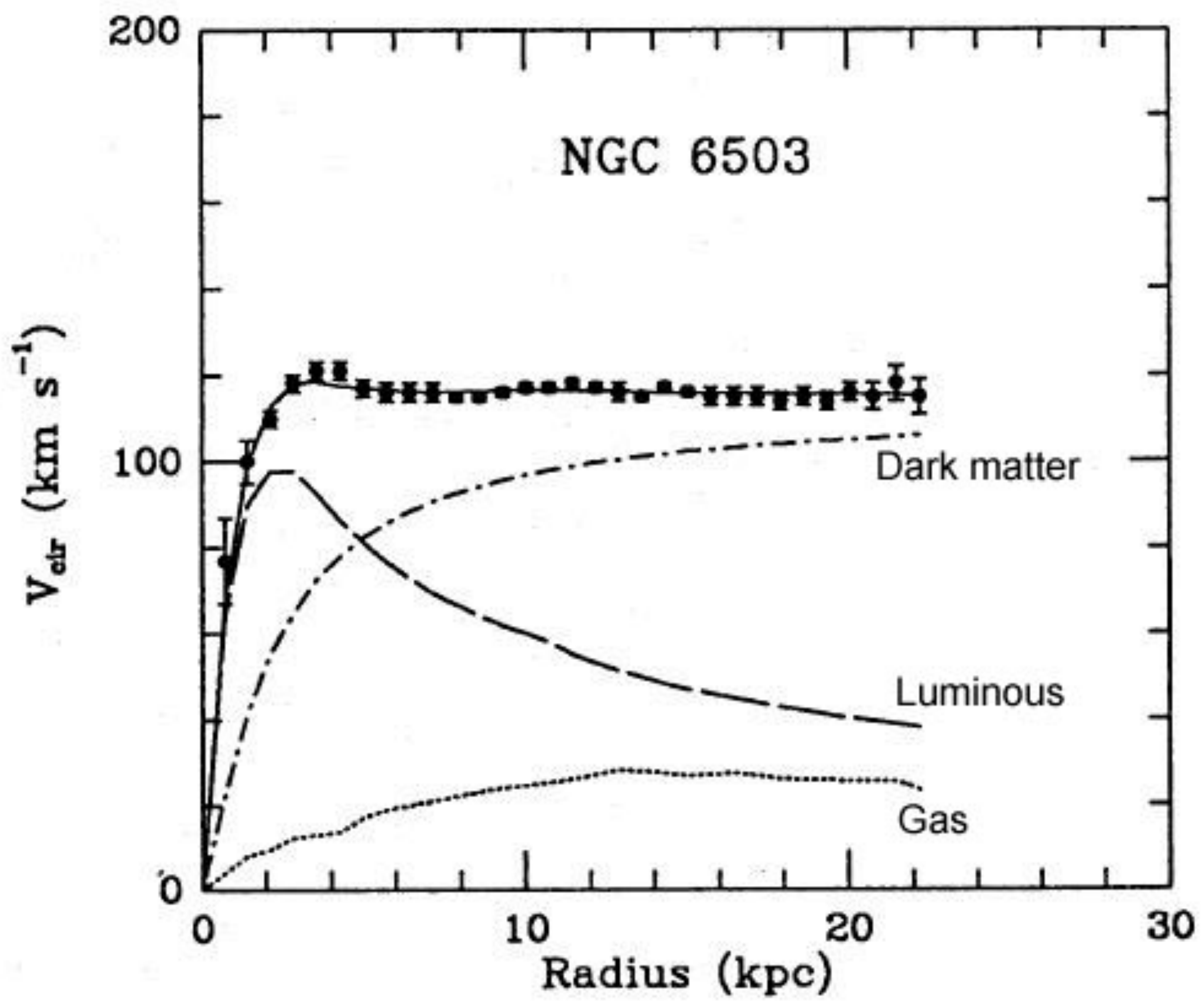
Machine Learning for Particle Astrophysics

Matthew R Buckley
Rutgers University

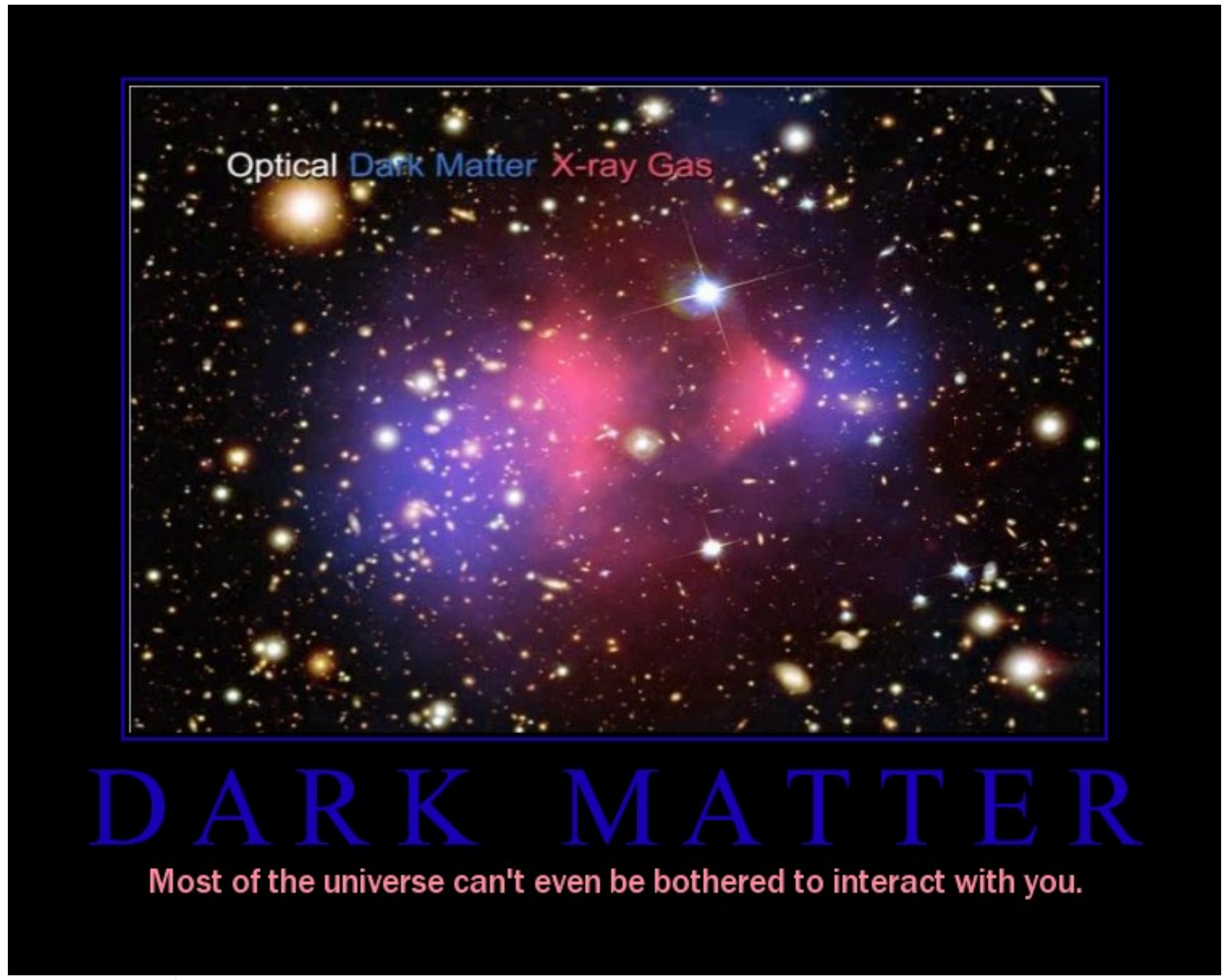
- David Shih (Rutgers), Lina Necib (MIT)
- Sung Hak Lim (Rutgers), Claudius Krause (Rutgers/Heidelberg)
- Eric Putney (Rutgers), Anna Hallin (Rutgers/Hamburg), John Tamanas (UCSC)
- Kailash Raman (Rutgers)

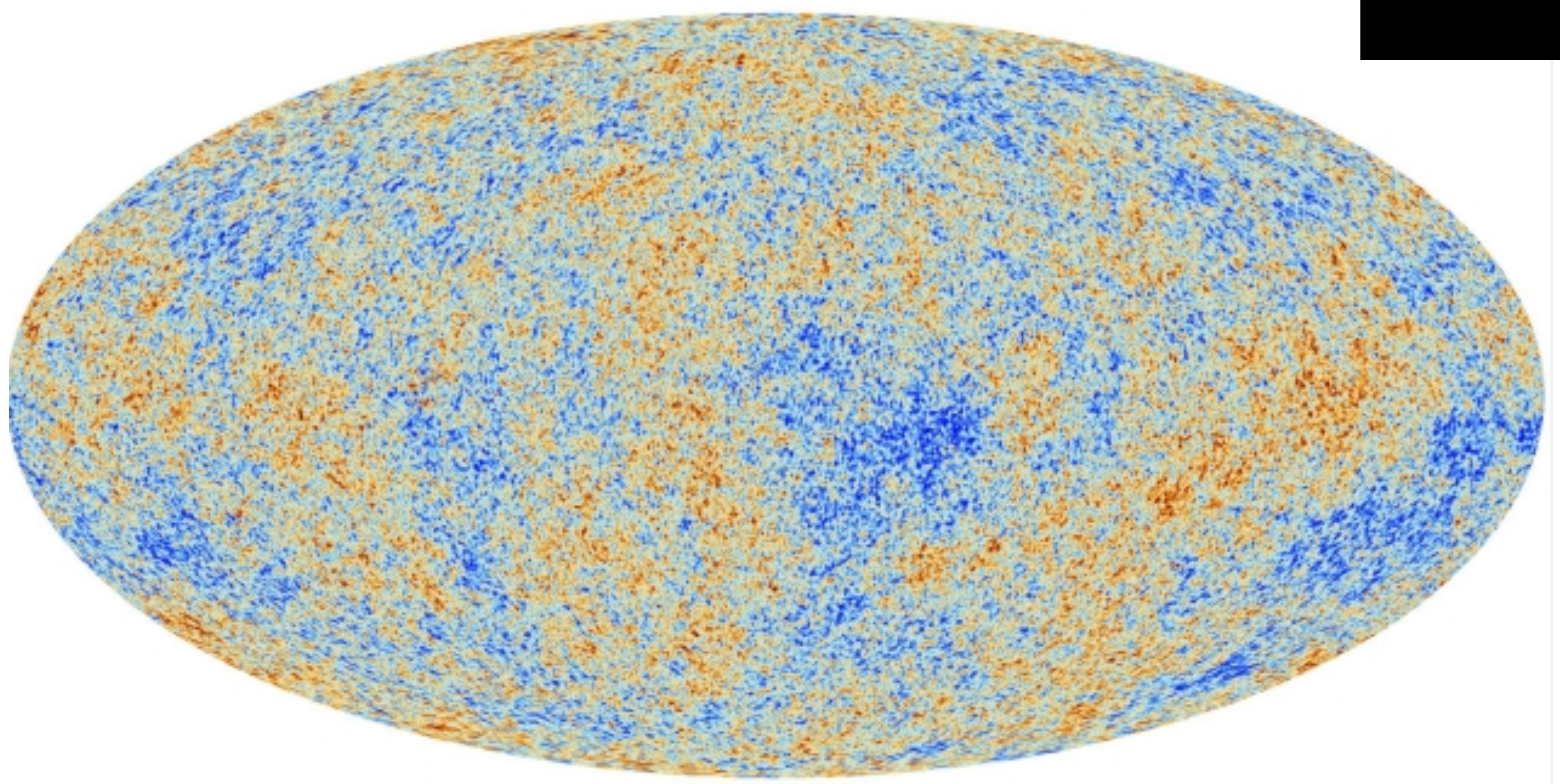
Theoretical Motivations

- We know dark matter exists, but our evidence is purely astrophysical:



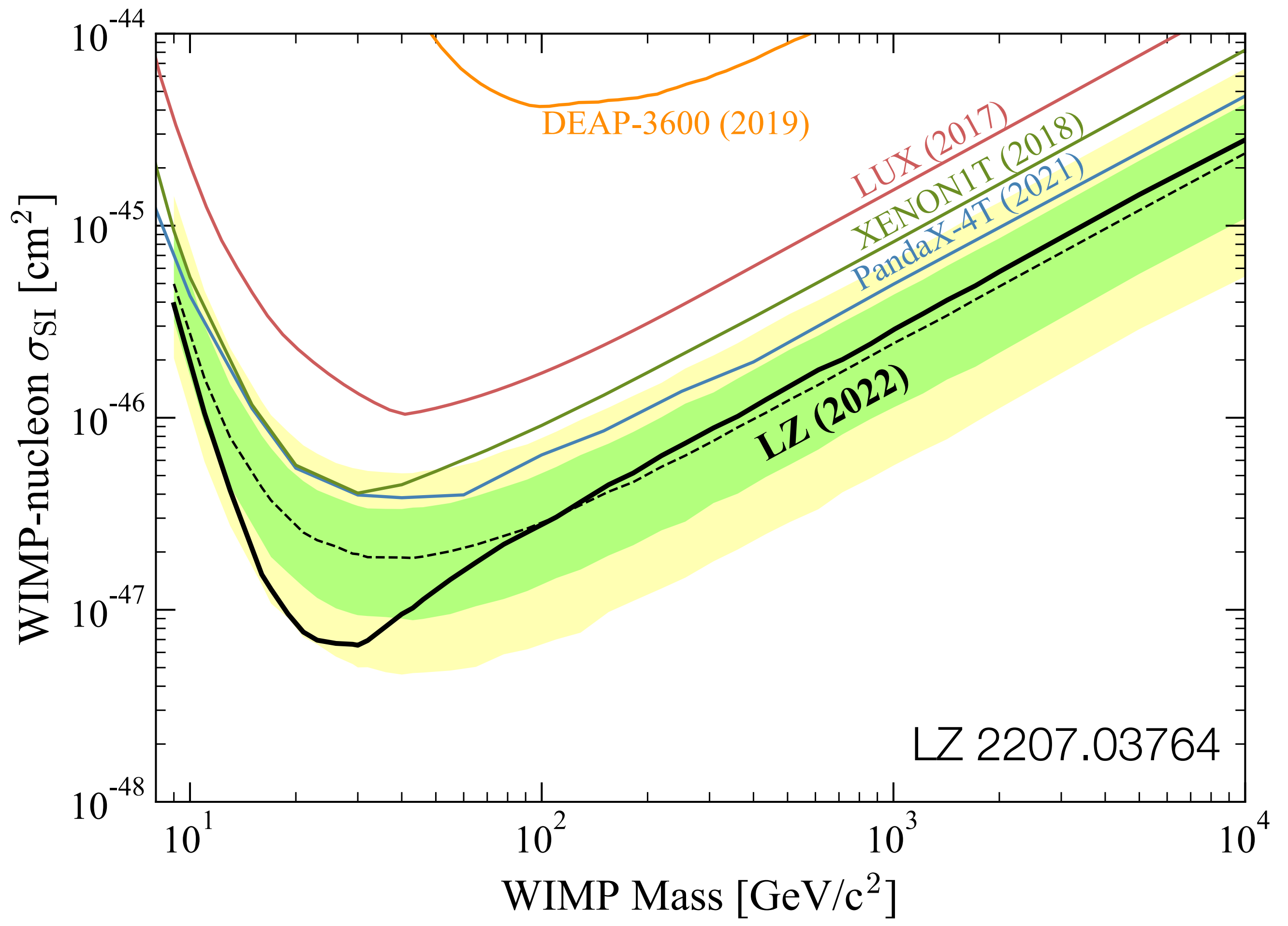
K.G. Begeman, A.H. Broels, R.H. Sanders. 1991. Mon.Not.RAS 249, 523.

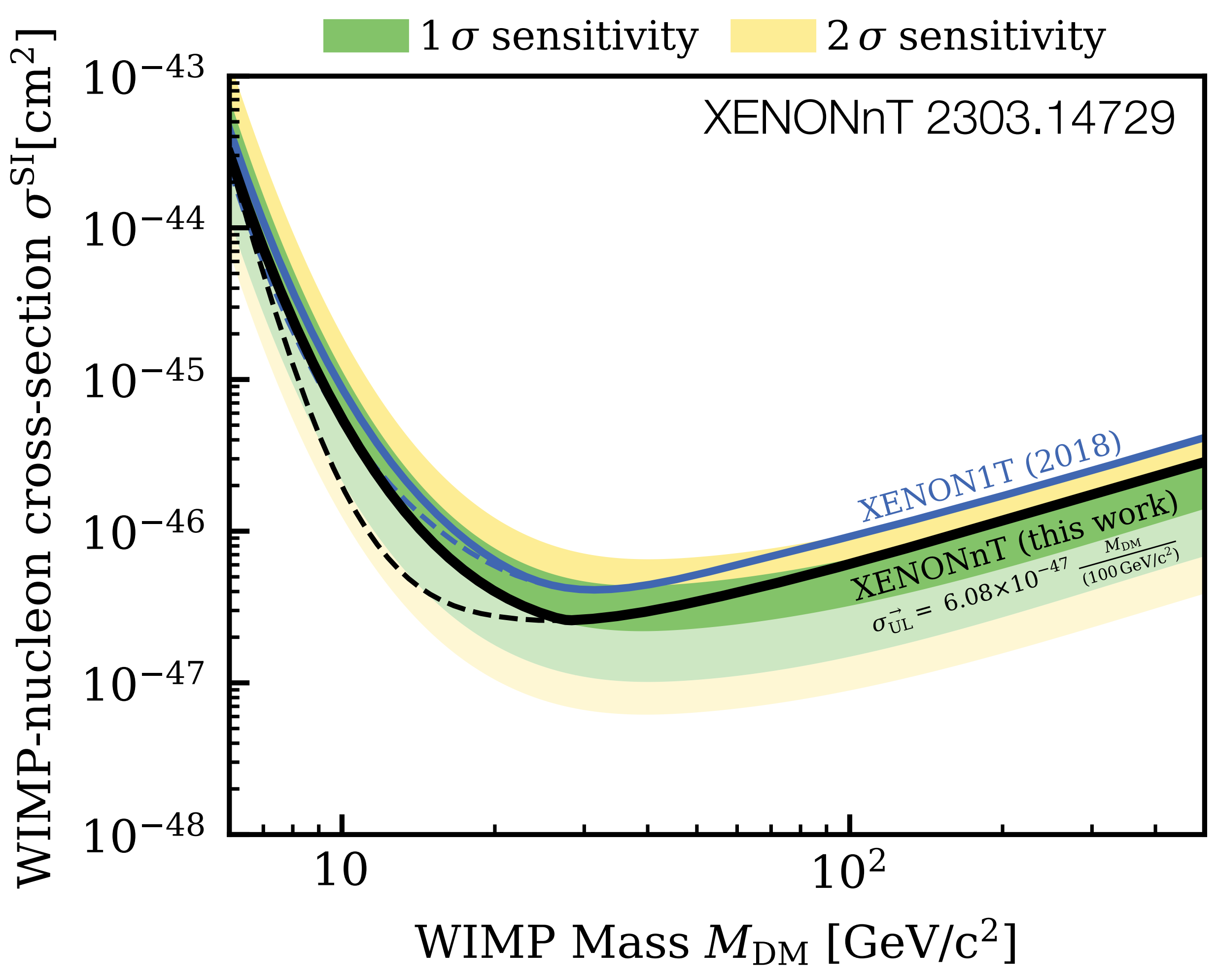




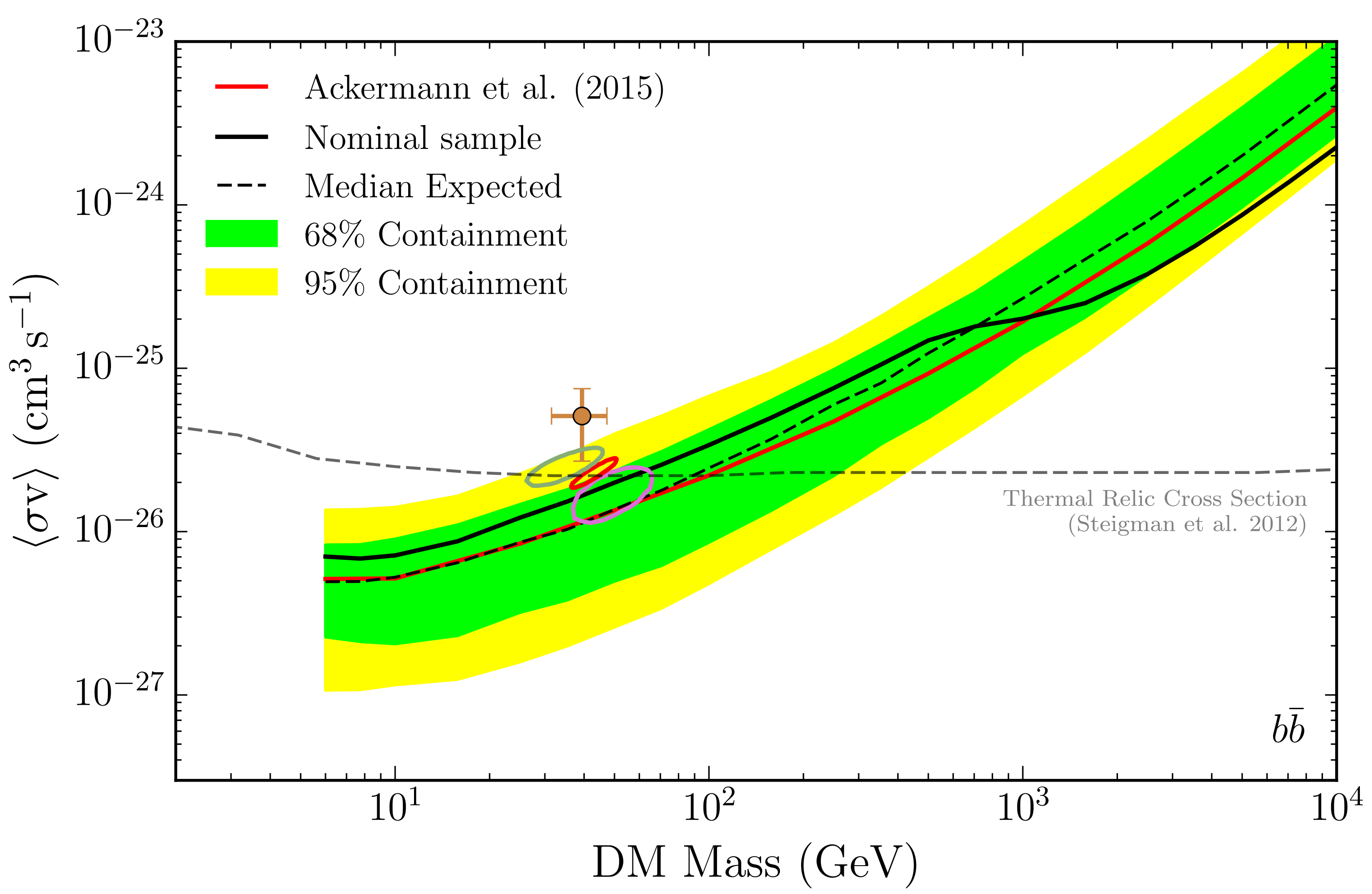
Theoretical Motivations

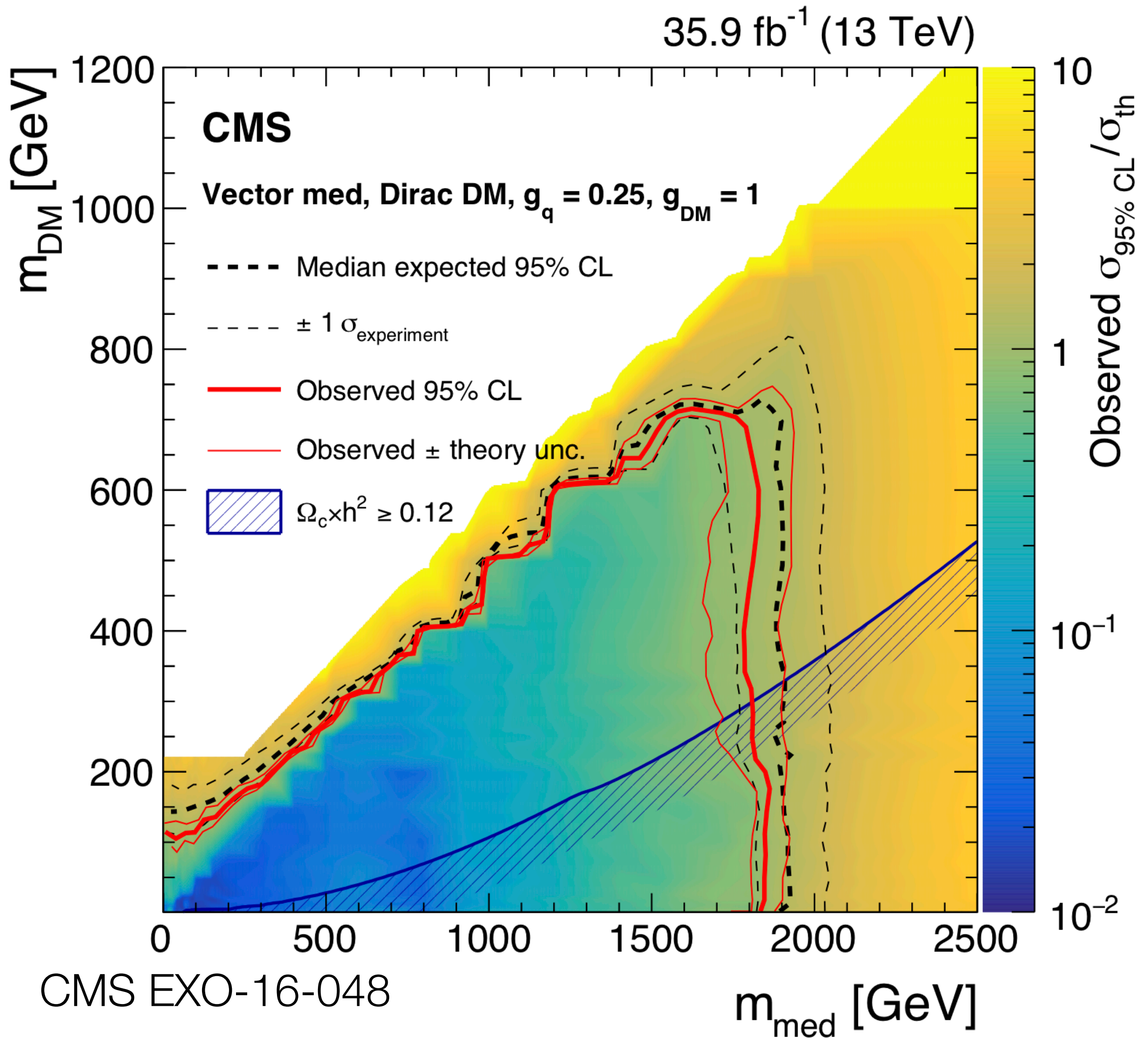
- Particle Physics experiments are motivated and important, but so far give only negative results



