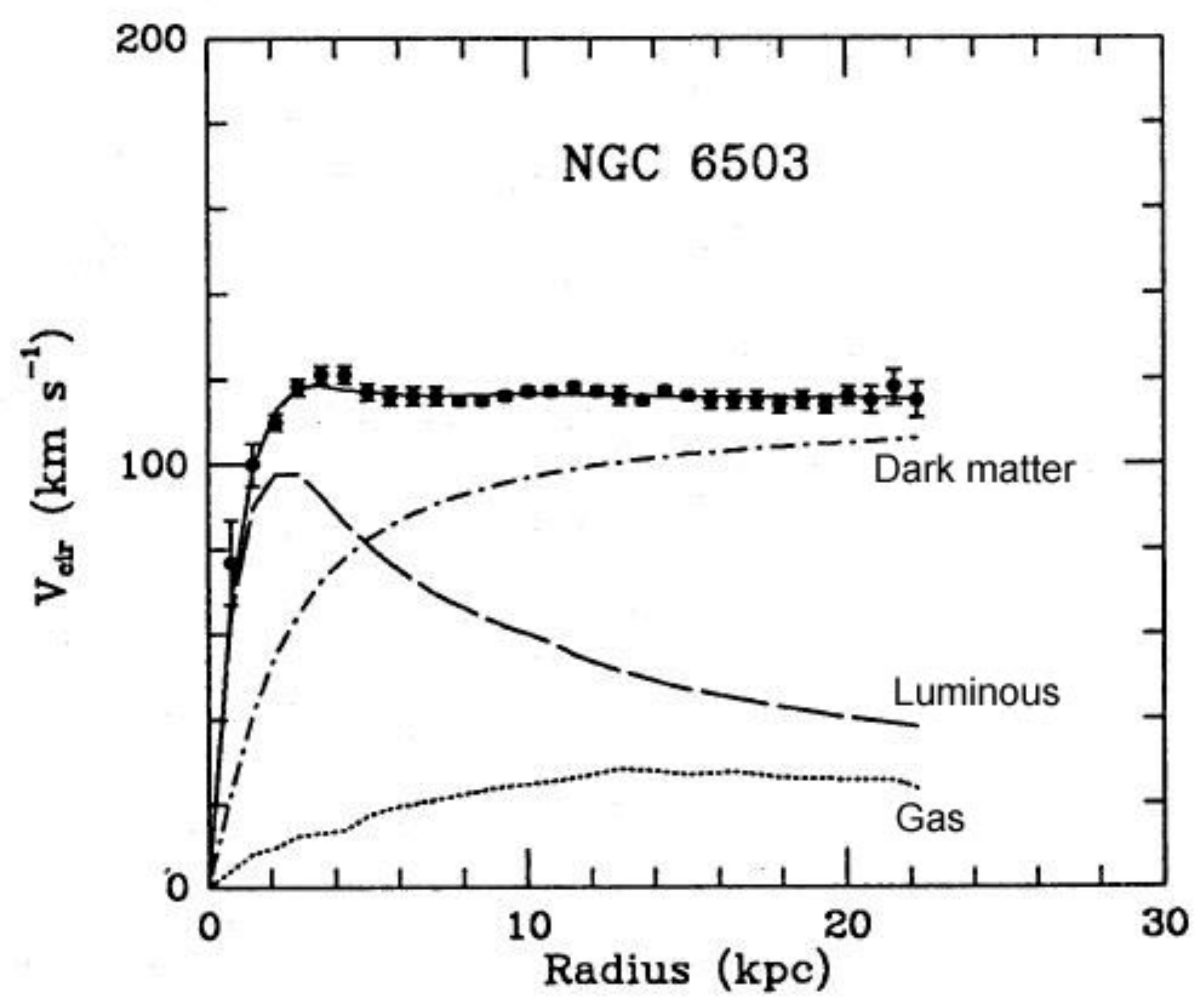
Overview of Machine Learning for Gaia

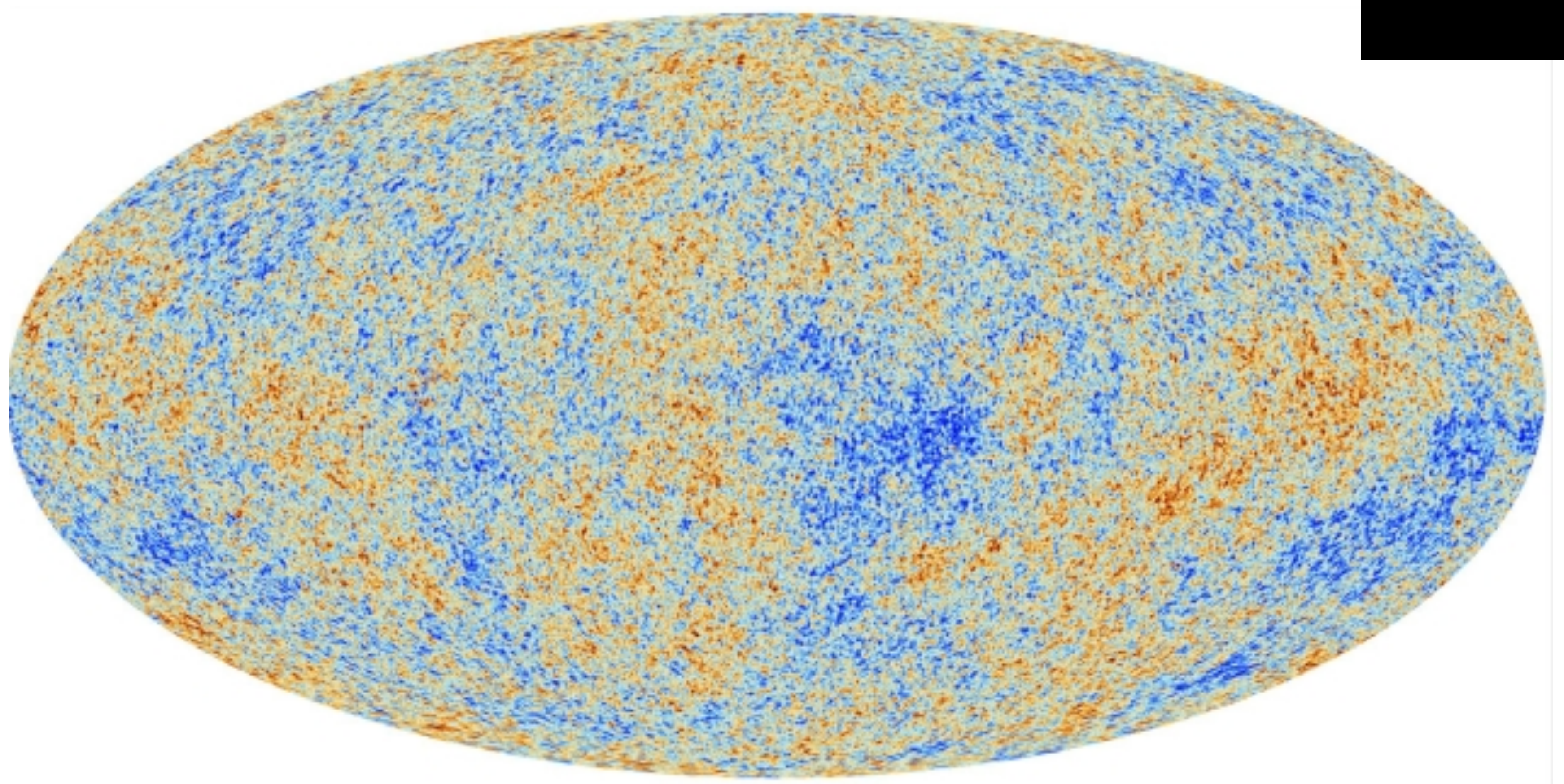
Matthew R Buckley
Rutgers University

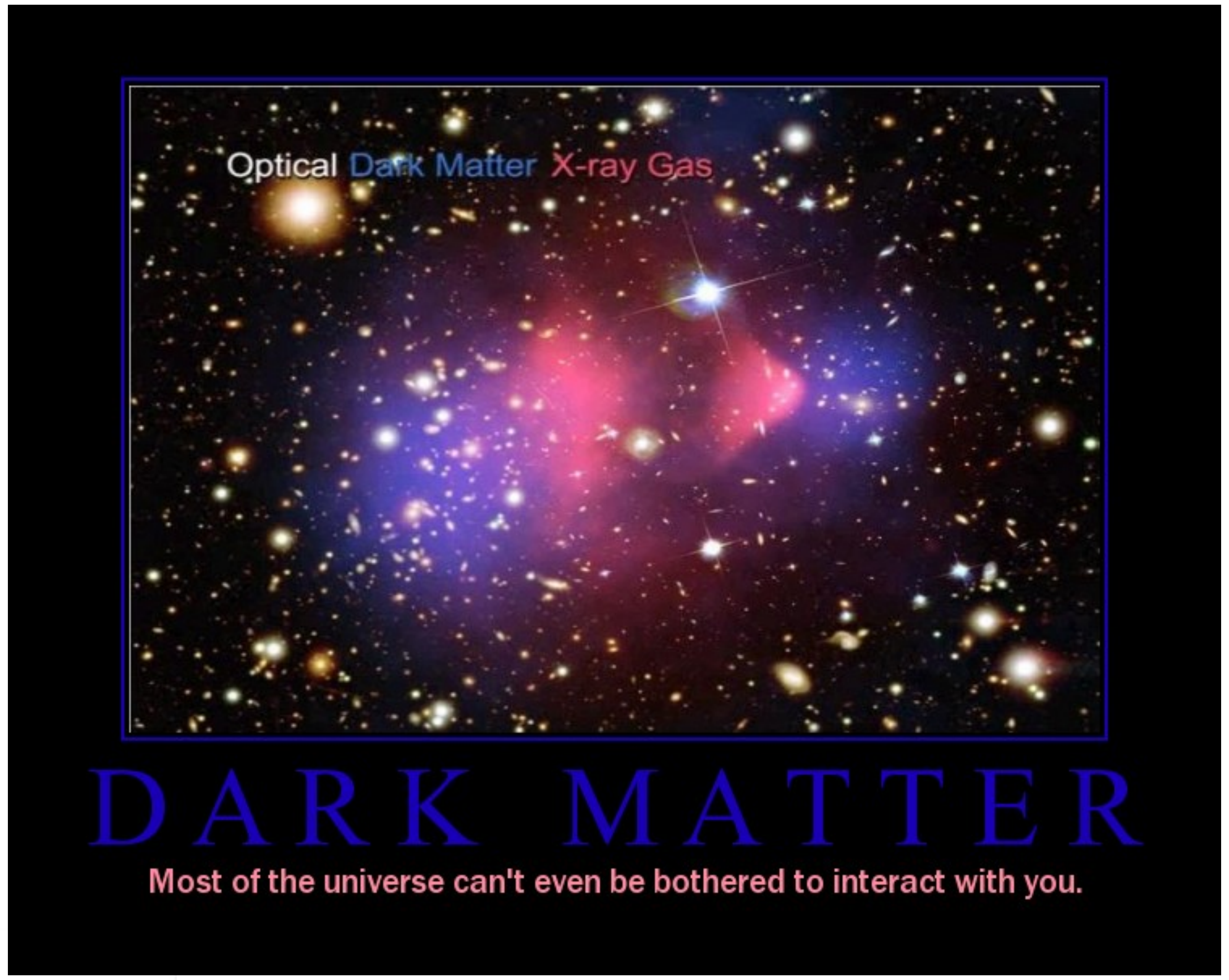
Theoretical Motivations

- 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.

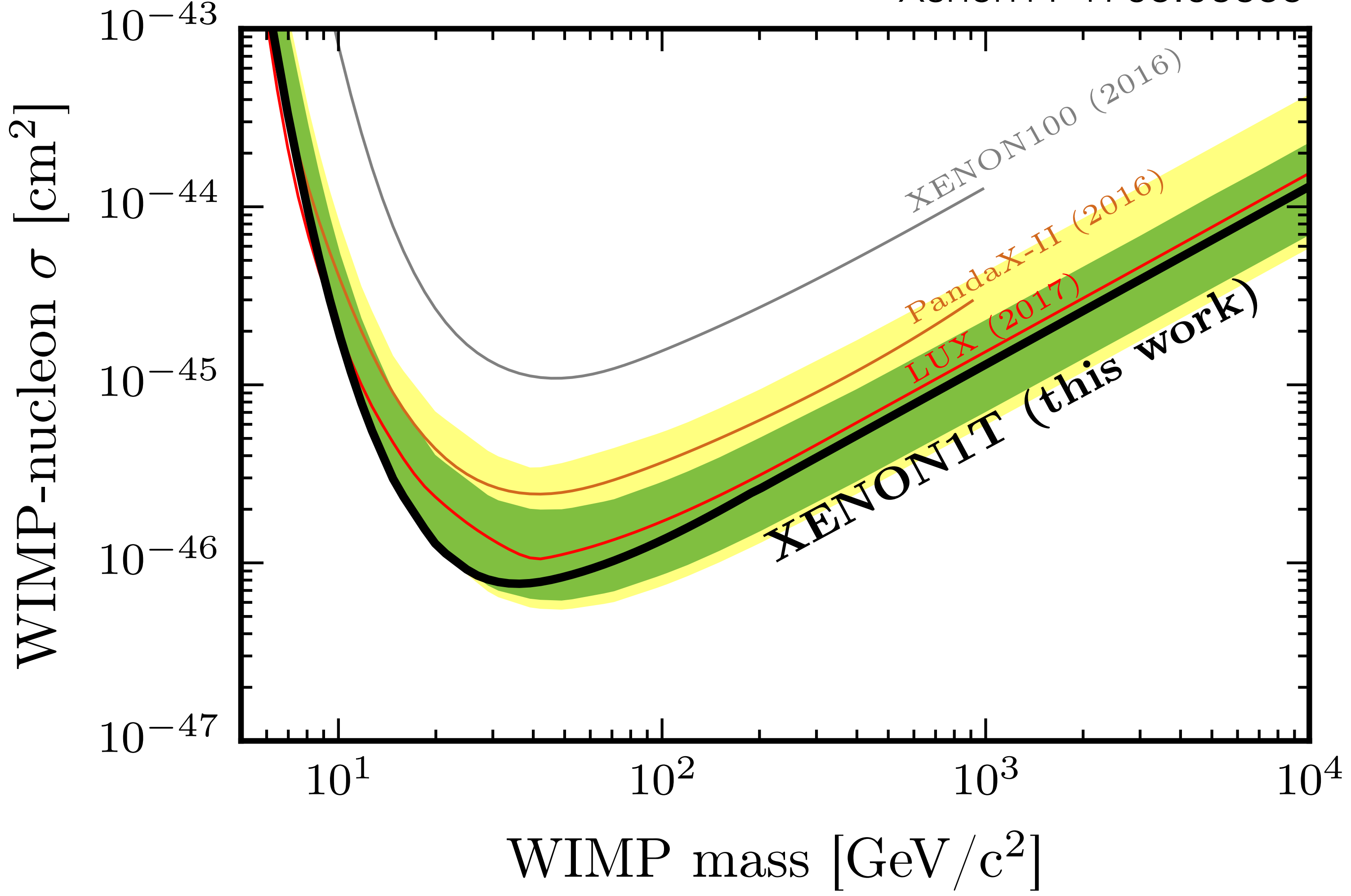




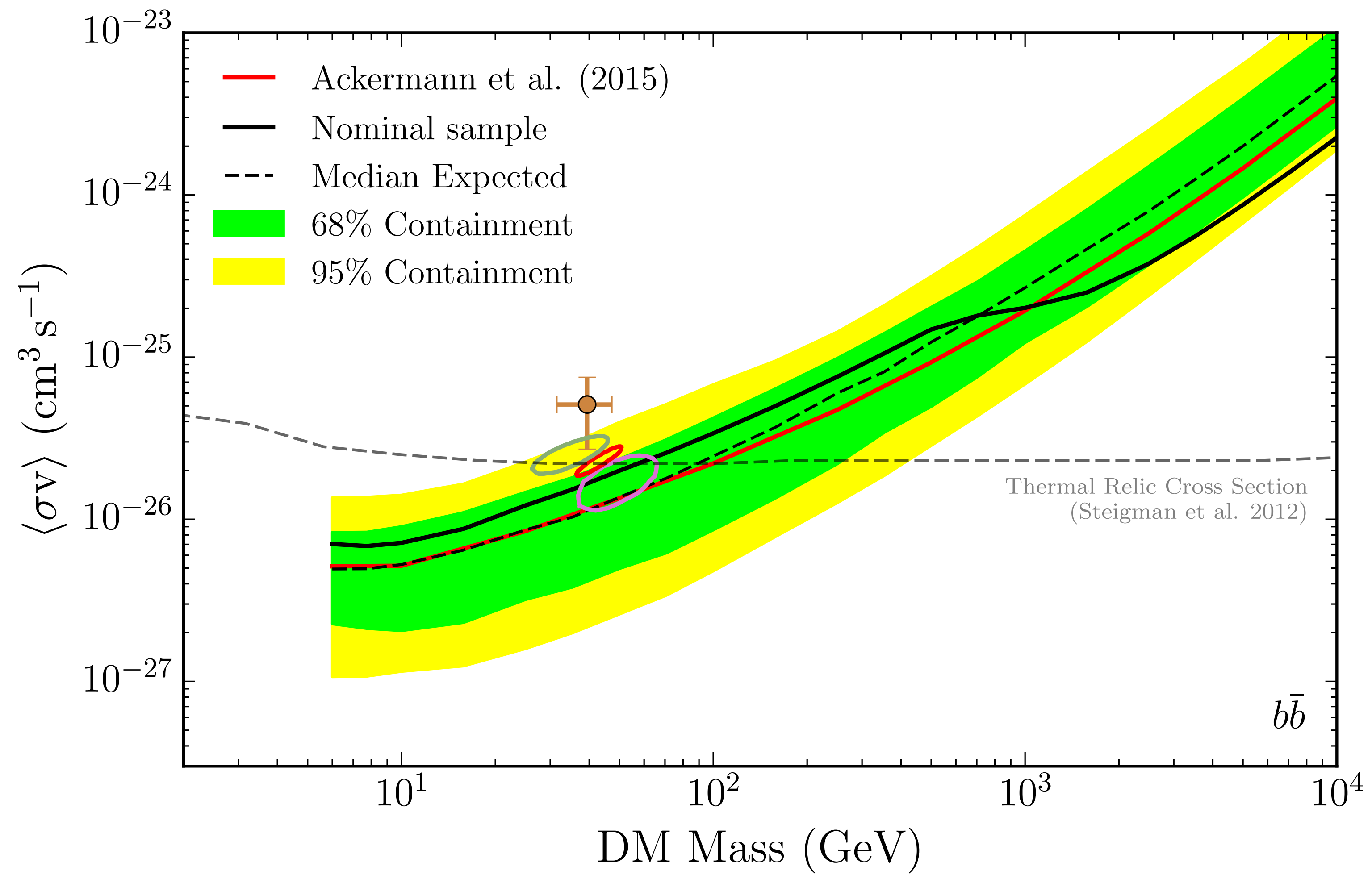
Theoretical Motivations

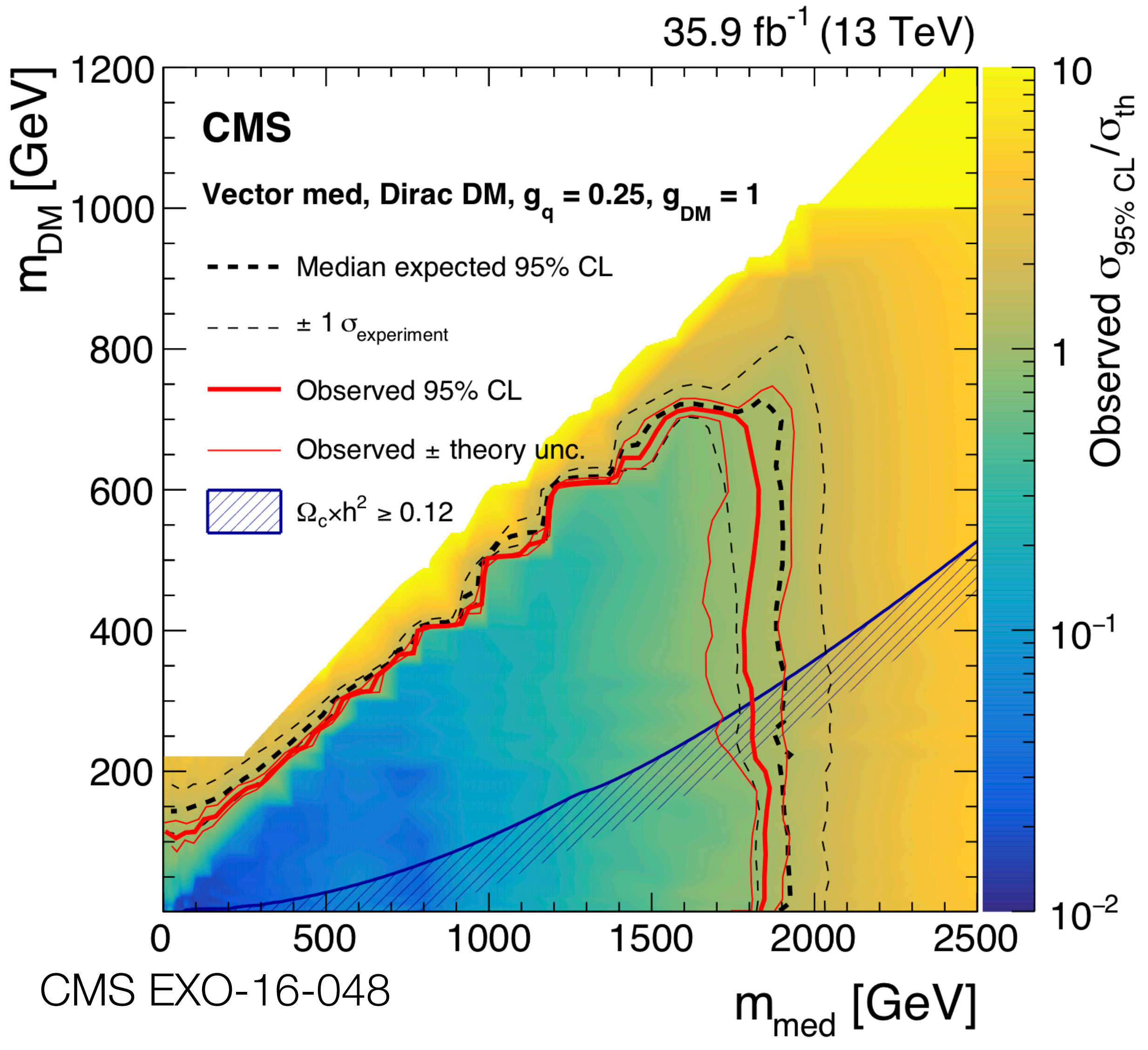
- Particle Physics experiments are motivated and important, but so far give only negative results

