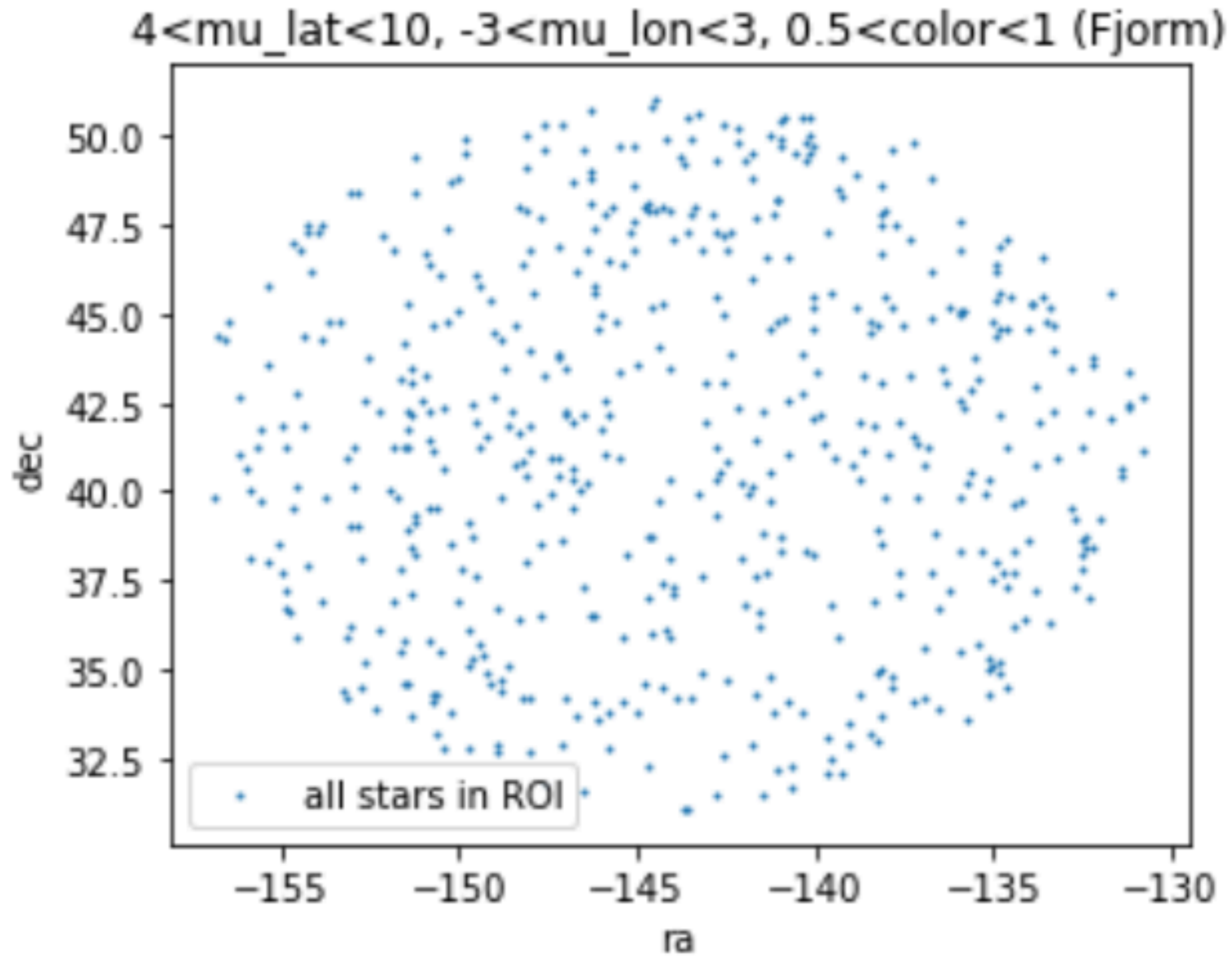


2104.12789

Via Machinae: Regions of Interest



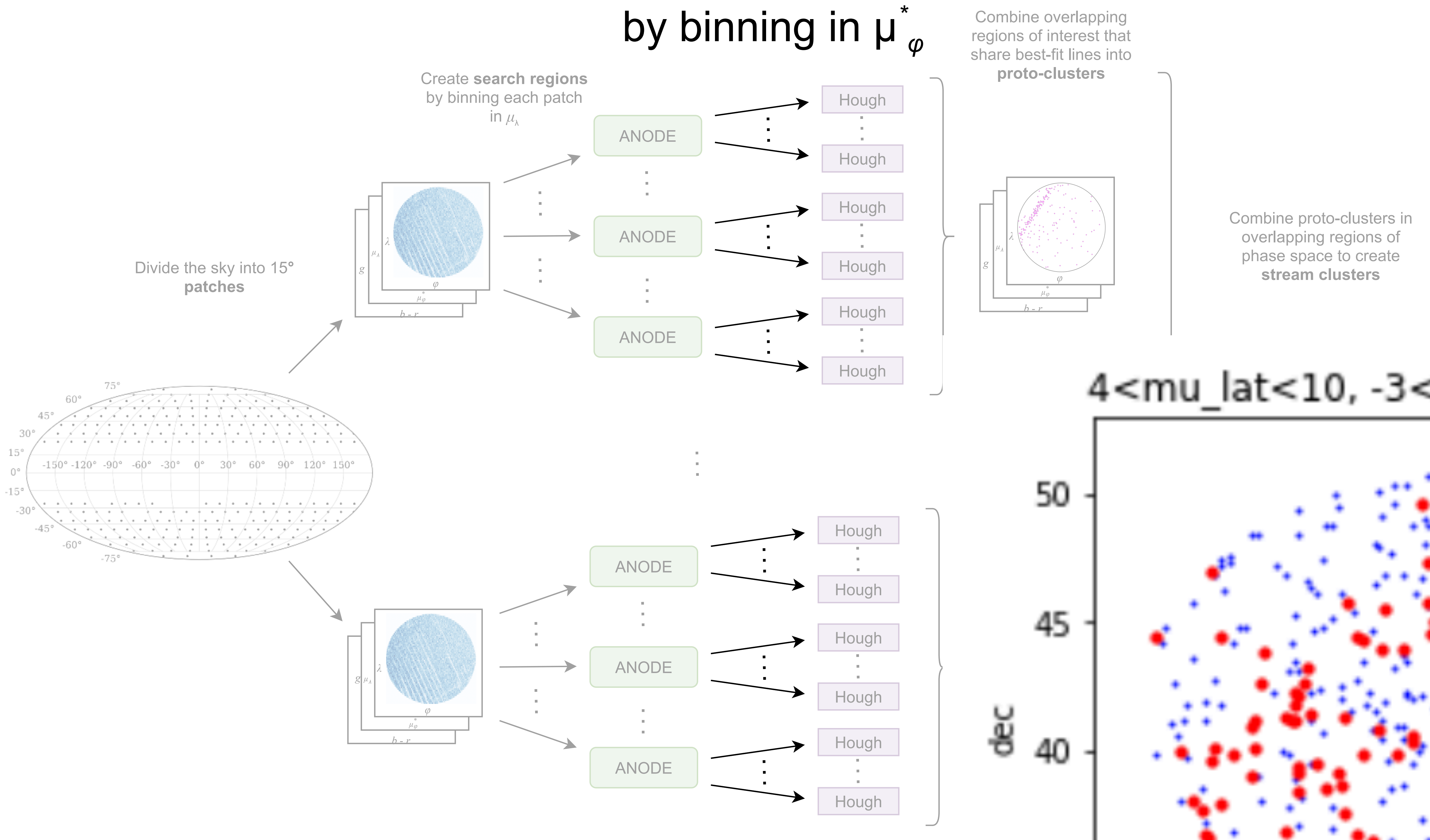
- For known streams less distinct than GD-1, we need further subdivide the search regions (which were defined using μ_λ).
- Each SR divided into overlapping *Regions of Interest* using μ^*_ϕ