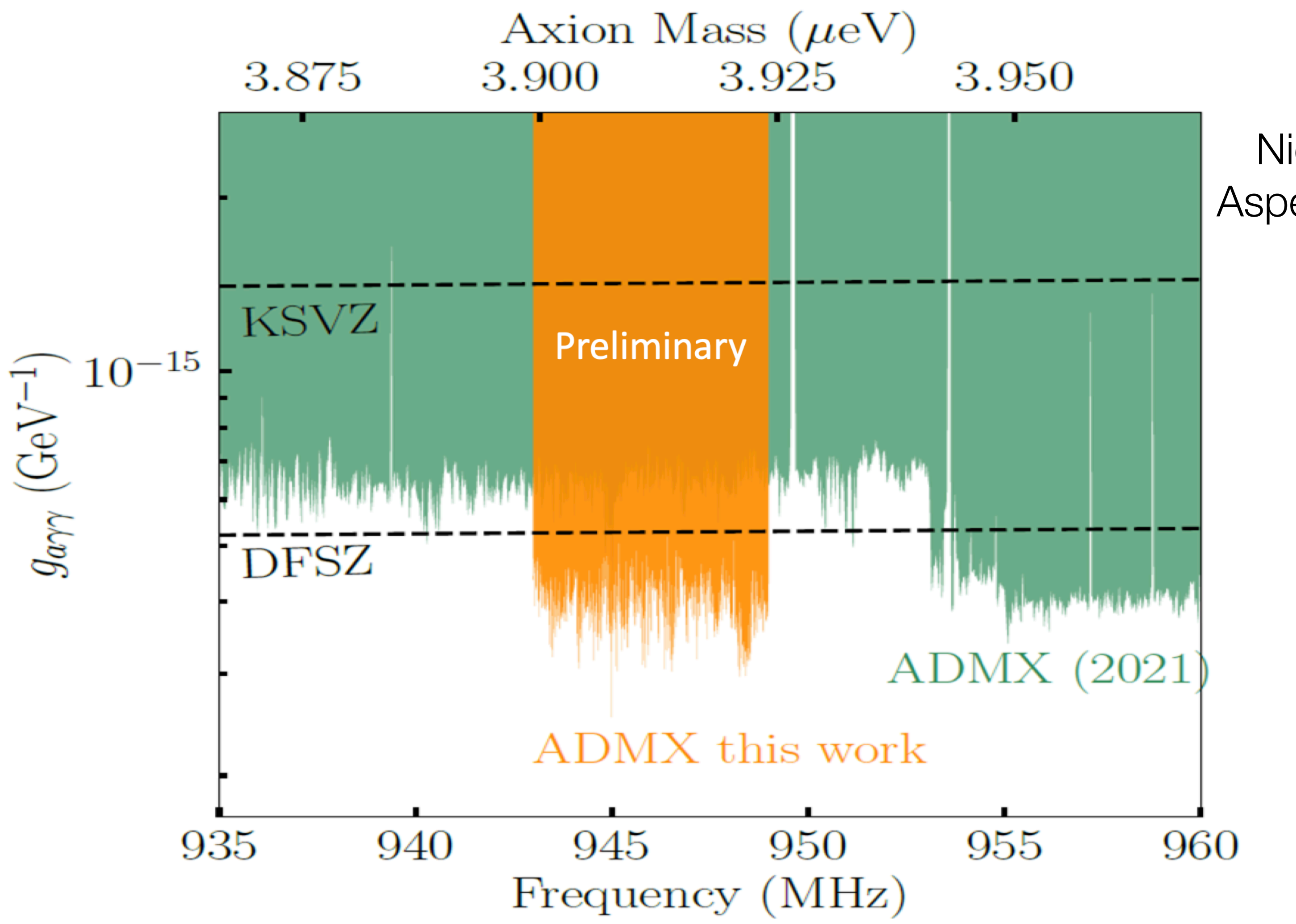


Fermi LAT 1611.03184





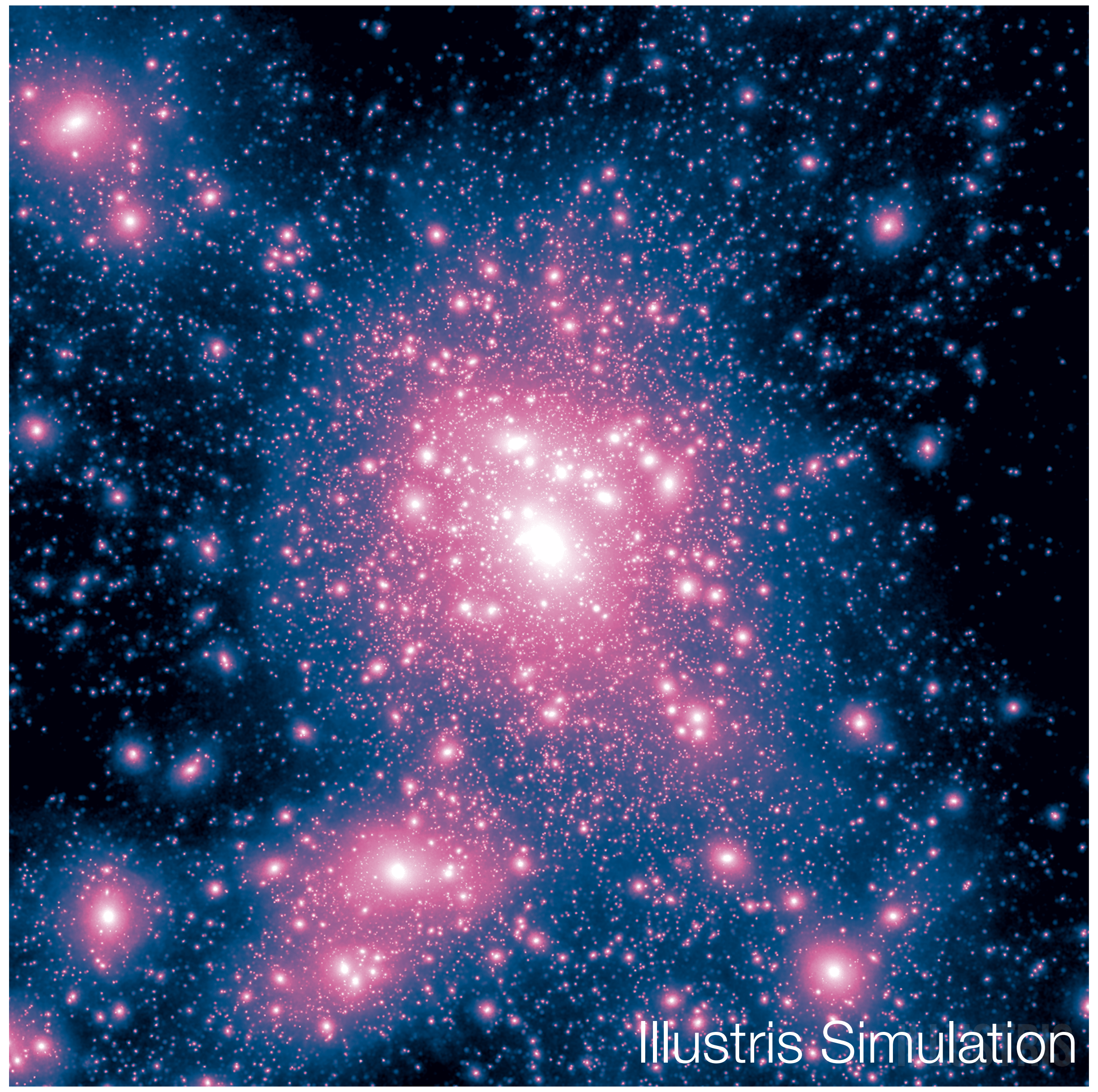
CMS EXO-16-048



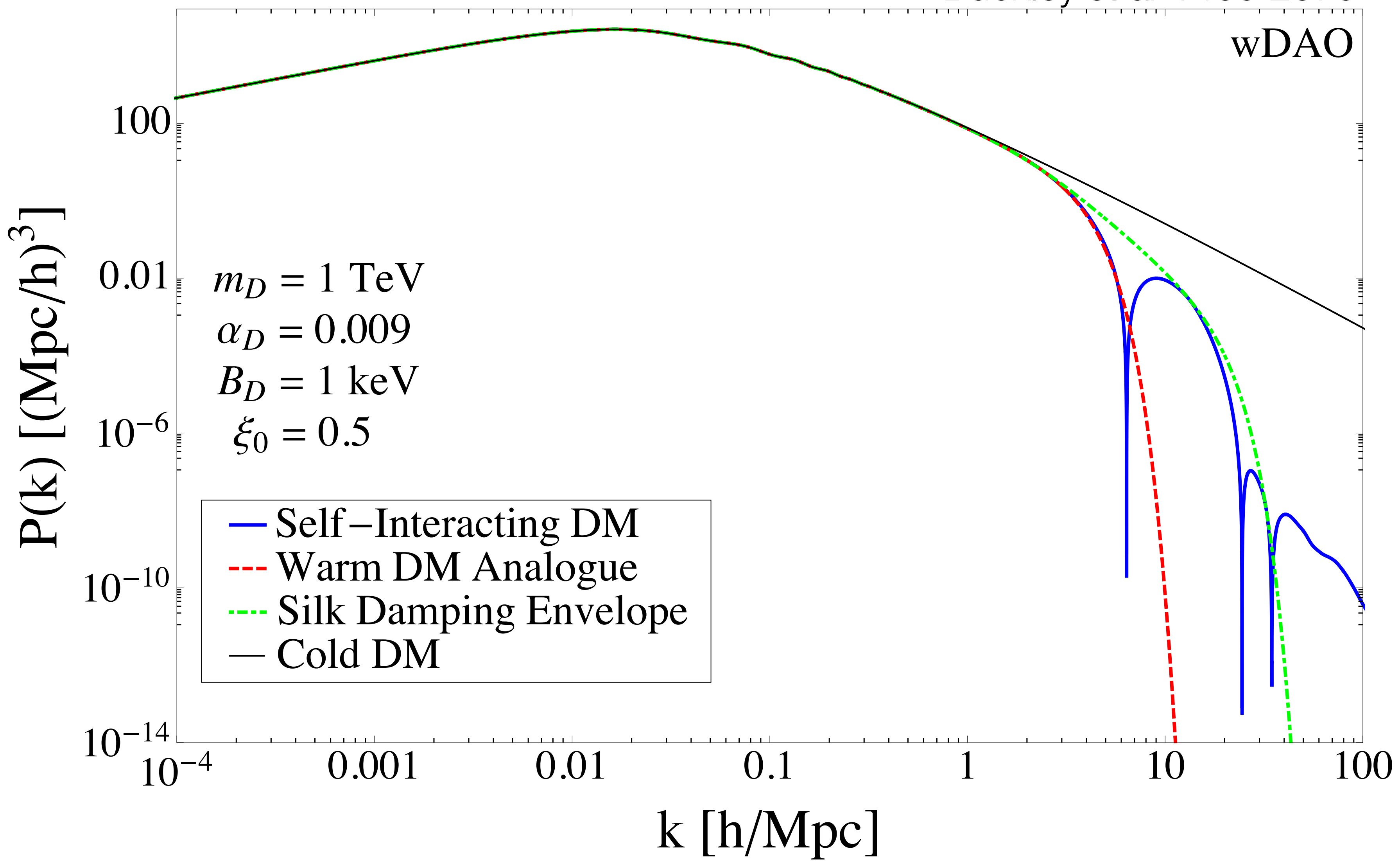
Nick Du
Aspen 2023

Theoretical Motivations

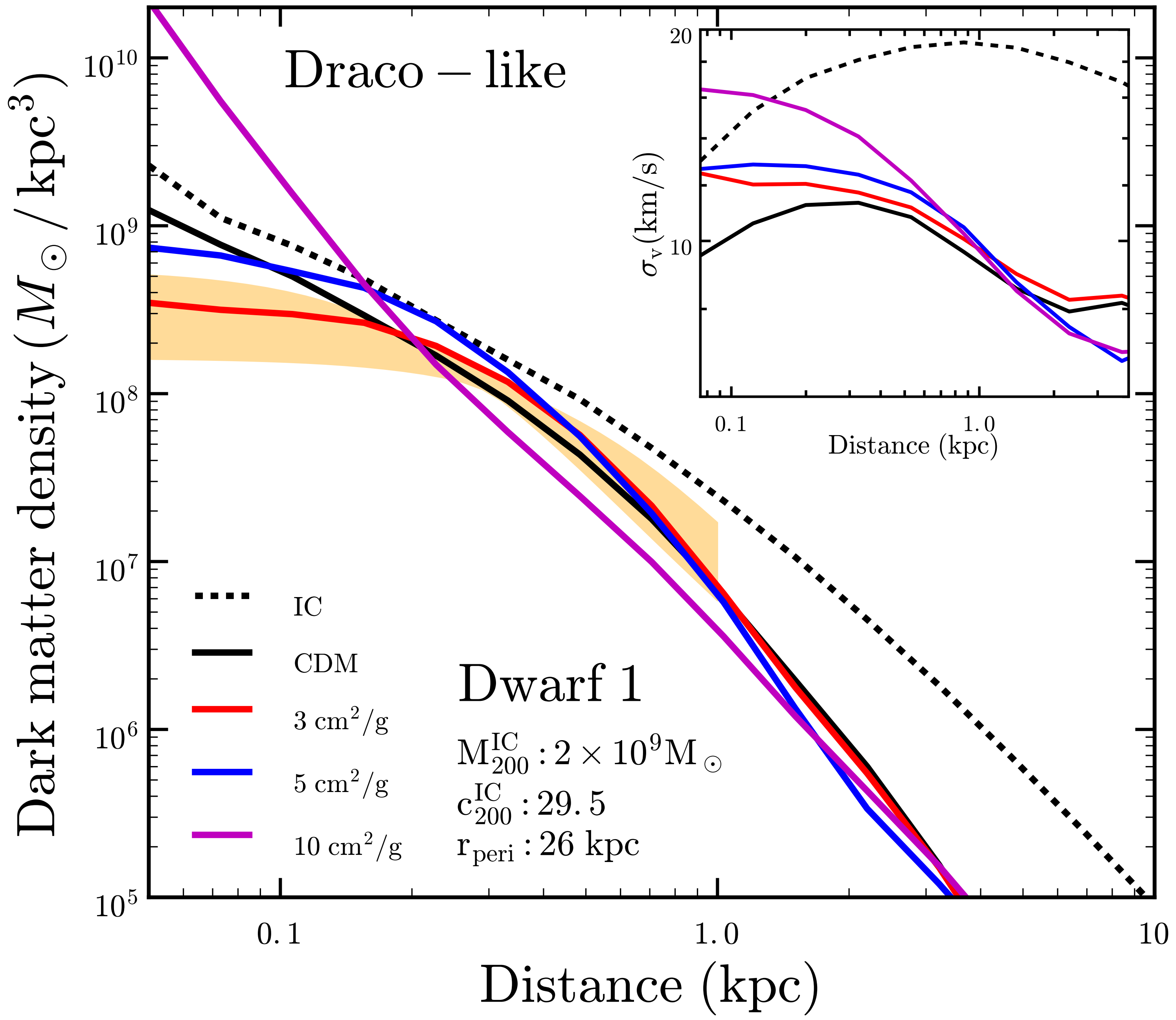
- Large-scale distribution of baryonic matter in the Universe and structure of galaxies can reveal hints of dark matter particle physics.



Buckley *et al* 1405.2075

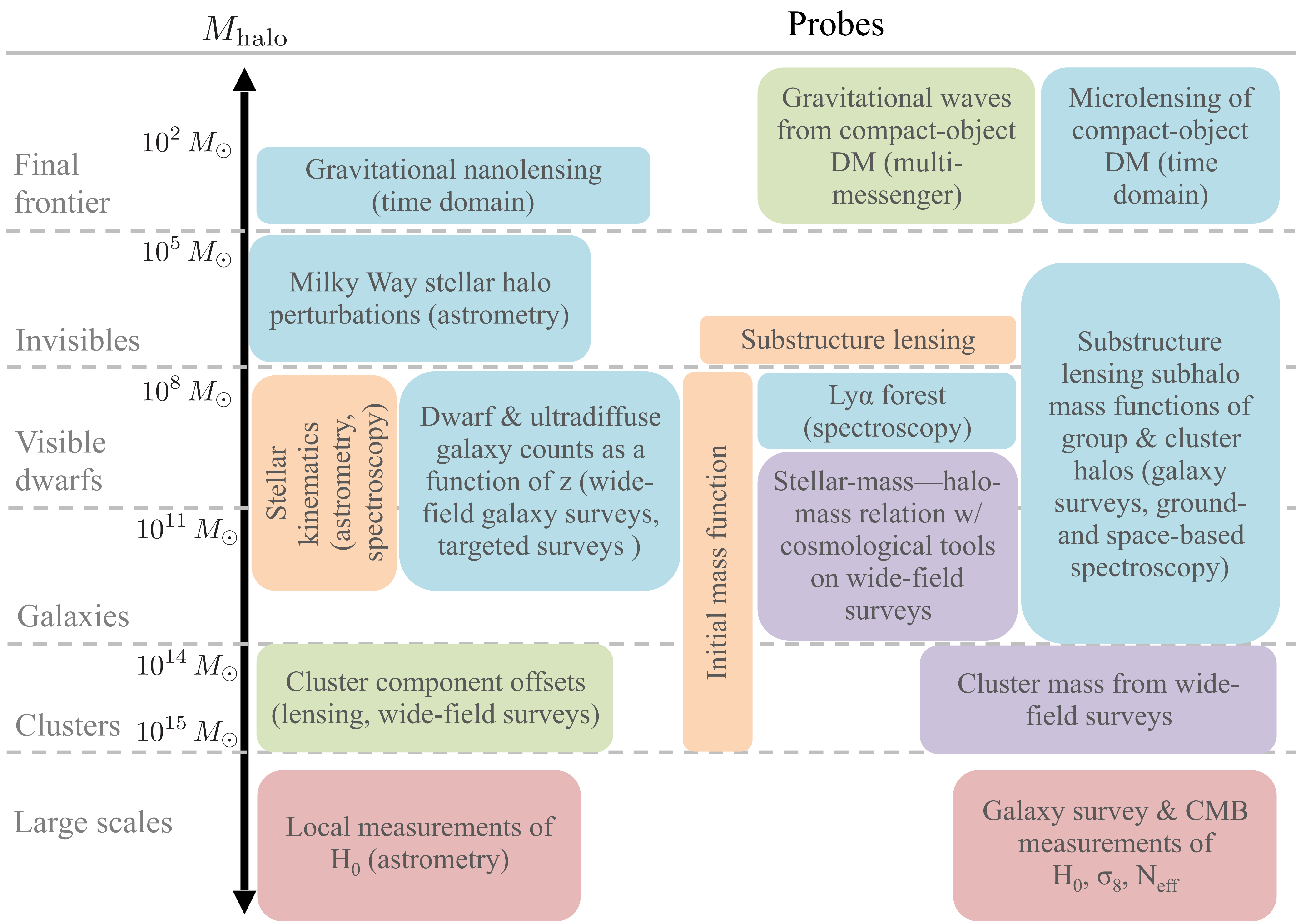


Sameie *et al* 1904.07872



Theoretical Motivations

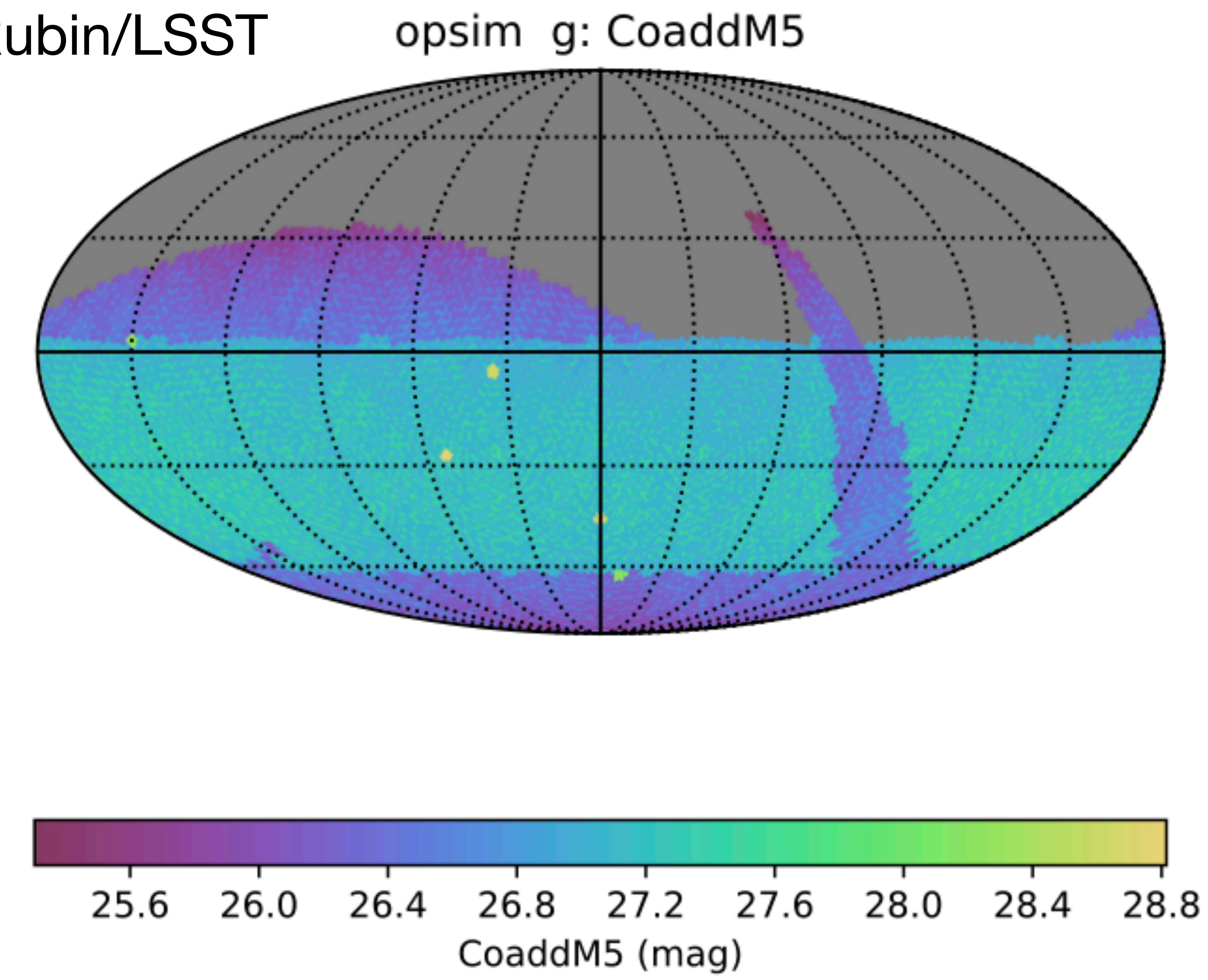
- Large-scale distribution of baryonic matter in the Universe and structure of galaxies can reveal hints of dark matter particle physics.