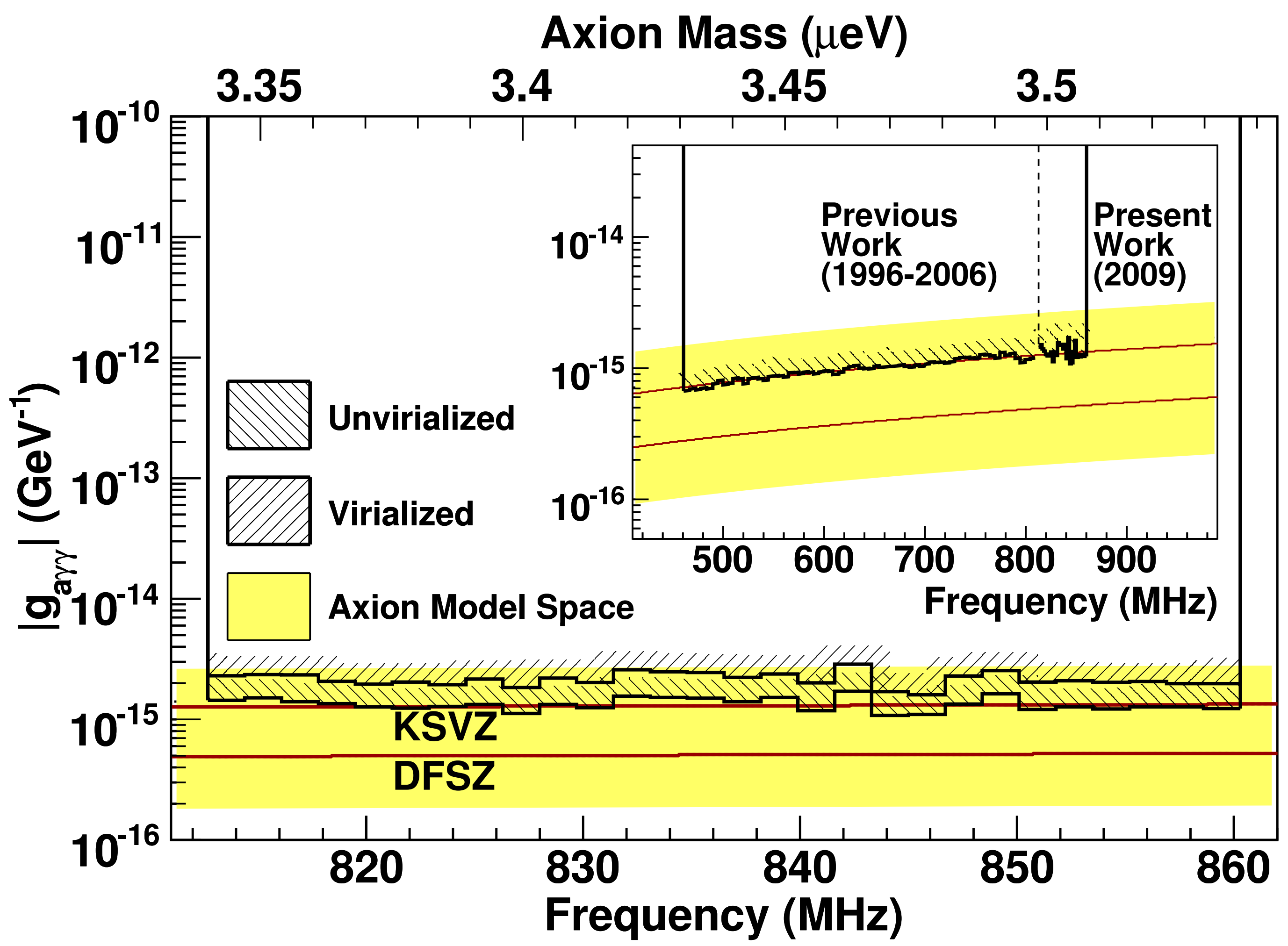
Xenon1T 1705.06655



Fermi LAT 1611.03184



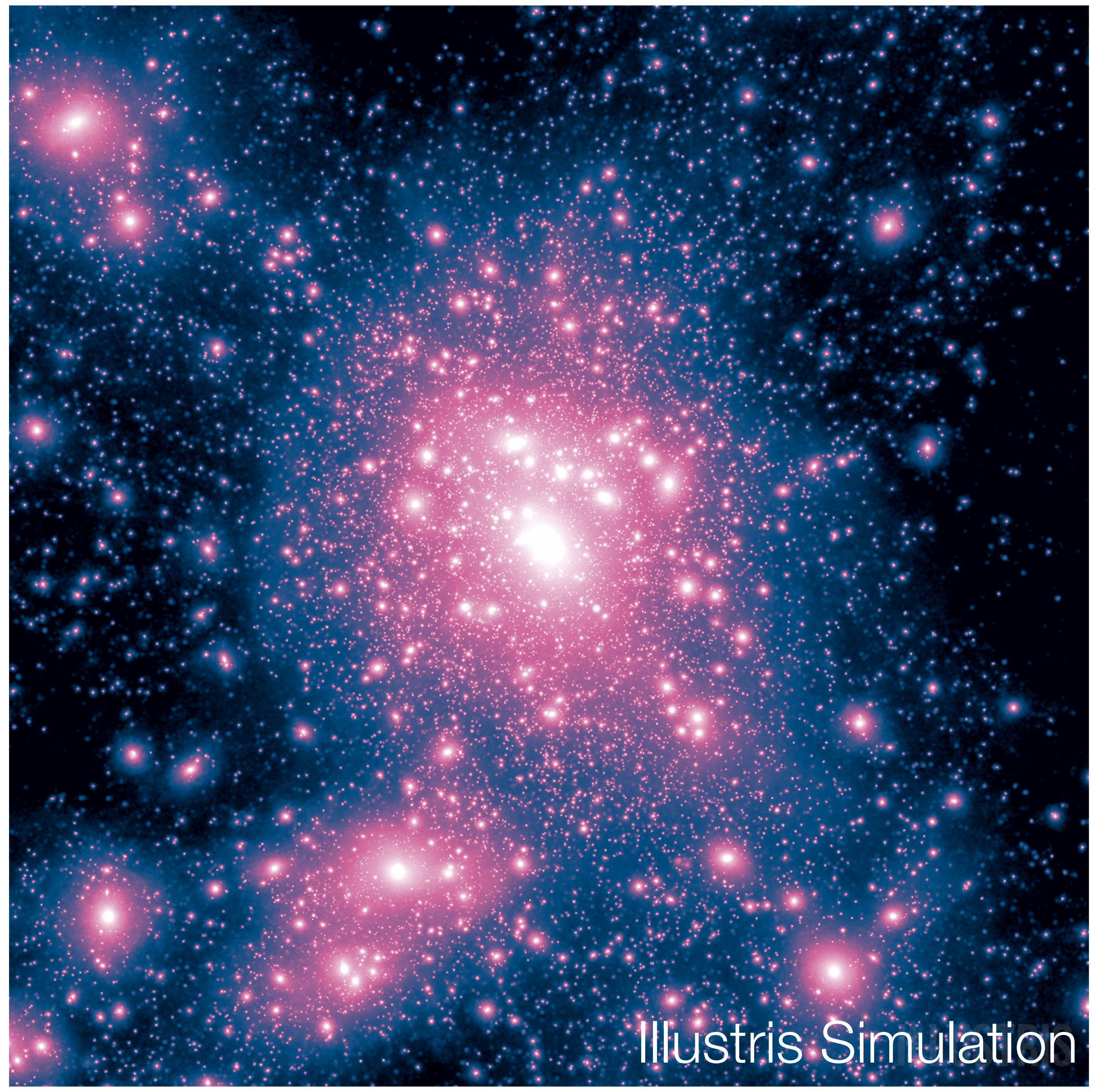




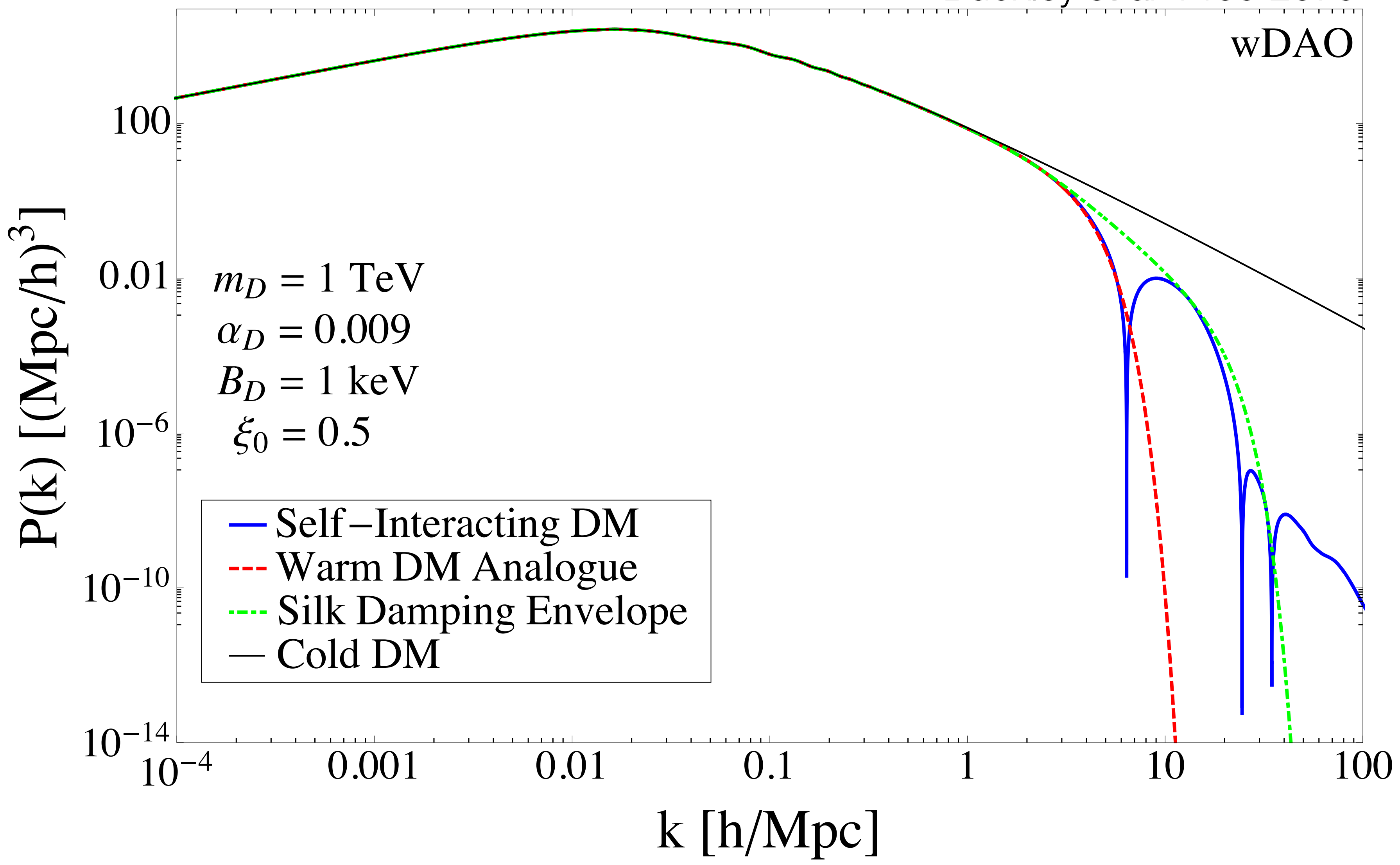
ADMX
0910.5914

Theoretical Motivations

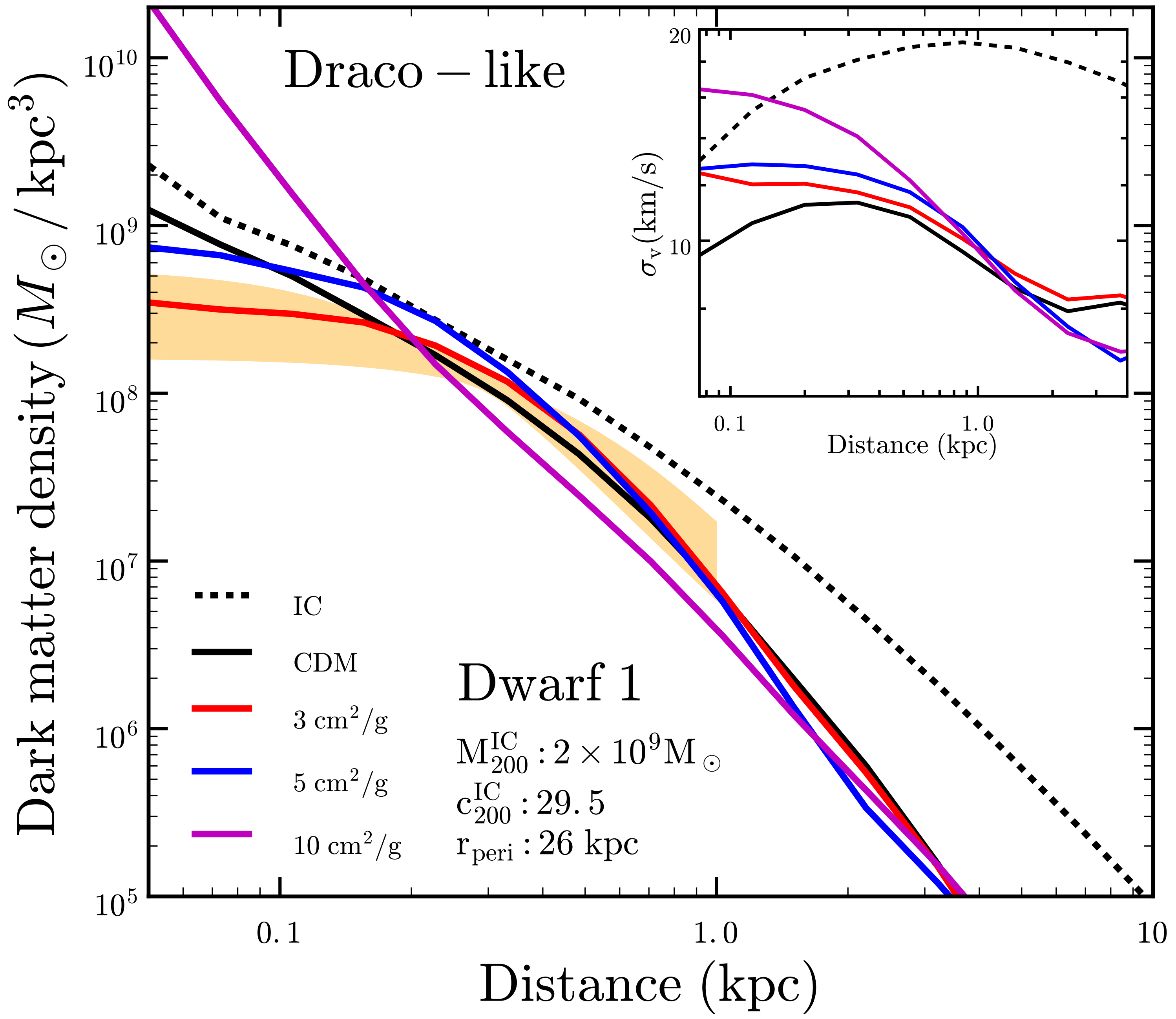
- Large-scale distribution of baryonic matter in the Universe and structure of galaxies can reveal hints of dark matter particle physics.



Buckley *et al* 1405.2075

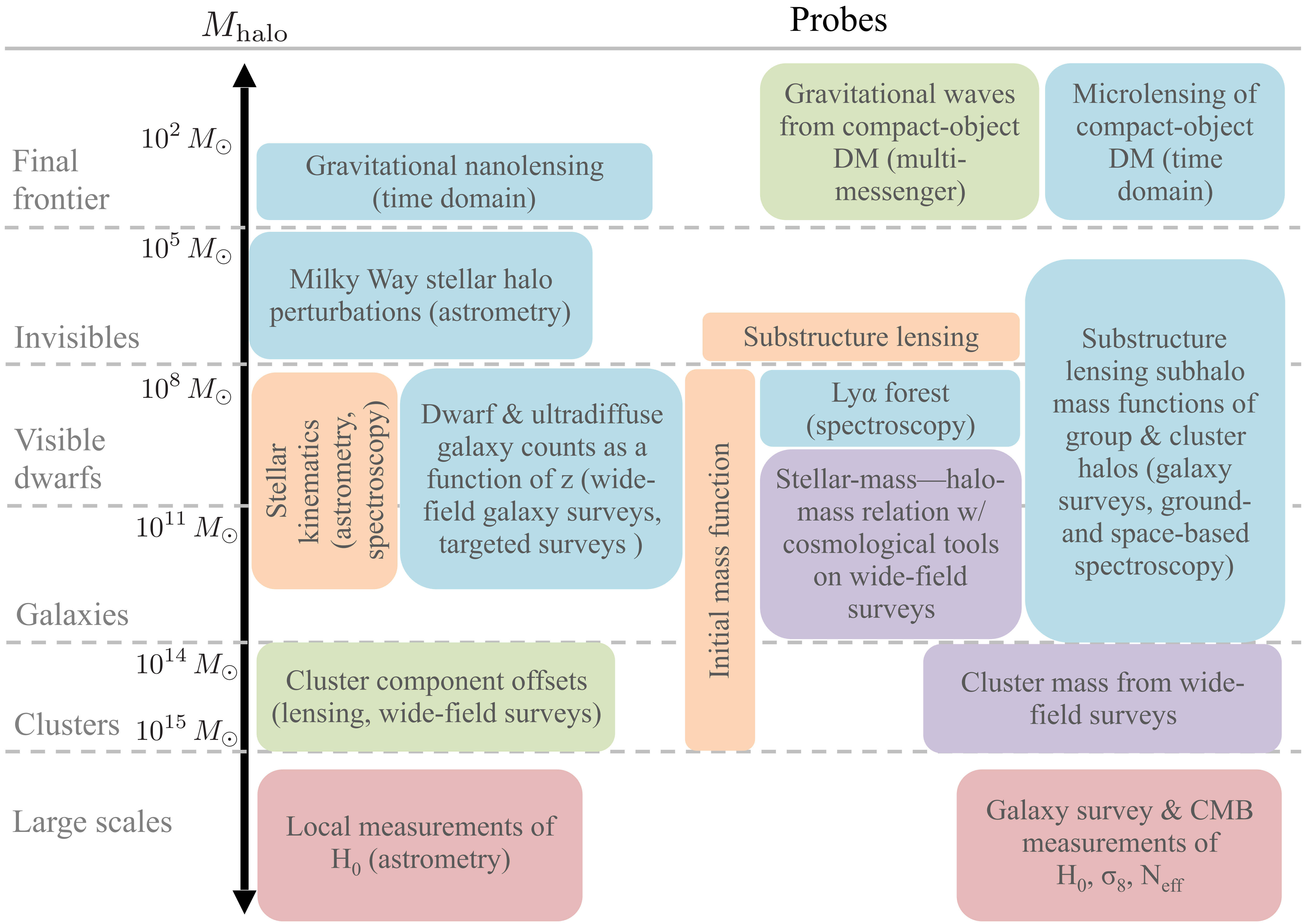


Sameie *et al* 1904.07872



Theoretical Motivations

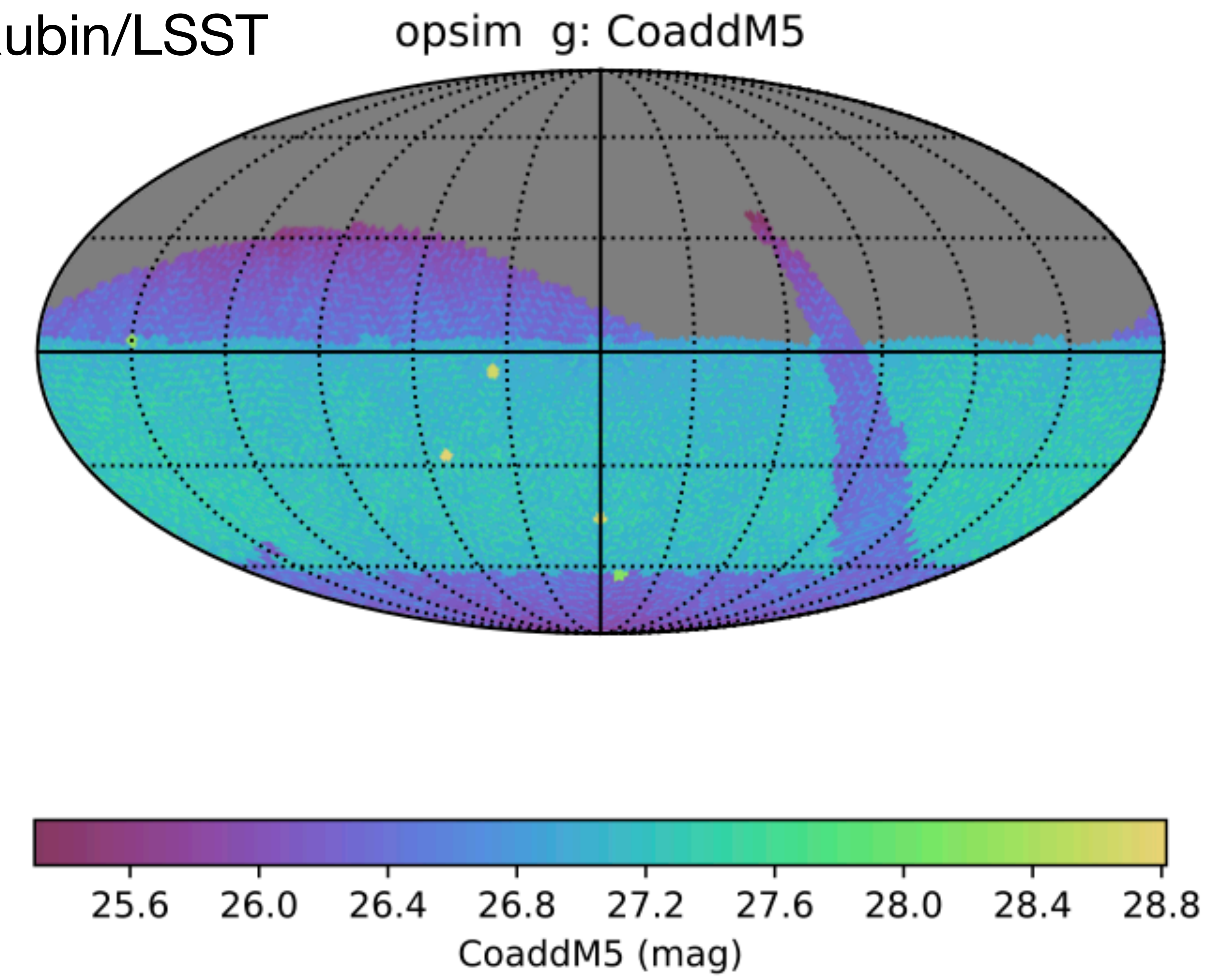
- Large-scale distribution of baryonic matter in the Universe and structure of galaxies can reveal hints of dark matter particle physics.