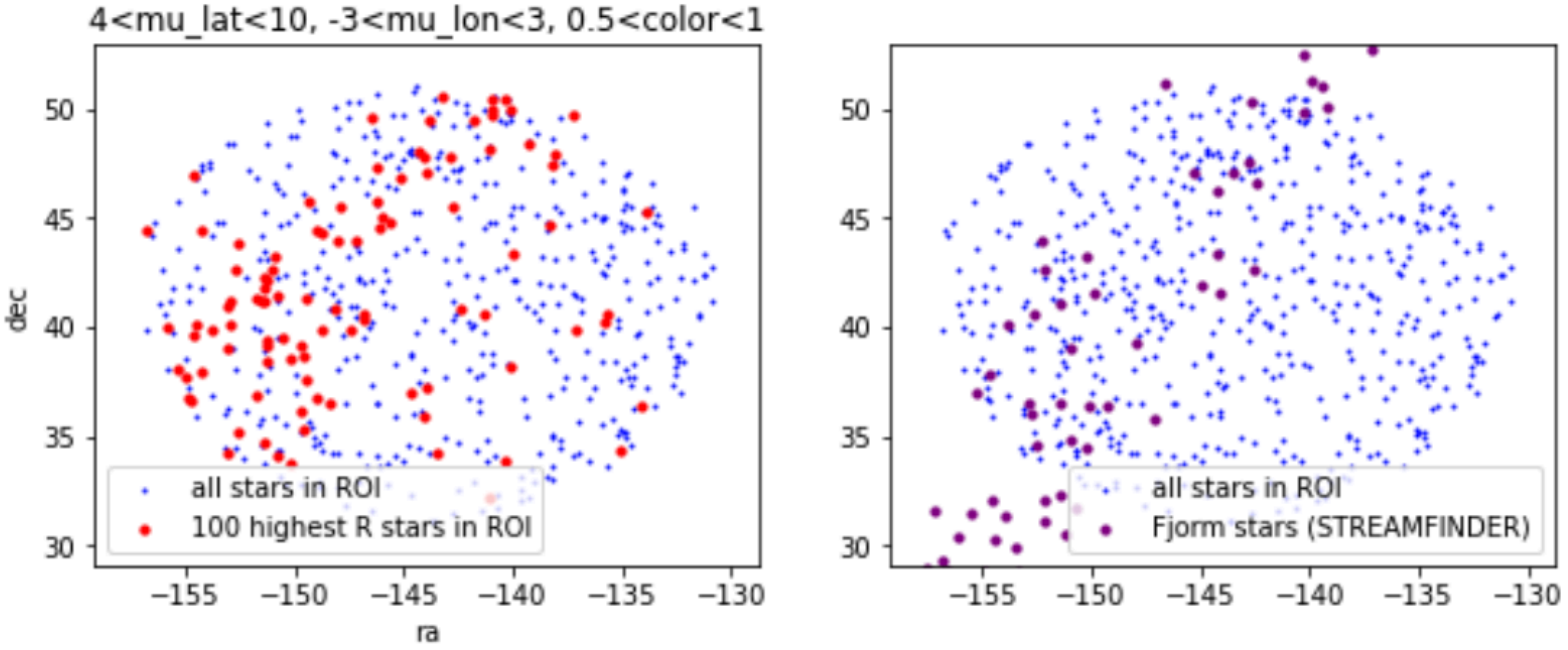
2104.12789

Via Machinae: Regions of Interest

Create **regions of interest** by binning in μ^*_φ

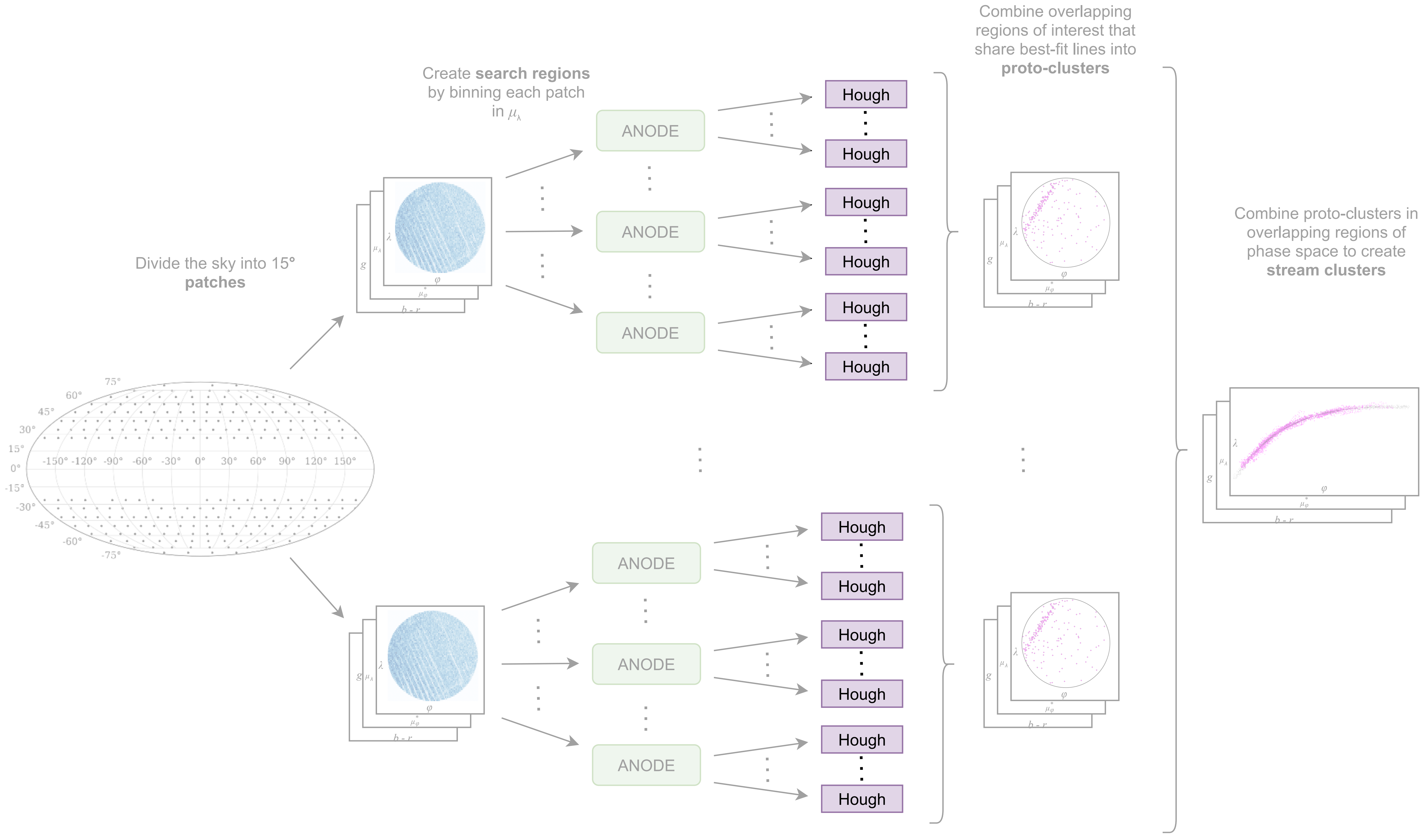


- For known streams less distinct than GD-1, we need further subdivide the search regions (which were defined using μ_λ).
- Each SR divided into overlapping *Regions of Interest* using μ^*_ϕ