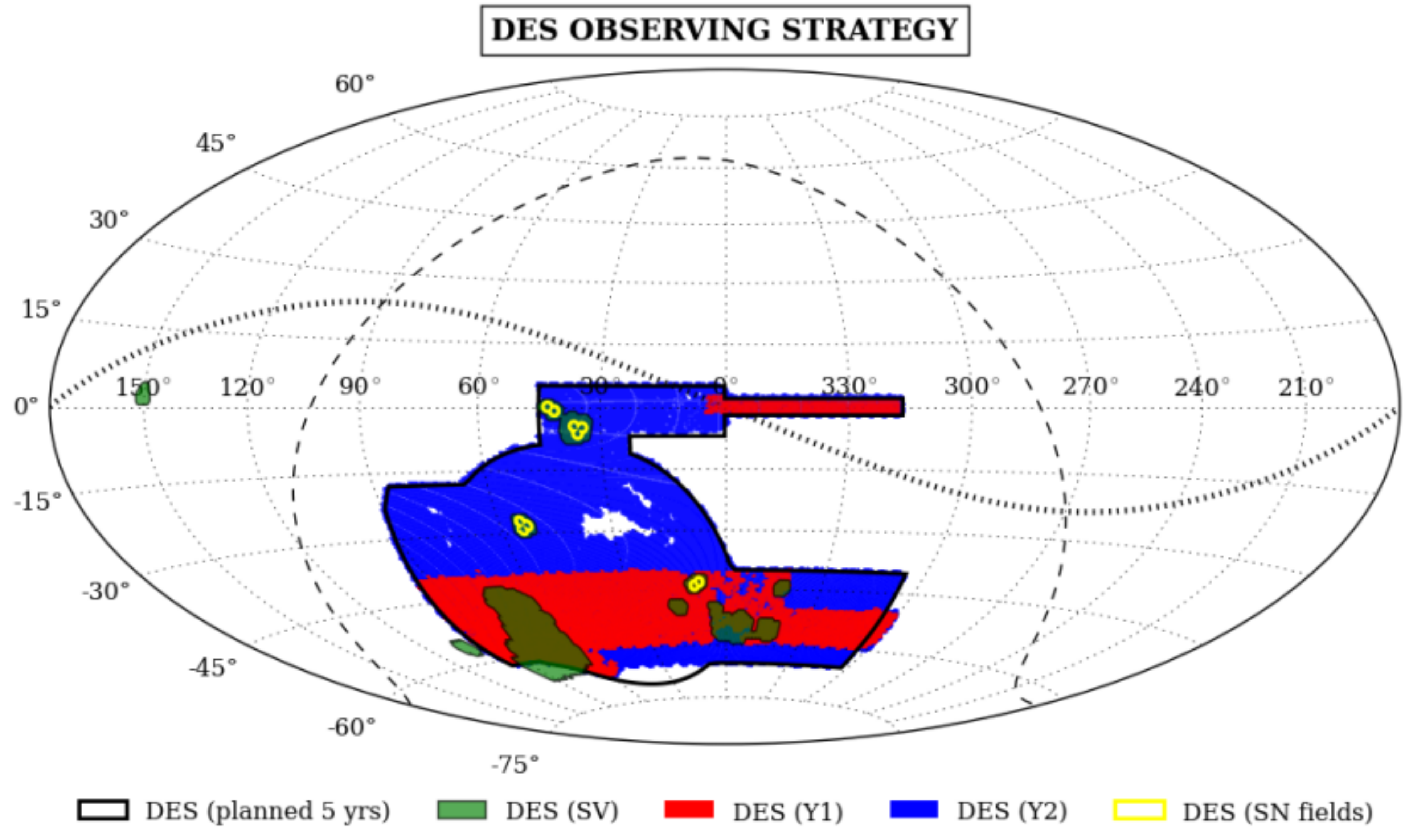


Buckley & Peter 1712.06615

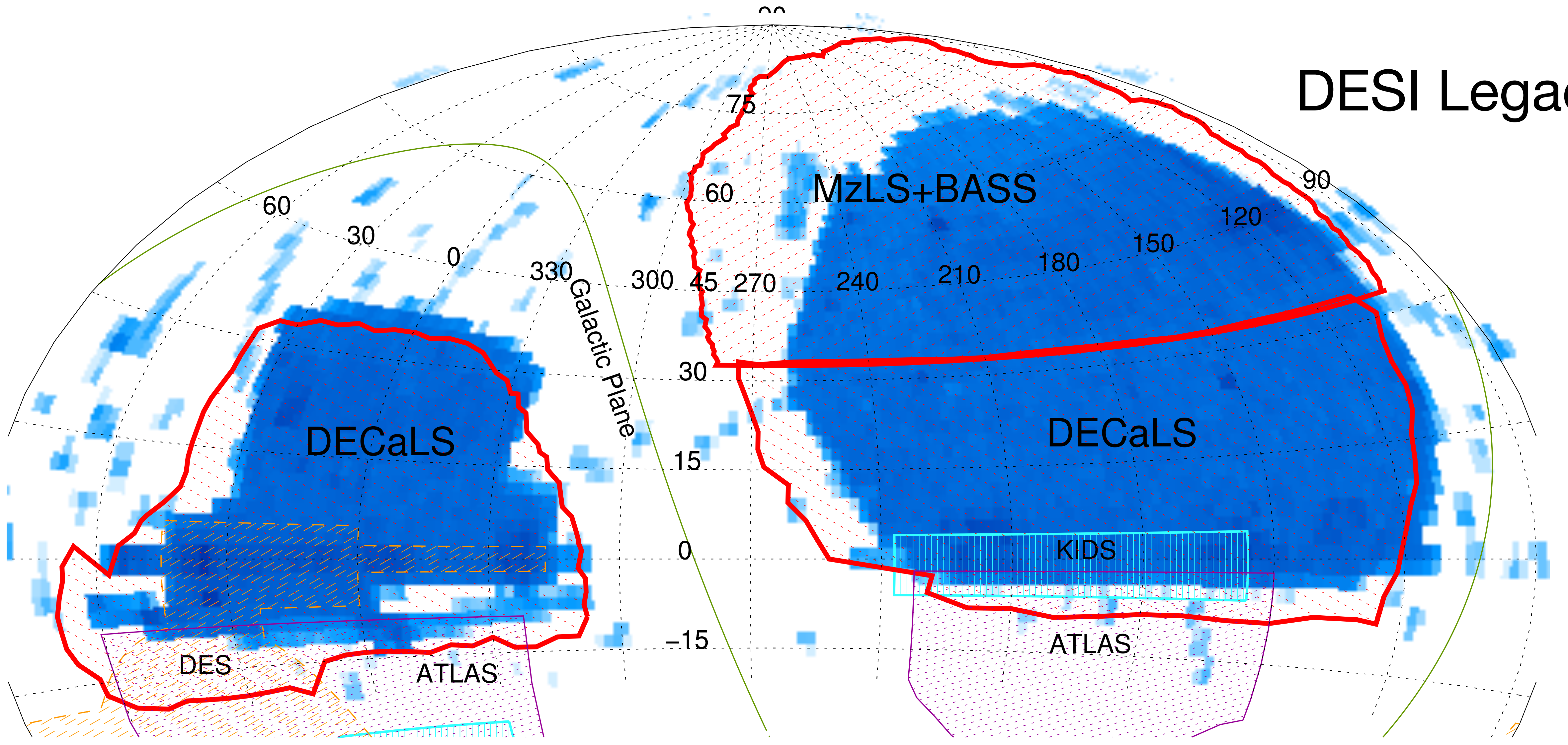
The Era of Big Astrophysical Data

Vera Rubin/LSST





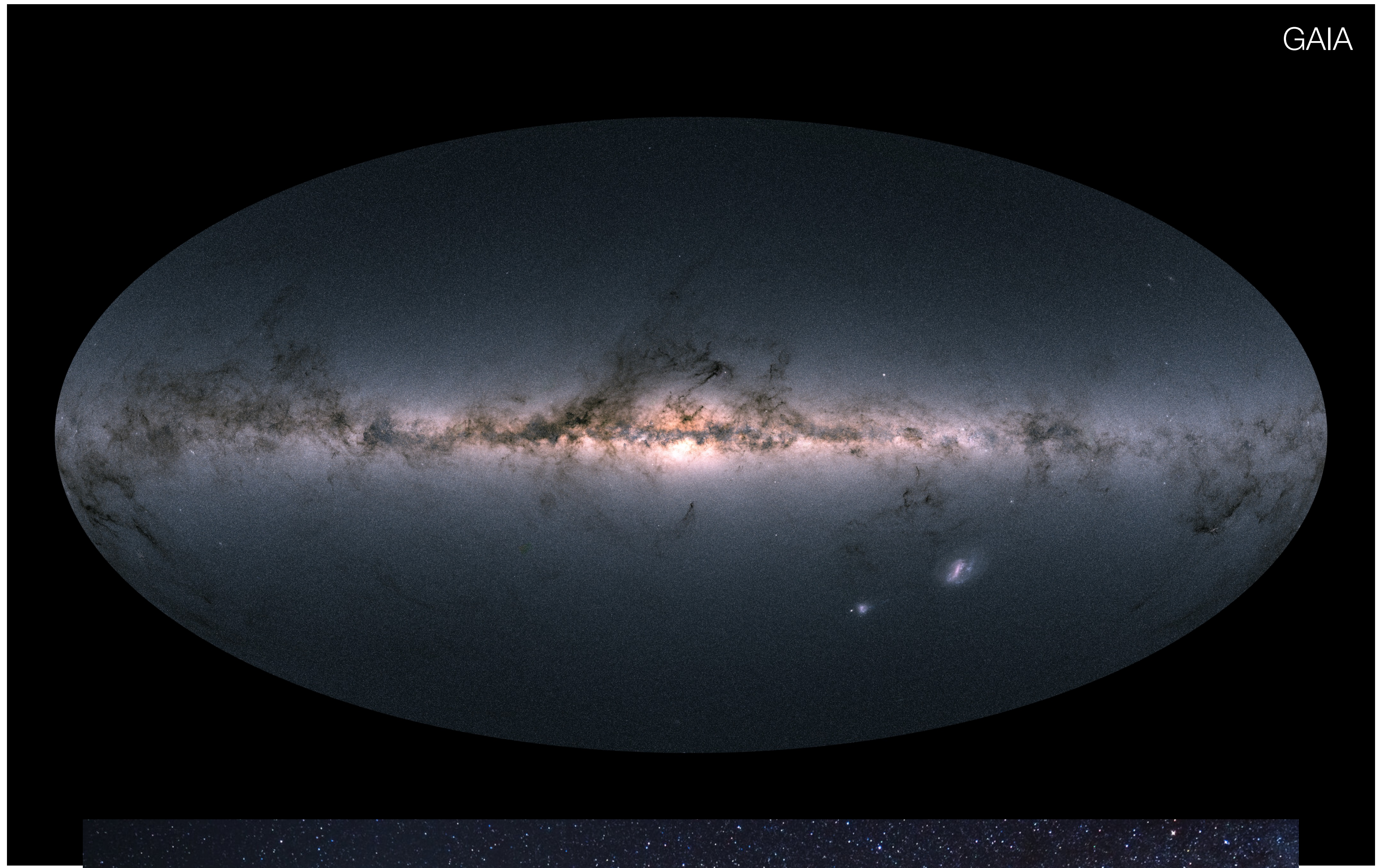
DESI Legacy

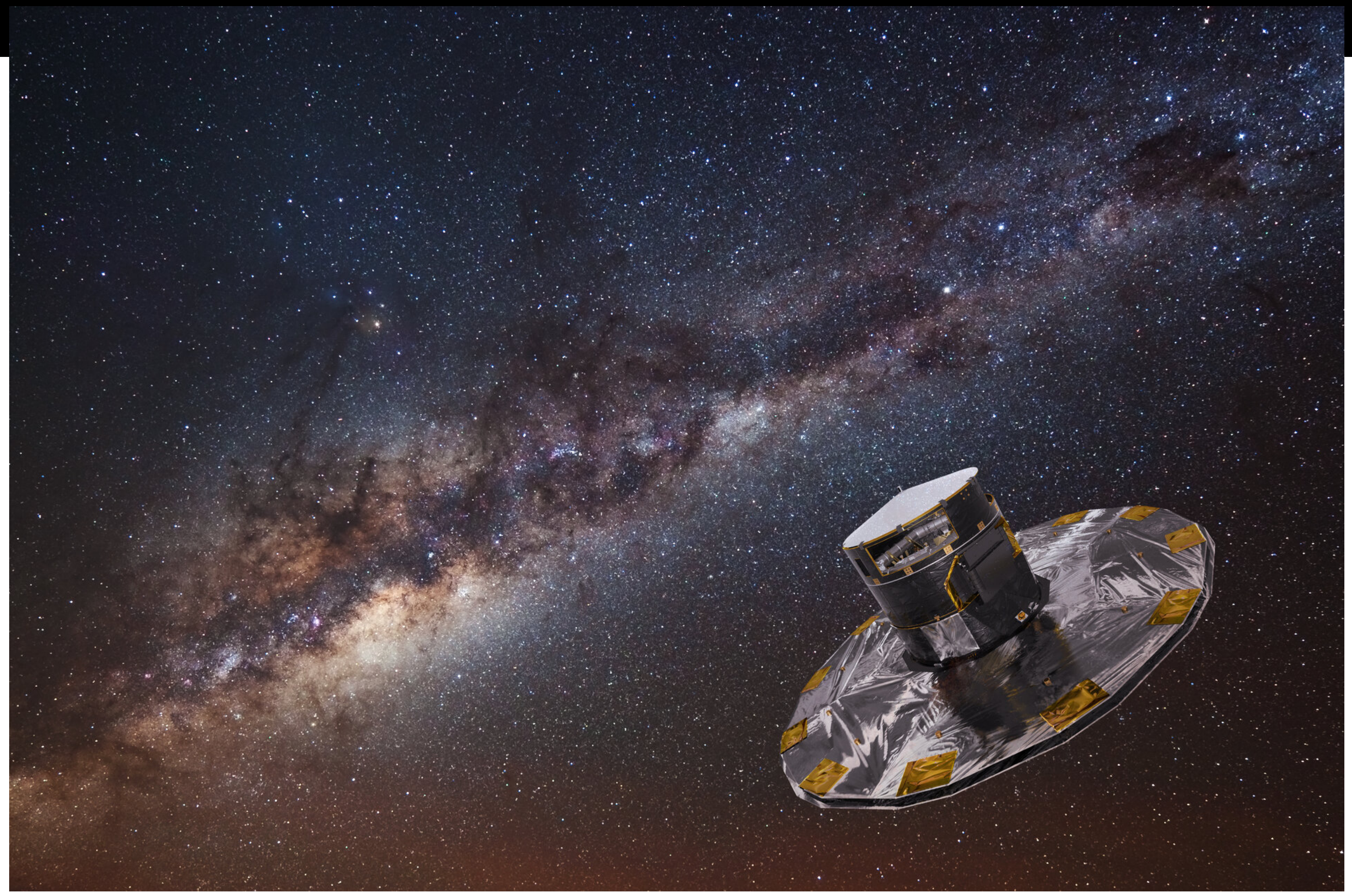


Gaia Space Telescope

- Gaia satellite measures the 3D positions and proper motions of ~1.5 billion stars in the Galaxy.
  - N.B: Gaia measures *parallax*, not *distance.*
  - Provides *photometry* (color and magnitude) and limited *spectroscopy*
  - Line-of-sight motion for ~34 million stars (DR3)
    - This will be ~150 million by end-of-mission

- A huge mine of data for the study of Galactic substructure.
- In this talk, I'm interested in Gaia data as processed locations of stars within 4/5/6D kinematic space — not as individual images/spectra (lots of analysis here!)





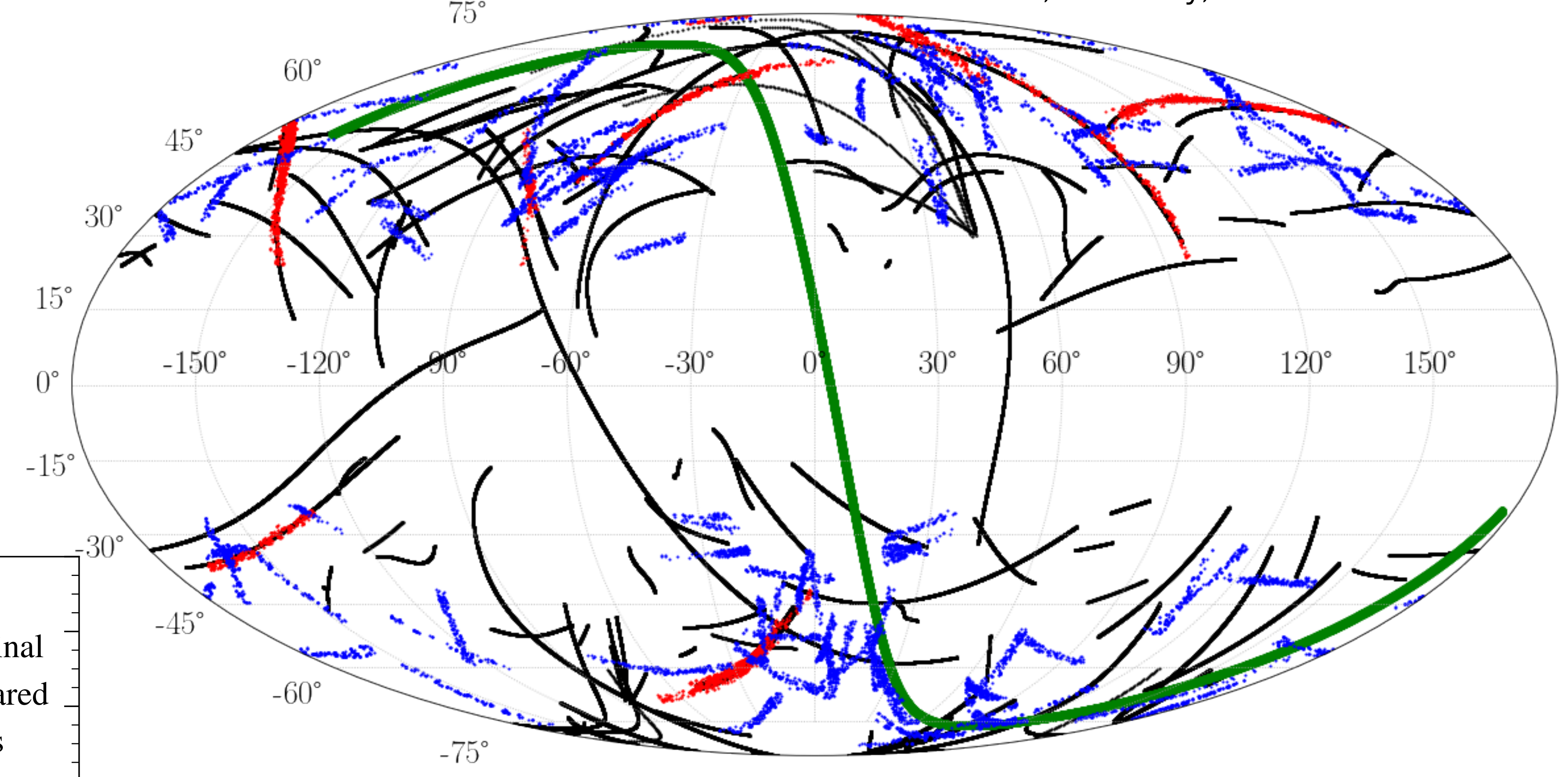
Gaia Space Telescope

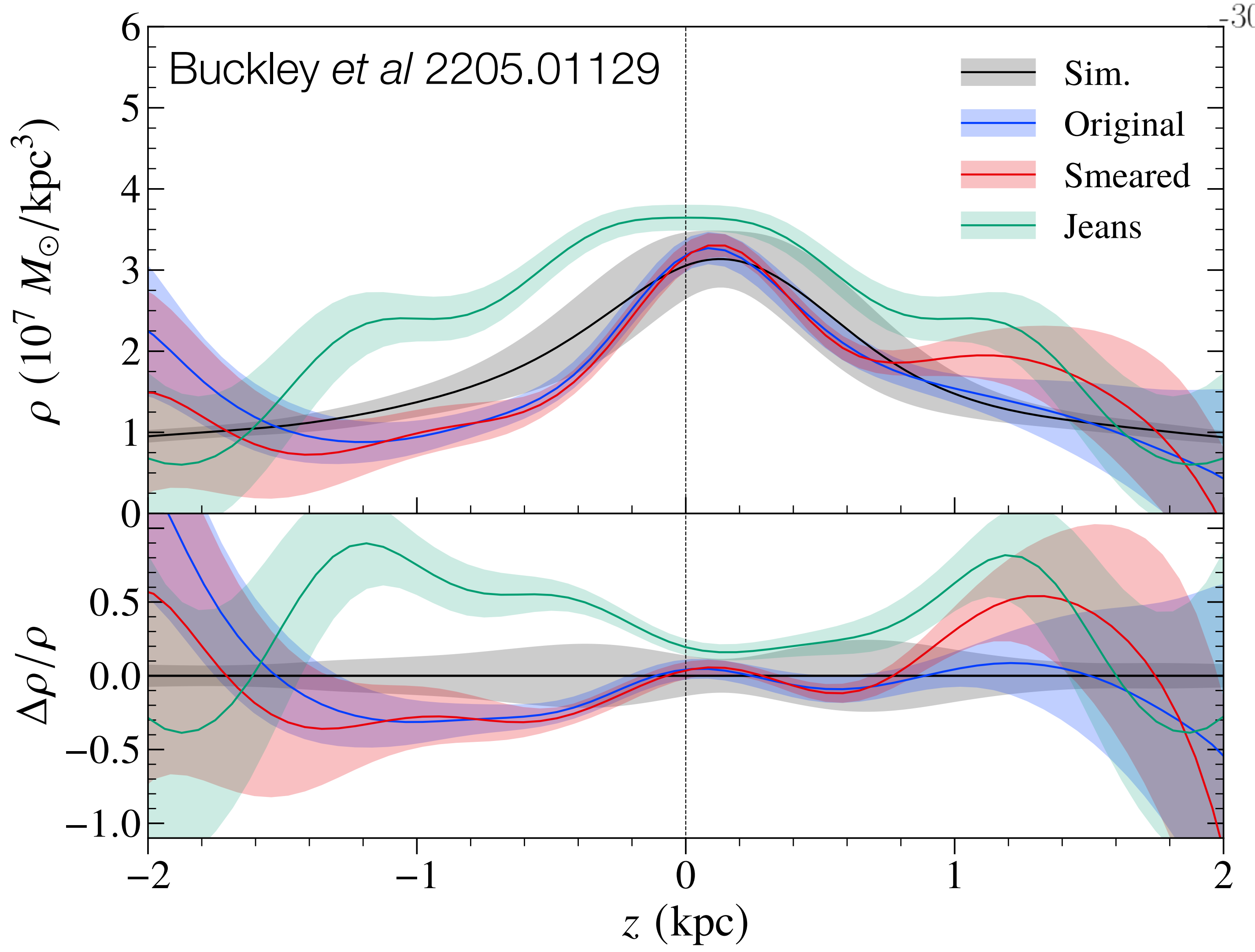
	# sources in Gaia DR3	# sources in Gaia DR2	# sources in Gaia DR1
Total number of sources	**1,811,709,771**	**1,692,919,135**	**1,142,679,769**
	Gaia Early Data Release 3		
Number of sources with full astrometry	1,467,744,818	1,331,909,727	2,057,050
Number of 5-parameter sources	585,416,709		
Number of 6-parameter sources	882,328,109		
Number of 2-parameter sources	343,964,953	361,009,408	1,140,622,719
Gaia-CRF sources	1,614,173	556,869	2191
Sources with mean G magnitude	1,806,254,432	1,692,919,135	1,142,679,769
Sources with mean G_{BP}-band photometry	1,542,033,472	1,381,964,755	-
Sources with mean G_{RP}-band photometry	1,554,997,939	1,383,551,713	-
	New in Gaia Data Release 3	Gaia DR2	Gaia DR1
Sources with radial velocities	33,812,183	7,224,631	-
Sources with mean G_{RVS}-band magnitudes	32,232,187	-	-
Sources with rotational velocities	3,524,677	-	-
Mean BP/RP spectra	219,197,643	-	-
Mean RVS spectra	999,645	-	-

ML Applications for Gaia

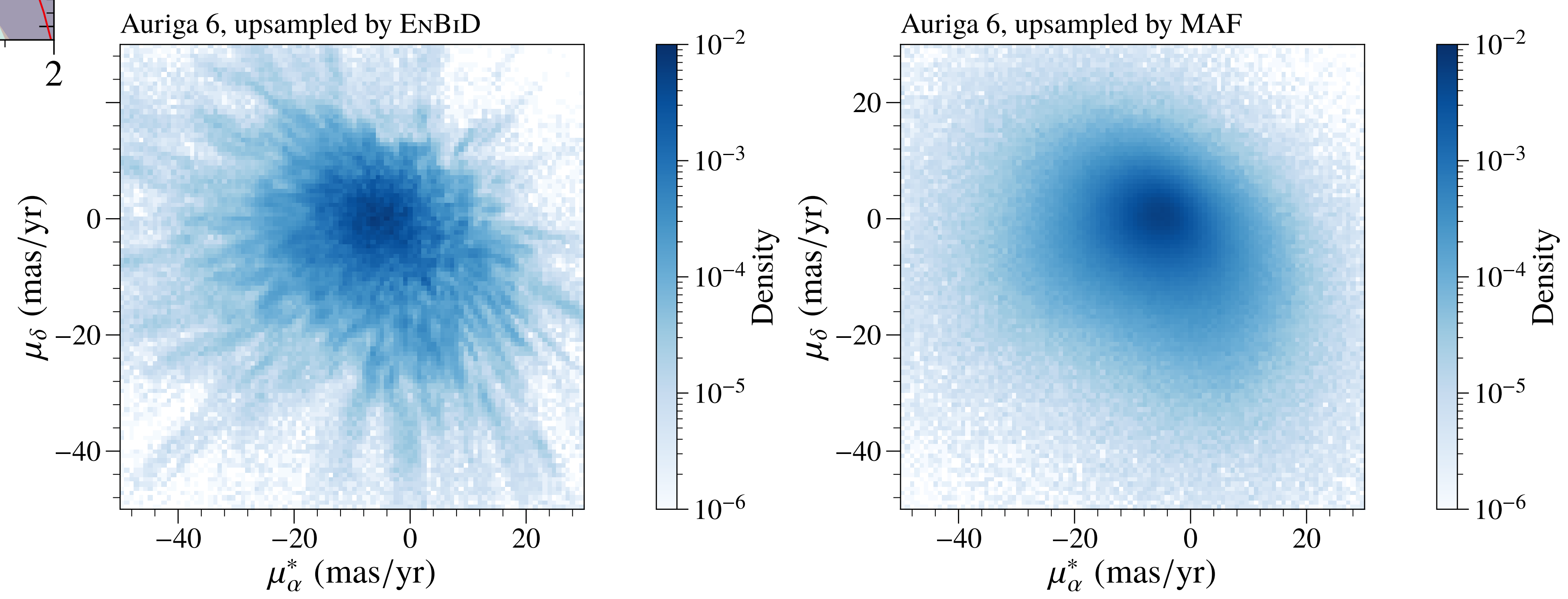
- The Milky Way's Mass Density
- Stellar Streams
  - Via Machinae (ANODE)
  - CATHODE
- Synthetic *Gaia* Observations

Shih, Buckley, Necib 2303.01529





Lim *et al* 2211.11765



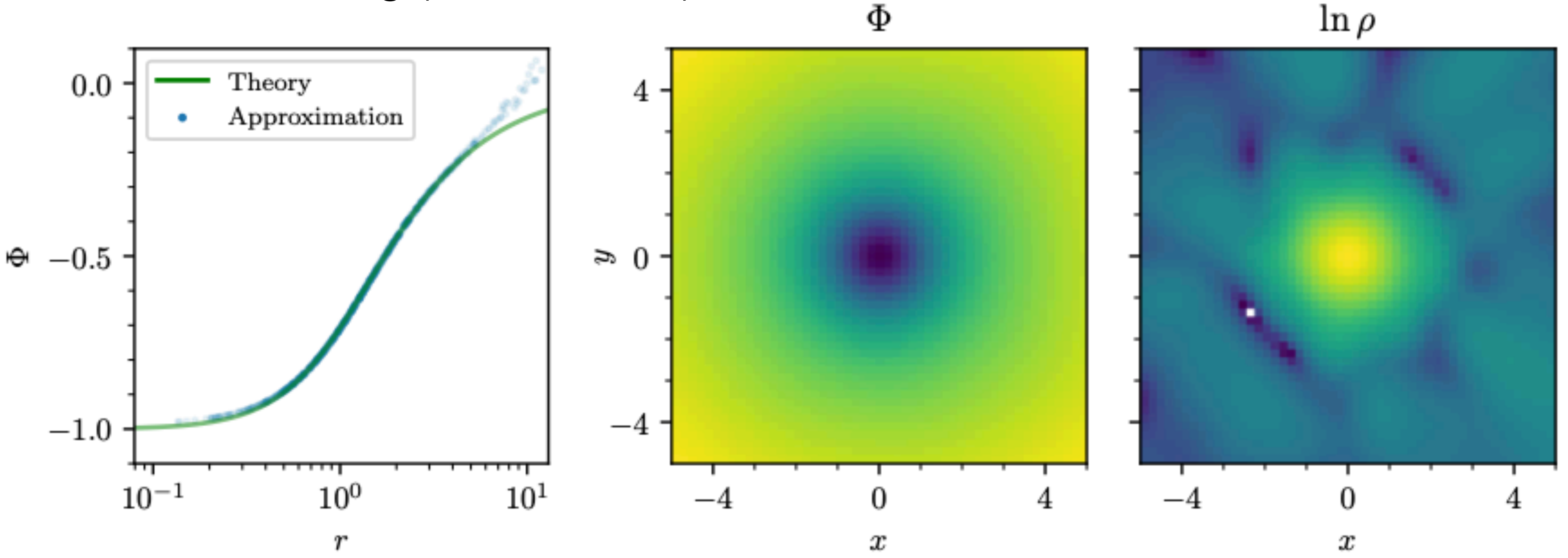
Dark Matter Density from Gaia

- The phase space density of stars in equilibrium is related to the underlying Galactic potential