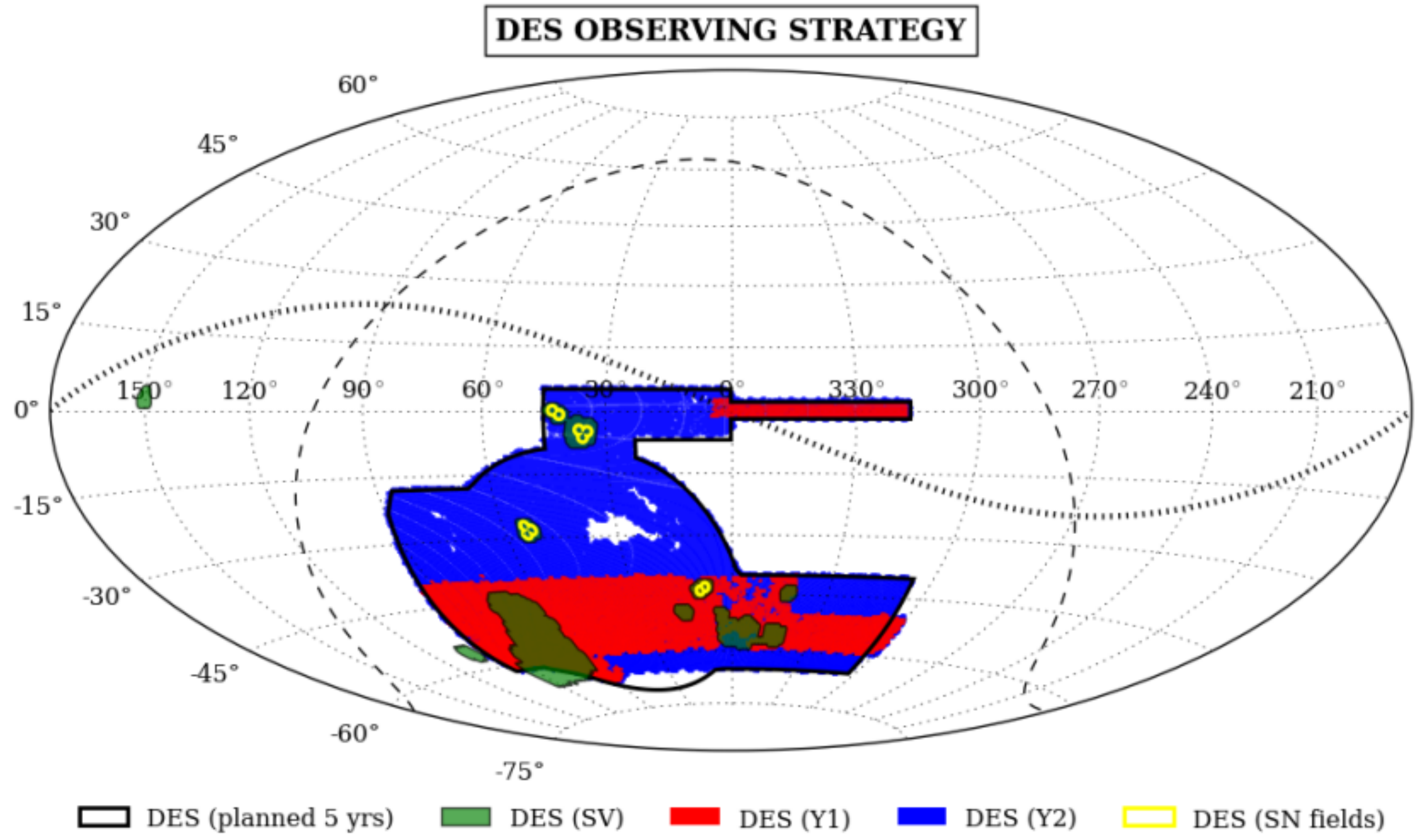


Buckley & Peter 1712.06615

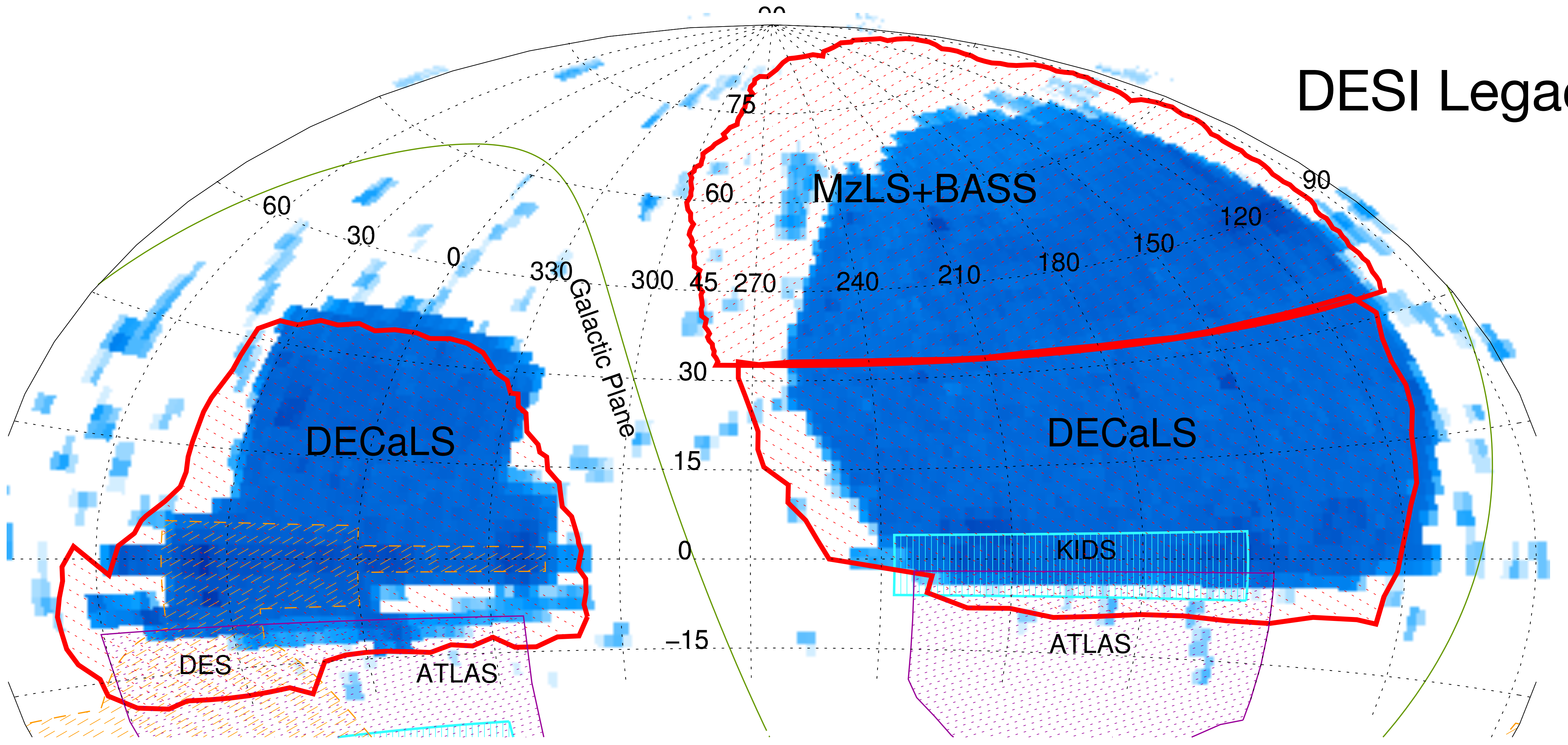
The Era of Big Astrophysical Data

Vera Rubin/LSST





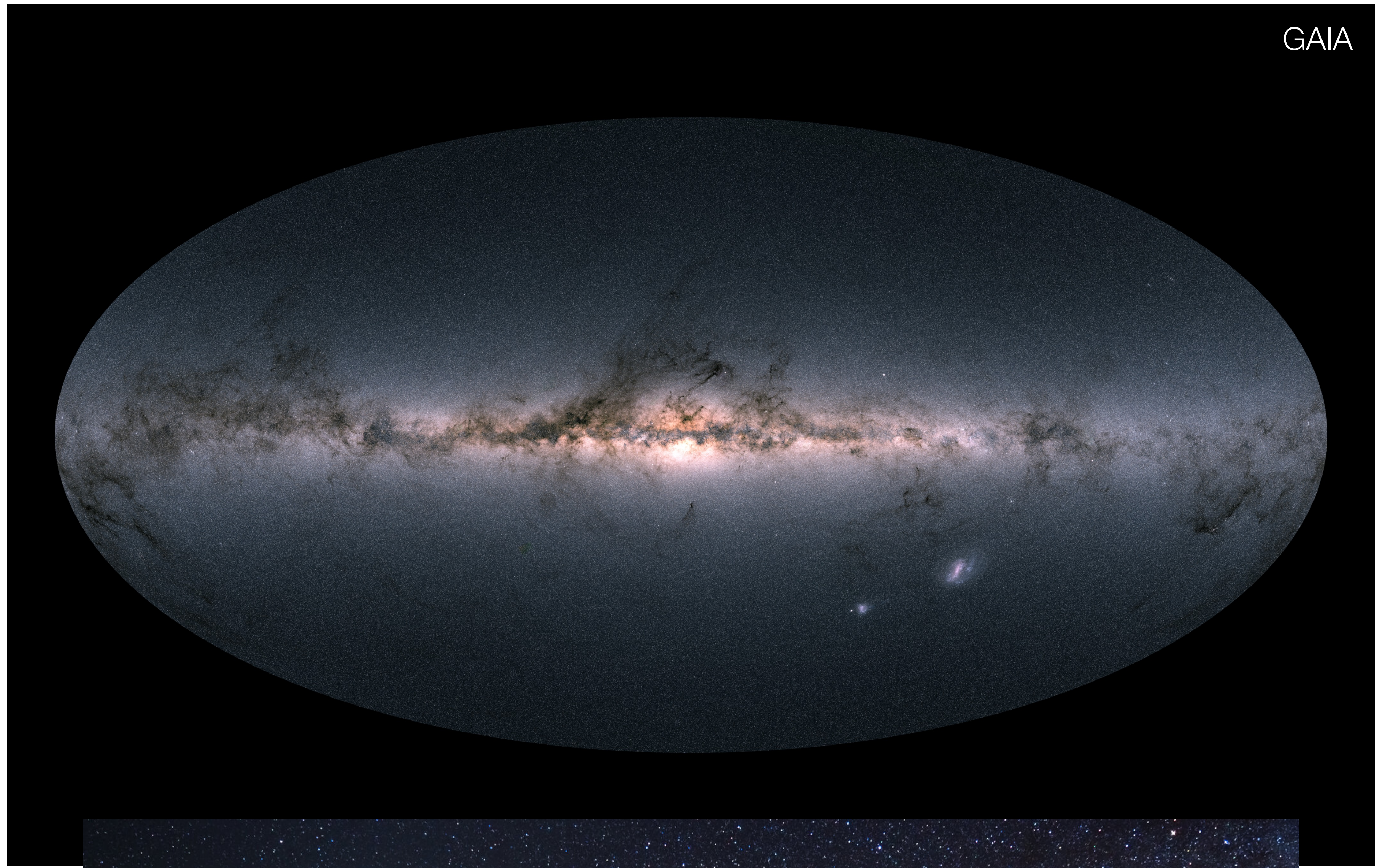
DESI Legacy

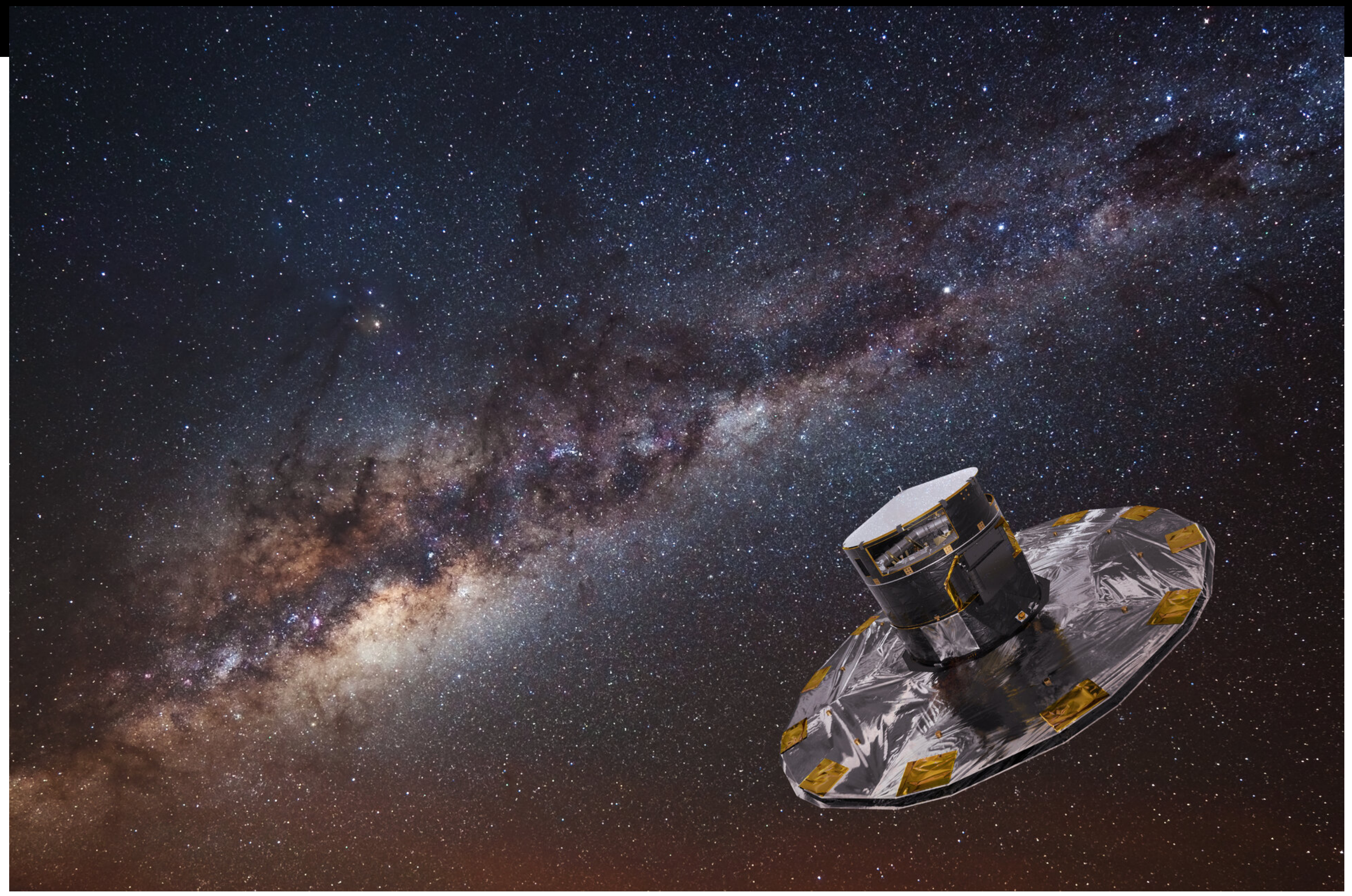


Gaia Space Telescope

- 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, we're interested in Gaia data as processed locations of stars within 4/5/6D kinematic space — not as individual images/spectra (lots of analysis here!)



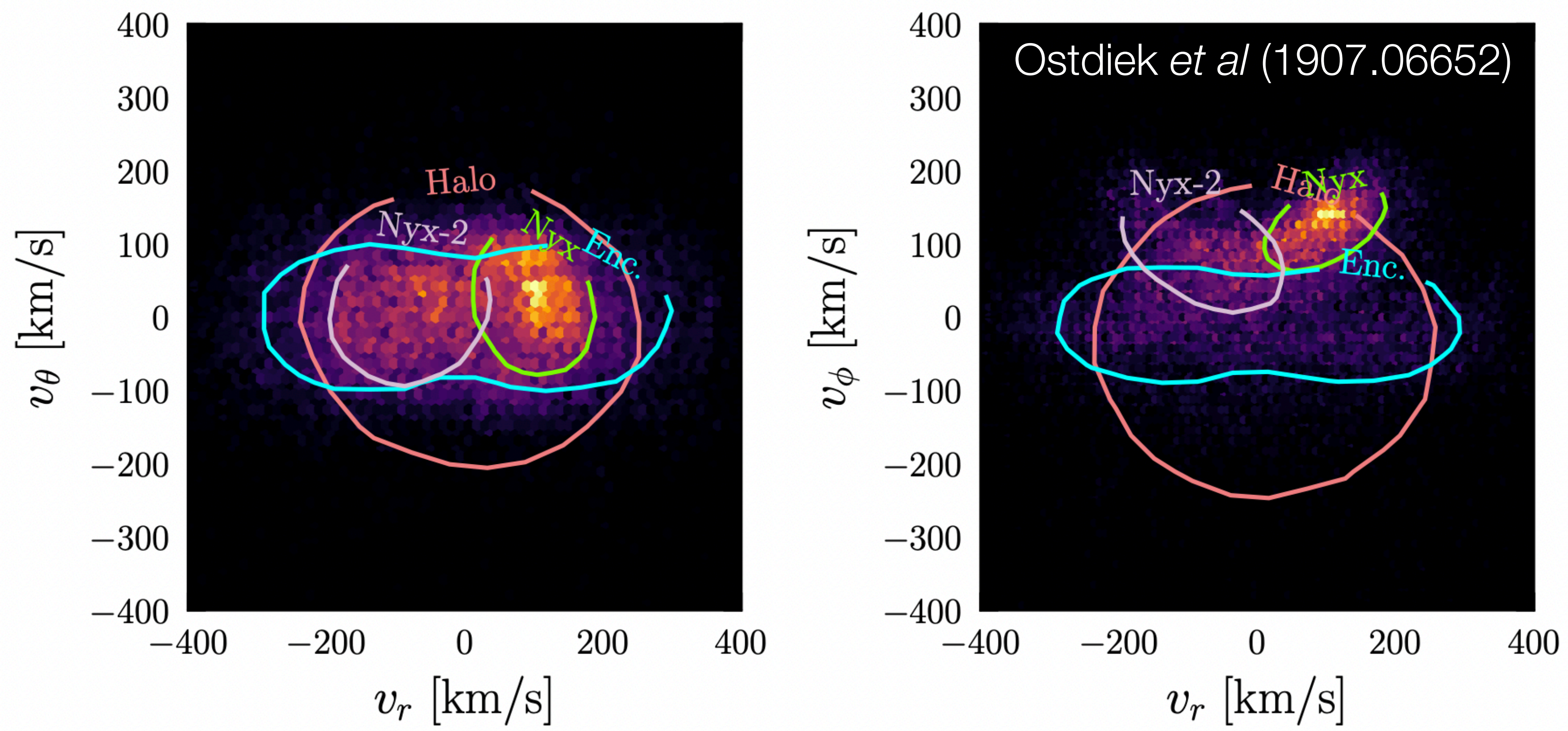


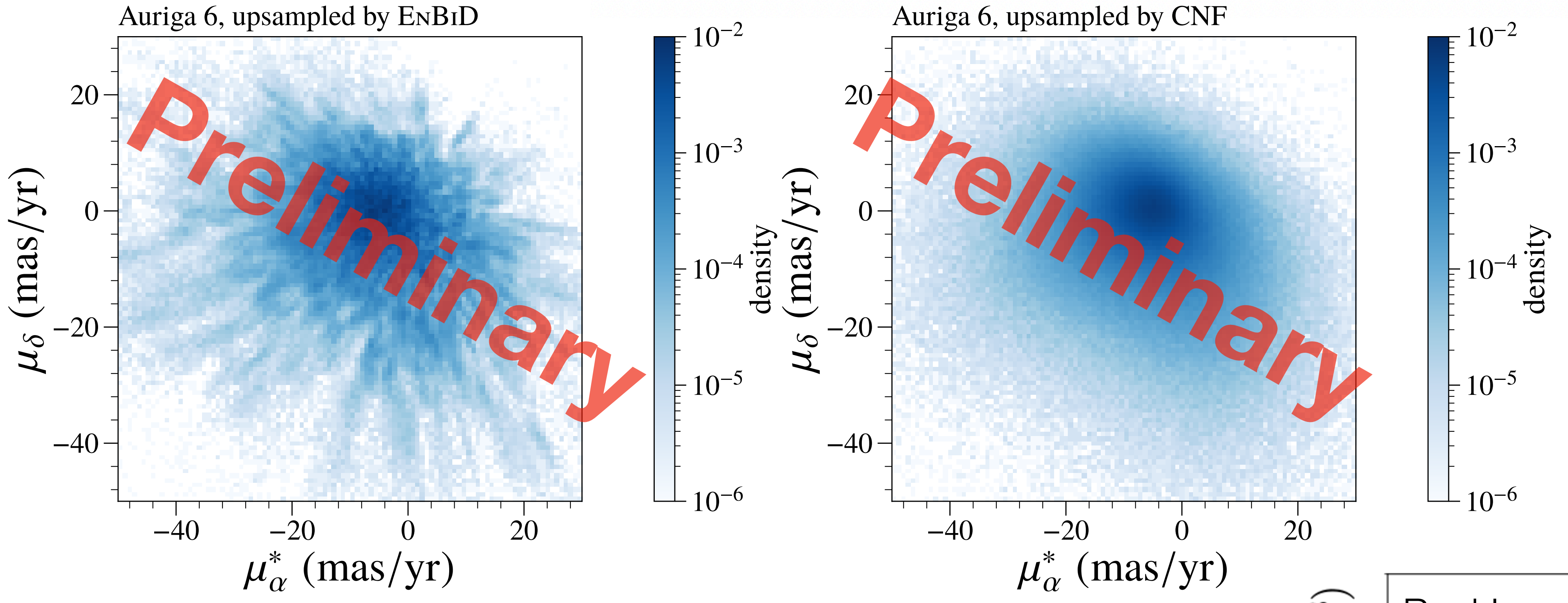
Gaia Space Telescope

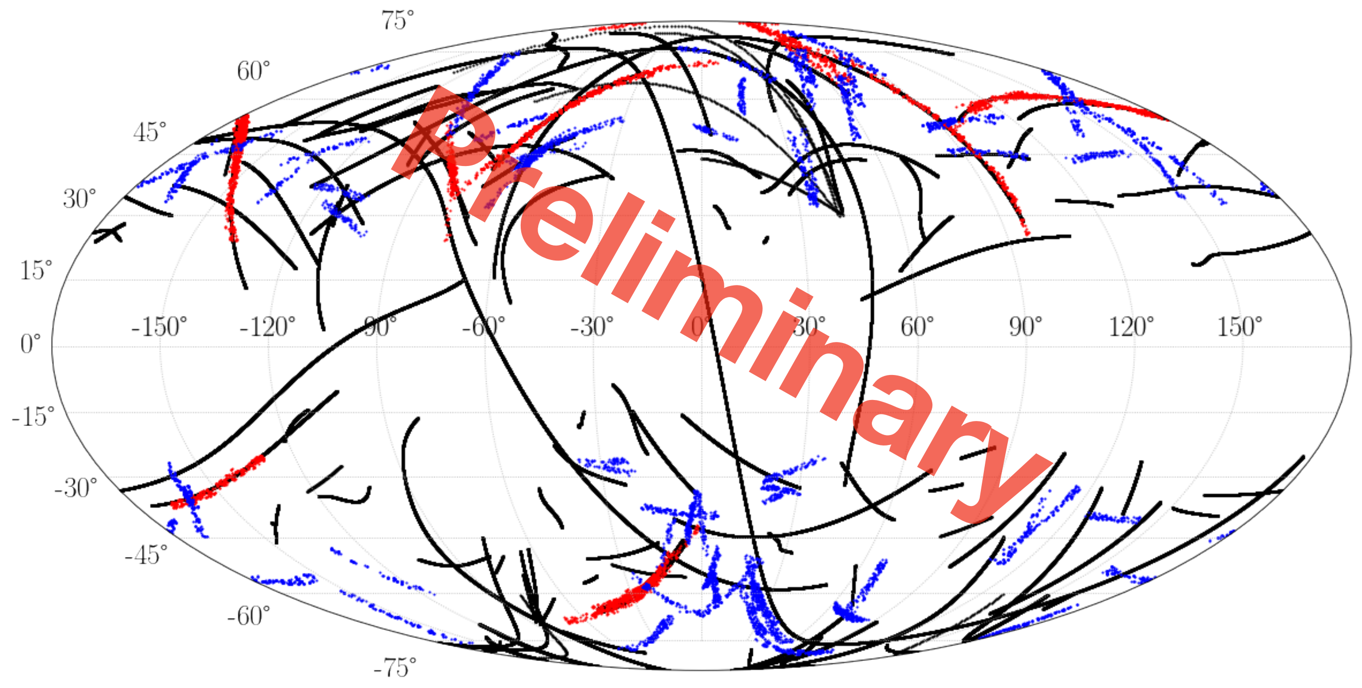
	# sources in Gaia DR3	# sources in Gaia DR2	# sources in Gaia DR1
Total number of sources	**1,811,709,771**	**1,692,919,135**	**1,142,679,769**
	Gaia Early Data Release 3		
Number of sources with full astrometry	1,467,744,818	1,331,909,727	2,057,050
Number of 5-parameter sources	585,416,709		
Number 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	-	-