2104.12789

Via Machinae: Line Finding

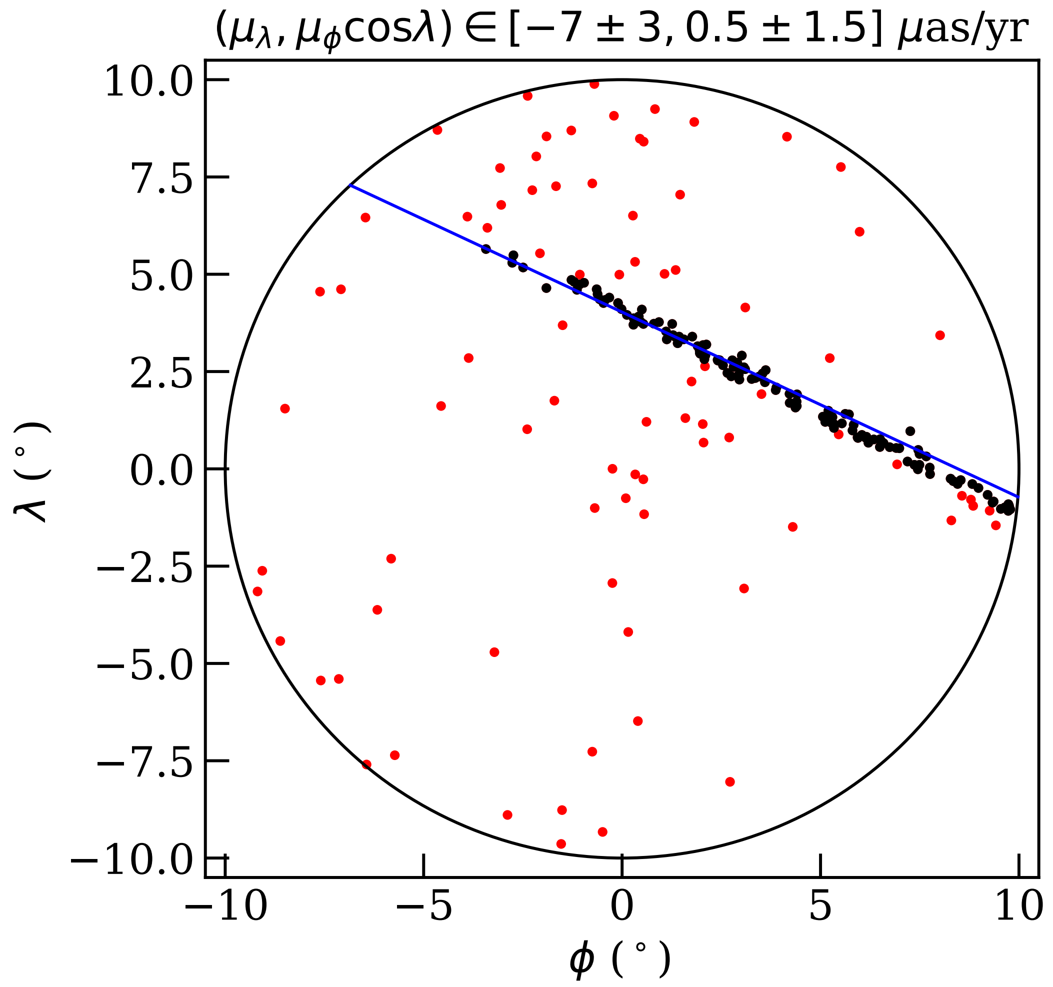


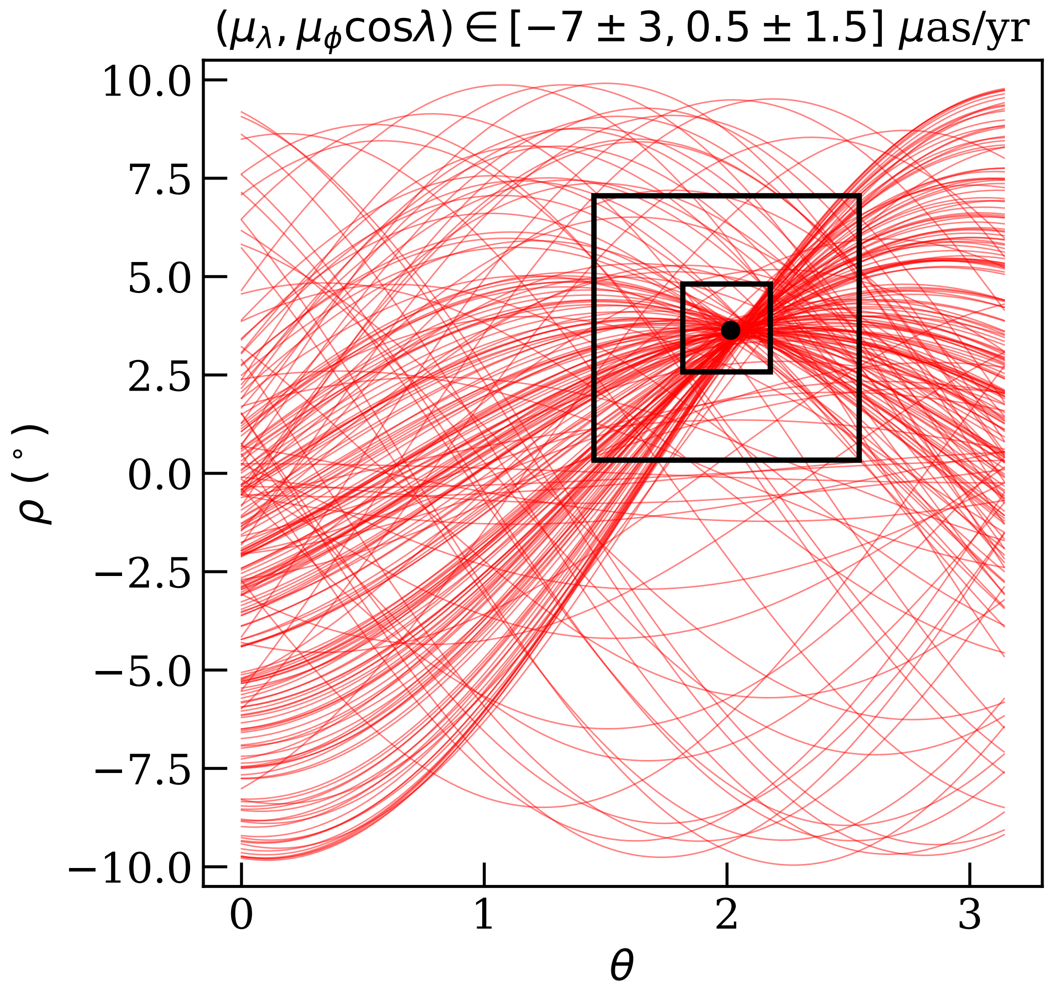
- Need to automate line-finding within the 140,000 ROIs.
- Use the Hough transform to convert line-finding to over-density finding.