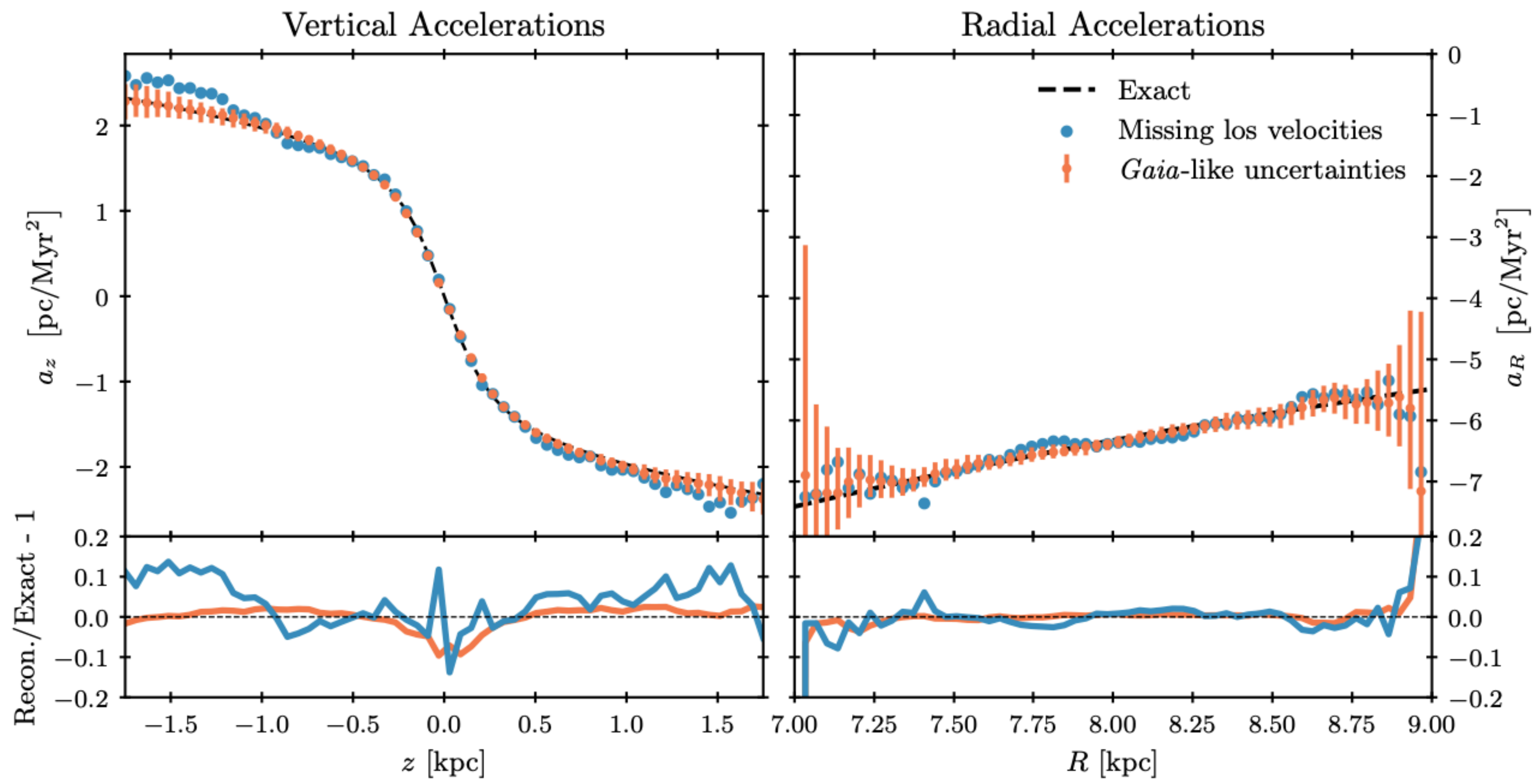
$$\frac{\partial f}{\partial t} + v_i \frac{\partial f}{\partial x_i} = \frac{\partial \Phi}{\partial x_i} \frac{\partial f}{\partial v_i}$$

- Curse of dimensionality makes it very hard to measure f and derivatives from stellar motions. Traditionally, take moments of the Boltzmann Equation and assume symmetries
- Normalizing flows can do a much better job in estimating f and its derivatives from the available data.

Green & Ting (2011.04673)



An *et al* (2106.05981) and Naik *et al* (2112.07657)

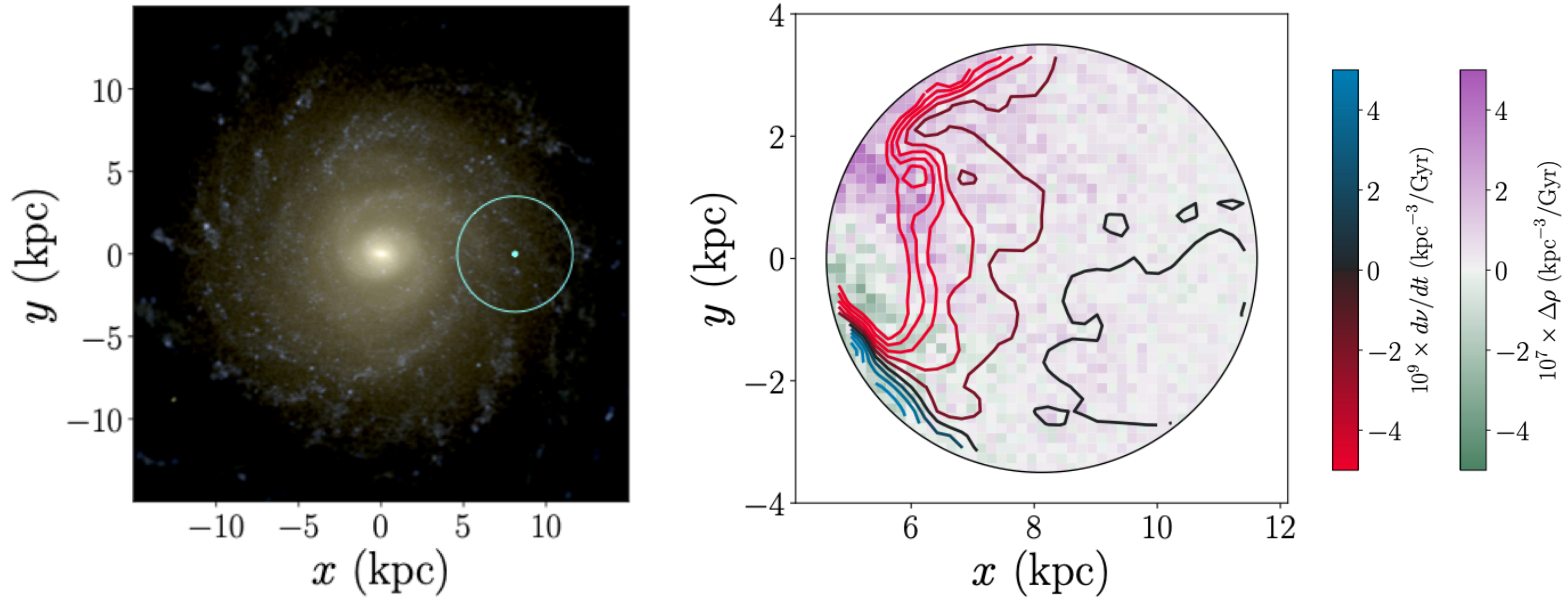


Dark Matter Density from Gaia

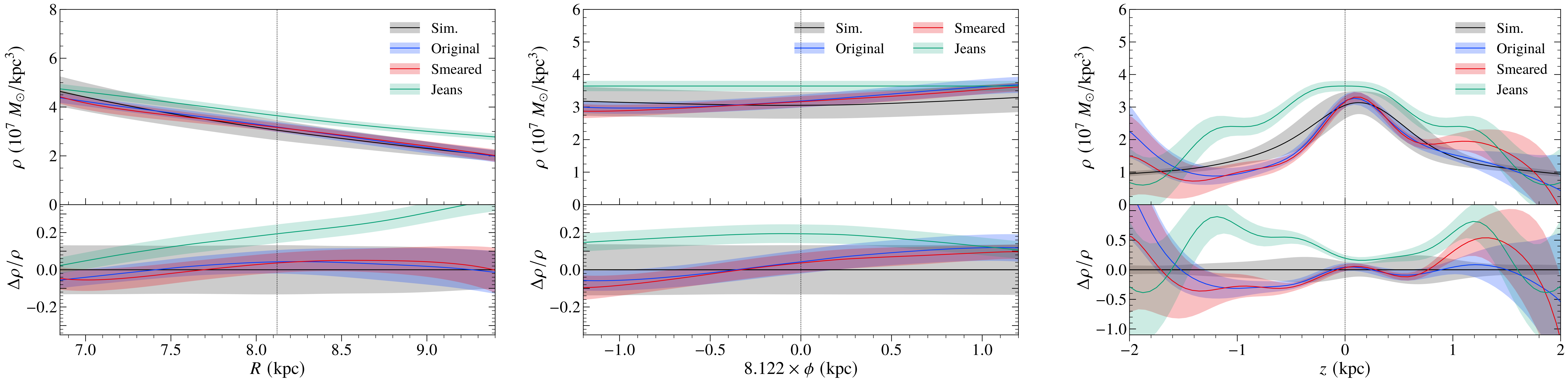
- The real Galaxy is not in equilibrium:

$$\frac{\partial f}{\partial t} \neq 0$$

- Is real data sufficiently precise to get good estimates of f?
- First with a simulated Milky Way-like galaxy:



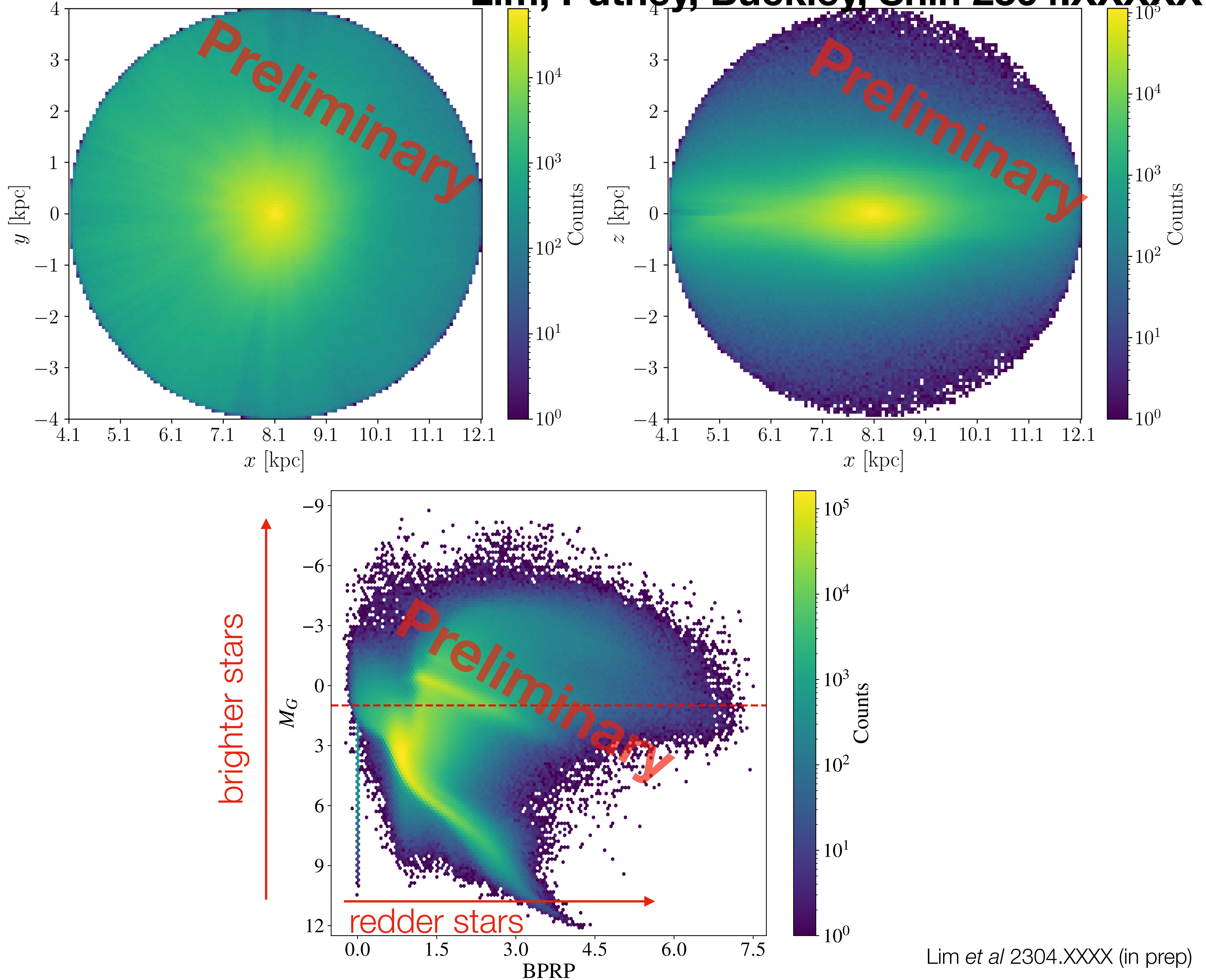
Buckley, Lim, Putney, Shih 2205.01129



Dark Matter Density from Gaia

- Can we do this with real Gaia data?
- Real data is complicated:
  - Observations are not complete, and this completeness varies as a function of distance
  - And with which kinematic parameters are measured, and/or stellar properties
- The goal: get low-error measurements off of the Galactic disk, to regions where dark matter dominates the mass density.

Lim, Putney, Buckley, Shih 2304.XXXXX

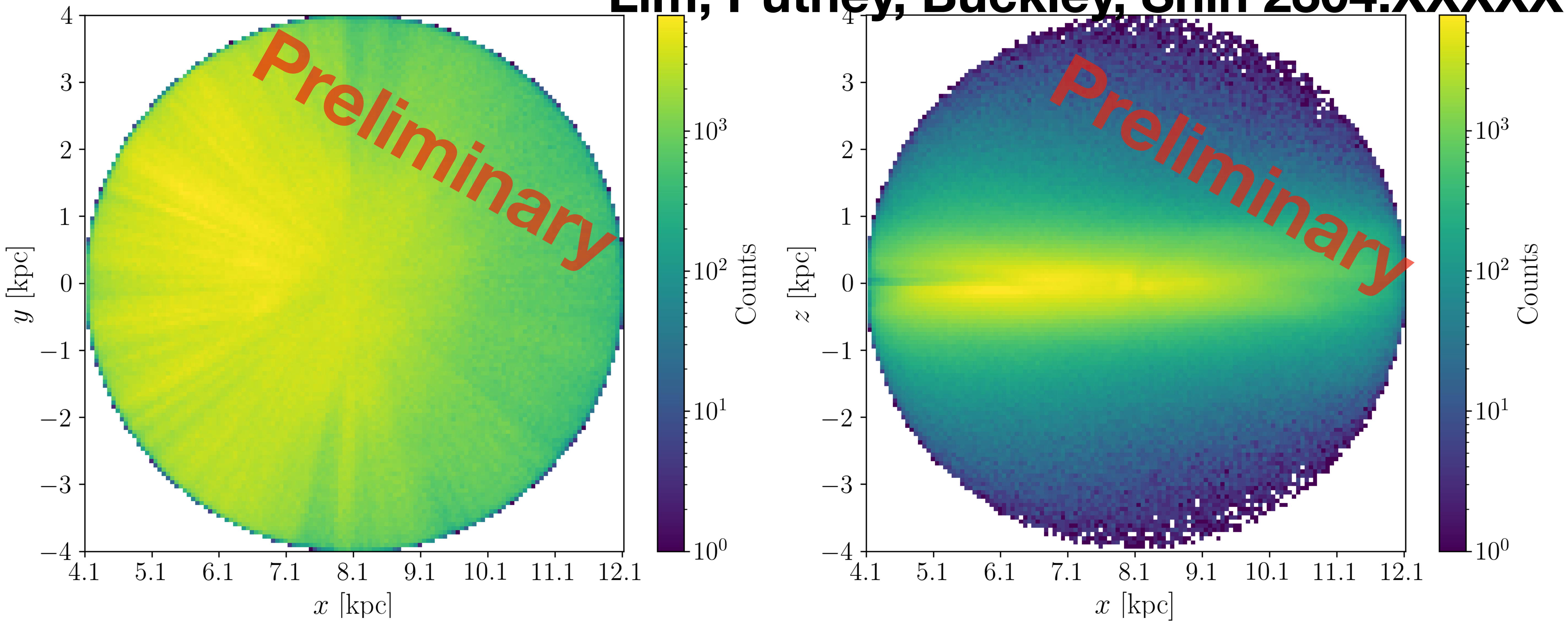


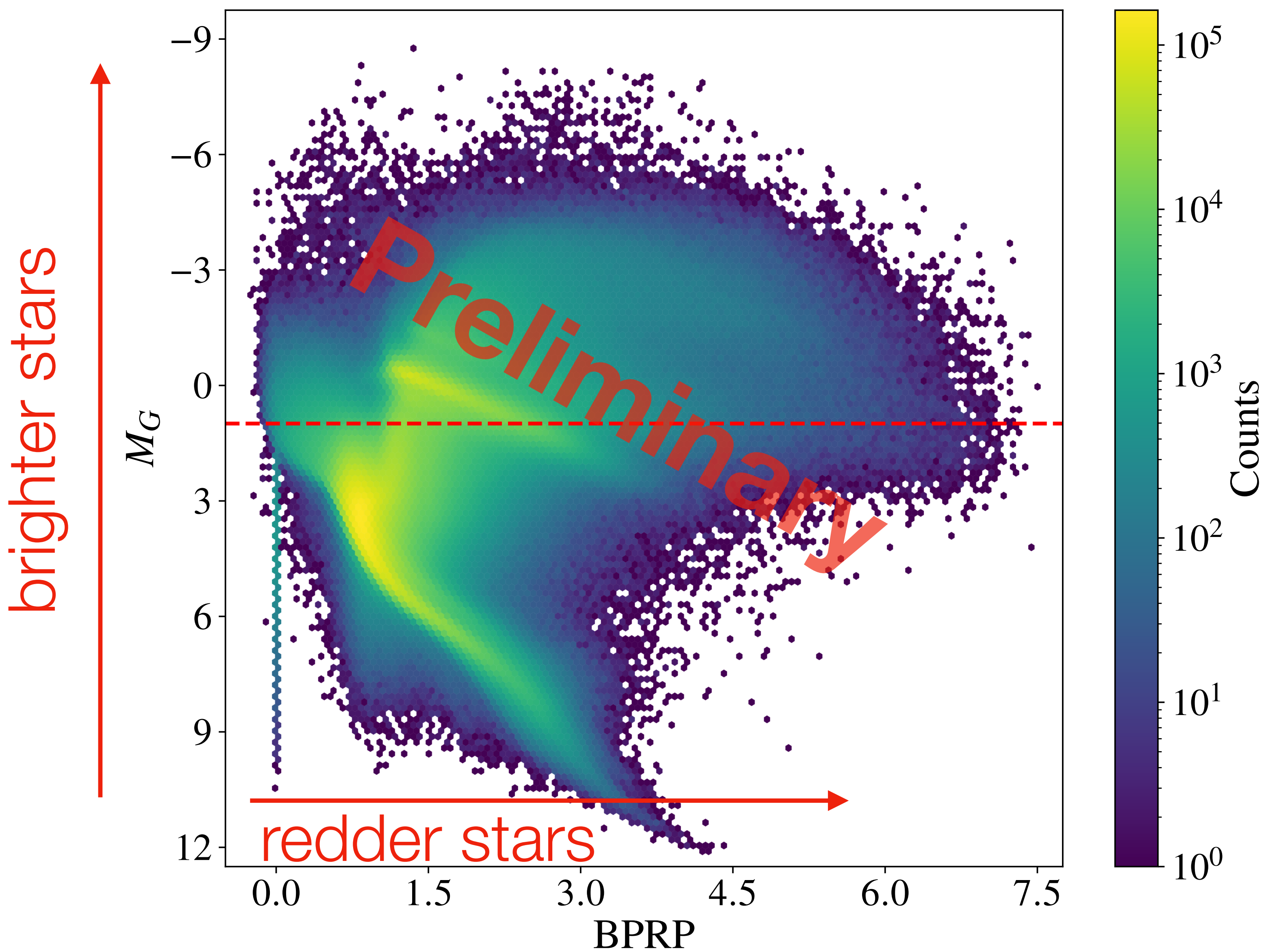
Lim *et al* 2304.XXXX (in prep)

Dark Matter Density from Gaia

- Can we do this with real Gaia data?
- Real data is complicated:
    - Observations are not complete, and this completeness varies as a function of distance
    - And with which kinematic parameters are measured, and/or stellar properties
- The goal: get low-error measurements off of the Galactic disk, to regions where dark matter dominates the mass density.

Lim, Putney, Buckley, Shih 2304.XXXXX





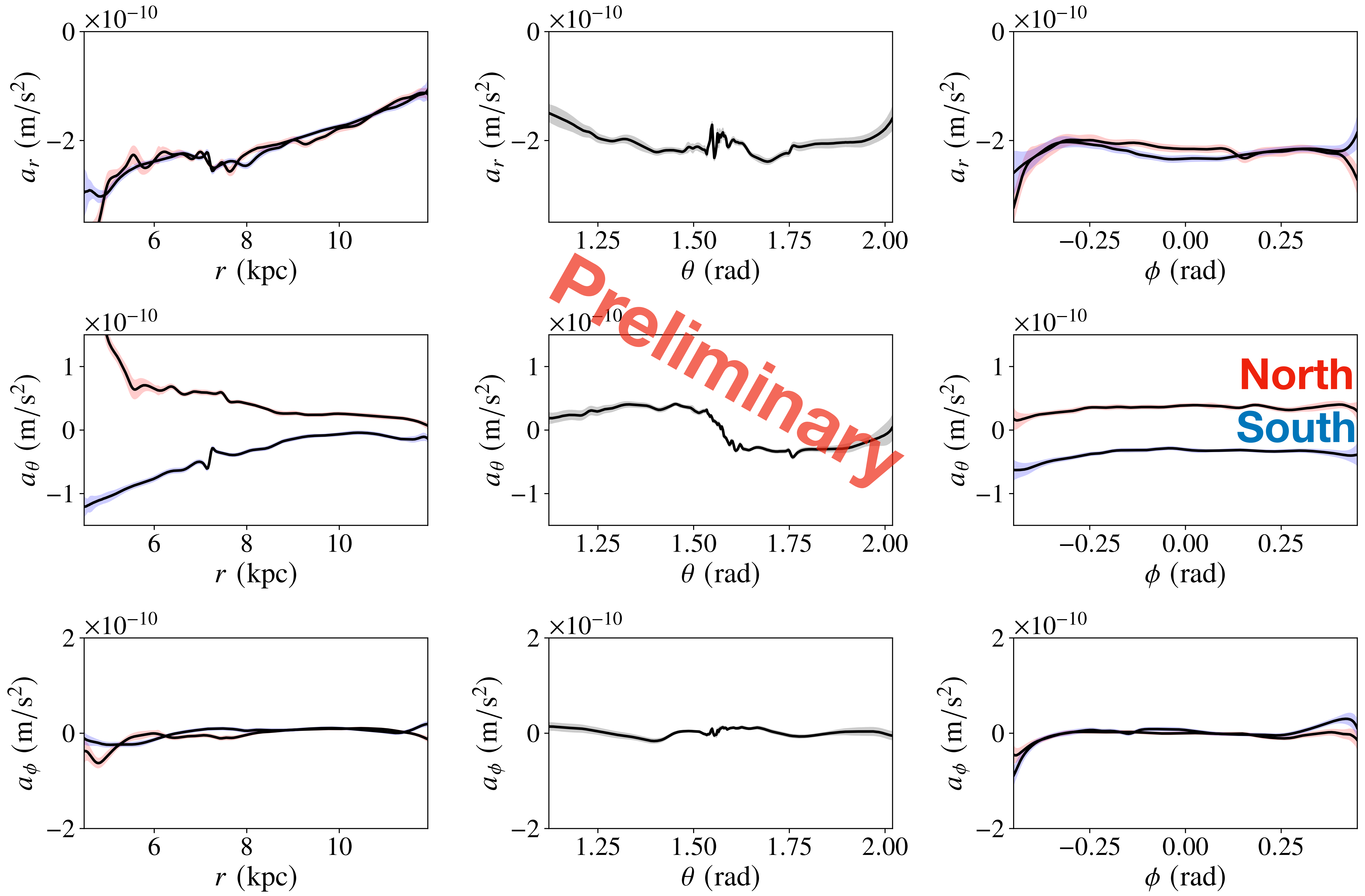
Lim *et al* 2304.XXXX (in prep)

Dark Matter Density from Gaia

- 1st: Calculate accelerations:

$$v_i \frac{\partial f}{\partial x_i} = \frac{\partial \Phi}{\partial x_i} \frac{\partial f}{\partial v_i}$$