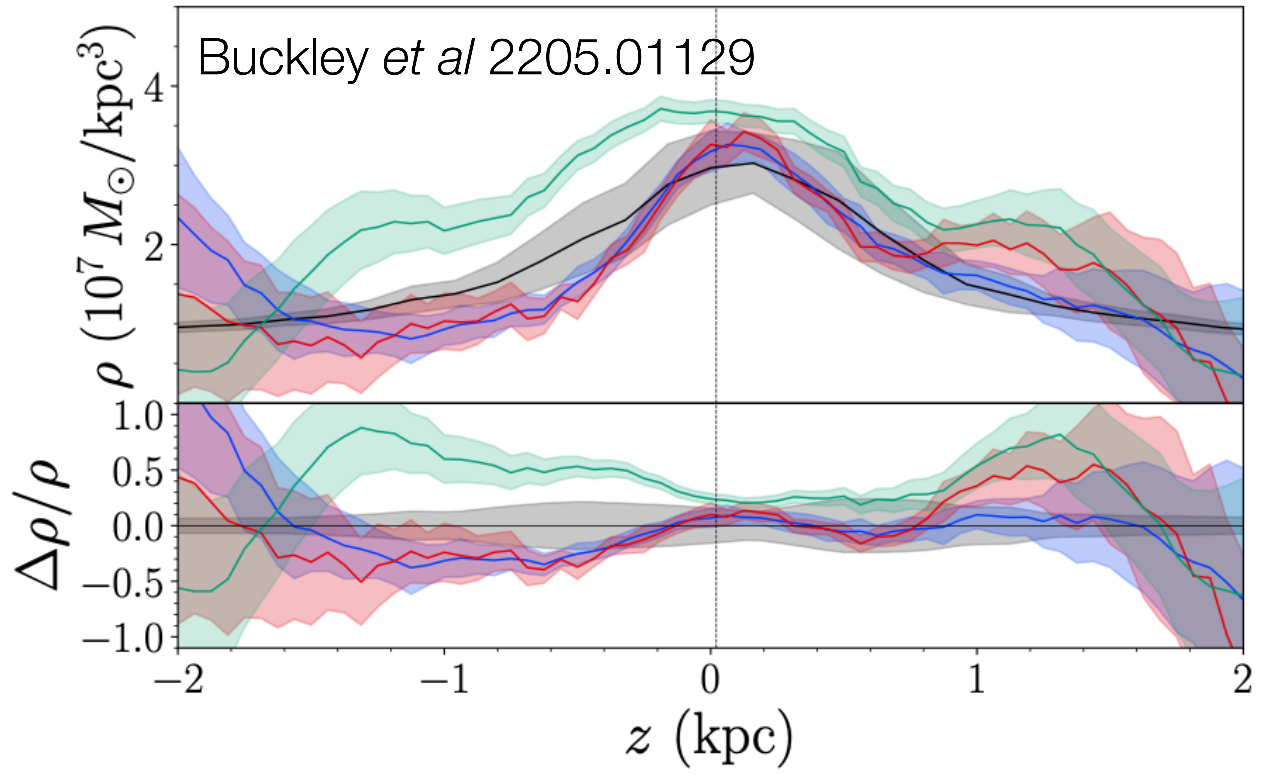
ML Applications for Gaia

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



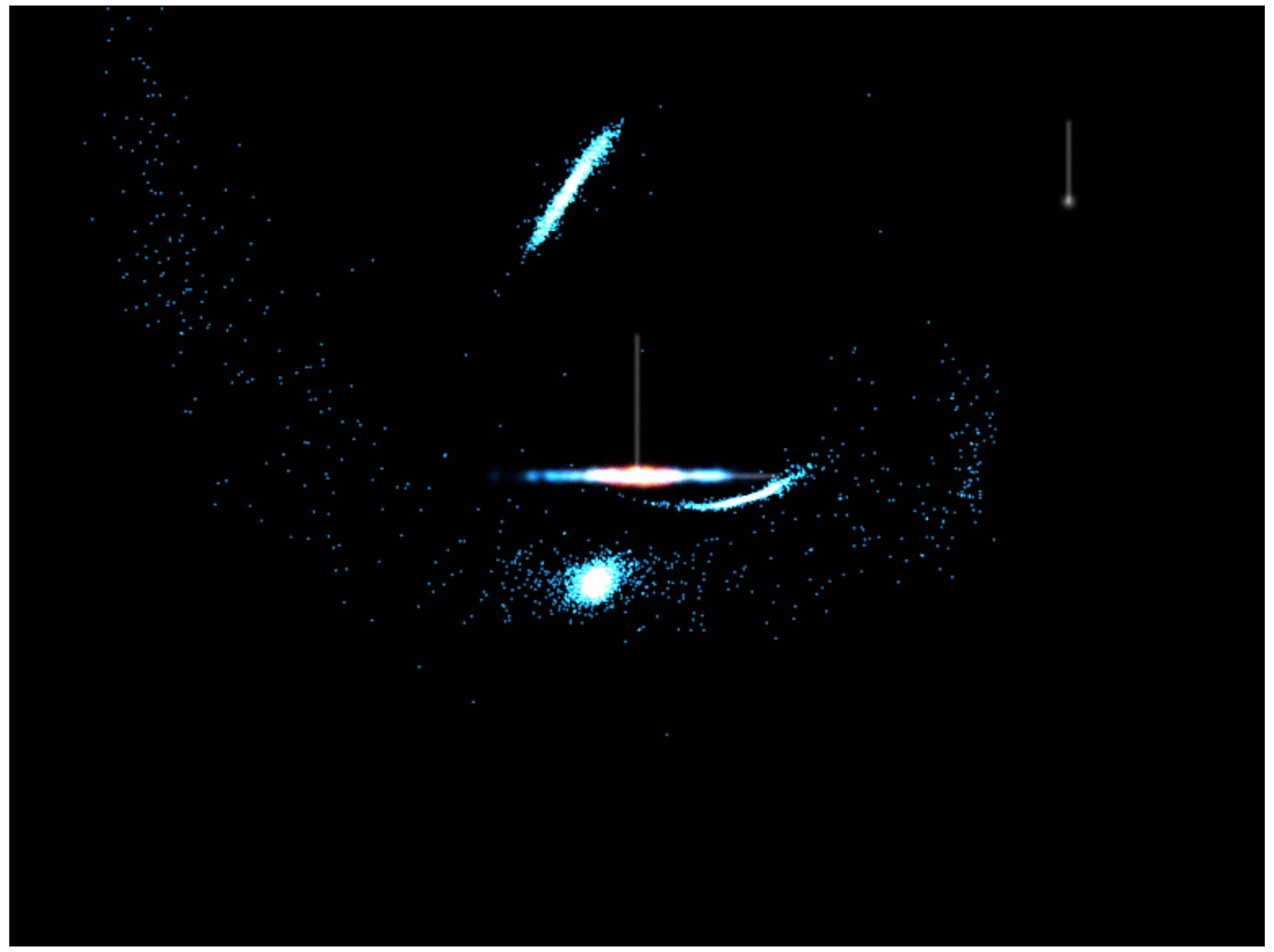




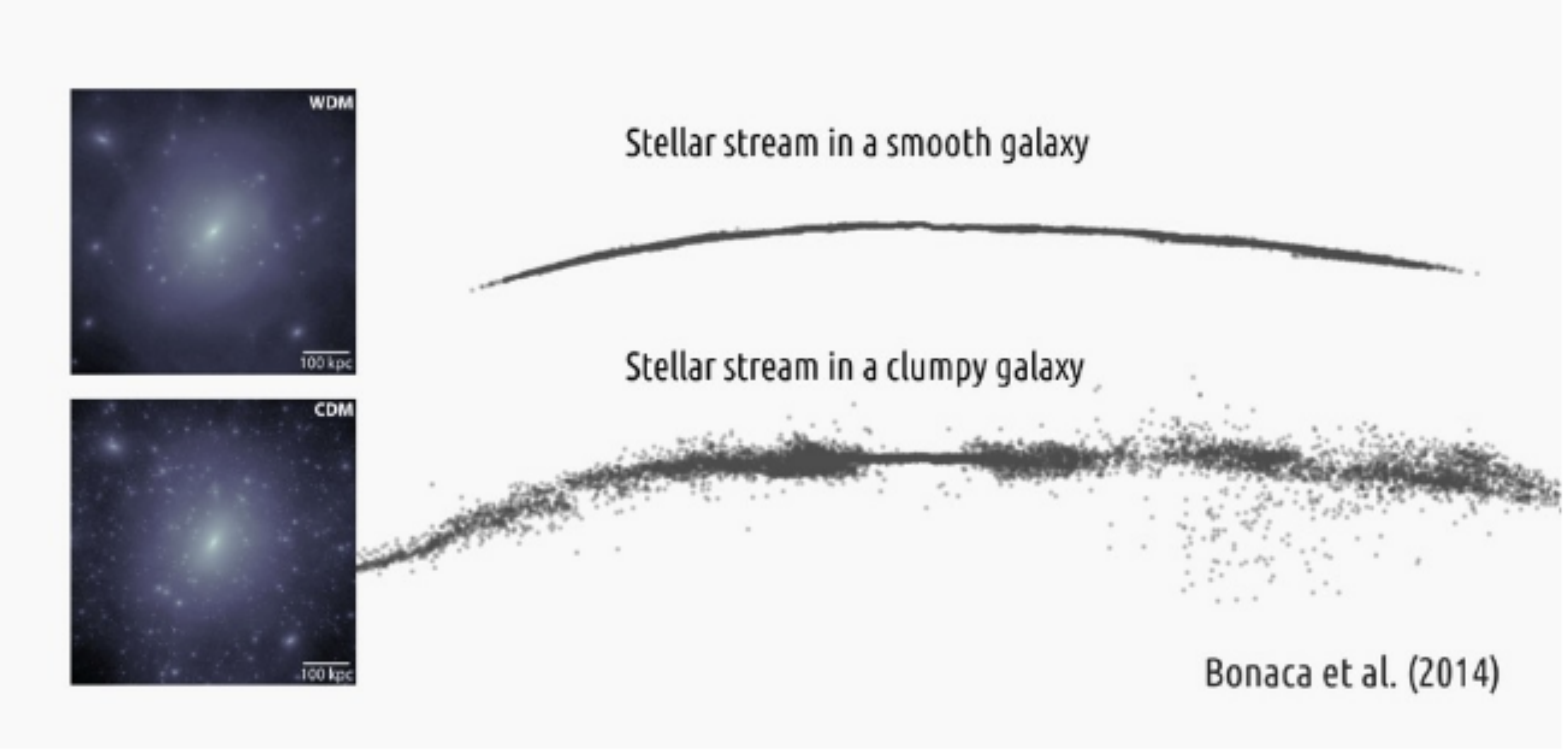


Mergers and Streams

- 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.



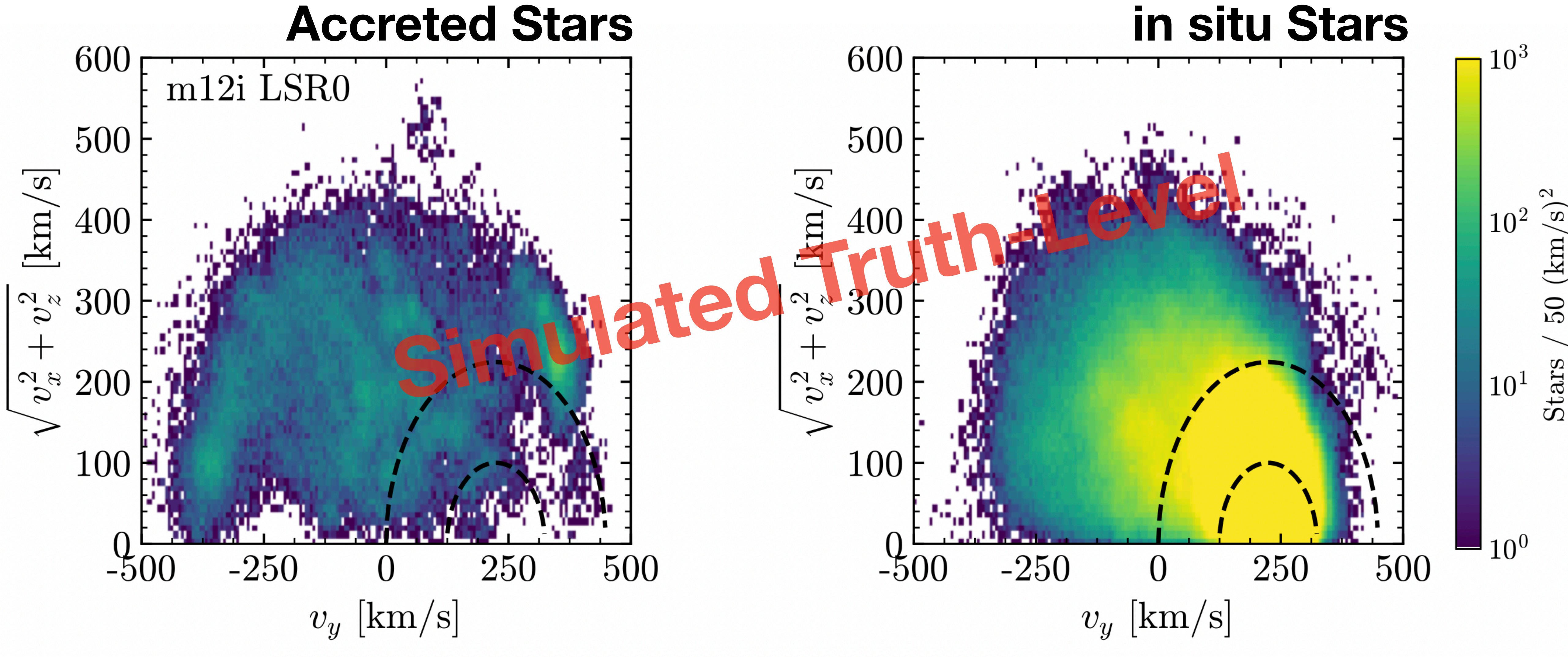
- Streams provide a probe into the Galactic potential through the stream's orbit.
  - Can reveal dark matter substructure through gravitational interactions with the stream itself.
- Both streams and debris give a glimpse into the Galaxy's merger history.

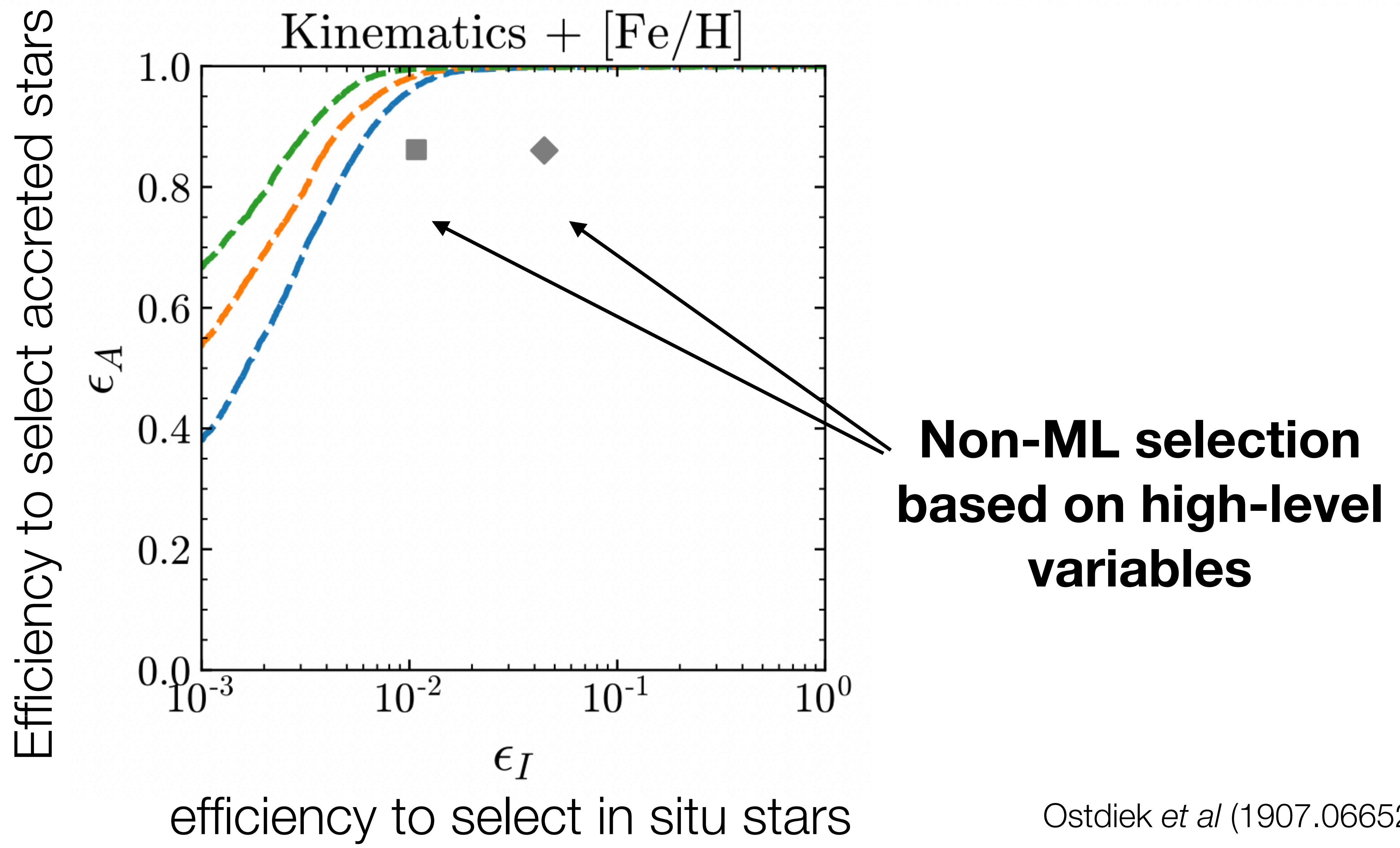


Bonaca et al. (2014)

Finding Tidal Debris with Machine Learning

- Stars that originate from dwarf galaxies will have different kinematics and metallicities, even after they are well-mixed into the Milky Way's halo in position space.
  - Ostdiek *et al*. (1907.06652) train a classifier on simulated Milky Way-like galaxies to distinguish halo stars that are formed in-situ versus accreted.
  - Trained on one simulated galaxy, demonstrated that network results transfer to 2nd simulated galaxy.
- Applied to Gaia DR2 (Necib *et al* 1907.07681), reidentifies known substructure within the halo, but also a new merger component: Nyx

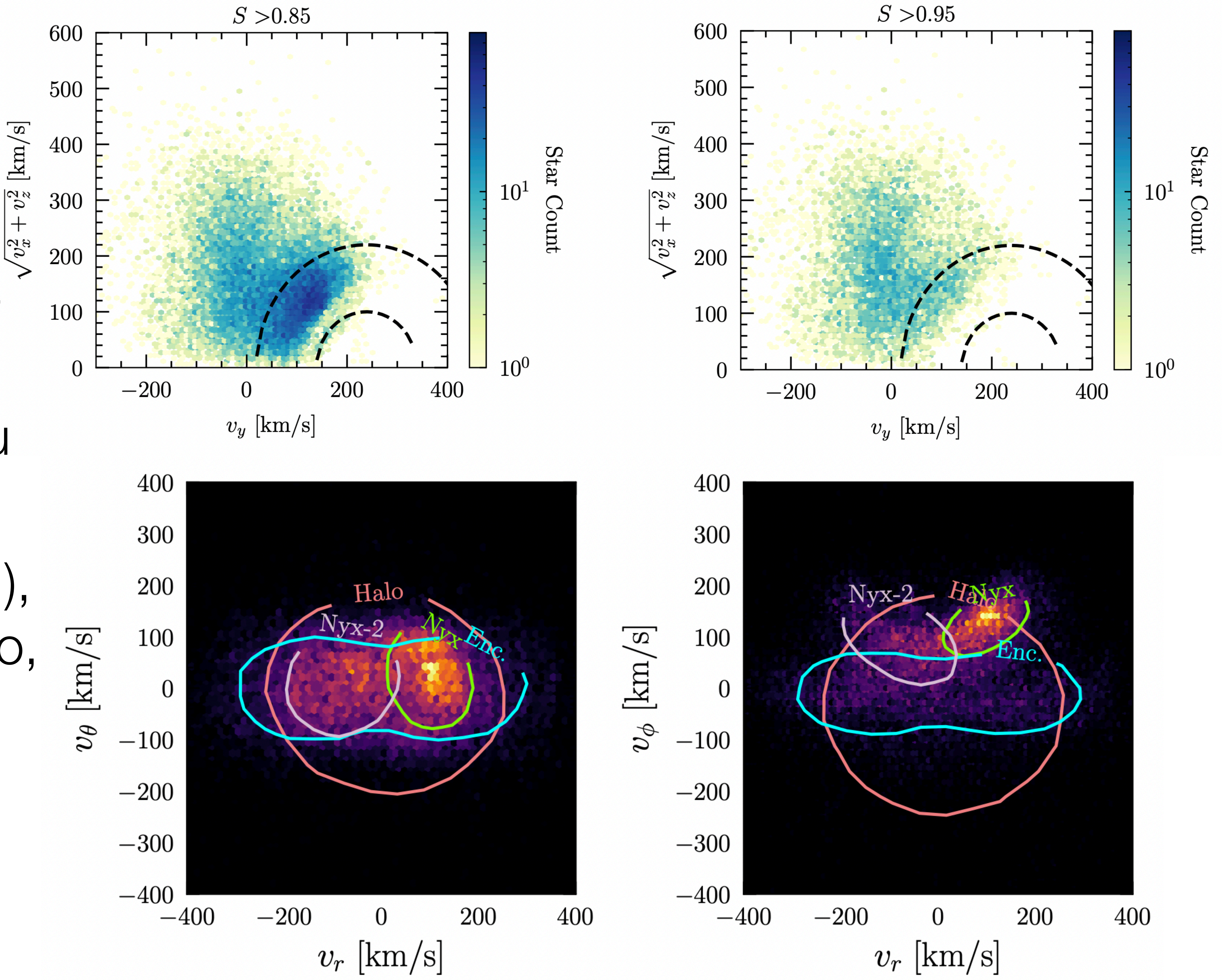




Ostdiek *et al* (1907.06652)