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

- Allows us to define a figure of merit for stream:

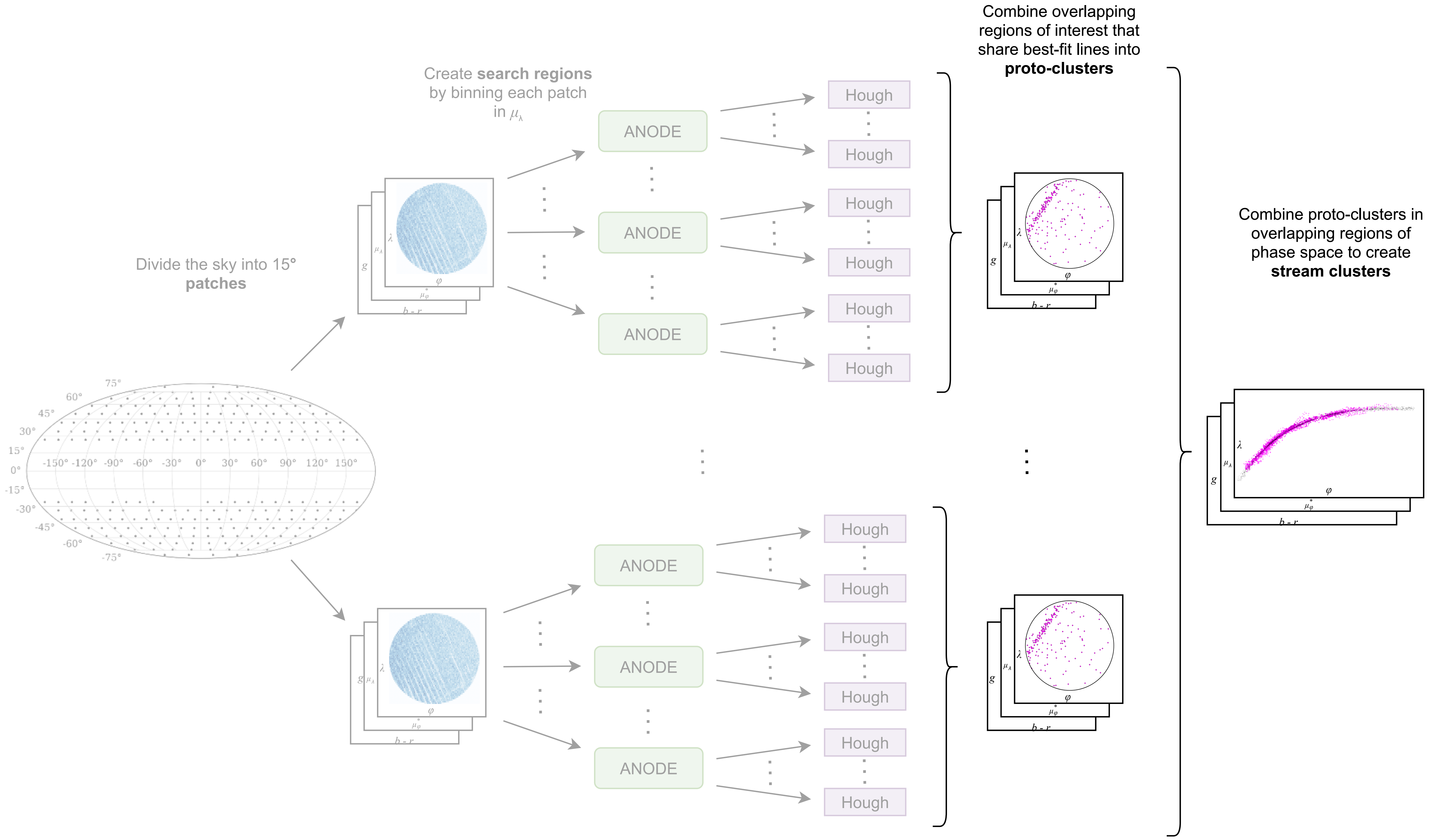
$$\sigma_L = \frac{N(\rho,\theta) - \bar{N}(\rho,\theta)}{\sqrt{\bar{N}(\rho,\theta)}}$$