  - Errors include multiple MAFs, bootstrap, measurement errors

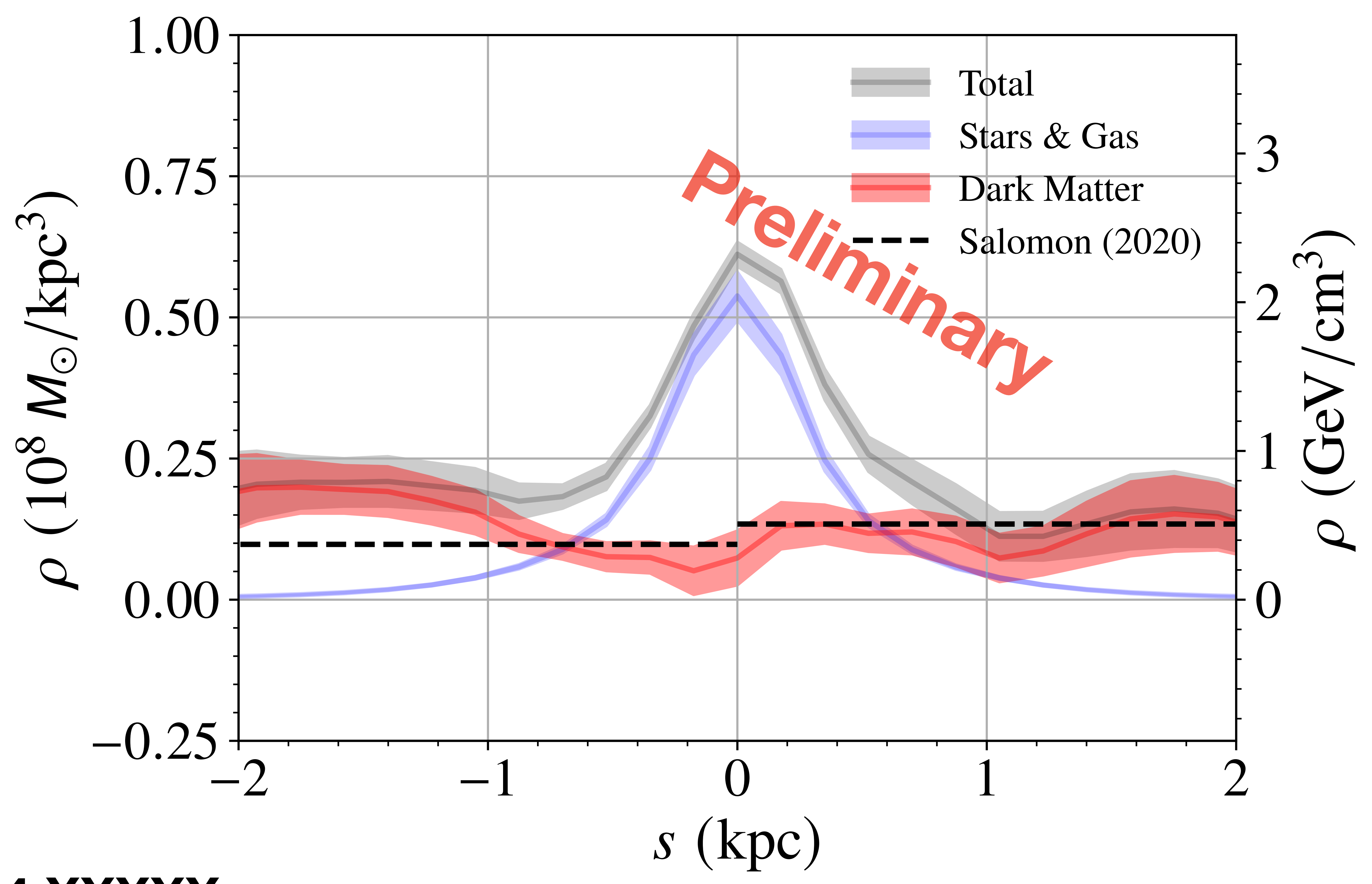


Lim, Putney, Buckley, Shih 2304.XXXXX

Dark Matter Density from Gaia

- Next, calculate mass density by integration by parts over a truncated Gaussian kernel
  - Baryonic model is a major source of uncertainty at the Solar location. Much less important away from the disk

$$\nabla^2\Phi = 4\pi G\rho$$

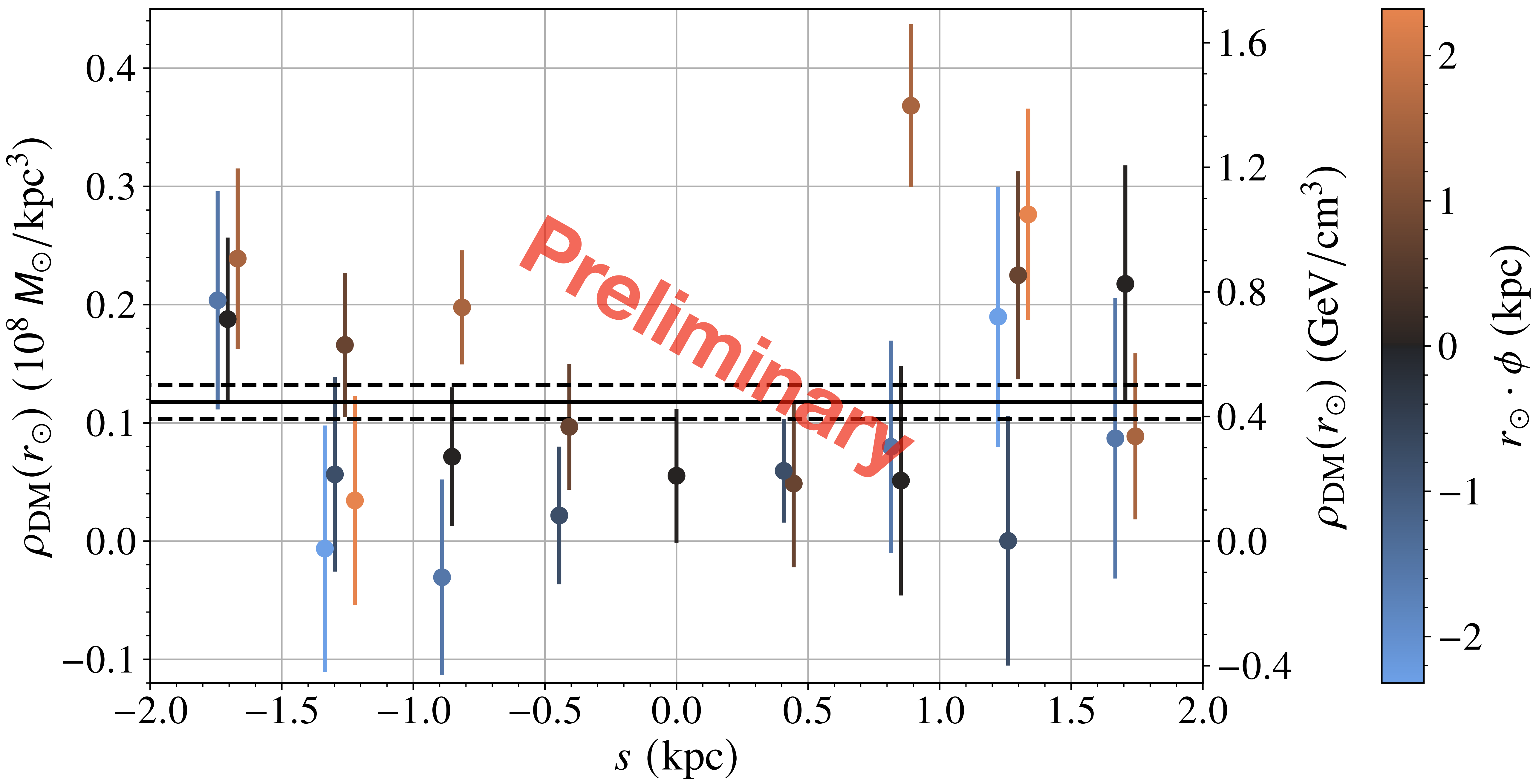


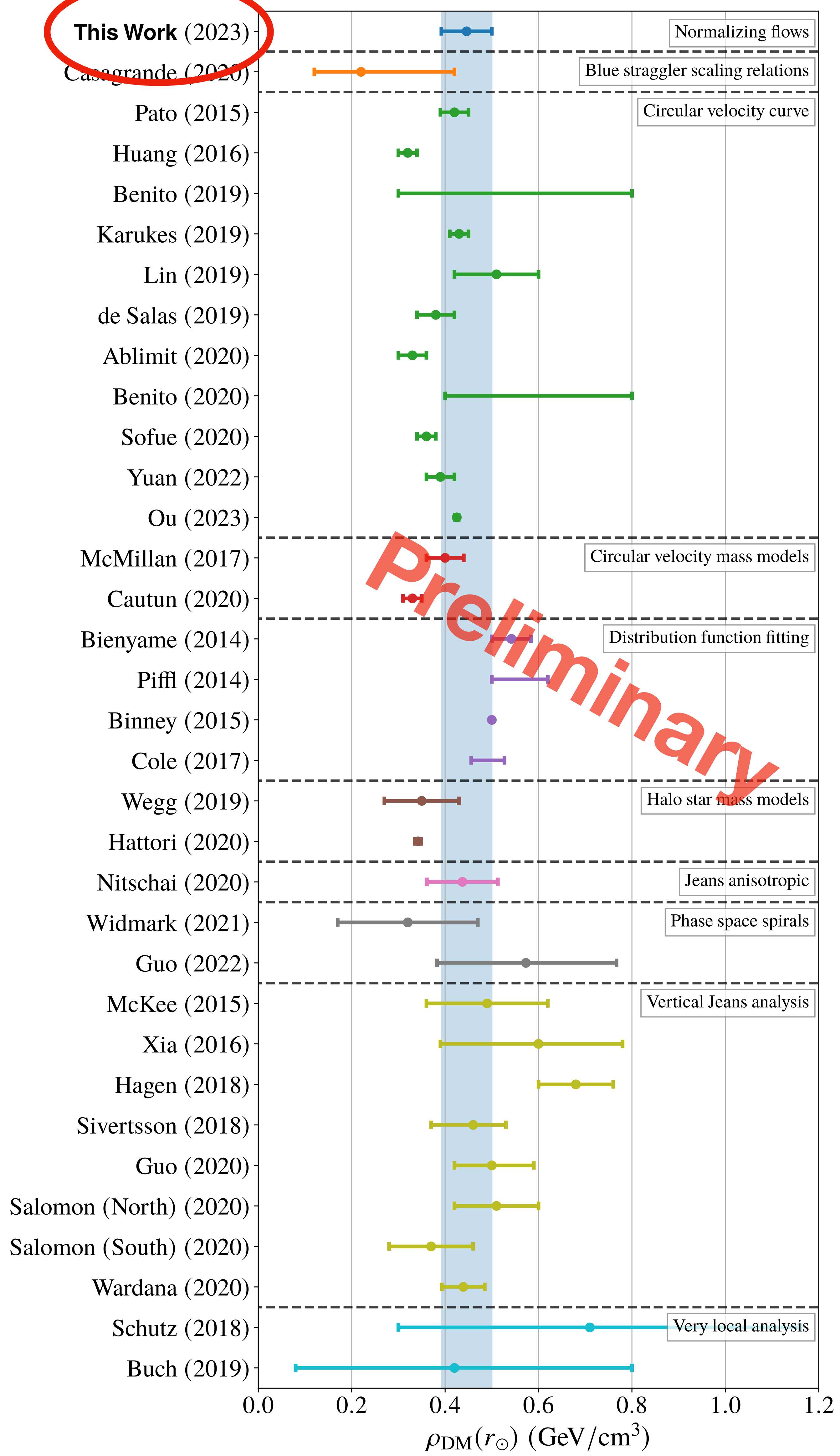
Lim, Putney, Buckley, Shih 2304.XXXXX

Dark Matter Density from Gaia

- Next, calculate mass density using finite differences (averaging over truncated Gaussian kernel)
  - Baryonic model is a major source of uncertainty at the Solar location. Much less important away from the disk

$$\rho_{\rm DM}(r = r_\odot) = 0.446 \pm 0.054\ {\rm GeV/cm^3}$$

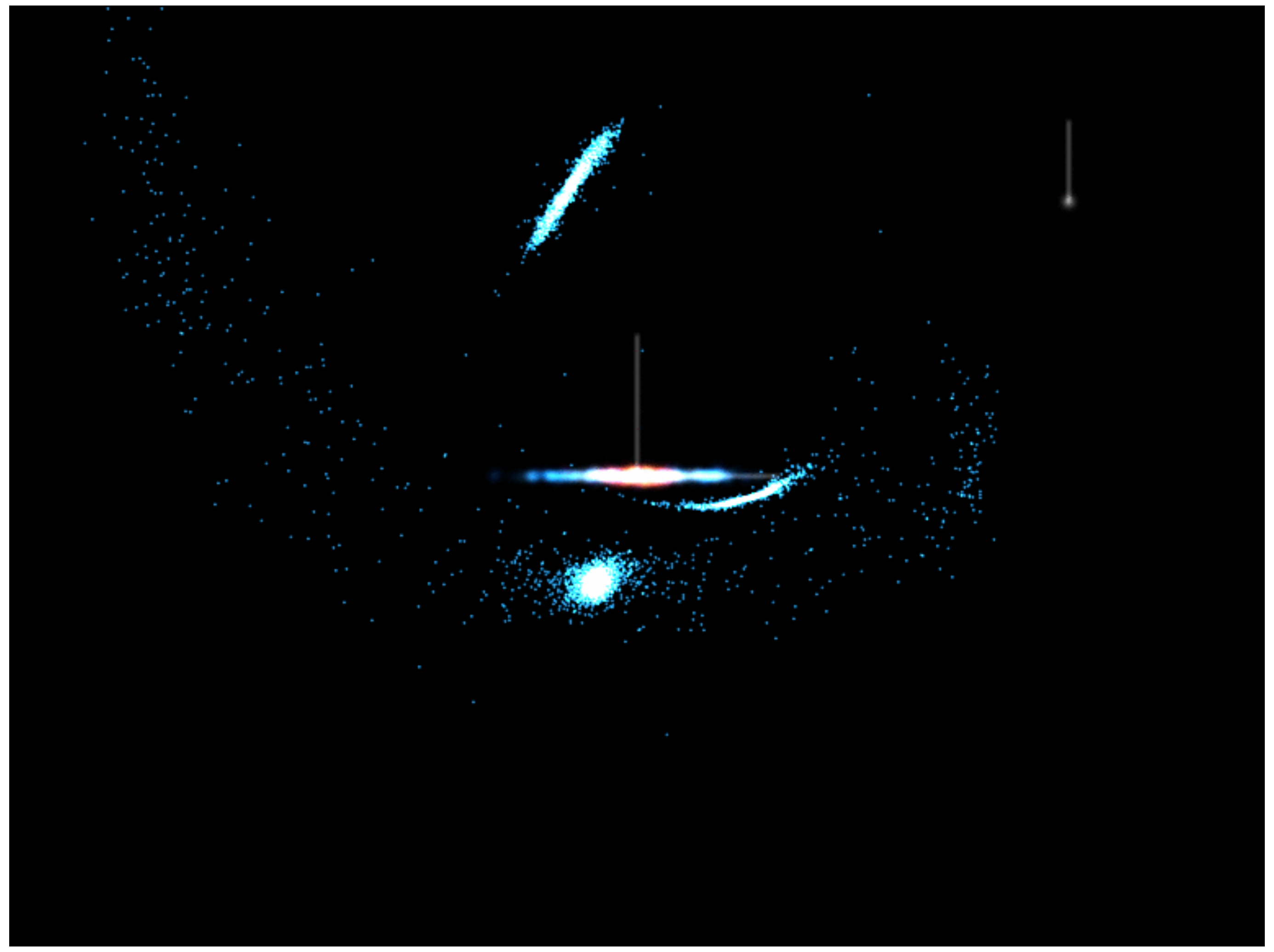




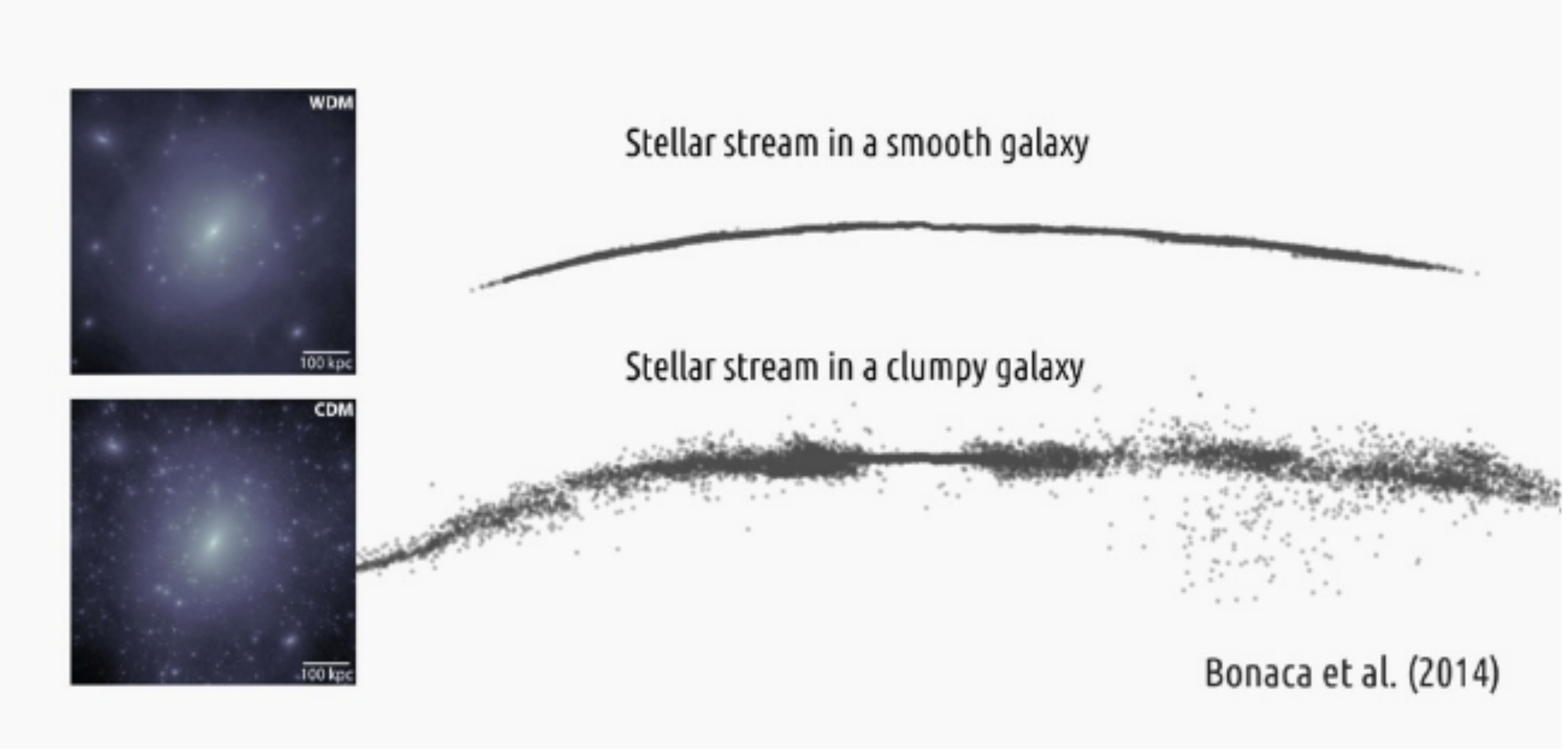
Lim, Putney, Buckley, Shih 2304.XXXXX

Mergers and 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**, then become **tidal debris**, before becoming completely mixed.

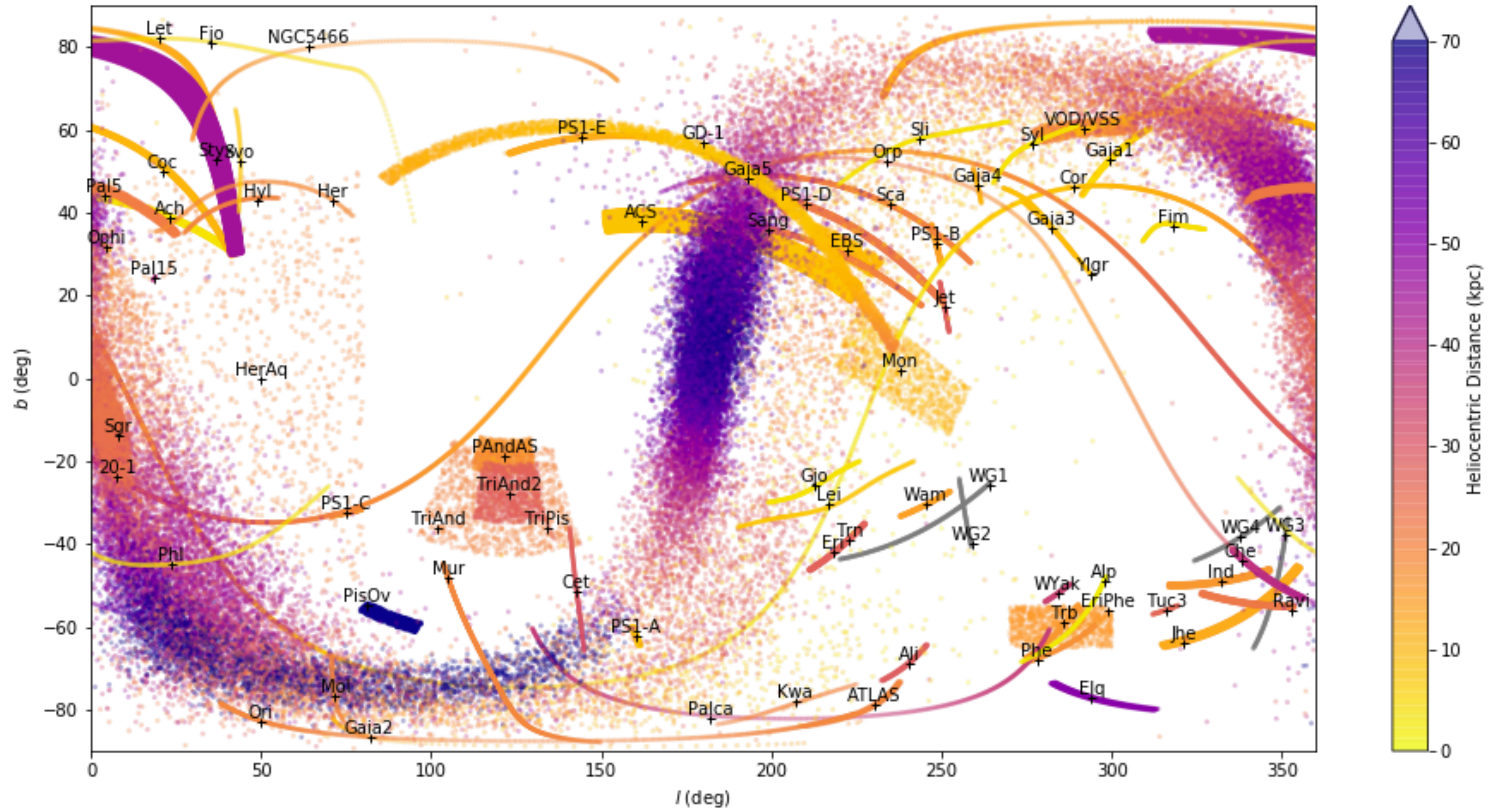


- Streams provide a probe into the Galactic potential through the stream's orbit.
  - Can reveal dark matter substructure through gravitational interactions with the stream itself.
- Both streams and debris give a glimpse into the Galaxy's merger history.

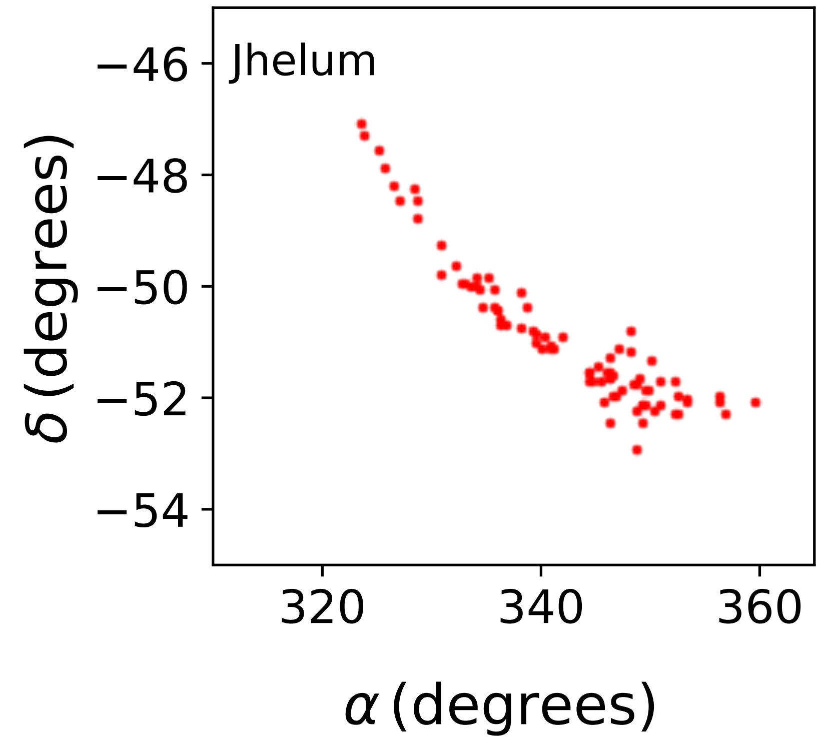


Finding Stellar Streams

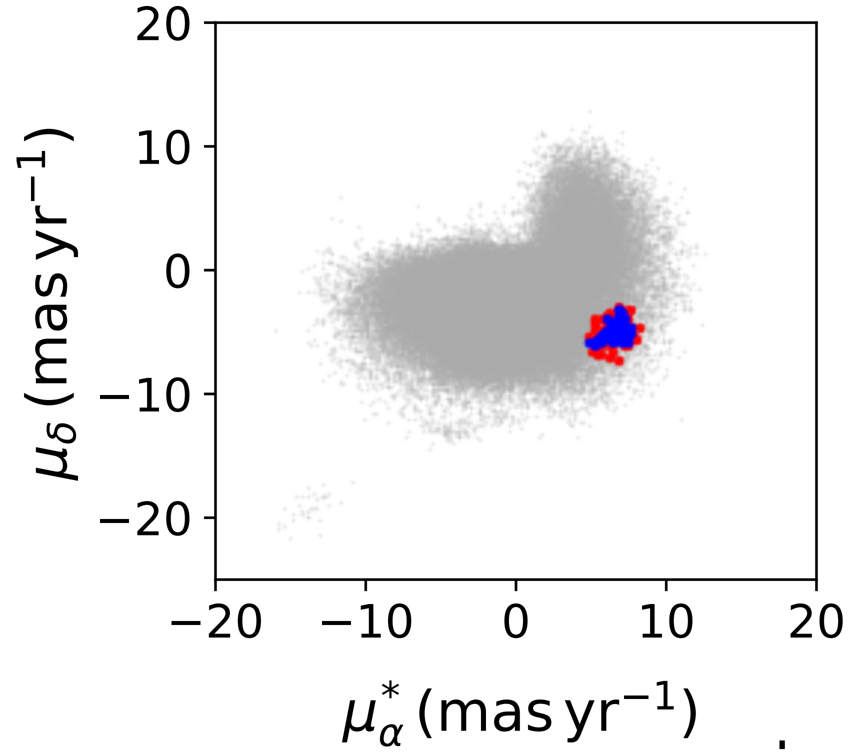
- Narrow & kinematically cold stellar streams are tracers of the Milky Way potential, merger history, imprint of dark matter substructure…
- A stellar stream is a narrow line of stars, compact in proper motion, and with all stars typically of similar age and composition.
- Use ML to build a stream-finding algorithm that:
  - Uses only Gaia data
  - Does not assume a Galactic potential or orbit
  - Does not assume stream stars lie on a particular isochrone.
  - Uses the fact that streams are compact in proper motion space.