Finding Dwarf Debris with Machine Learning

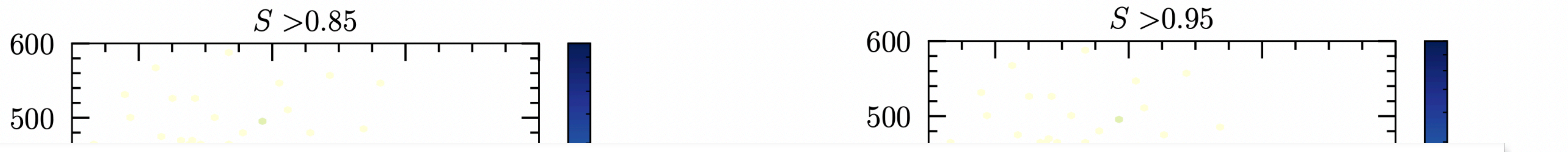
- Stars that originate from dwarf galaxies will have different kinematics and metallicities, even after they are well-mixed into the Milky Way's halo in position space.
  - Ostdiek *et al*. (1907.06652) train a classifier on simulated Milky Way-like galaxies to distinguish halo stars that are formed in-situ versus accreted.
- Applied to Gaia DR2 (Necib *et al* 1907.07681), reidentifies known substructure within the halo, but also a new merger component: Nyx
  - A very large stellar stream/debris (Necib *et al*. 1907.07190)



Ostdiek *et al* (1907.06652)

Finding Tidal Debris with Machine Learning

- Stars that originate from dwarf galaxies will have different kinematics and metallicities



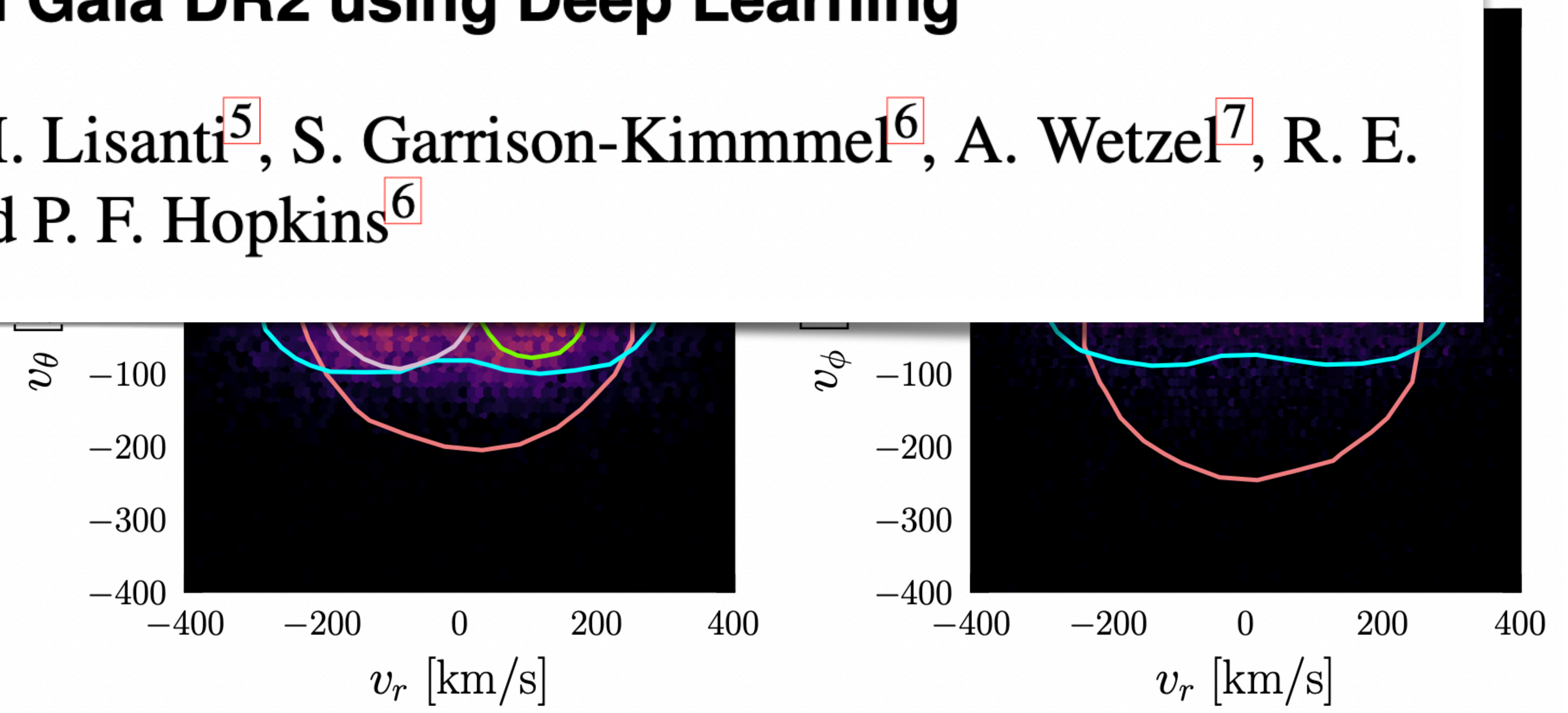
Chasing Accreted Structures within Gaia DR2 using Deep Learning

Lina Necib,[1,2] Bryan Ostdiek,[3] Mariangela Lisanti,[4] Timothy Cohen,[3] Marat Freytsis,[5,6] and Shea Garrison-Kimmel[7]

Cataloging Accreted Stars within Gaia DR2 using Deep Learning

B. Ostdiek ⋆[1], L. Necib[2], T. Cohen[1], M. Freytsis[3,4], M. Lisanti[5], S. Garrison-Kimmmel[6], A. Wetzel[7], R. E. Sanderson[8,9], and P. F. Hopkins[6]

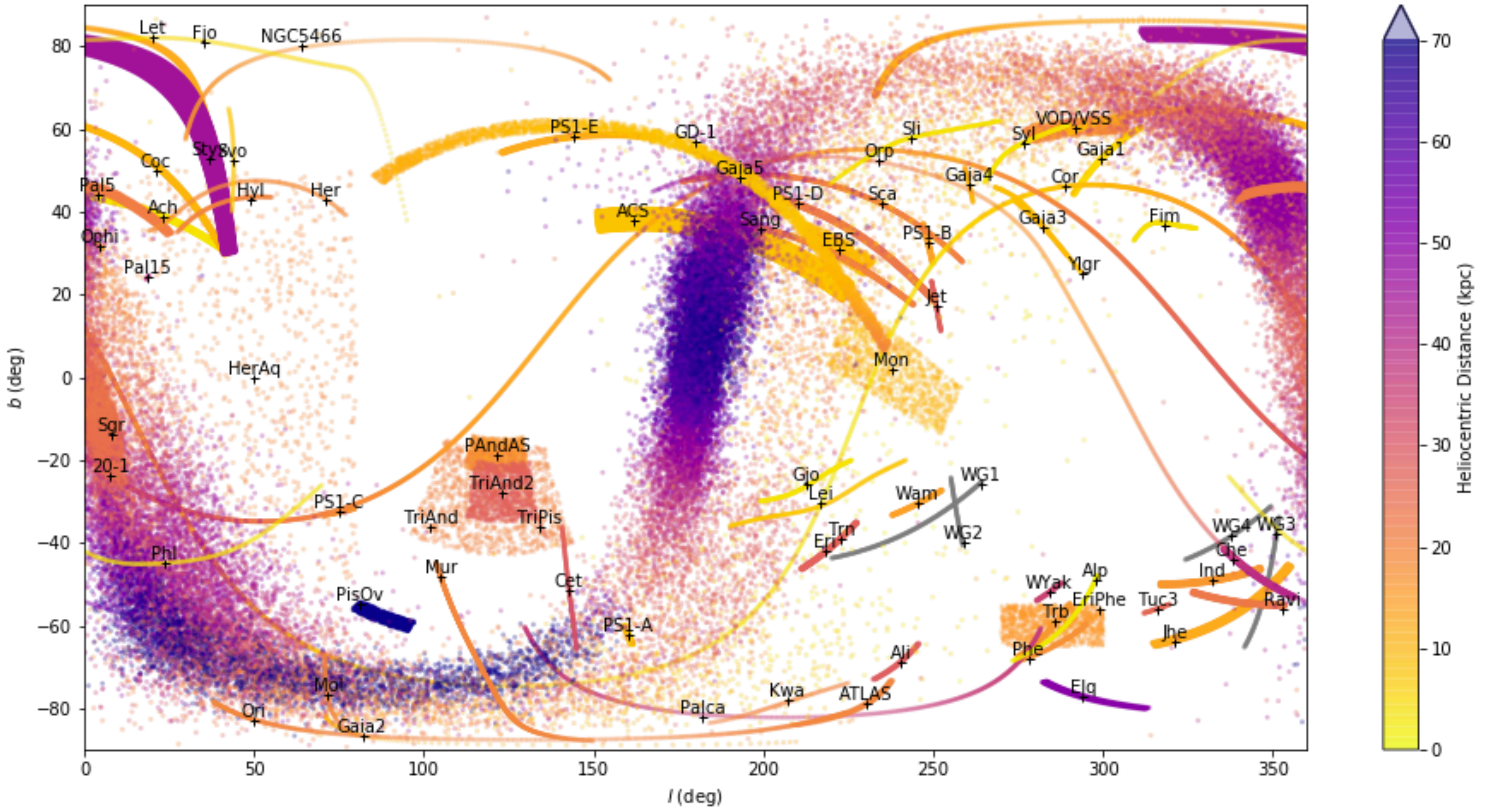
- but also a new merger component: Nyx
  - A very large stellar stream/debris (Necib *et al*. 1907.07190)



Ostdiek *et al* (1907.06652)

Finding Stellar Streams

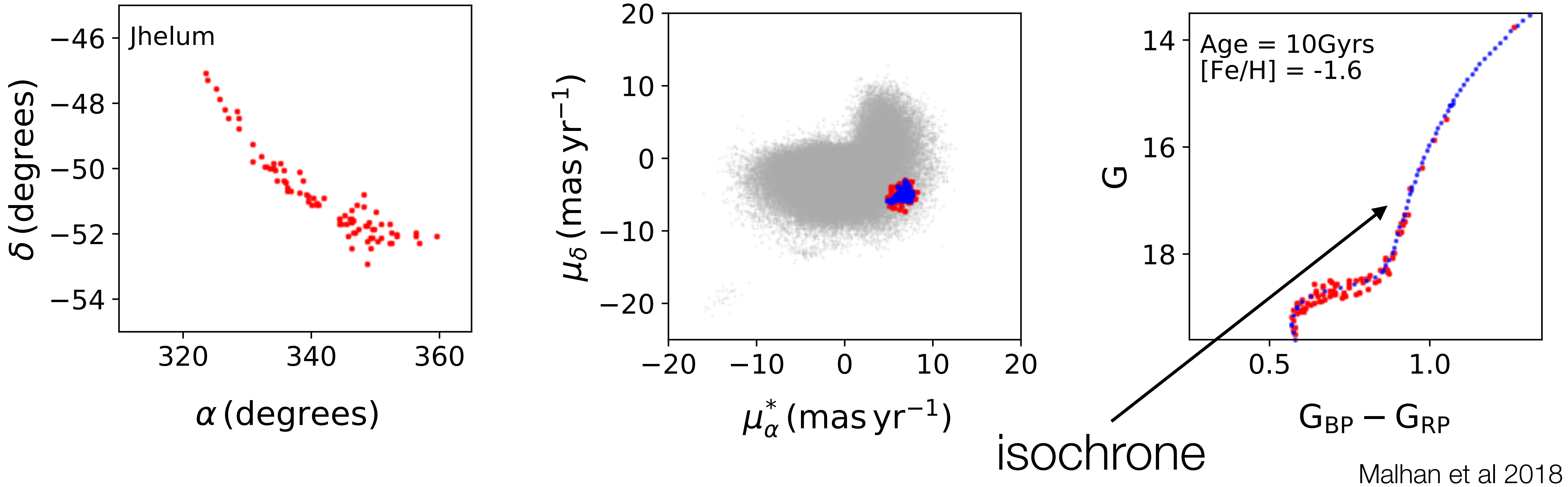
- Narrow & kinematically cold stellar streams are tracers of the Milky Way potential, merger history, imprint of dark matter substructure…
- A stellar stream is a narrow line of stars, compact in proper motion, and with all stars typically of similar age and composition.
- Use ML to build a stream-finding 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

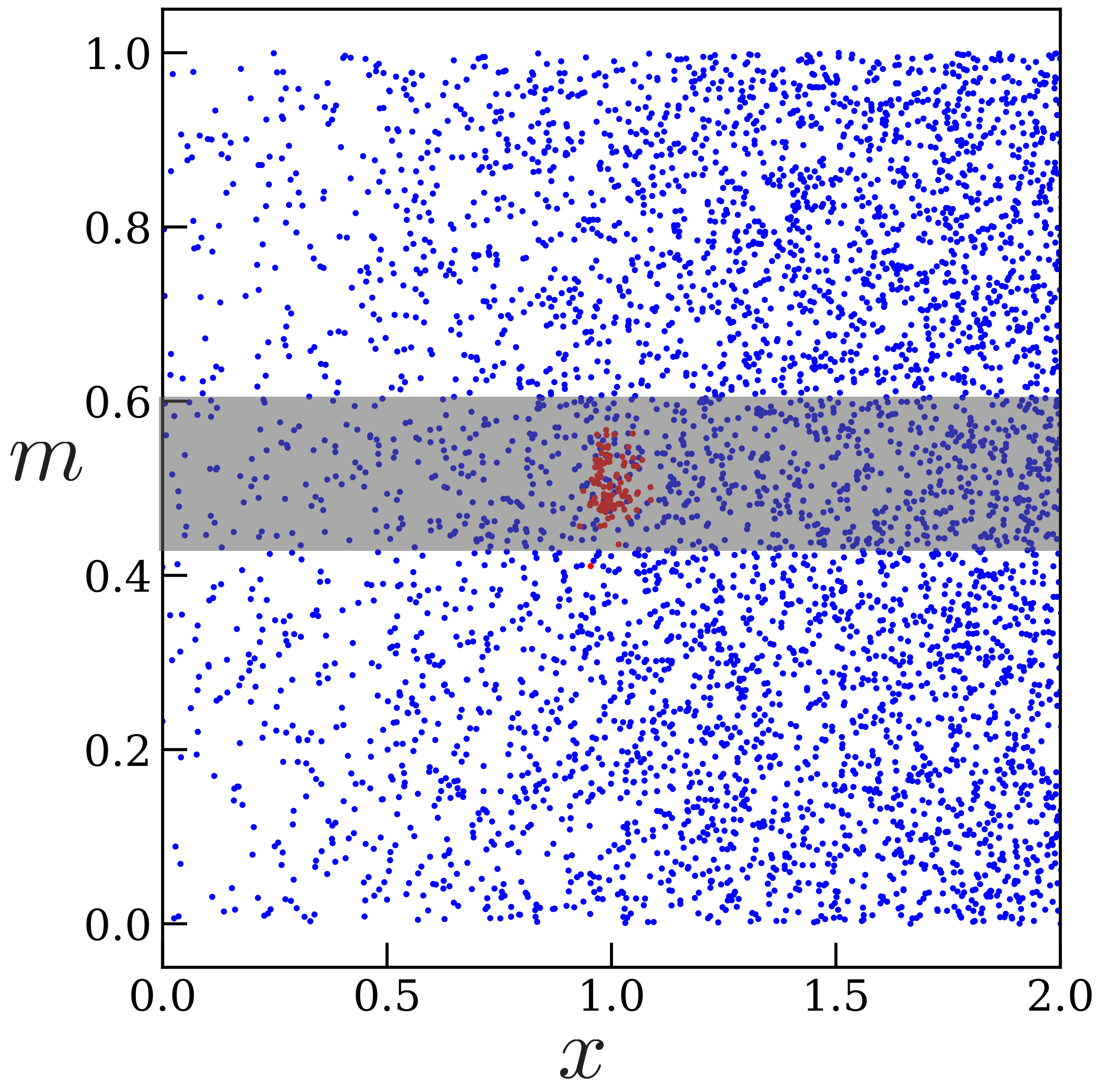
Stellar brightness and color



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.

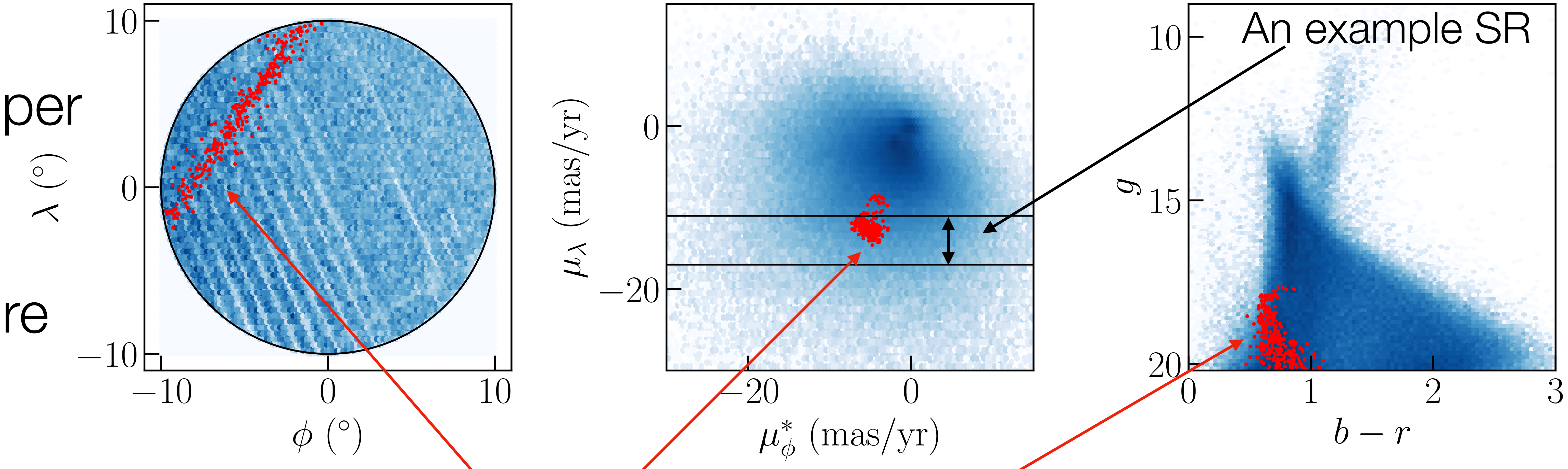


$$R(\vec{x}|m \in \mathrm{SR}) = \frac{P(\vec{x}|m \in \mathrm{SR})}{P_{\rm CR}(\vec{x}|m \in \mathrm{SR})}$$

Shih *et al* (2104.12789)

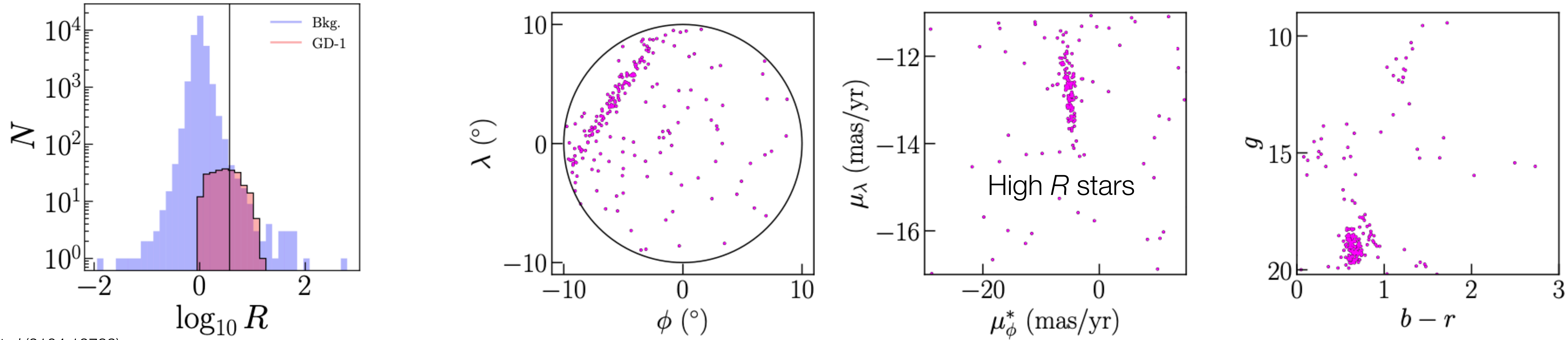
Via Machinae: Unsupervised Stream Finding

- 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)



Shih, Buckley, Necib, Tamanas (2104.12789)

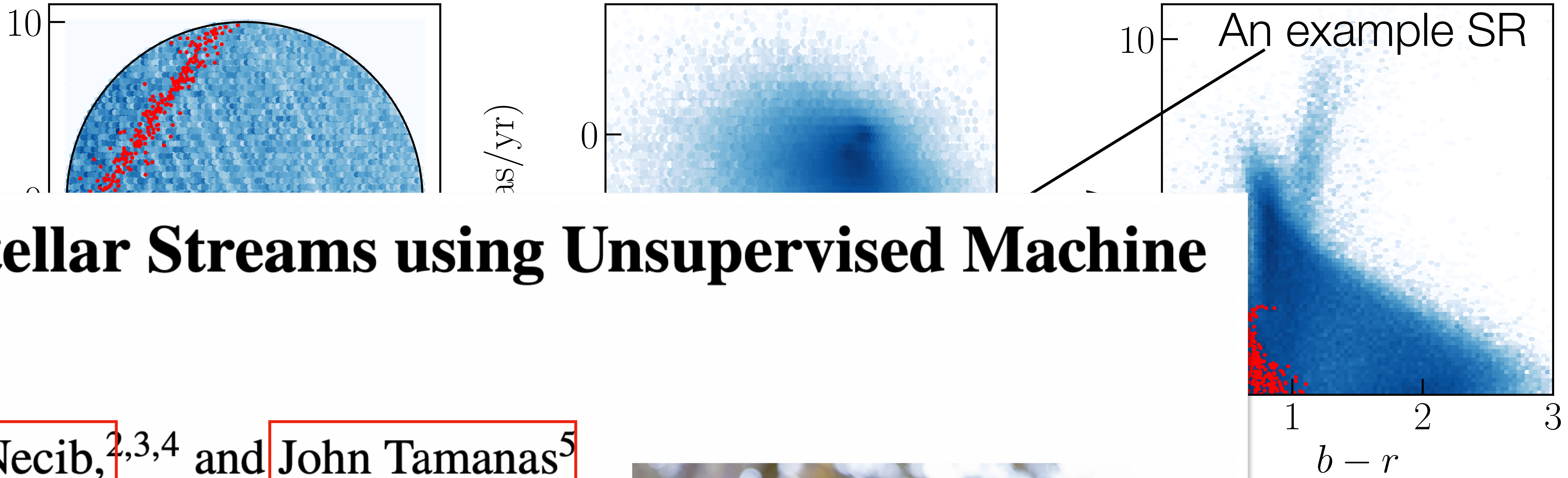
Stars identified as likely GD-1 members by Price-Whelan & Bonaca



Shih *et al* (2104.12789)

Via Machinae: Unsupervised Stream Finding

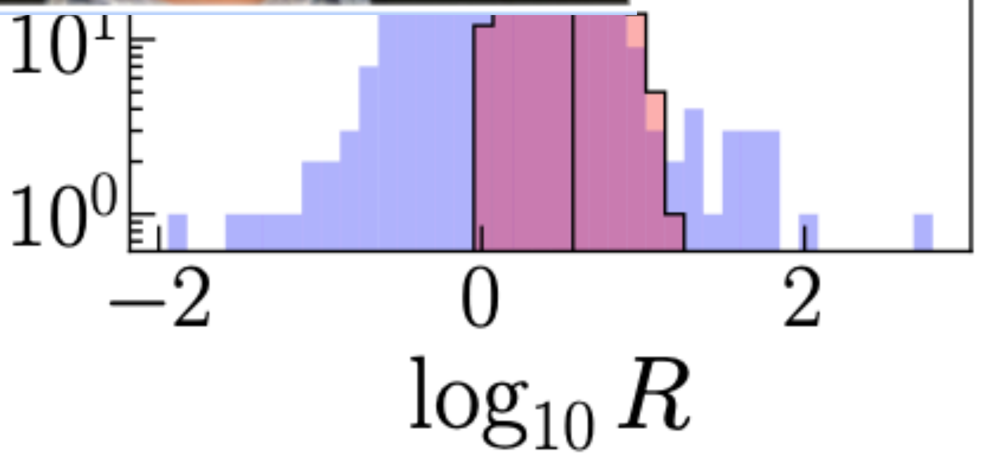
- Want to find stars that are anomalous based on their position in position, proper motion, and photometry. Use ANODE ... anomal... 2001.0... *R* for st... Regions...