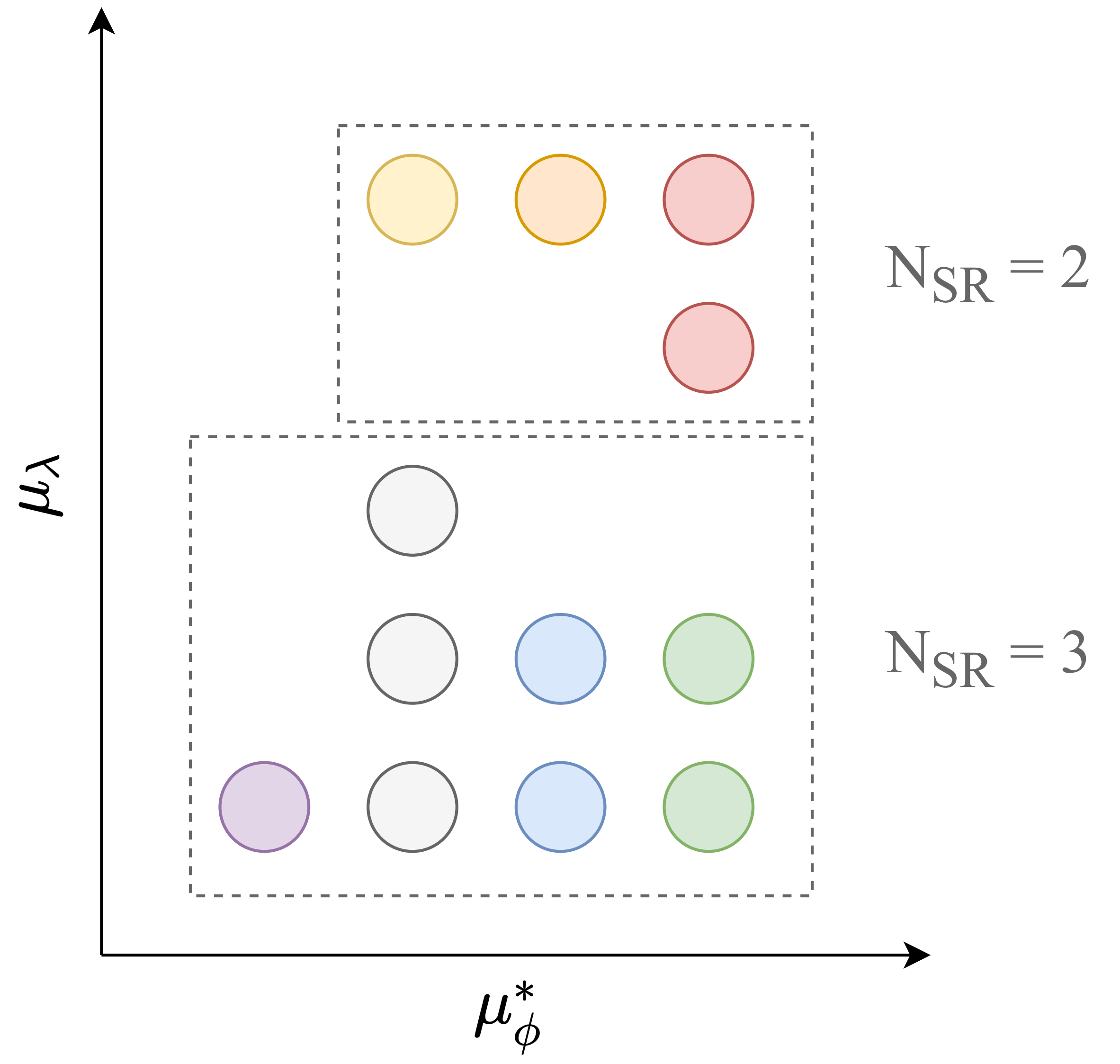


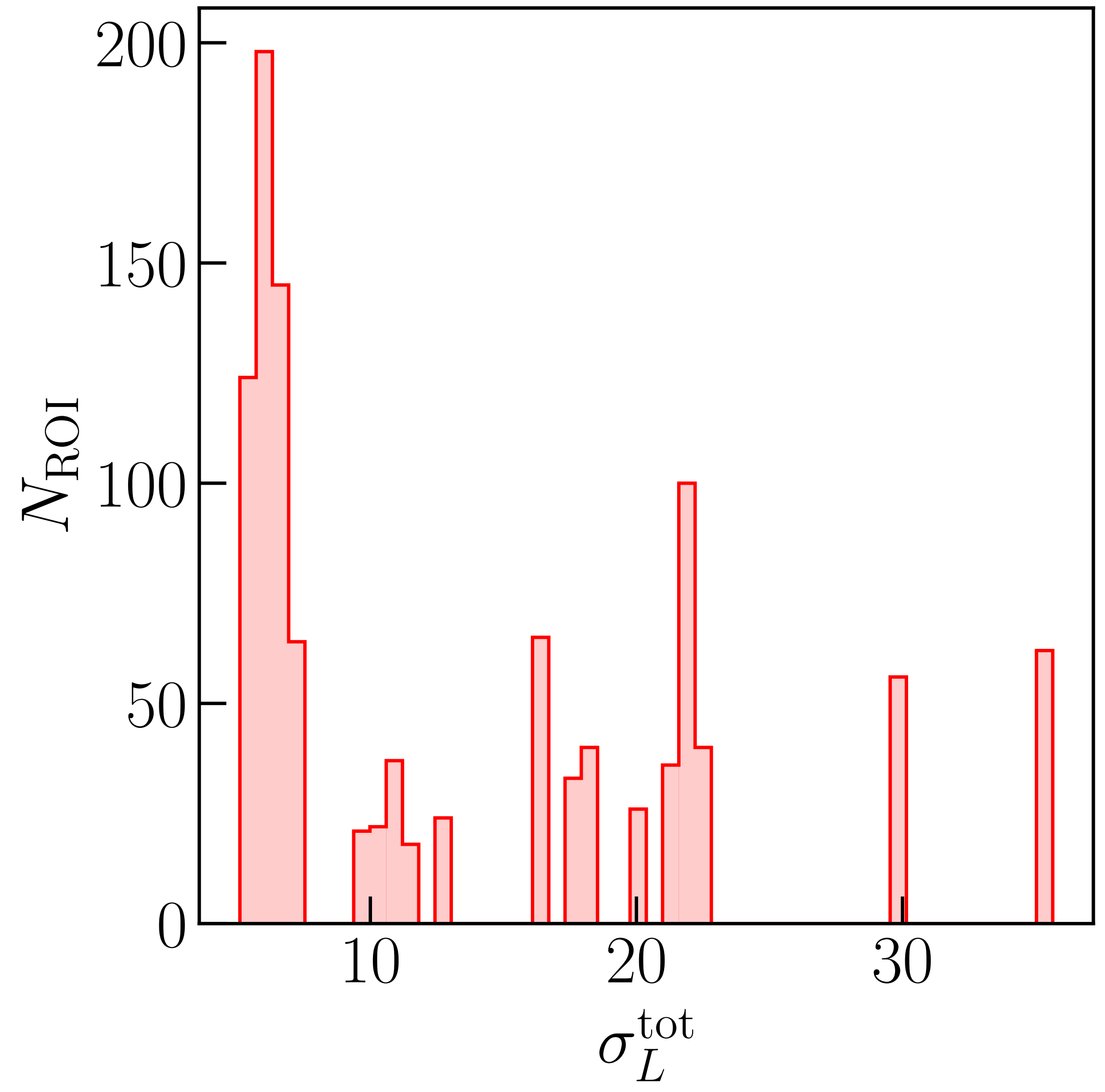
2104.12789

Via Machinae: Combining Results



- Combine line-candidates which appear in adjacent ROIs.
- Number of SRs and combined line-significance allow us to select high-confidence stream candidates.