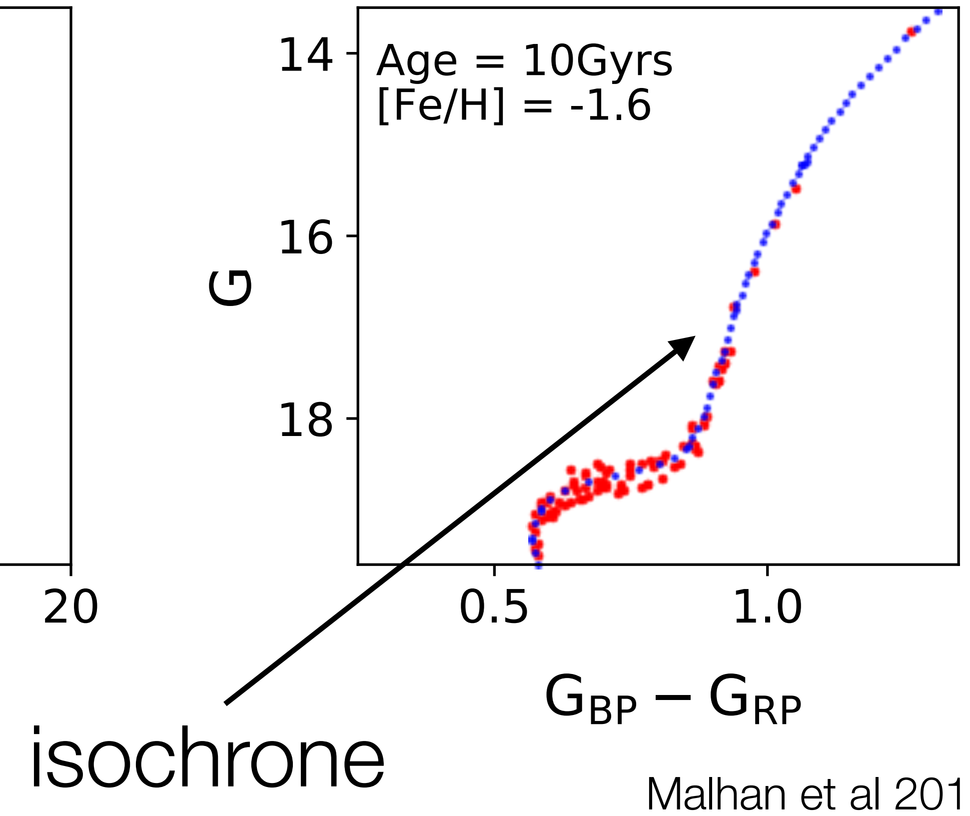
Angular position on sky



Angular motion on sky



Stellar brightness and color

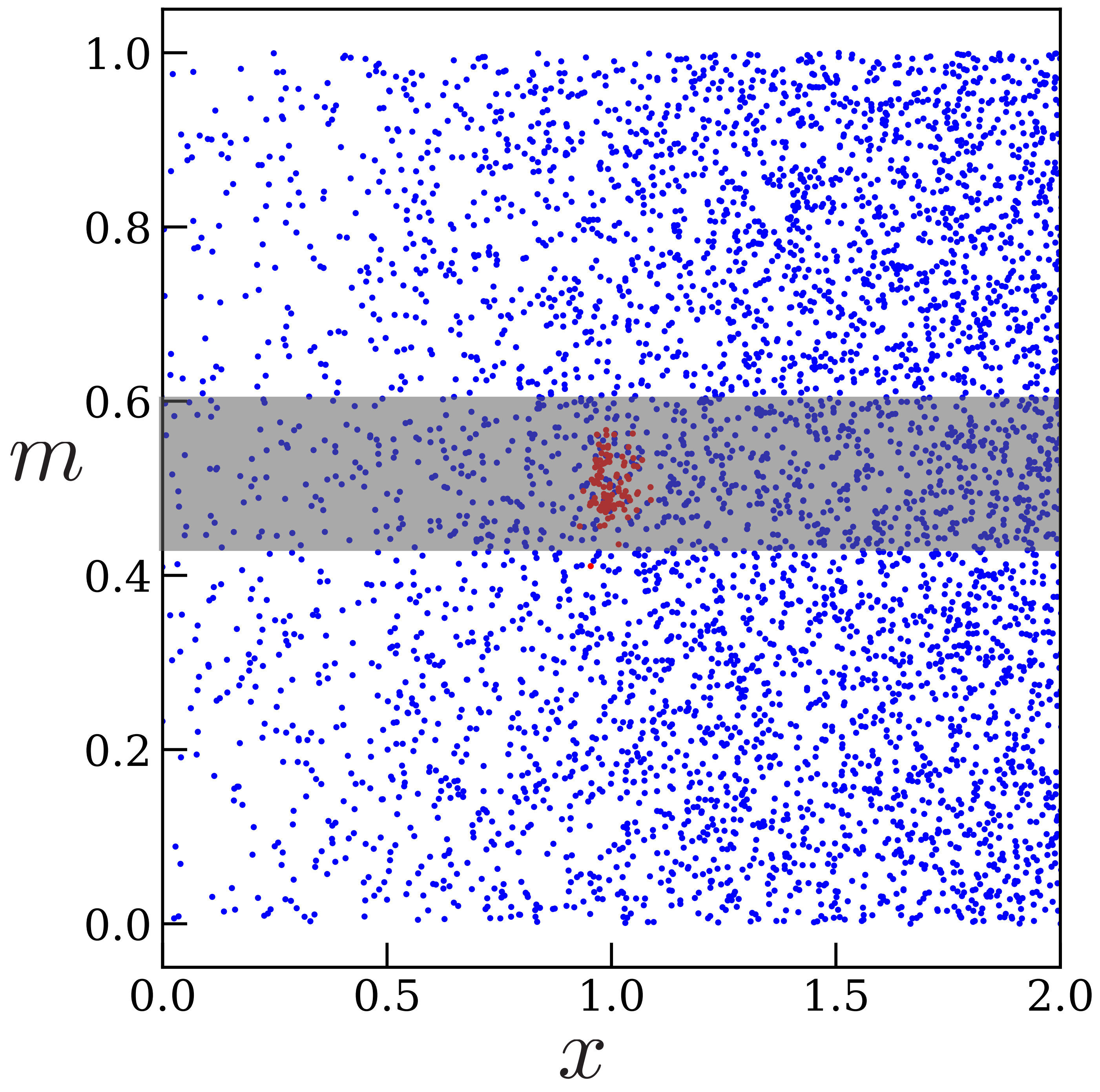


Malhan et al 2018

Via Machinae: Unsupervised Stream Finding

- Want to find stars that are anomalous based on their position in position, proper motion, and photometry. Use ANODE anomaly detection (Nachman & Shih 2001.04990) to calculate anomaly score *R* for stars in proper motion Search Regions (SRs)
- 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 in a control region, then interpolating in: $\approx P_{\rm bkg}(\vec{x}|m)$
- Allows direct estimation of the ratio *R* inside the SR.

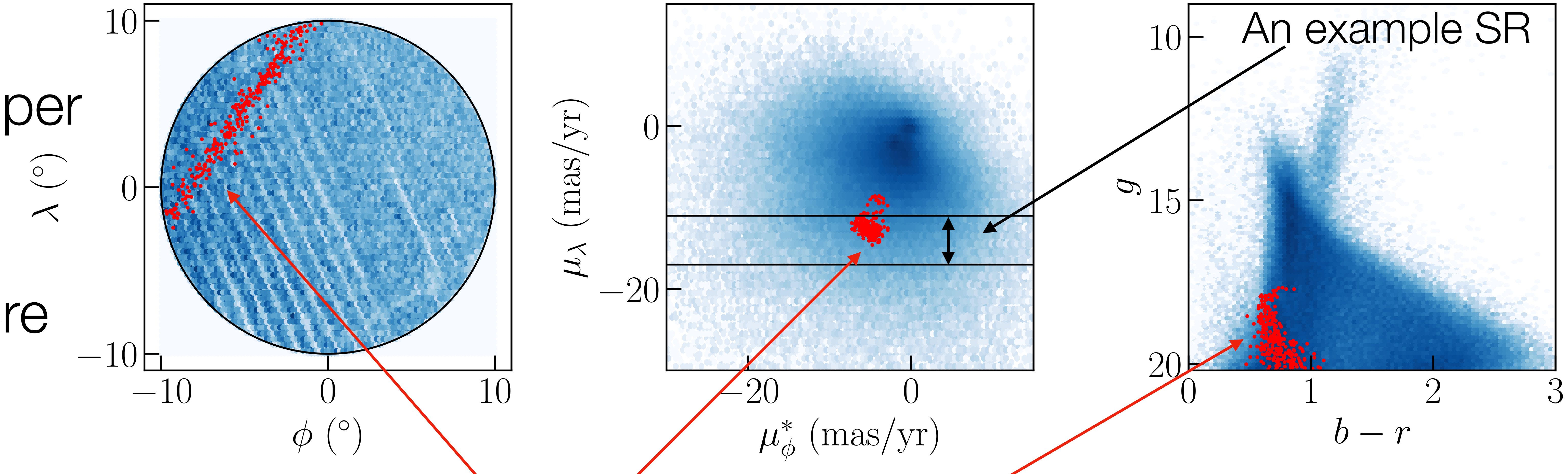
$$R(\vec{x}|m \in \mathrm{SR}) = \frac{P(\vec{x}|m \in \mathrm{SR})}{P_{\rm CR}(\vec{x}|m \in \mathrm{SR})}$$



Shih *et al* (2104.12789)

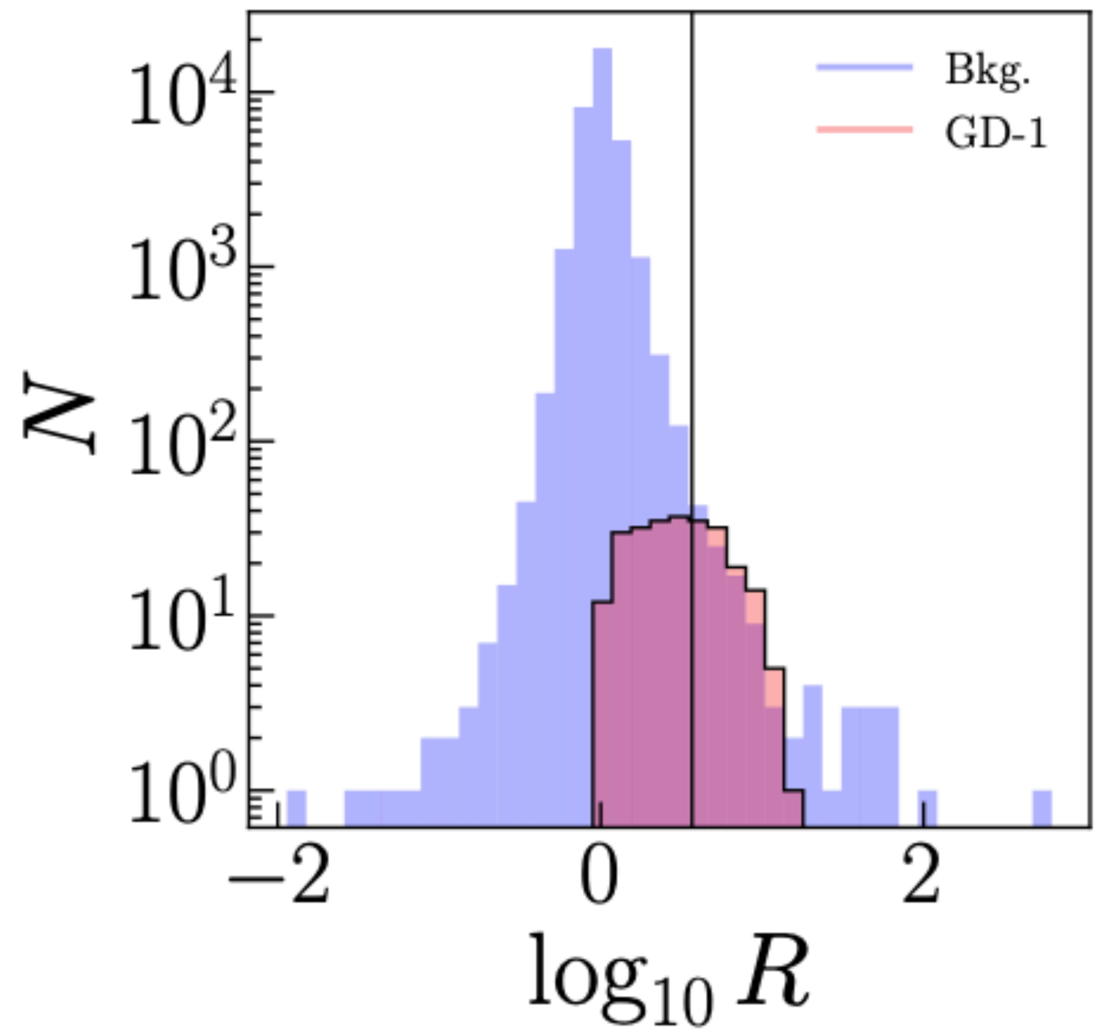
Via Machinae: Unsupervised Stream Finding

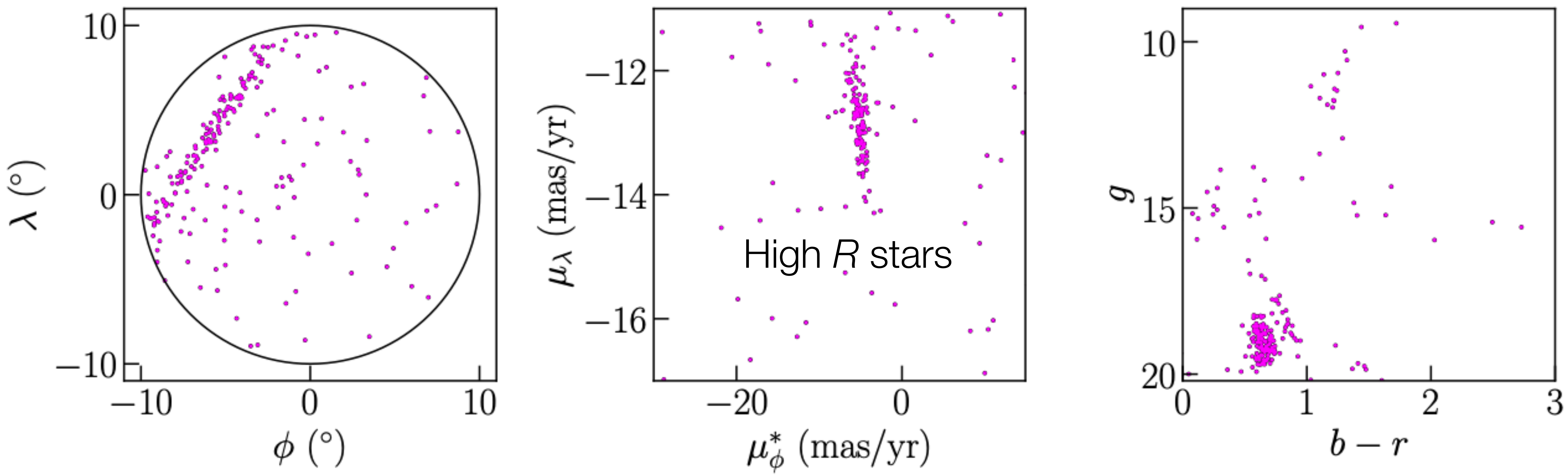
- Want to find stars that are anomalous based on their position in position, proper motion, and photometry. Use ANODE anomaly detection (Nachman & Shih 2001.04990) to calculate anomaly score R for stars in proper motion Search Regions (SRs)



Stars identified as likely GD-1 members by Price-Whelan & Bonaca

Shih, Buckley, Necib, Tamanas (2104.12789)



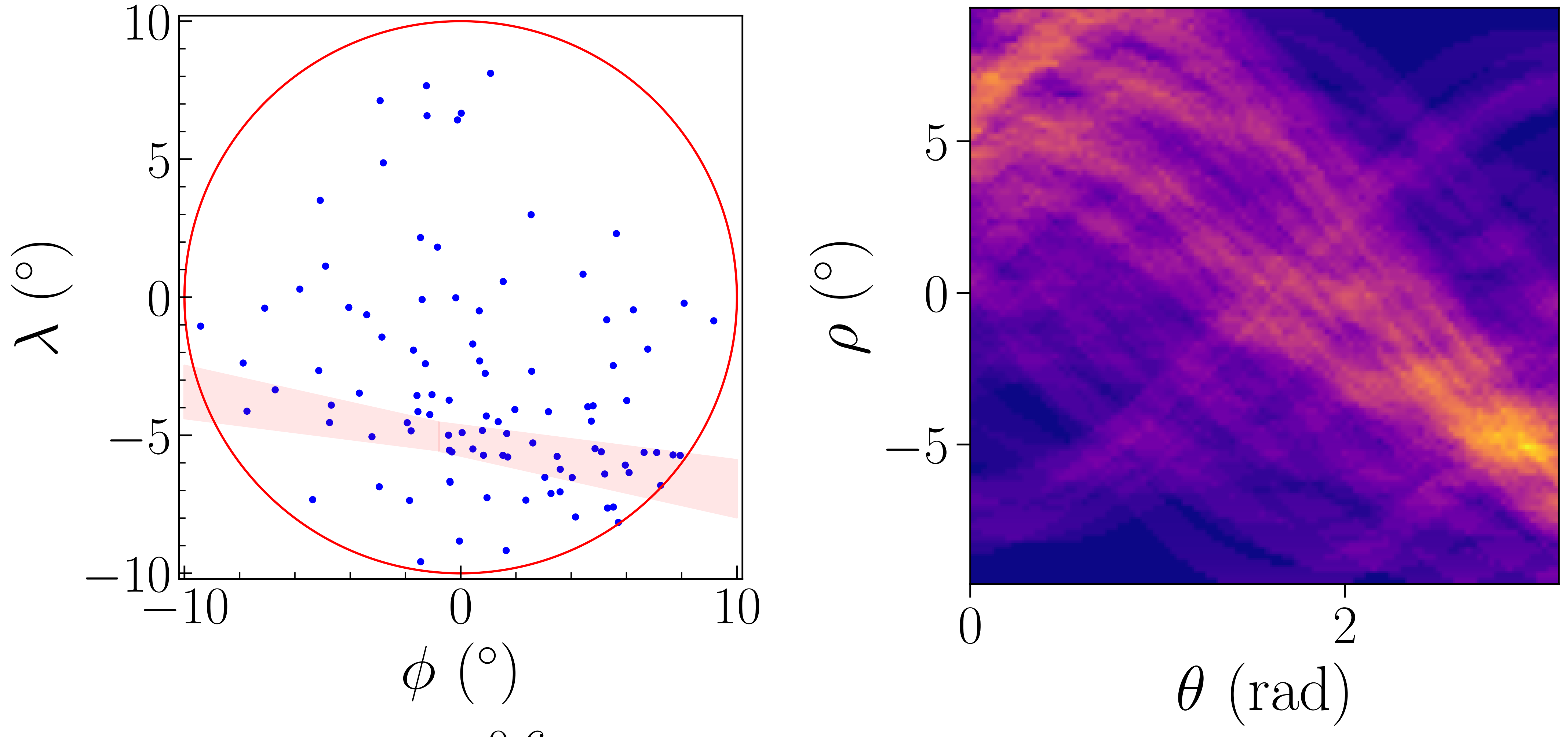


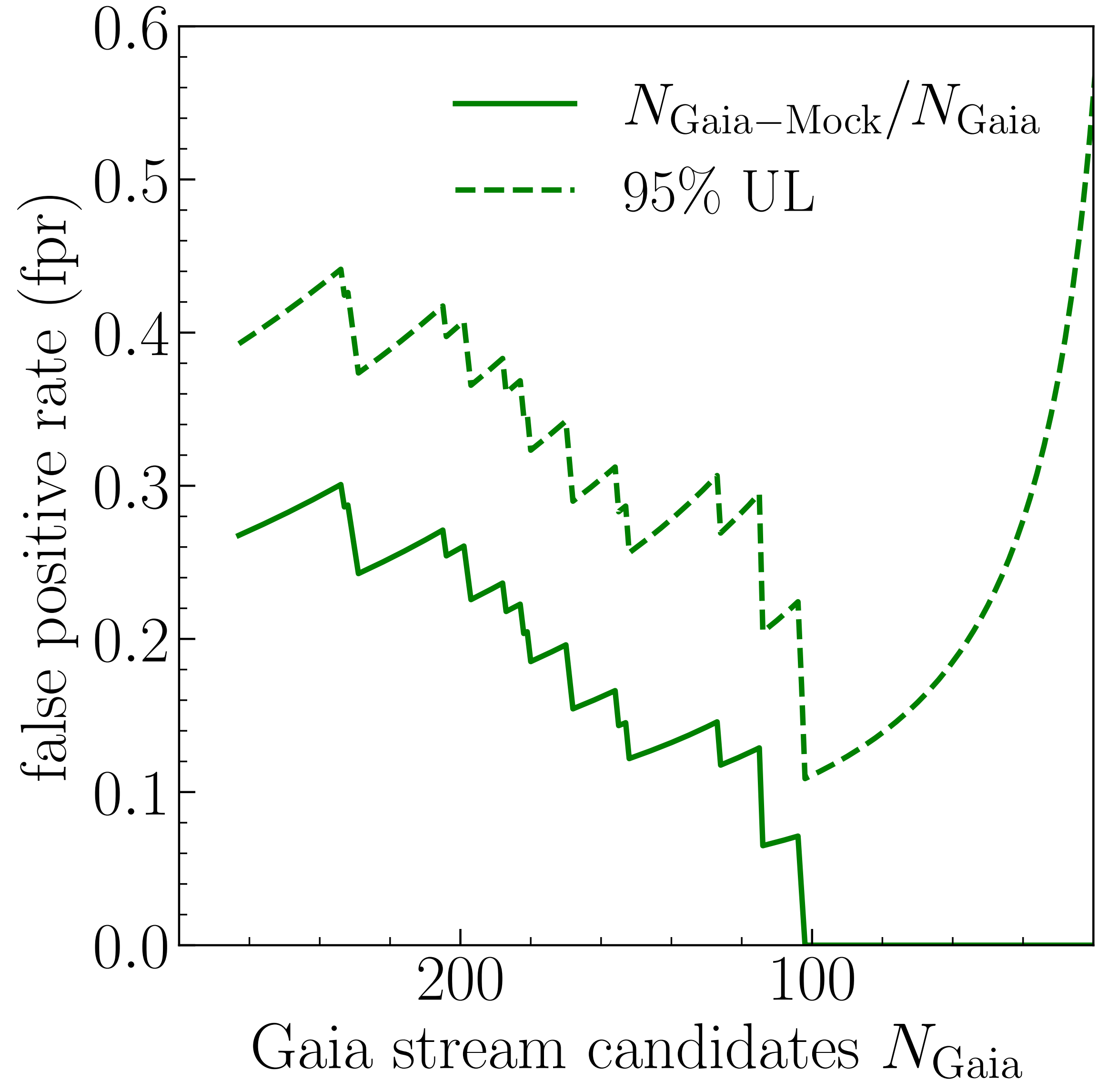
Via Machinae: Unsupervised Stream Finding

- There are a *lot* of stars in Gaia. Lots of reasons for them to be anomalous.
  - Dust lanes, globular clusters, disk stars…
- The ML anomaly score is only one part, need to automatically identify line-like features in overlapping regions of positions and proper motion.
  - Many hyperparameters needed identify stellar streams at high confidence
- Use a smooth analytic simulation of the Milky Way (totally devoid of streams) to build an estimate of a false positive rate