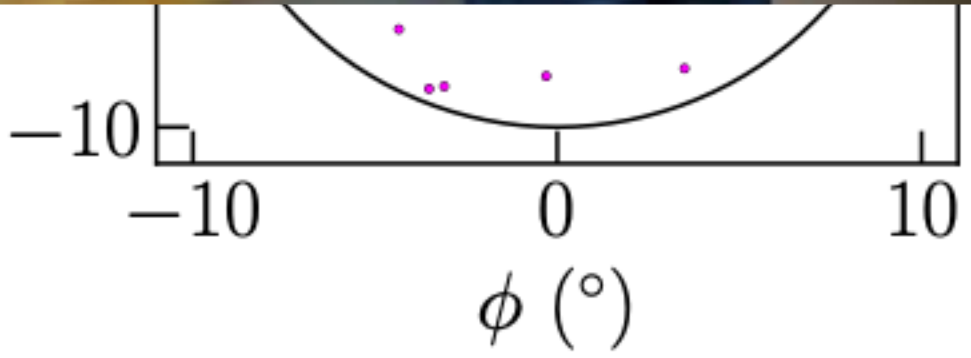


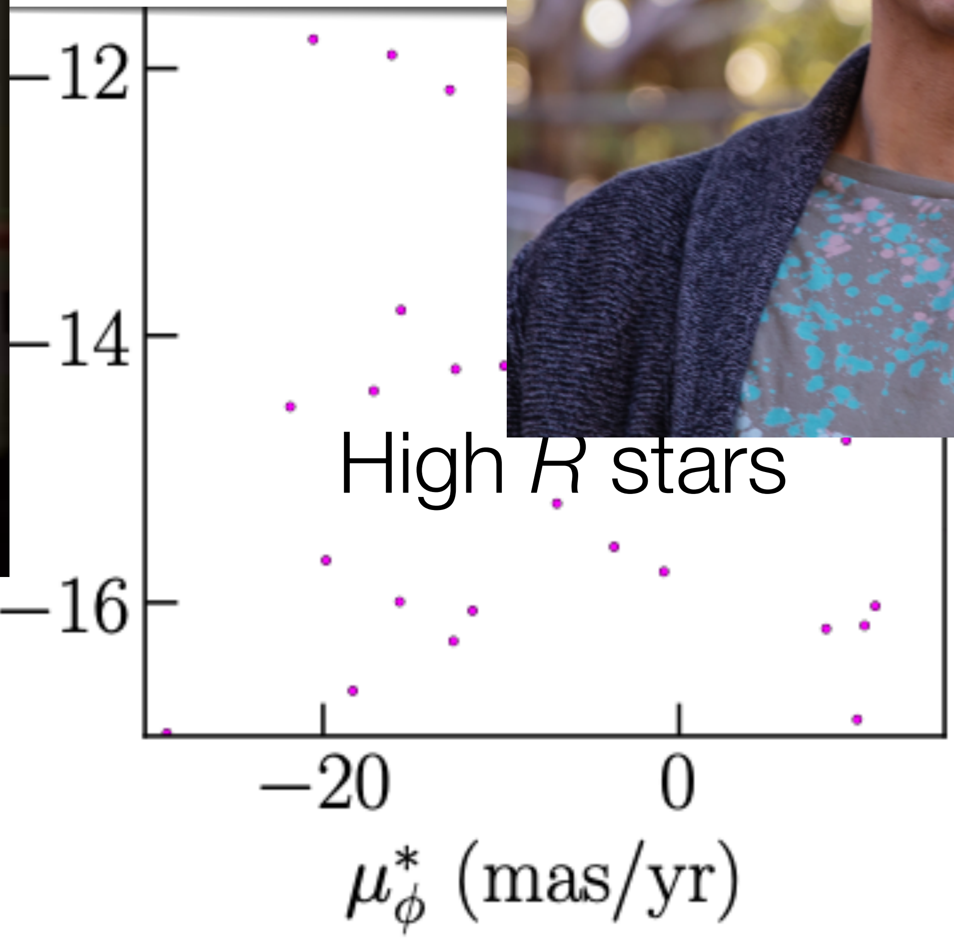
Via Machinae: Searching for Stellar Streams using Unsupervised Machine Learning

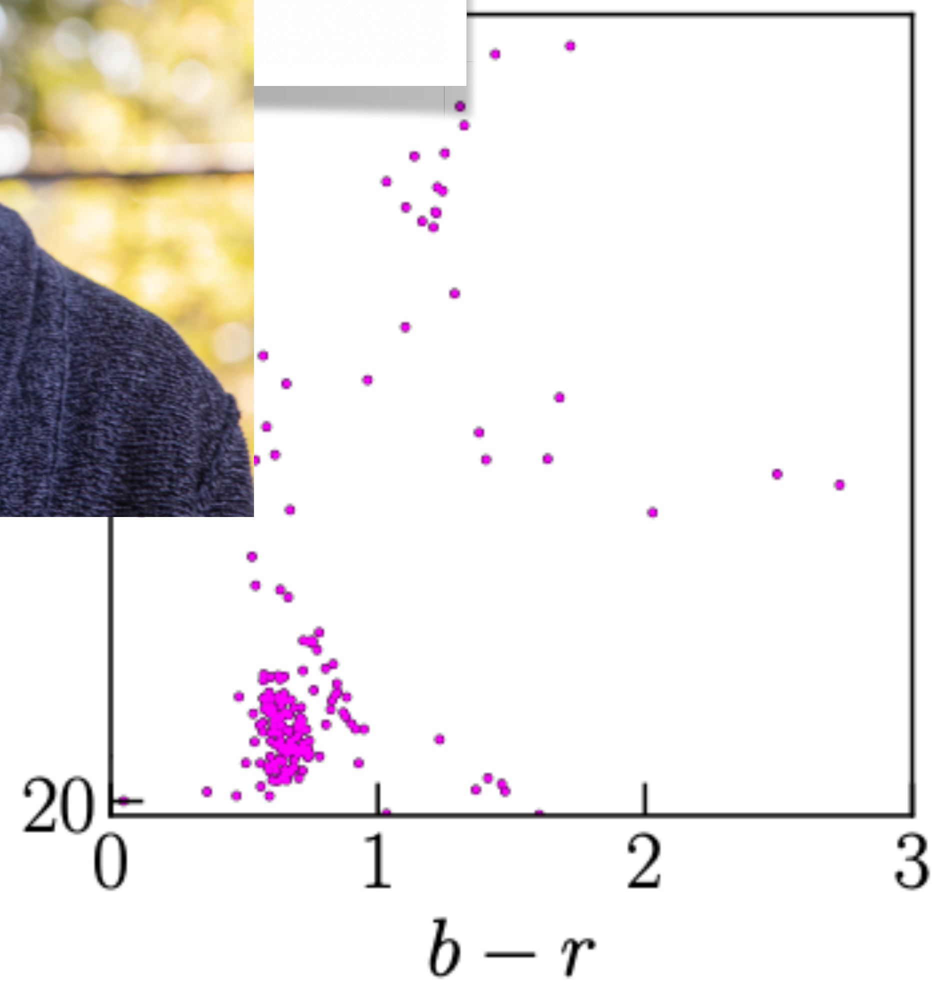
David Shih,[1★] Matthew R. Buckley,[1] Lina Necib,[2,3,4] and John Tamanas[5]
[1]NHETC, Dept. of Physics and Astronomy, Rutgers, Piscataway, NJ 08854, USA
... Institute for Theoret... CA 91125, USA
...smology, Departme... e, CA 92697, USA
...s of the Carnegie Ins... 91101, USA
...of Physics, University... California 95064, US...

...n & Bonaca







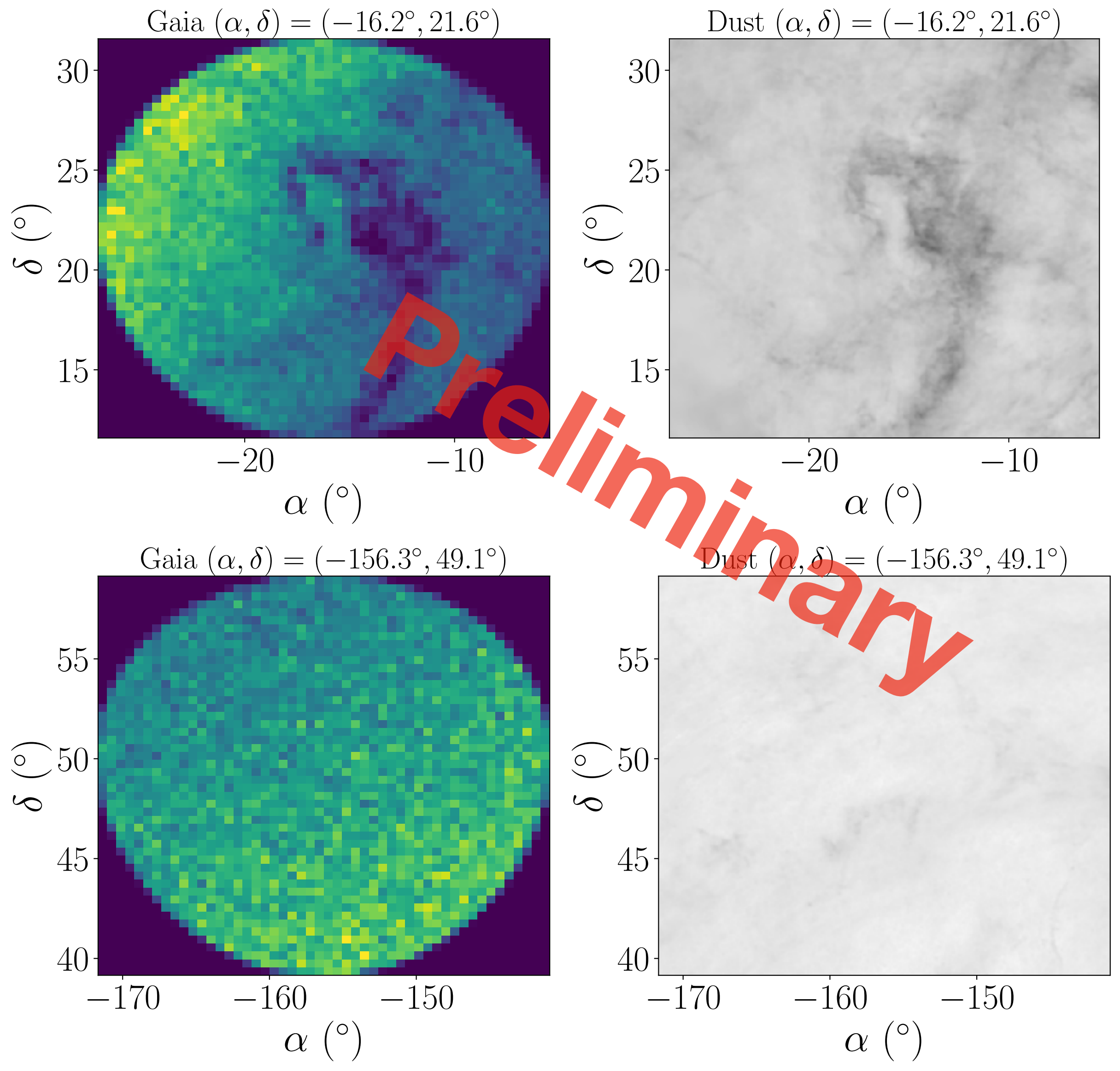


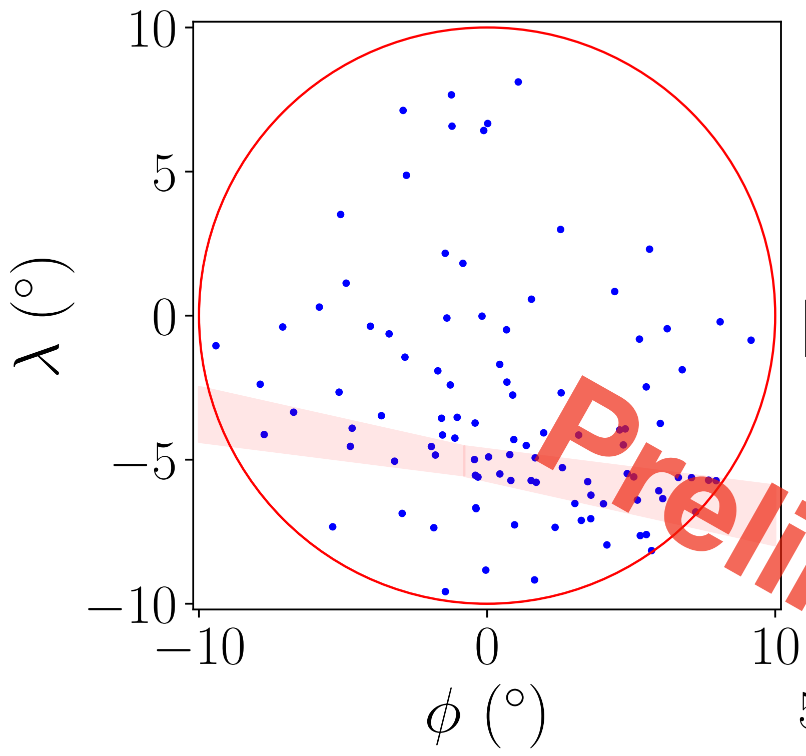
Shih *et al* (2104.12789)

Via Machinae: Unsupervised Stream Finding

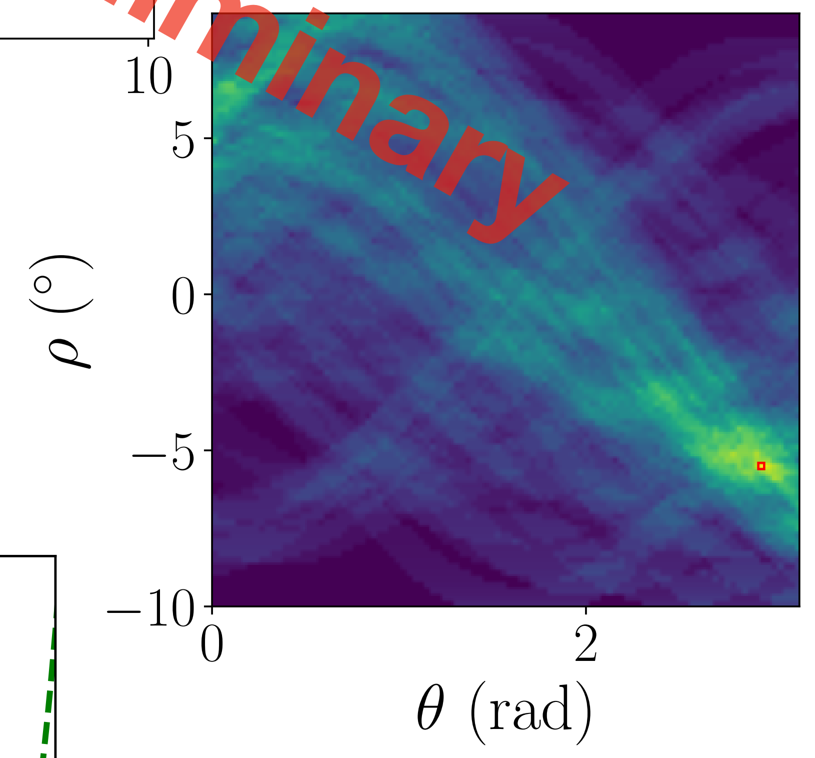
- There are a *lot* of stars in Gaia. Lots of reasons for them to be anomalous.
  - Dust lanes, globular clusters, disk stars…
- The ML anomaly score is only one part, need to automatically identify line-like features in overlapping regions of positions and proper motion.
  - Many hyperparameters needed 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

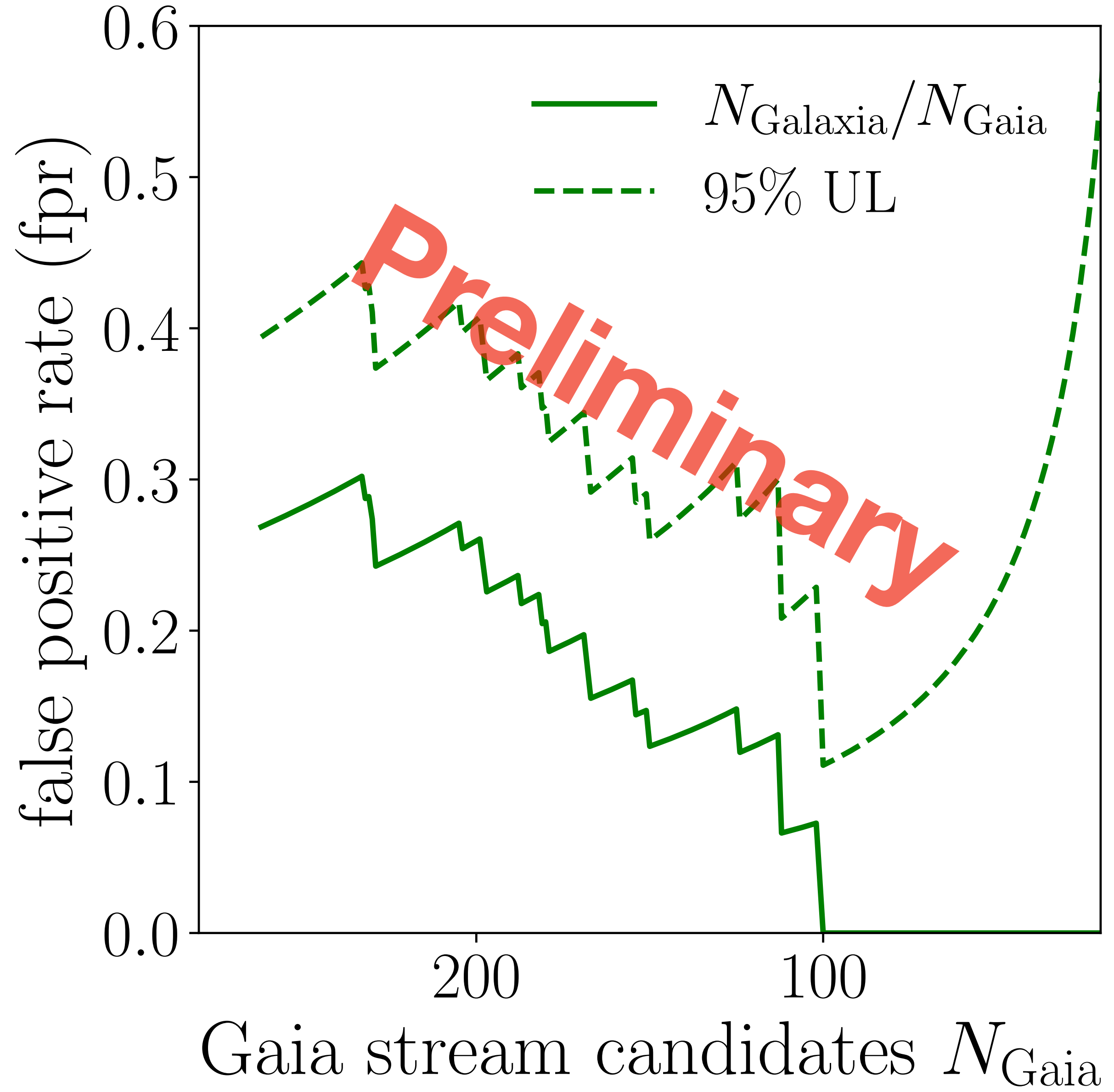
Shih, Buckley, Necib, Tamanas (in prep)