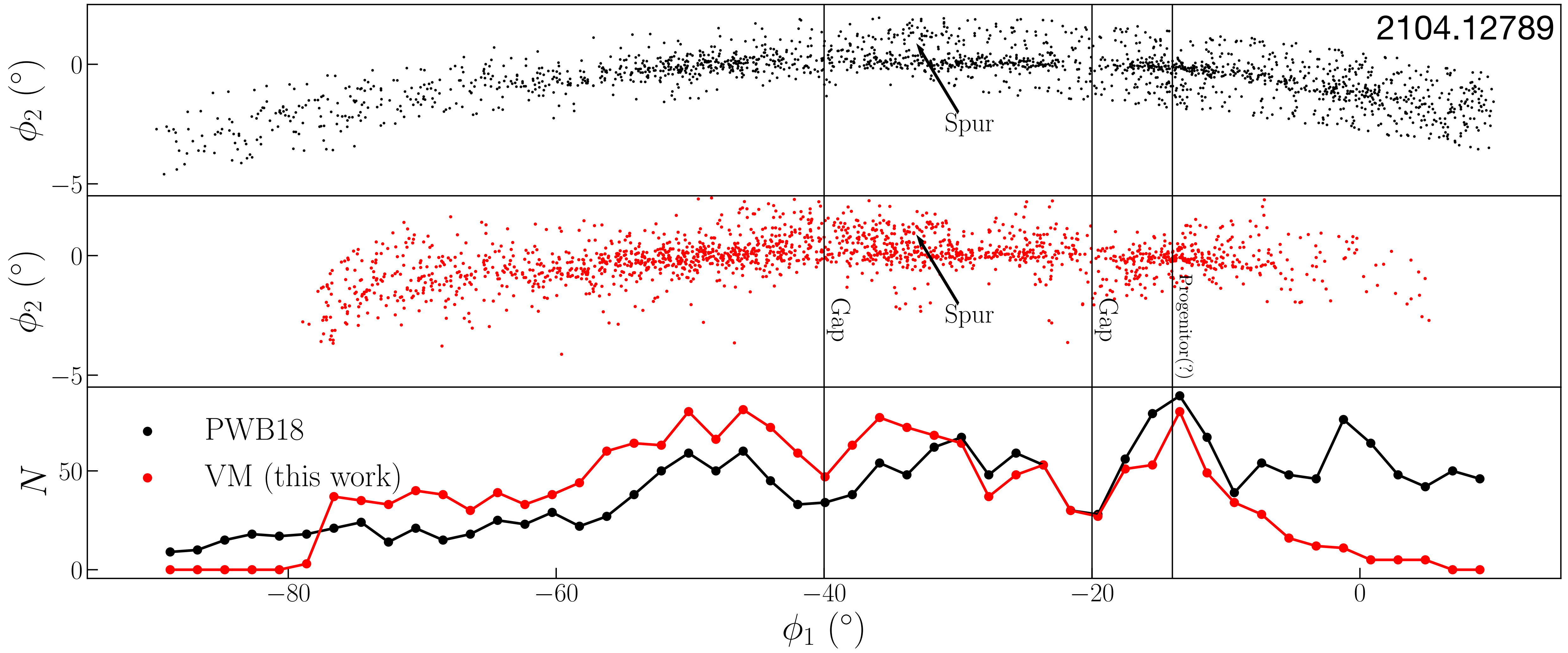
2104.12789

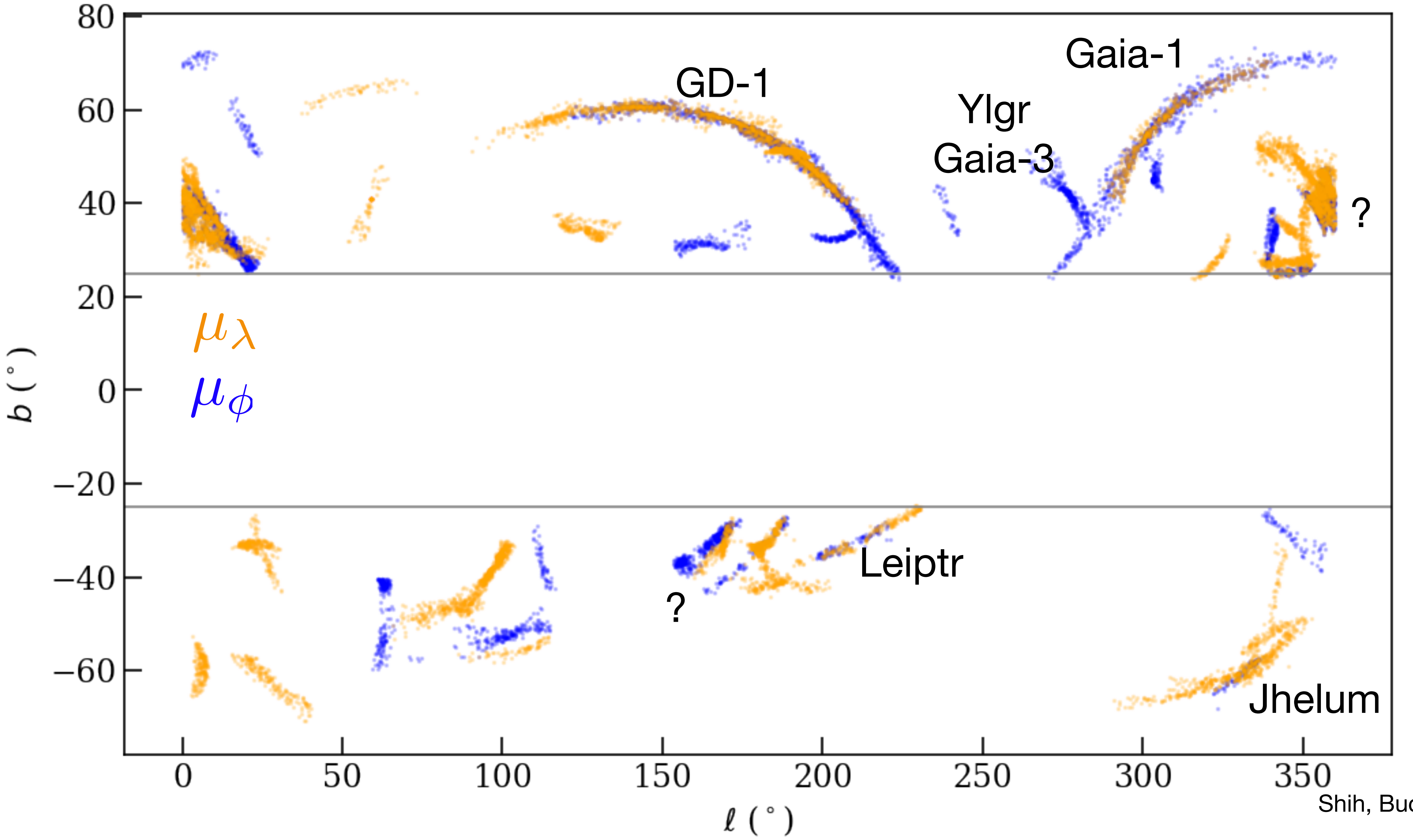
Results

- Via Machinae reproduces the GD-1 stream and identifies several non-trivial and astrophysically important structures.
    - SRs defined by μ_λ have difficulty with streams near $\mu_\lambda \sim 0$, due to larger number of background stars.
- Many post-ANODE hyperparameters determine the sensitivity to other streams — Hough width, $\sigma_L^{\rm tot}$, $N_{\rm SR}$
    - First pass: “narrow” streams with high significance in SRs defined by either μ_λ or μ_ϕ^*

$$\sigma_L^{\rm tot} \geq 8, \;\; N_{\rm SR} \geq 3$$

- How can we improve our density estimator?





Preliminary
Shih, Buckley, Necib, Tamanas