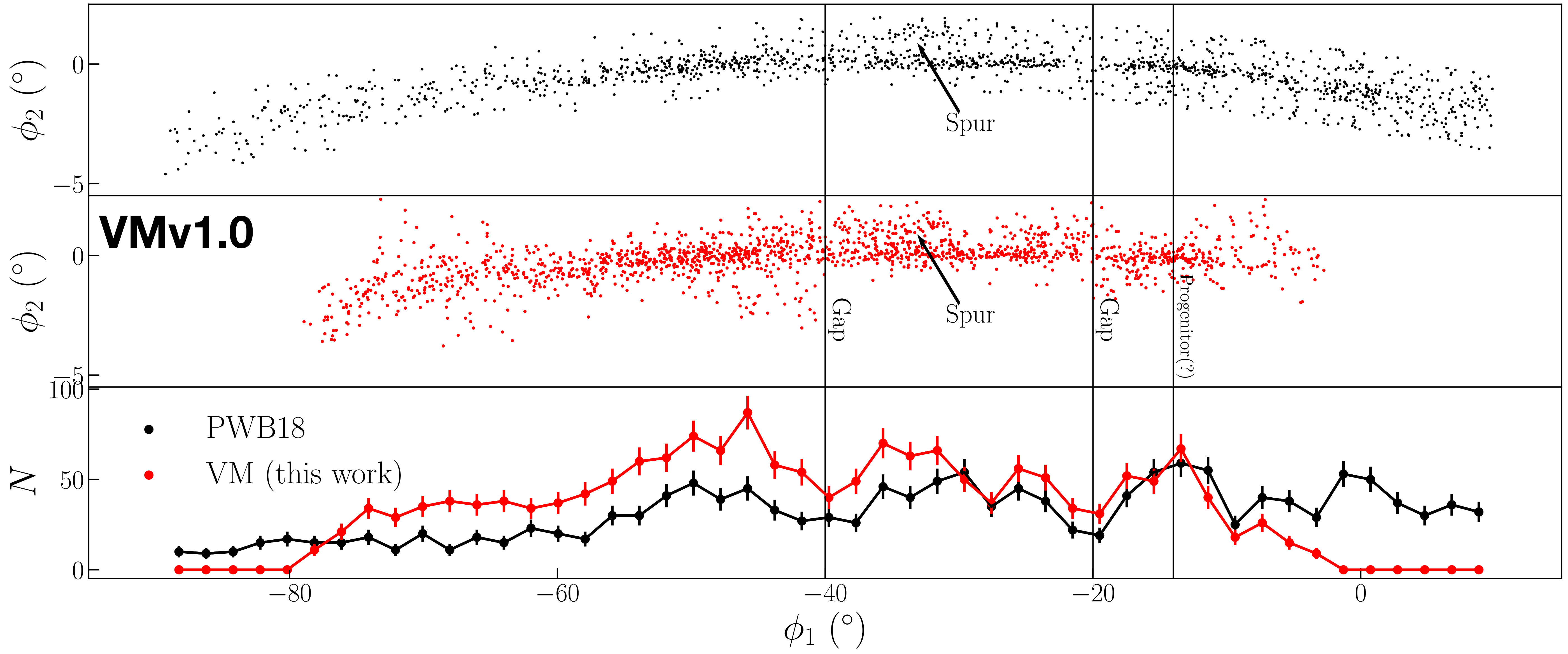
Shih, Buckley, and Necib 2303.01529

Hough transform for line-finding

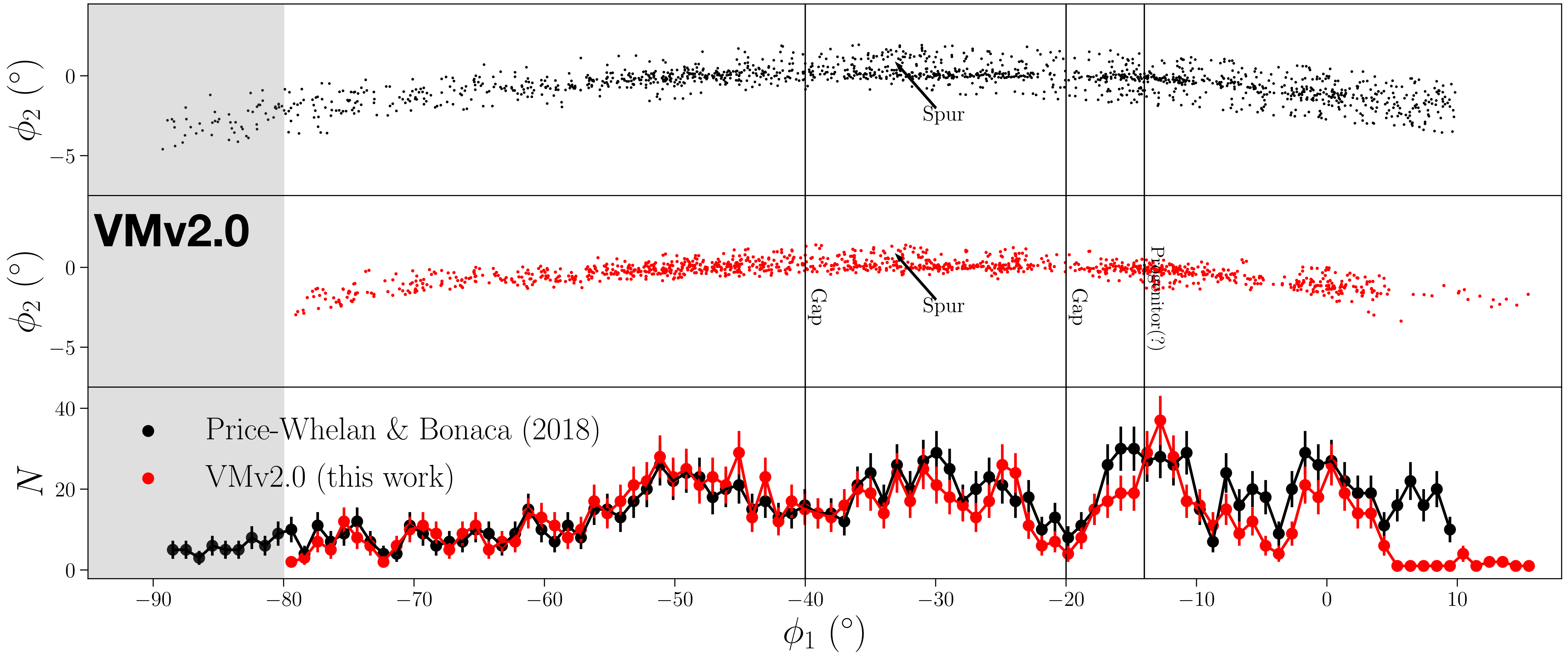




Via Machinae: Unsupervised Stream Finding



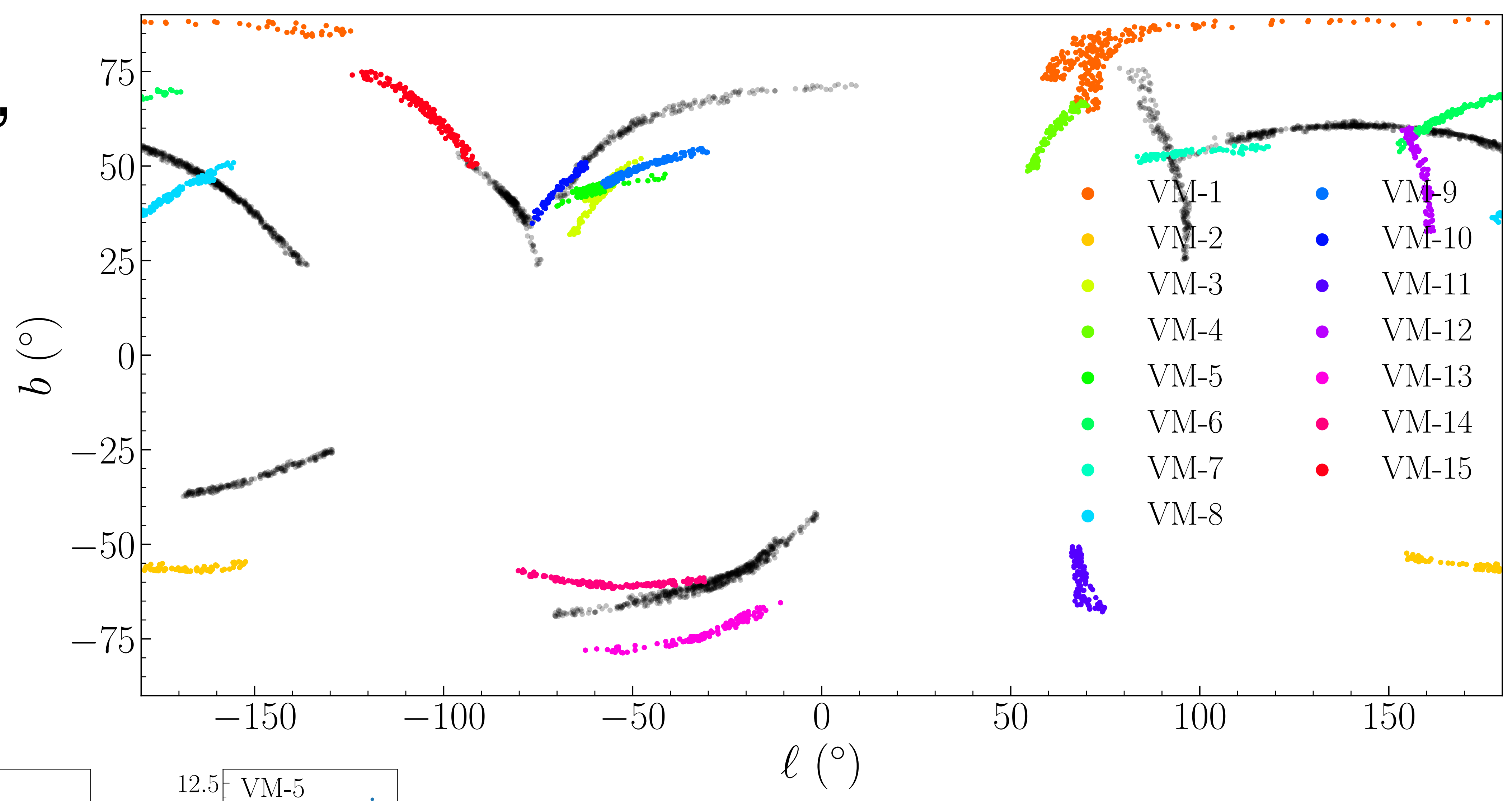
Shih, Buckley, Necib, and Tamanas 2104.12789

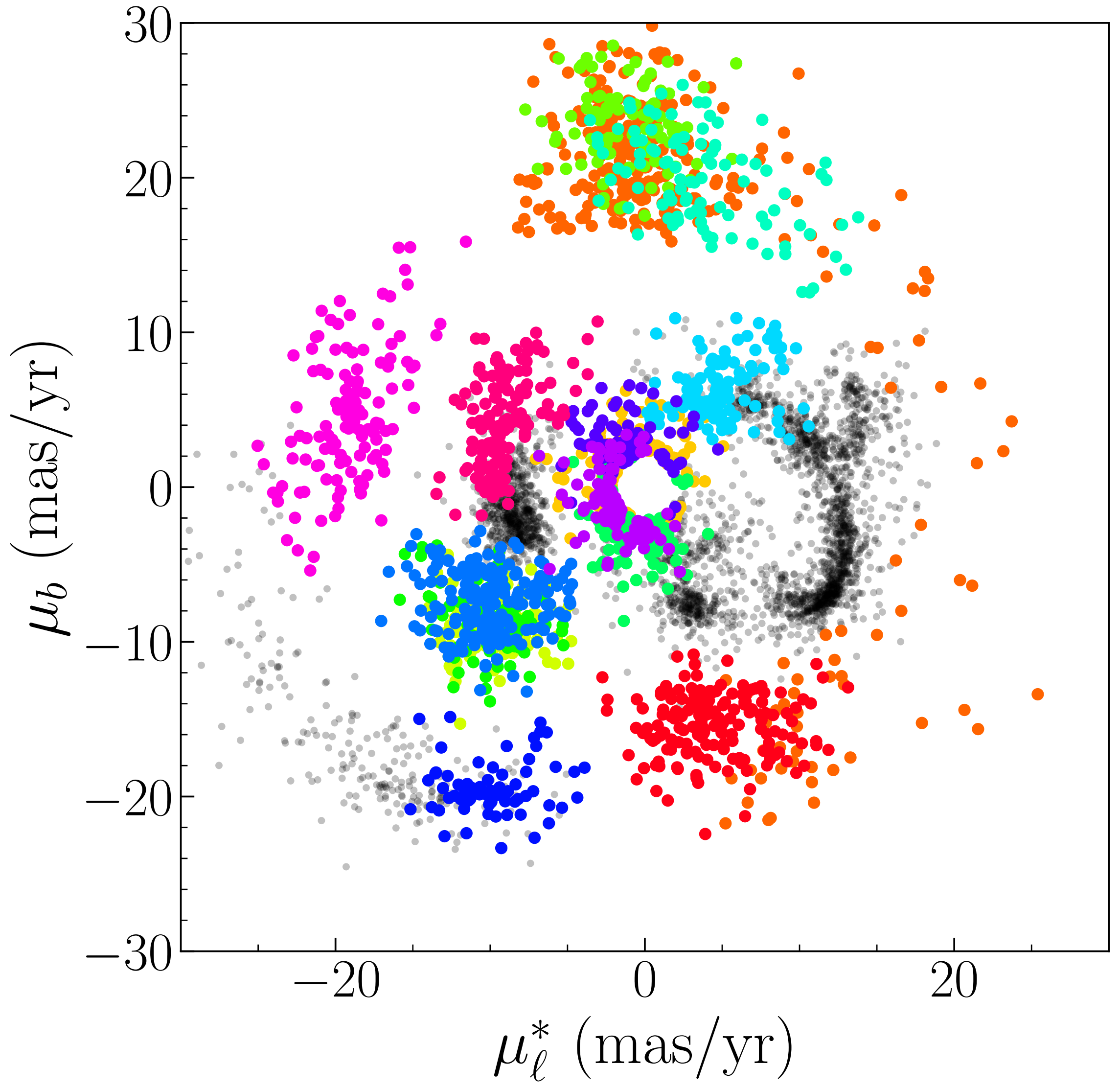


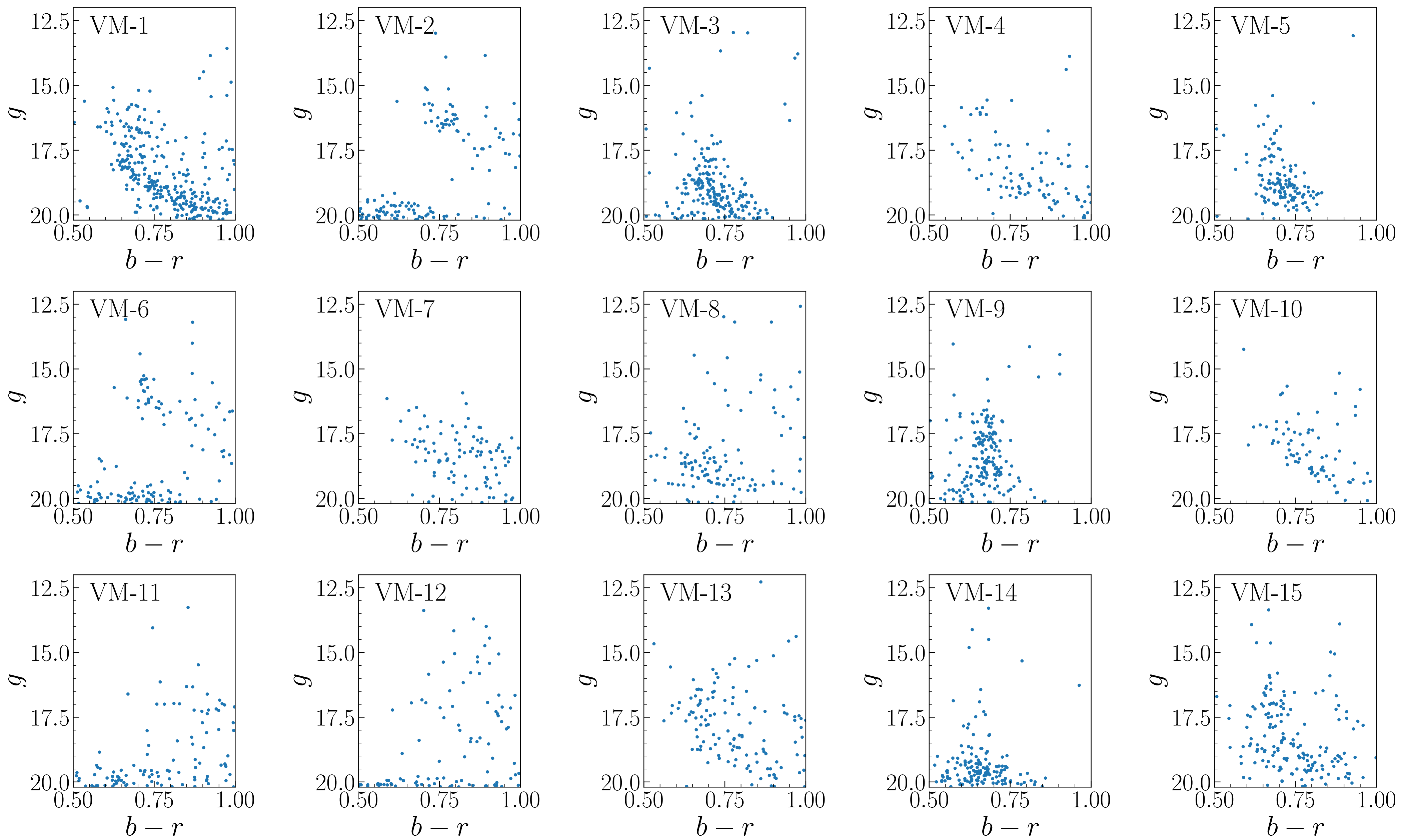
Shih, Buckley, and Necib 2303.01529

Via Machinae: Unsupervised Stream Finding

- We identify 82 stream candidates, expect a false-positive rate of ~10%.
  - Here are the top 15.





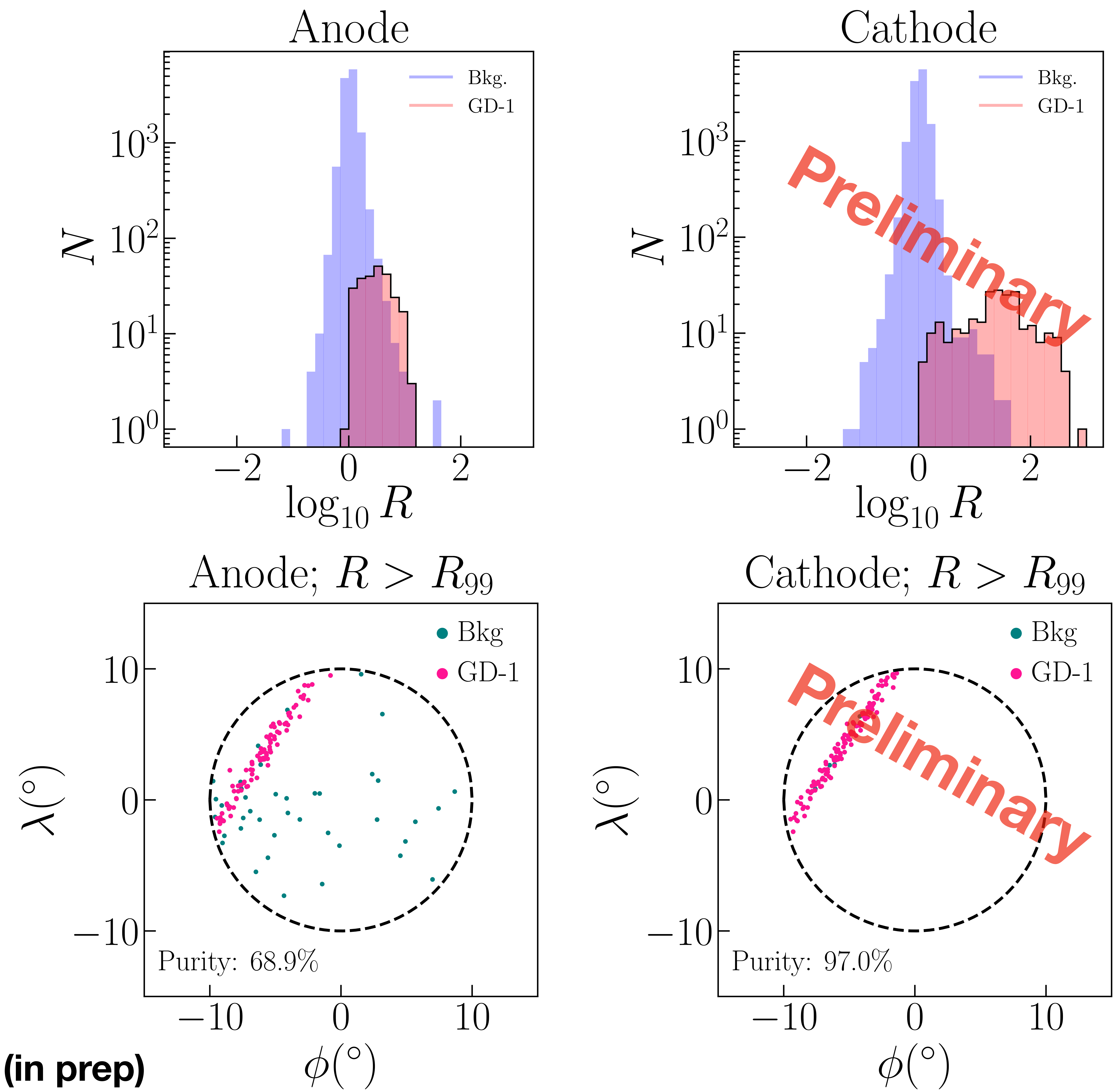


- How to confirm stellar streams?
  - Spectroscopic follow-ups with other telescopes.
  - Do the stars have consistent metallicity, age, distance, radial velocity…?

Shih, Buckley, and Necib 2303.01529

Via Machinae: Unsupervised Stream Finding

- The input for the stream-finding is the ML-derived anomaly score *R*
  - Existing version from ANODE, using normalizing flows to learn conditional probabilities in proper motion SR and backgrounds from control regions.
- What if we could do this better?
  - CATHODE (Hallin *et al* 2109.00546)
  - Train a classifier to distinguish events generated in signal region from density estimator trained on control-region.
  - Use this as input for rest of Via Machinae

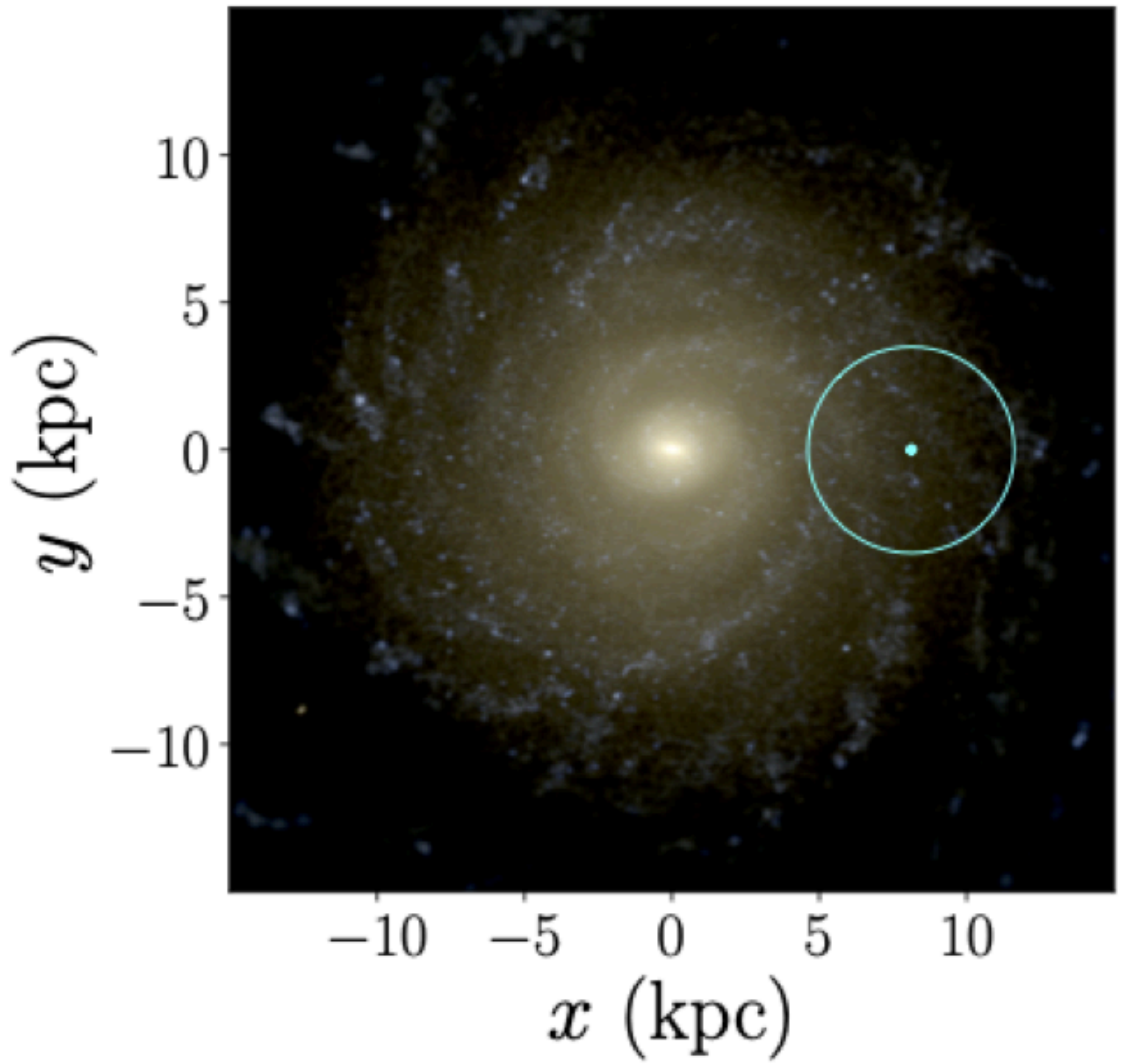


Hallin *et al* (in prep)

Upsampling Simulations

- Tools exist that can create "theorist-level" simulation for LHC machine learning.
- Much trickier for astrophysics. Can either:
  - Create by-hand analytic smooth models of the Galaxy or,
  - Use *N*-body hydrodynamical simulations
- But in the latter case, there complications:
  - Every galaxy is unique.
  - Simulations work on the level of tens of millions of "star particles," not hundreds of billions of *stars*.
- Upsampling required!