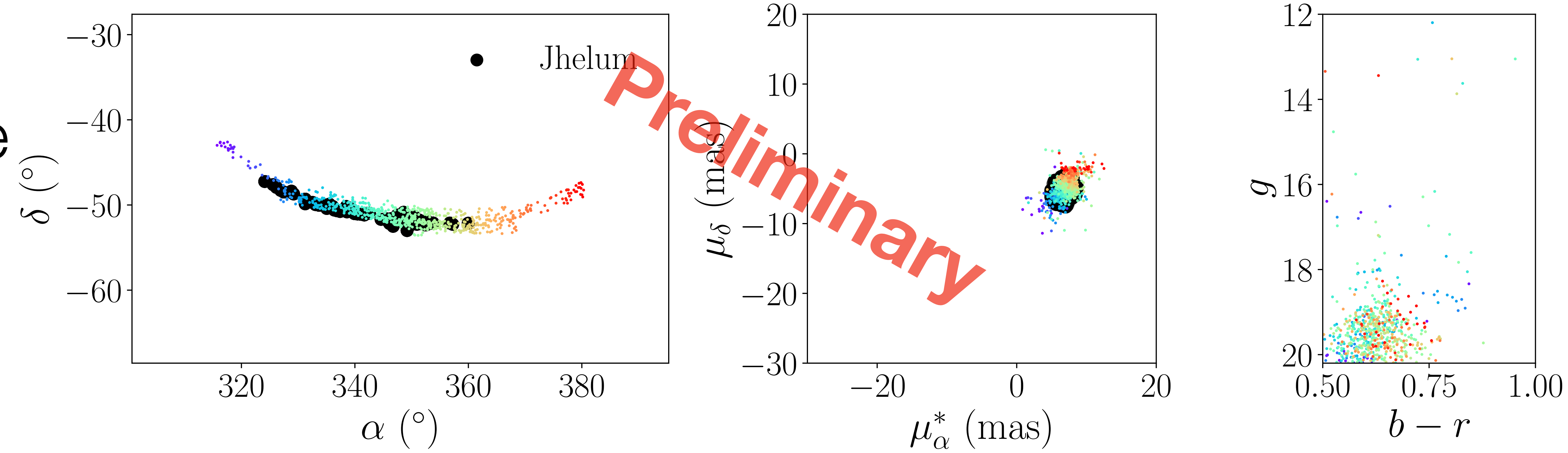
Hough transform for line-finding

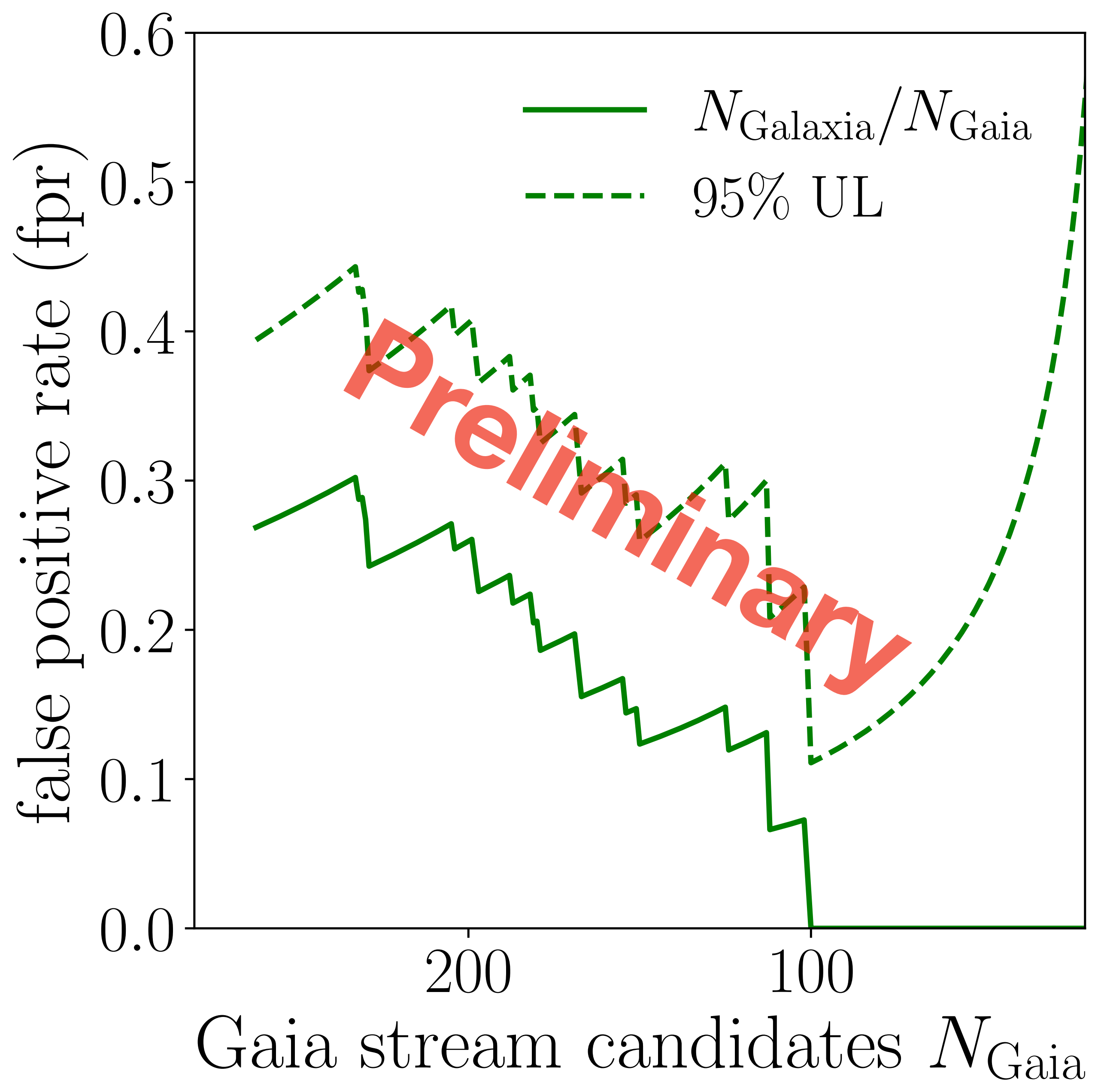


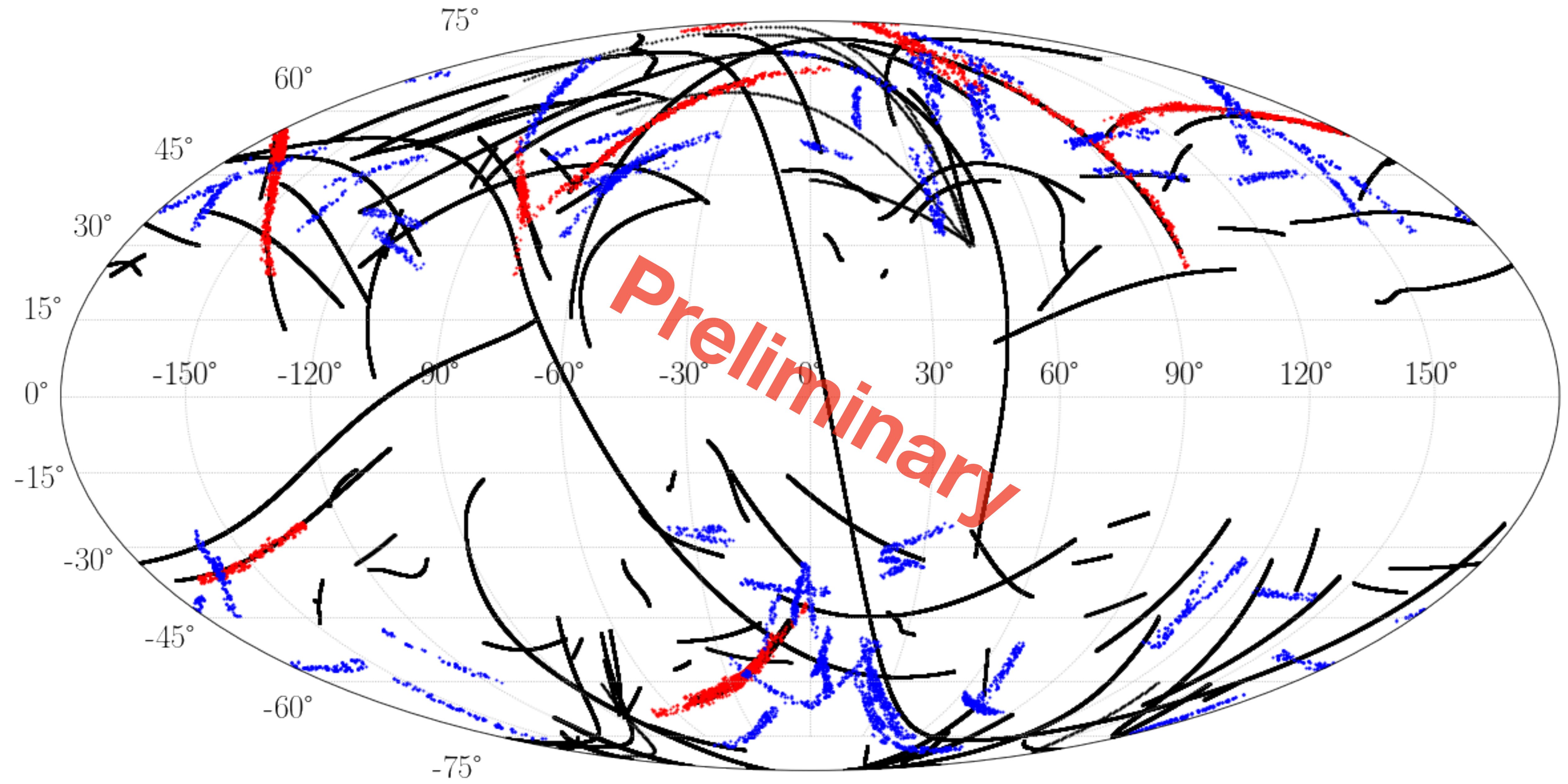


Via Machinae: Unsupervised Stream Finding

- Full-sky stream search in prep.
- We have 82 stream candidates which are more significant the most significant false positive in simulation.
  - ~20% false positive rate estimated



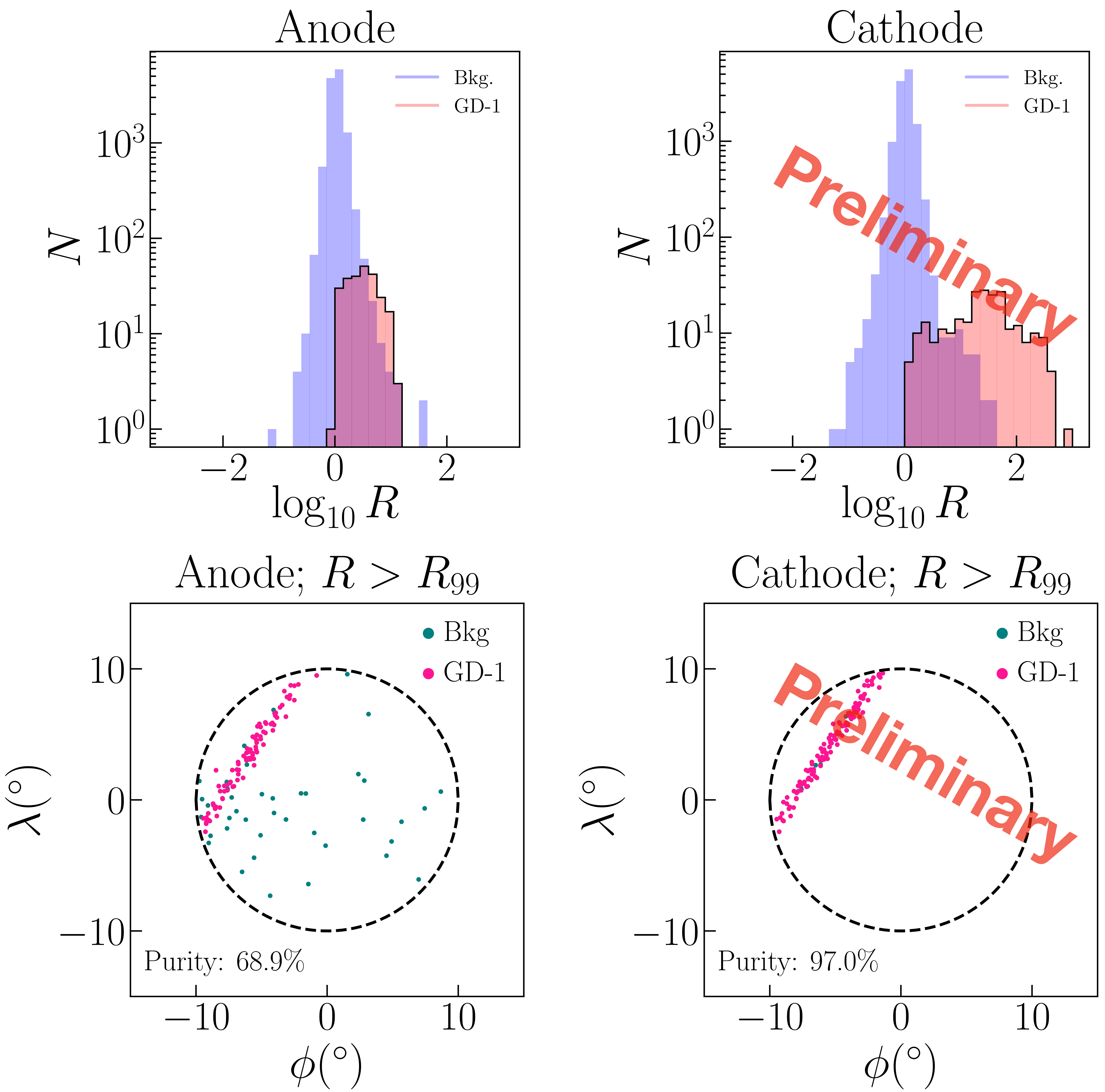




Shih, Buckley, Necib, Tamanas (in prep)

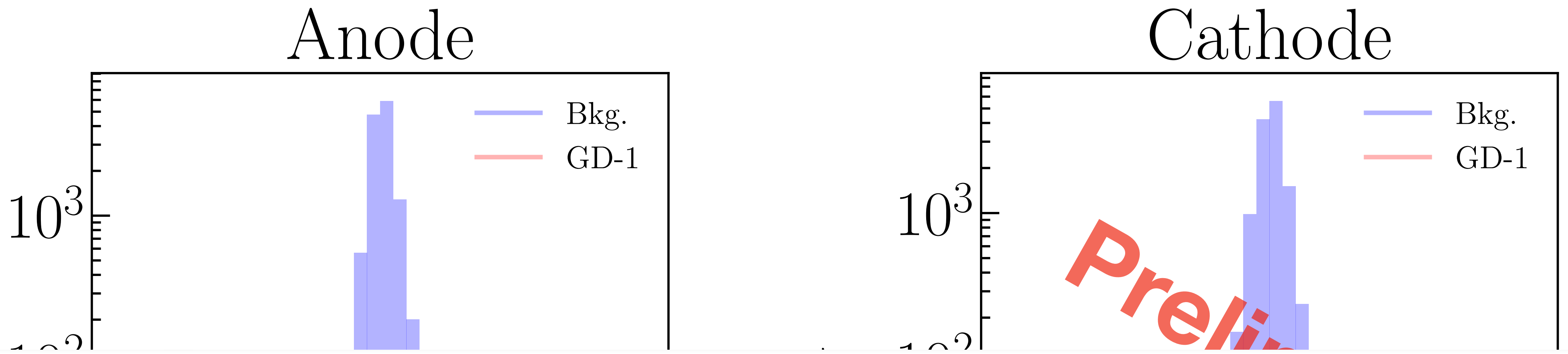
Via Machinae: Unsupervised Stream Finding

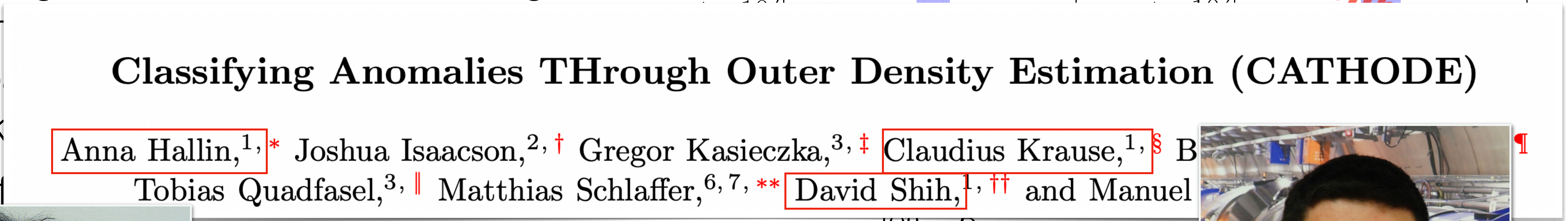
- The input for the stream-finding is the ML-derived anomaly score *R*
    - Existing version from ANODE, using normalizing flows to learn conditional probabilities in proper motion SR and backgrounds from control regions.
- What if we could do this better?
    - CATHODE (Hallin *et al* 2109.00546)
    - 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



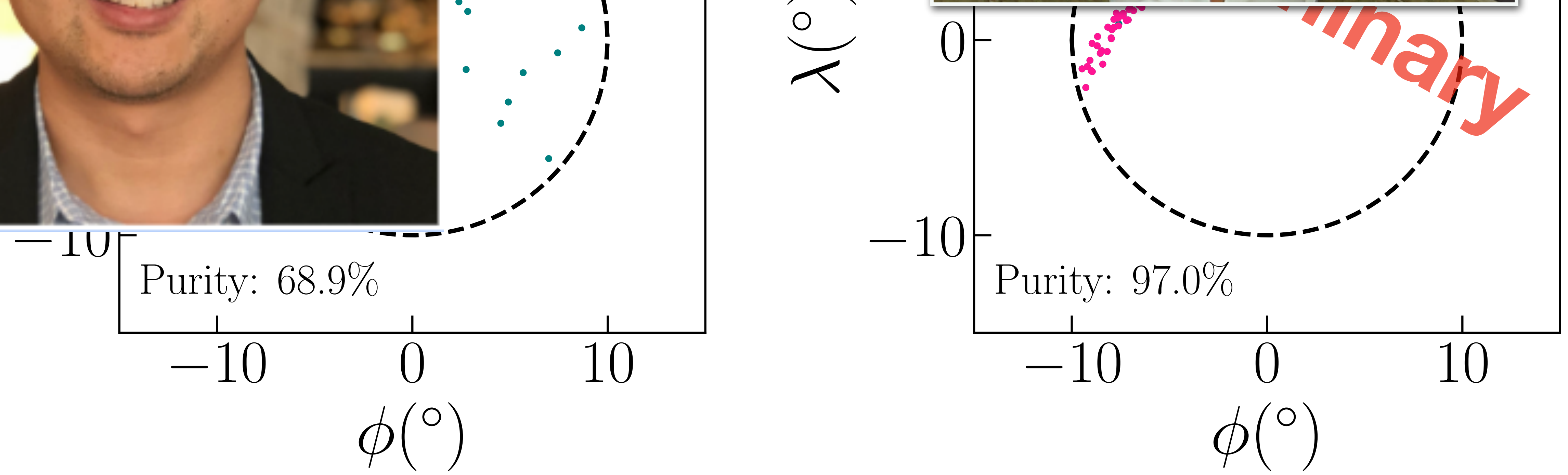
Via Machinae: Unsupervised Stream Finding

- The input for the stream-finding is the ML-derived anomaly score *R*
  - Existing version from ANODE, using norm… prob… back…
- What i… …in *et al* 2109.00546)
  - …r to distinguish events …gnal region from density …d on control-region.
  - …ut for rest of Via Machinae




Classifying Anomalies THrough Outer Density Estimation (CATHODE)

Anna Hallin,[1,*] Joshua Isaacson,[2,†] Gregor Kasieczka,[3,‡] Claudius Krause,[1,§] B… [¶]
Tobias Quadfasel,[3,‖] Matthias Schlaffer,[6,7,**] David Shih,[1,††] and Manuel …



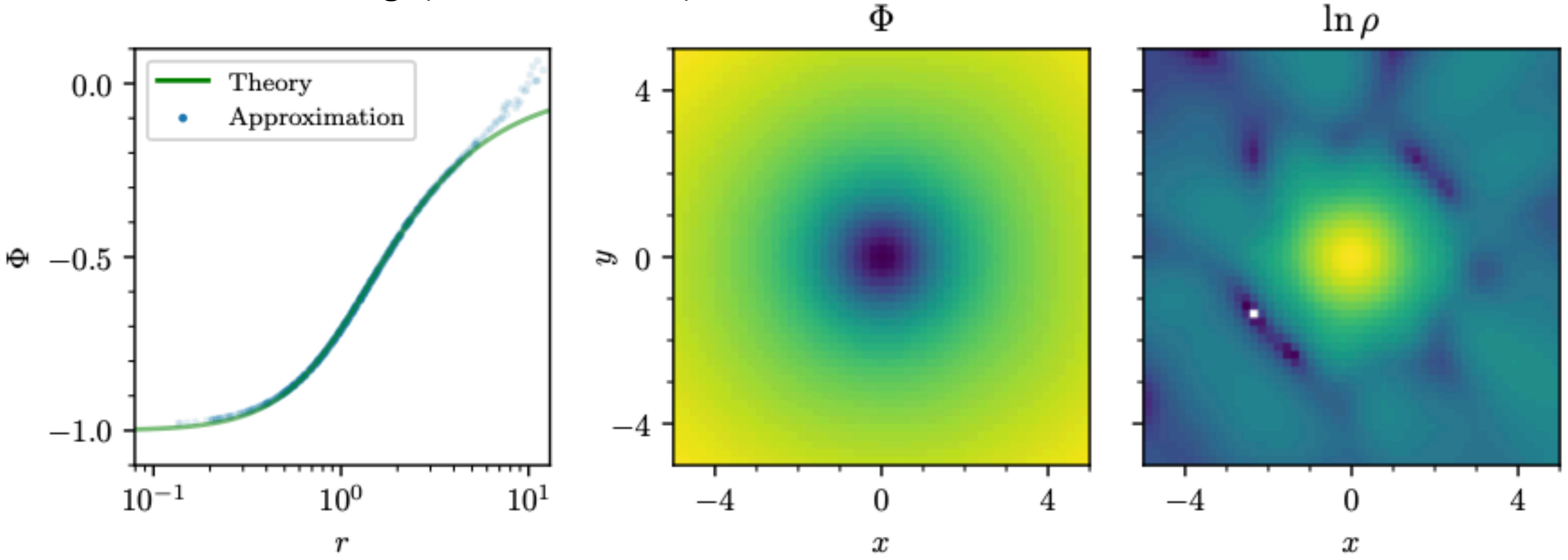
Dark Matter Density from Gaia

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

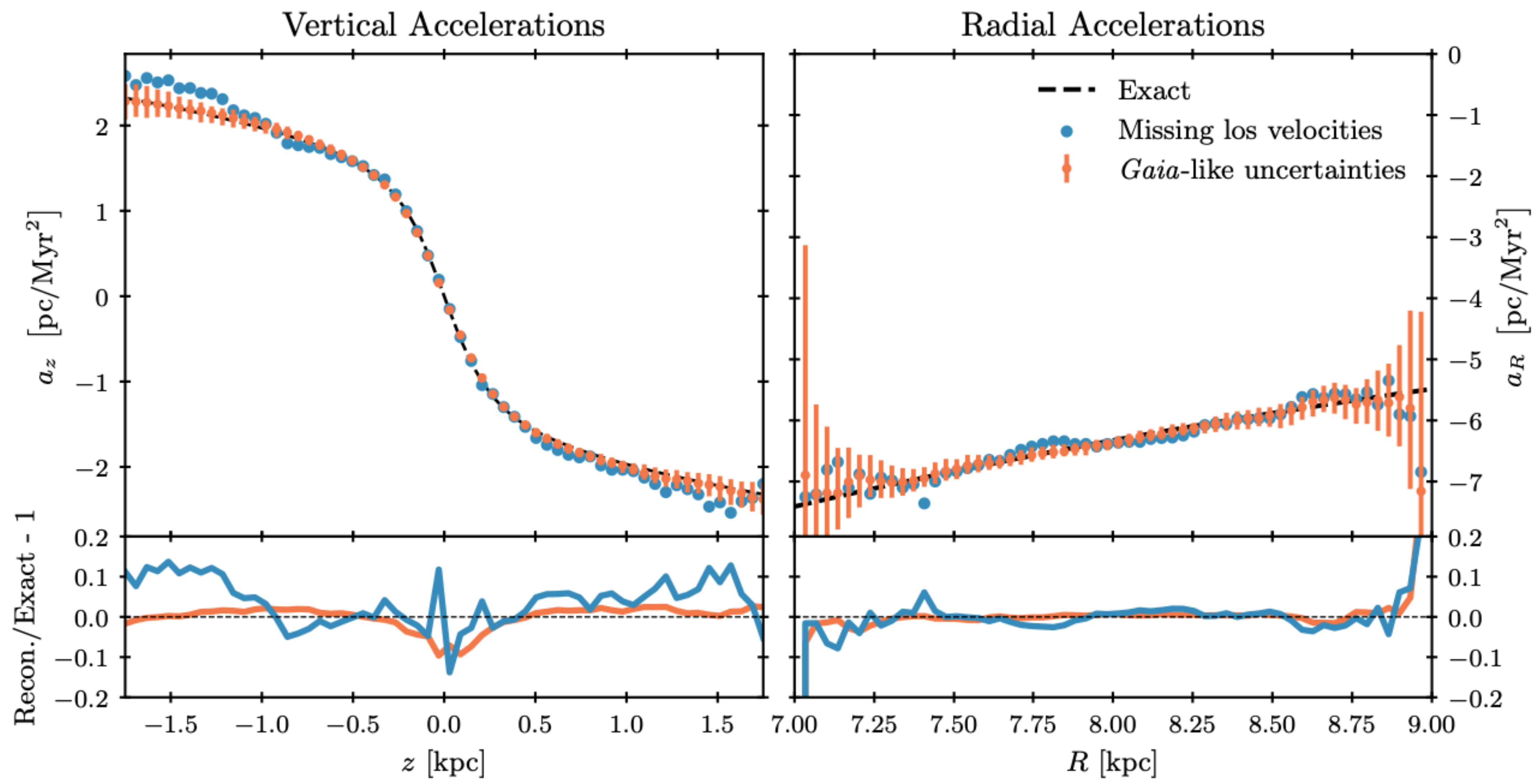
$$\frac{df}{dt} + v_i \frac{\partial f}{\partial x_i} = \frac{\partial \Phi}{\partial x_i}\frac{\partial f}{\partial v_i}$$