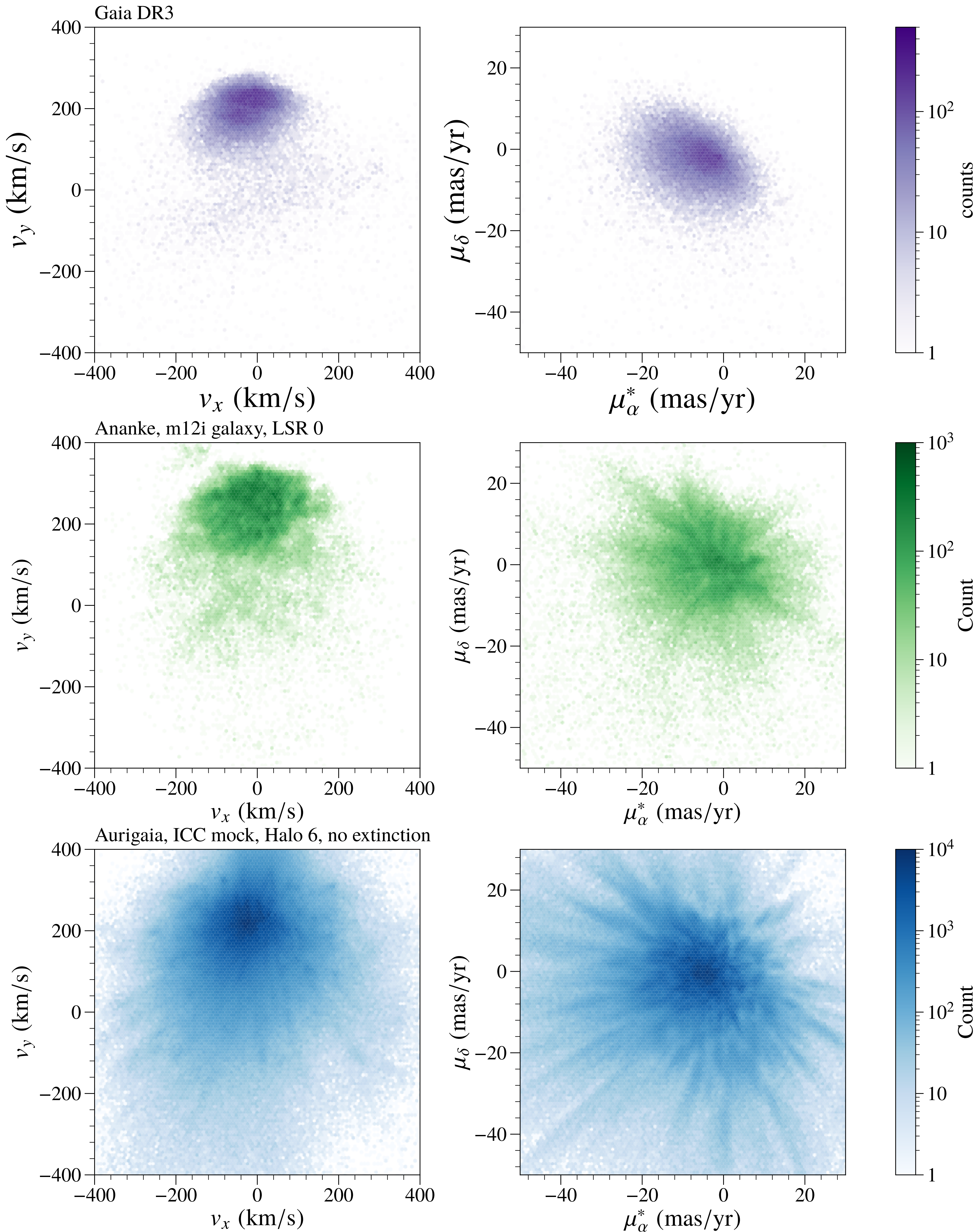
Galaxy h277 (N-Body Shop)



Lim *et al* 2211.11765

Upsampling Simulations

- Tools exist that can create "theorist-level" simulation for LHC machine learning.
- Much trickier for astrophysics. Can either:
  - Create by-hand analytic smooth models of the Galaxy or,
  - Use *N*-body hydrodynamical simulations
- But in the latter case, there complications:
  - Every galaxy is unique.
  - Simulations work on the level of tens of millions of "star particles," not hundreds of billions of *stars*.
- Upsampling required!
  - But existing upsamplers are "clumpy"

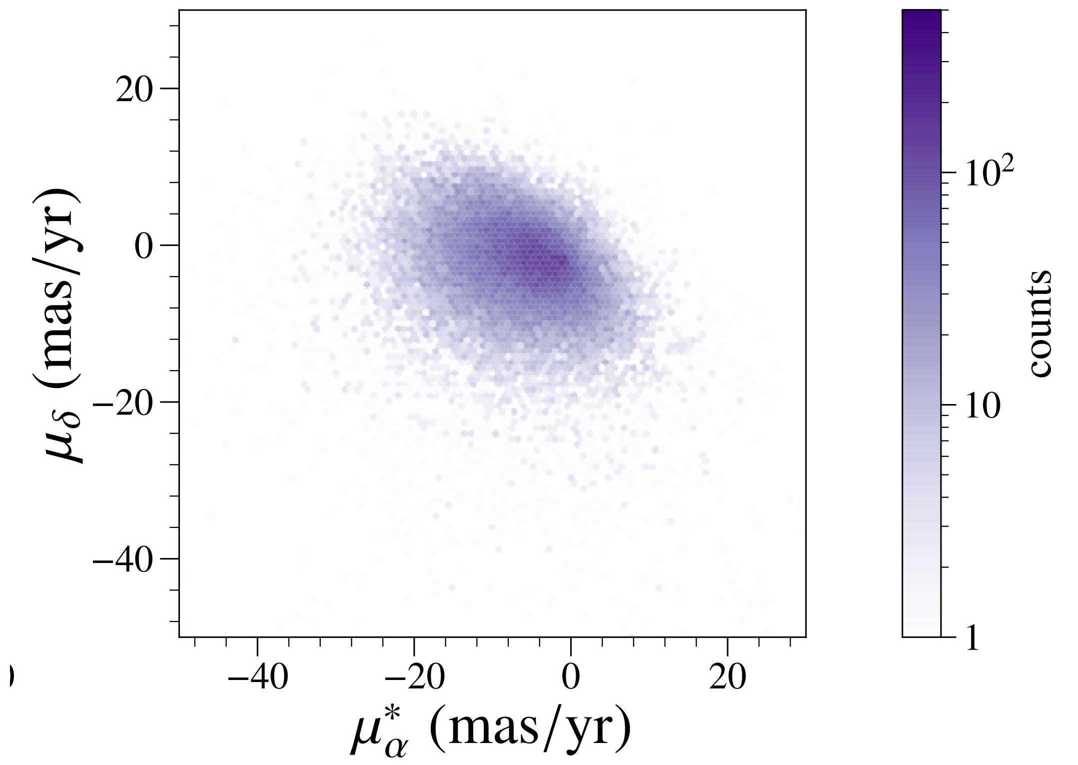


Lim *et al* 2211.11765

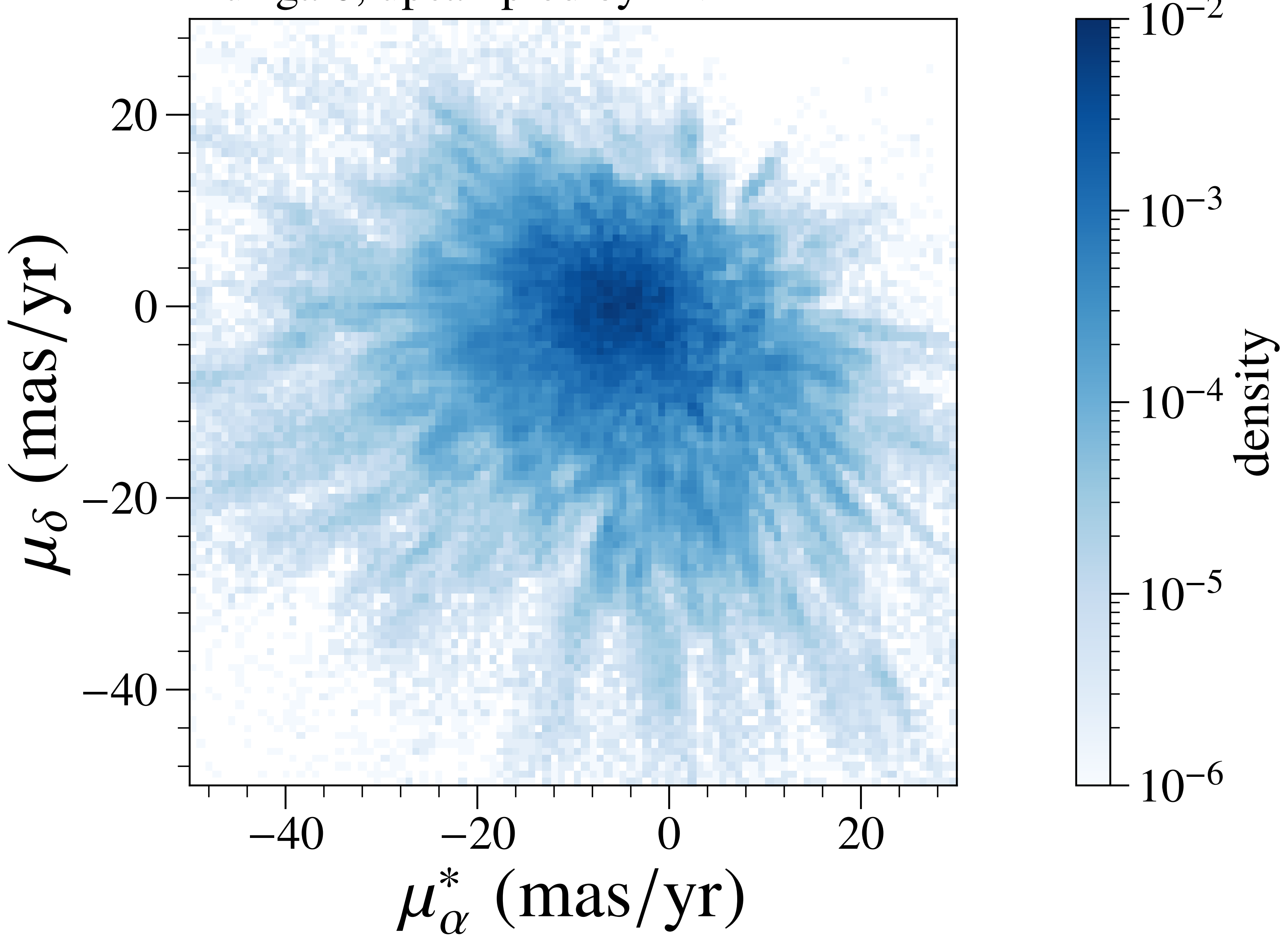
Upsampling Simulations

- Use normalizing flows (CNFs) to learn the density distribution of simulation star particles, then generate synthetic stars from the flow.
  - Demonstrating with stars near the "Sun"
  - Much smoother than stars drawn from existing upsamplers (EnBid)
  - Confirmed with classifier tests comparing CNF and EnBid

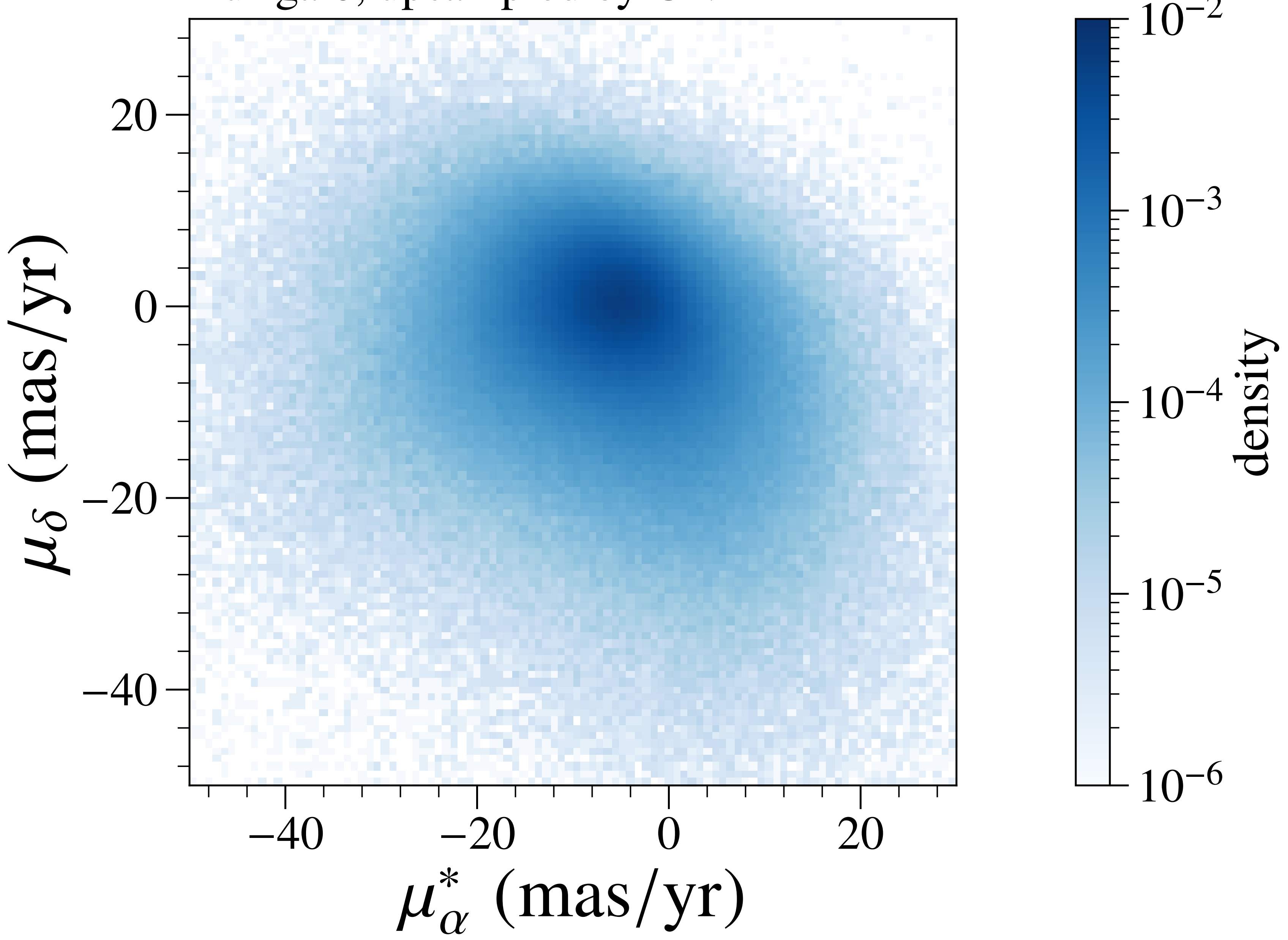
Real Gaia Data



Auriga 6, upsampled by EnBiD



Auriga 6, upsampled by CNF

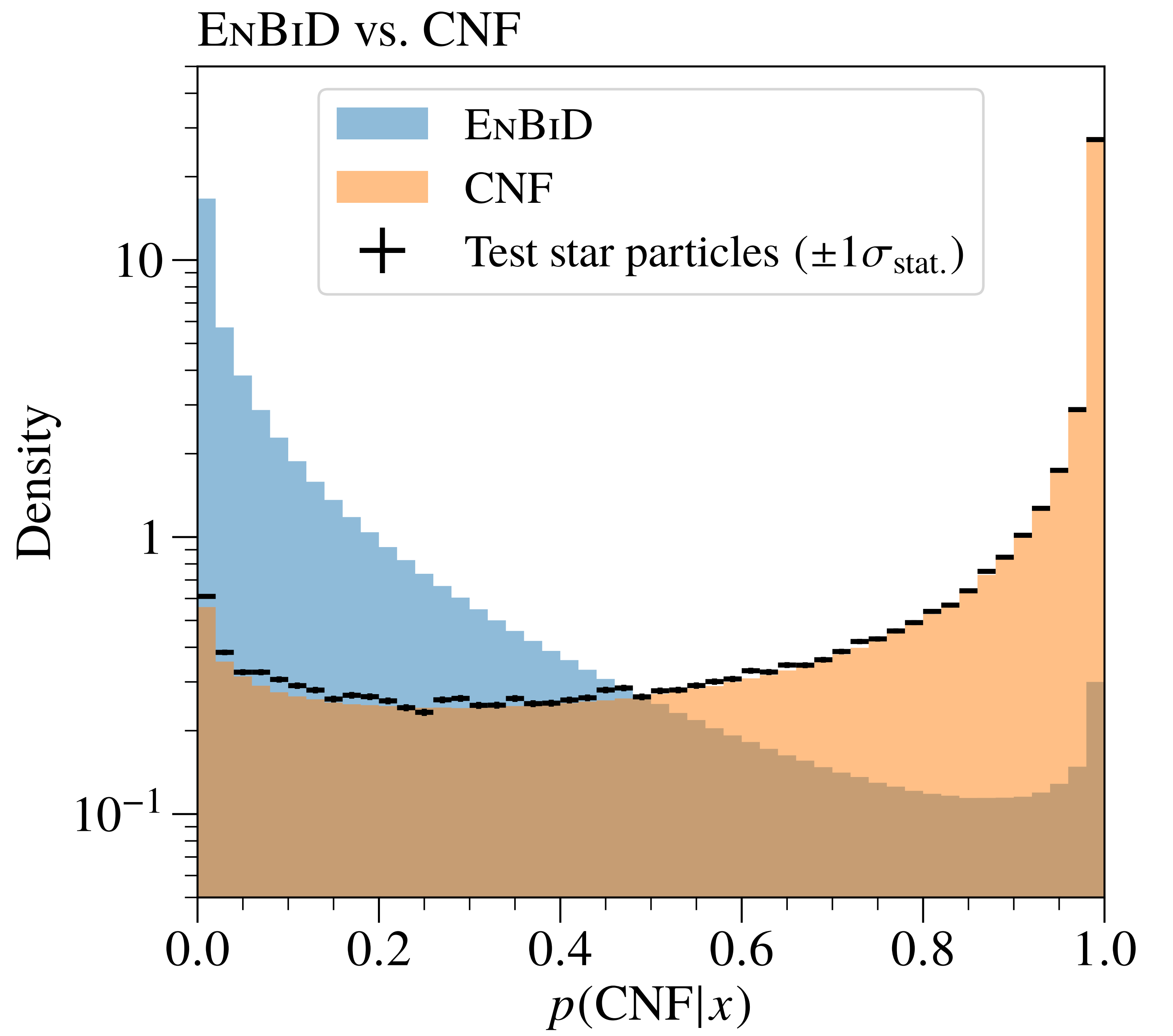


Lim *et al* 2211.11765

Upsampling Simulations

- 3-sample classifier: we are statistics-limited on the star particles
  - Construct CNF and EnBid datasets from a training subset of the star particles, reserving some star particles for validation
  - Train classifier between a subset of the CNF and EnBid datasets
  - Compare validation star particles with CNF and with EnBid separately

network	classification target	AUC
trained on	EnBiD vs. CNF	0.952
applied to	EnBiD vs. Star particles	0.950
	Star particles vs. CNF	0.508

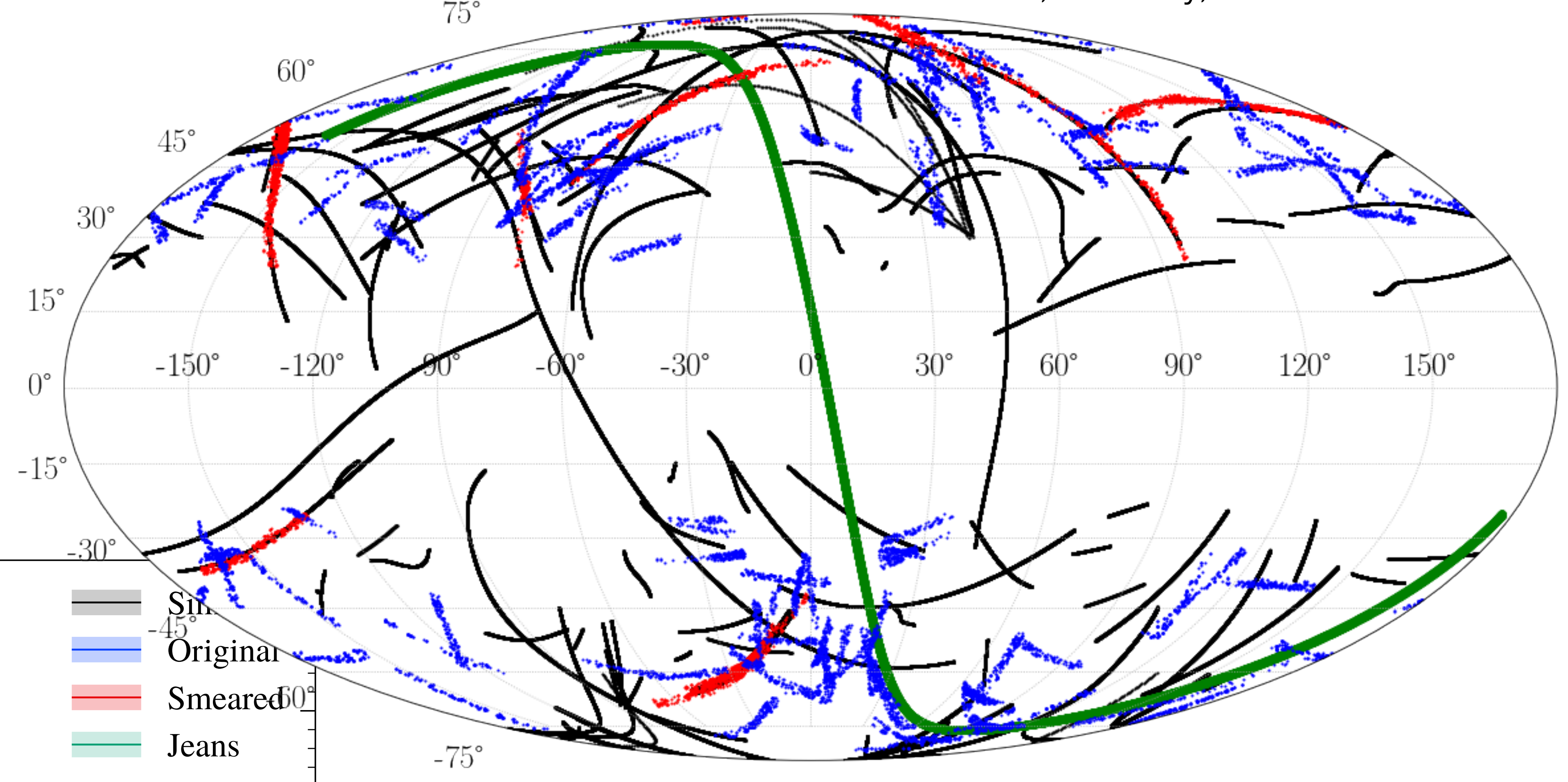


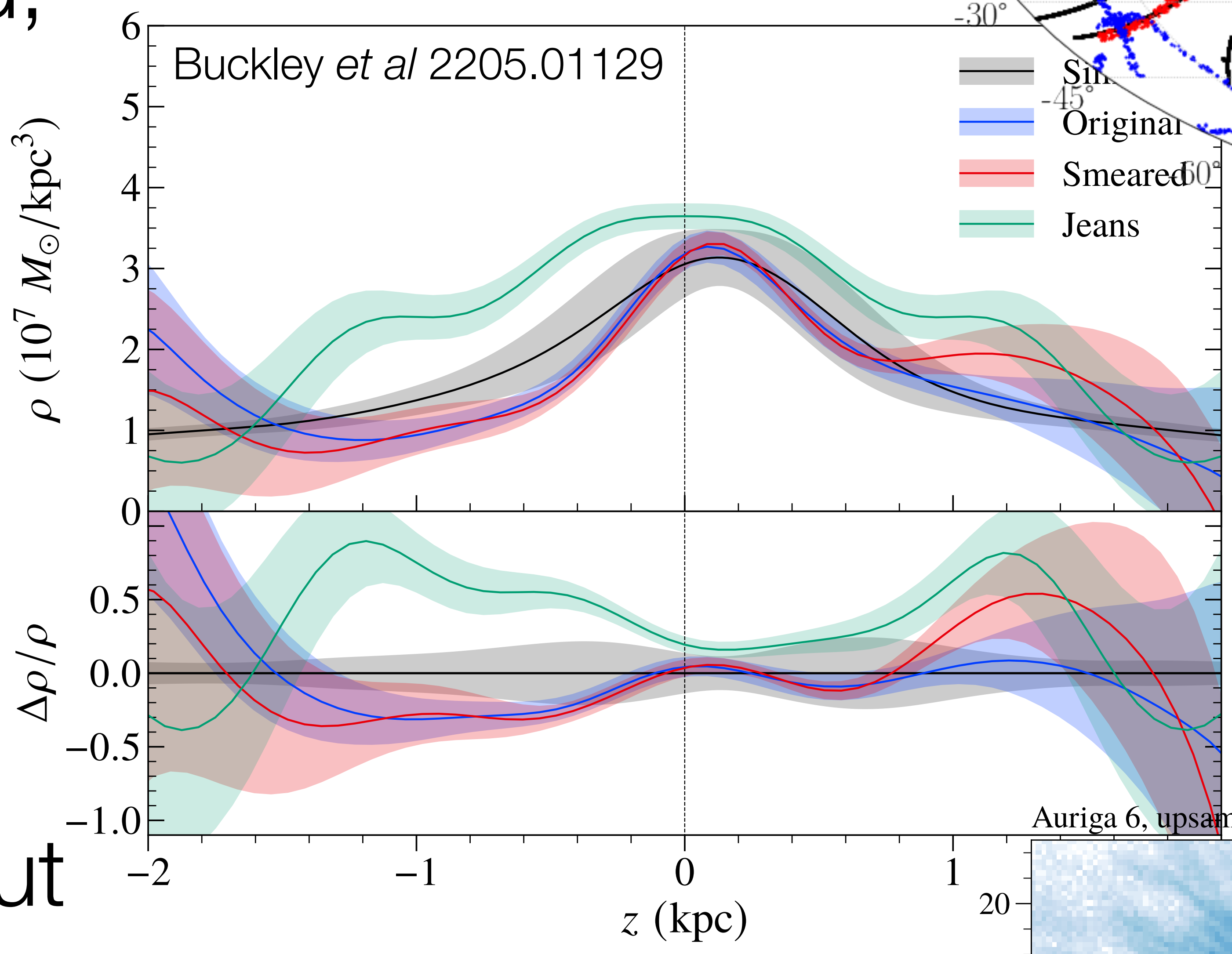
Lim *et al* 2211.11765

Conclusions

- Astrophysical datasets contain information relevant to particle physics questions
  - …and intrinsically interesting on their own merits!
- The datasets are massive and complicated, with lots of systematic effects to deal with.
  - Often harder to simulate exactly what you'd need to test your technique. Interesting ML problems here in transfer learning, generation, quantifying errors.
  - Unsupervised techniques very useful.
- *Gaia* data in particular has lots to say about dark matter and Galaxy structure/history.
  - Lots of need for new techniques, opportunities for ML to help!

Shih, Buckley, Necib 2303.01529





Lim *et al* 2211.11765