- 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)

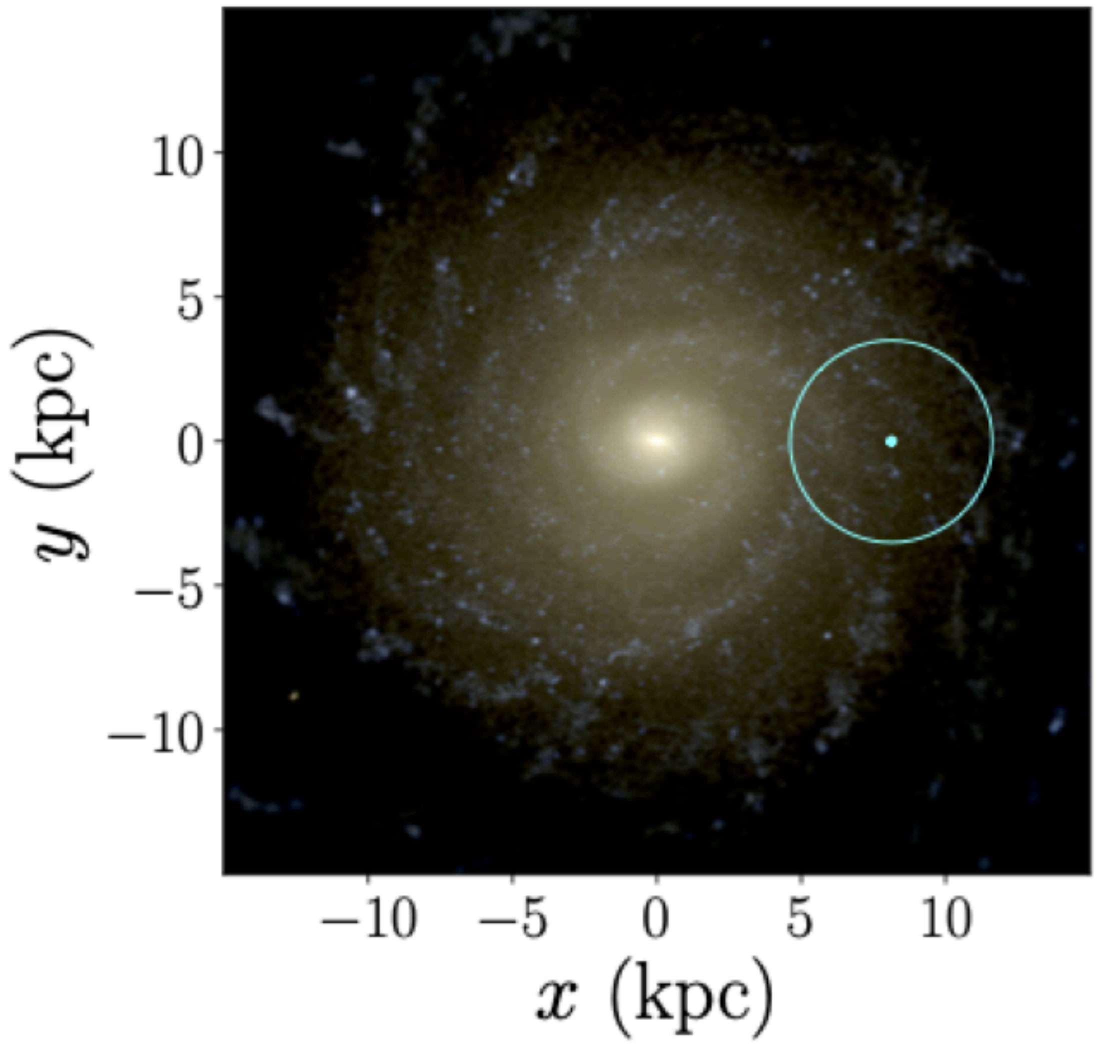


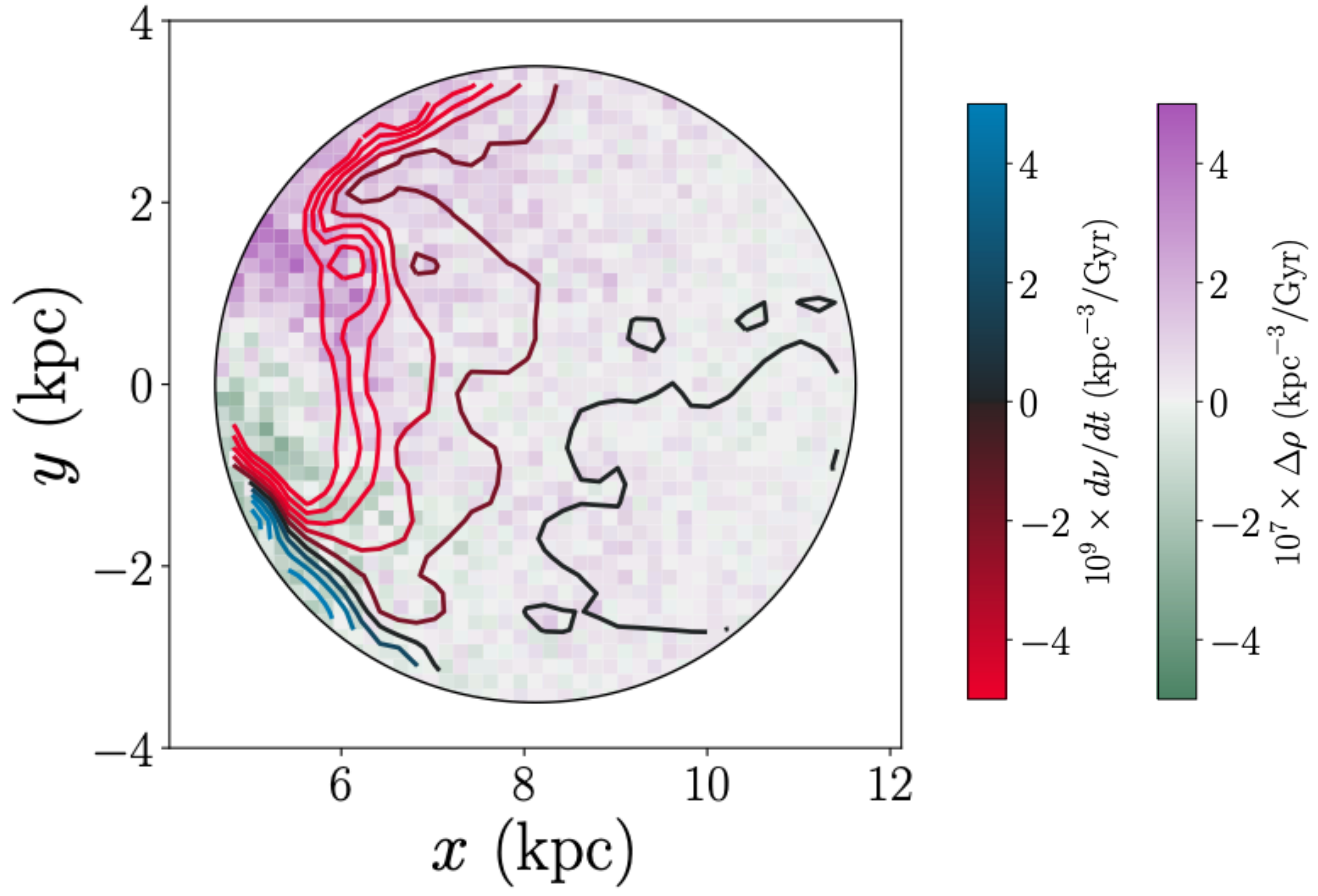
Dark Matter Density from Gaia

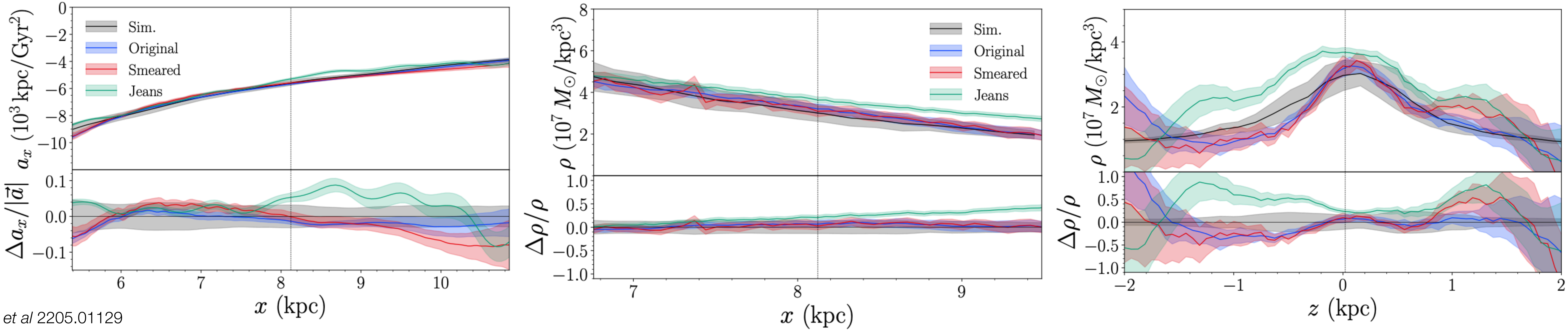
- The real Galaxy is not in equilibrium:

$$\frac{df}{dt} \neq 0$$

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







Buckley *et al* 2205.01129

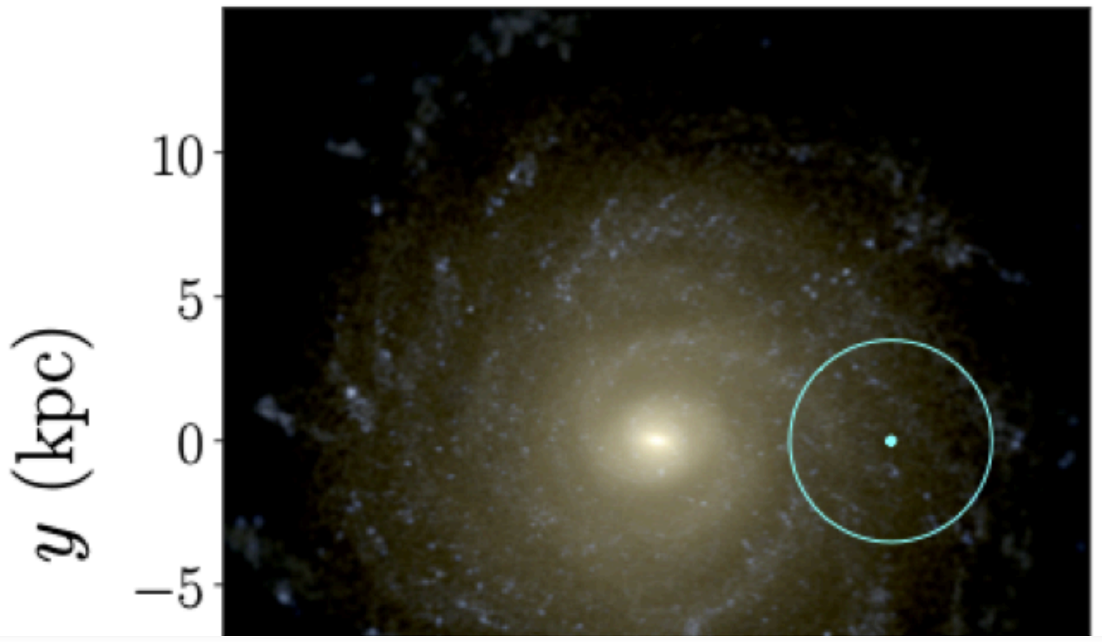
Dark Matter Density from Gaia

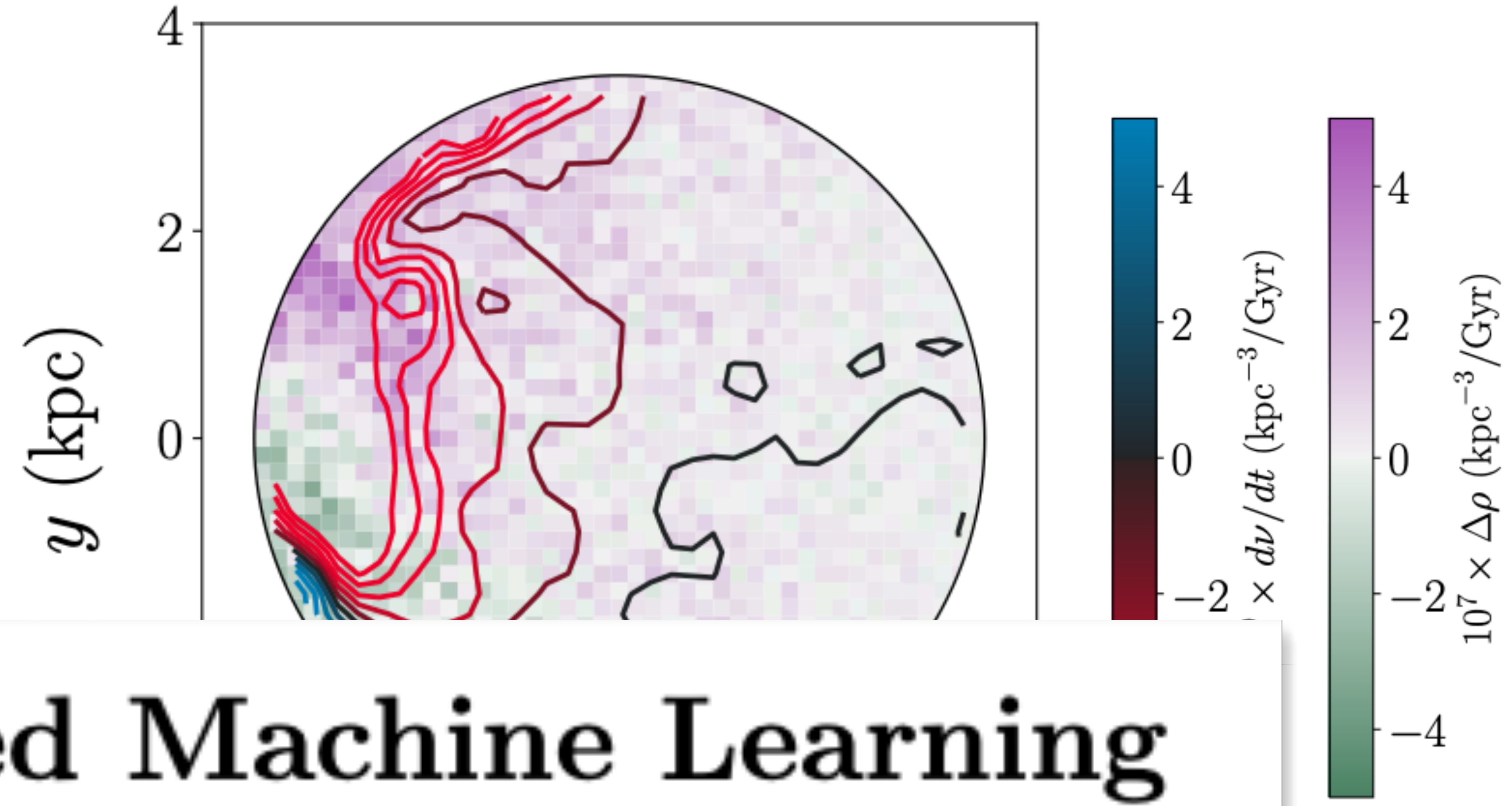
- The real Galaxy is not in equilibrium:

$$\frac{df}{dt} \neq 0$$

- Is real data sufficiently precise to get good

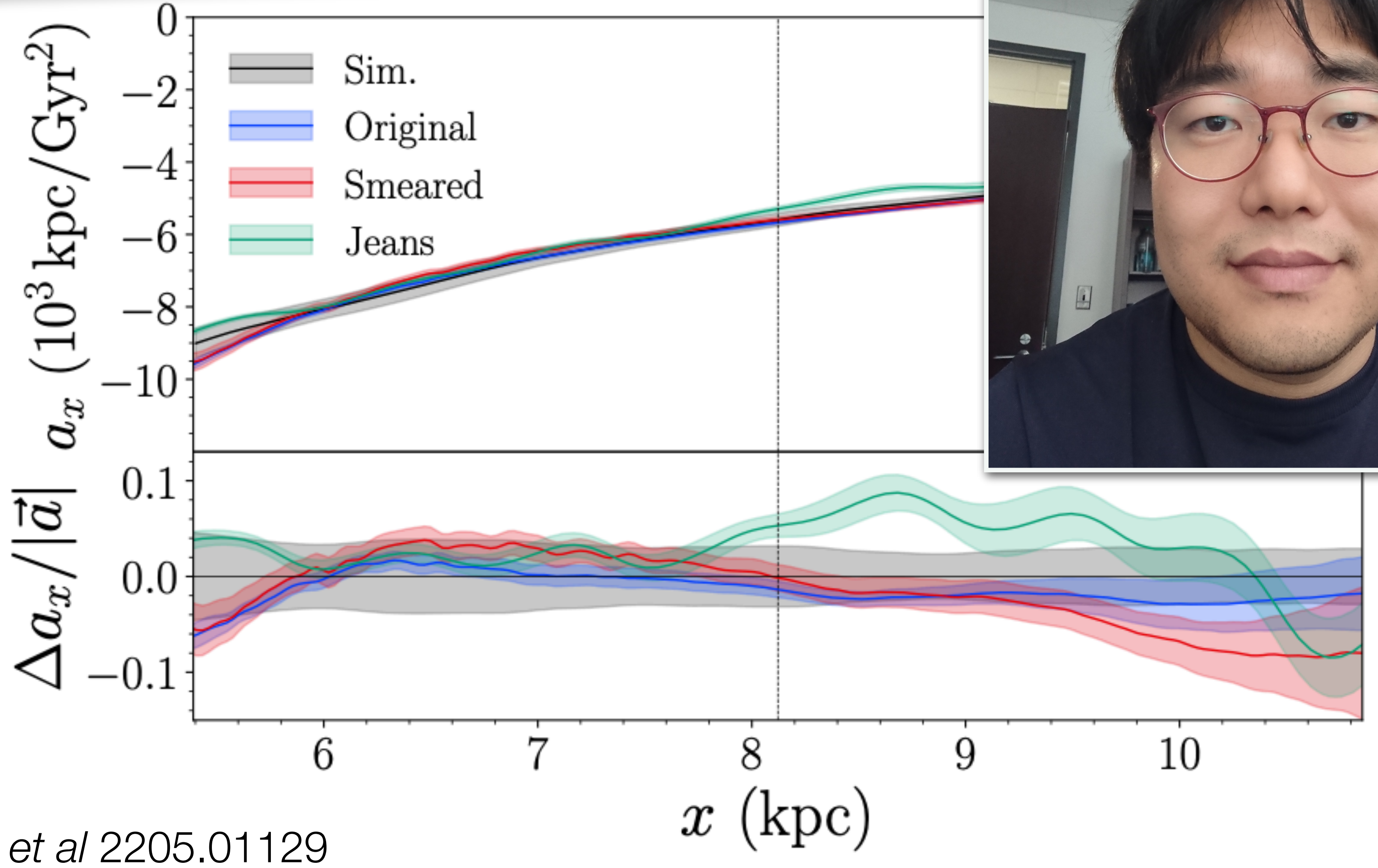
-

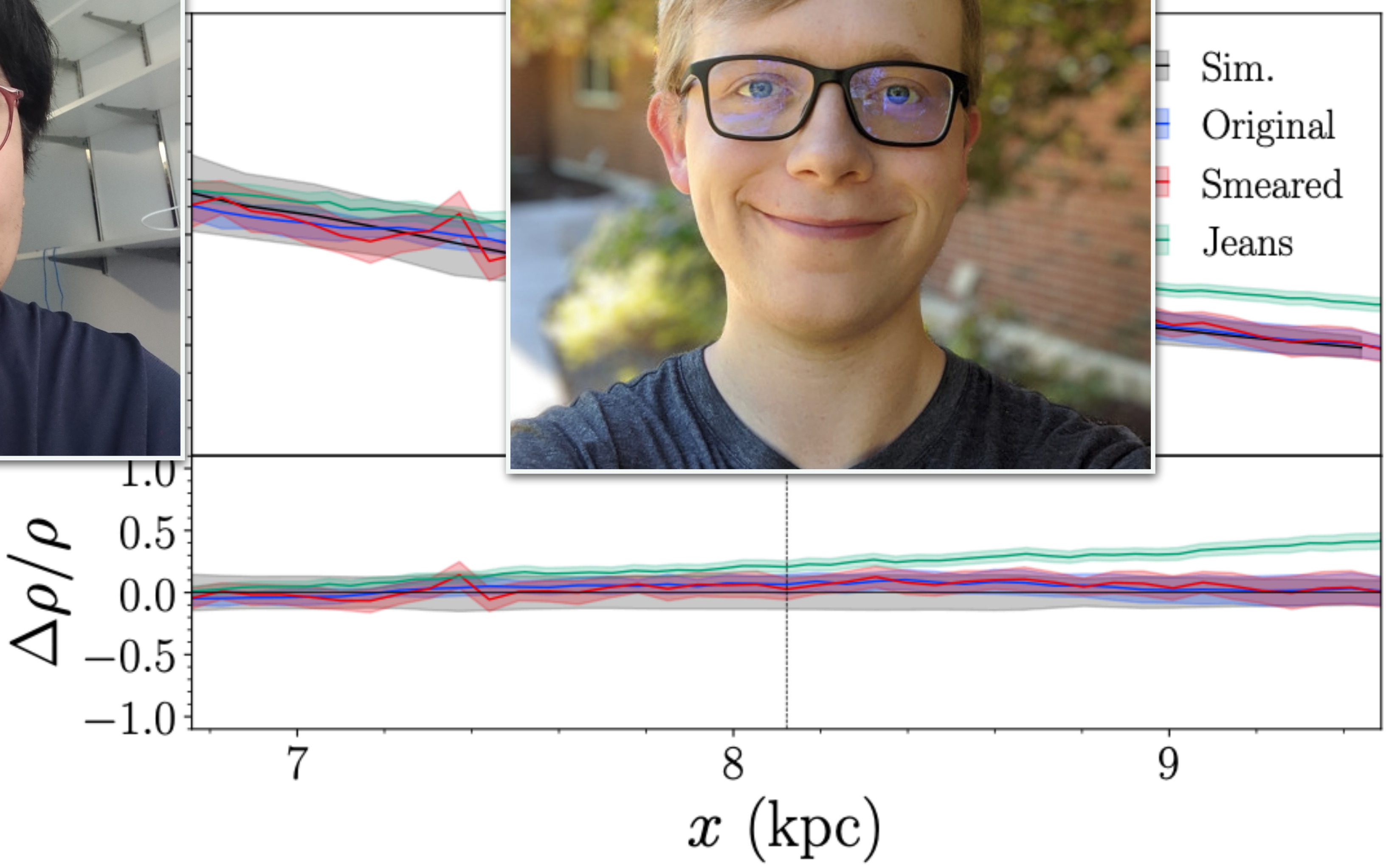


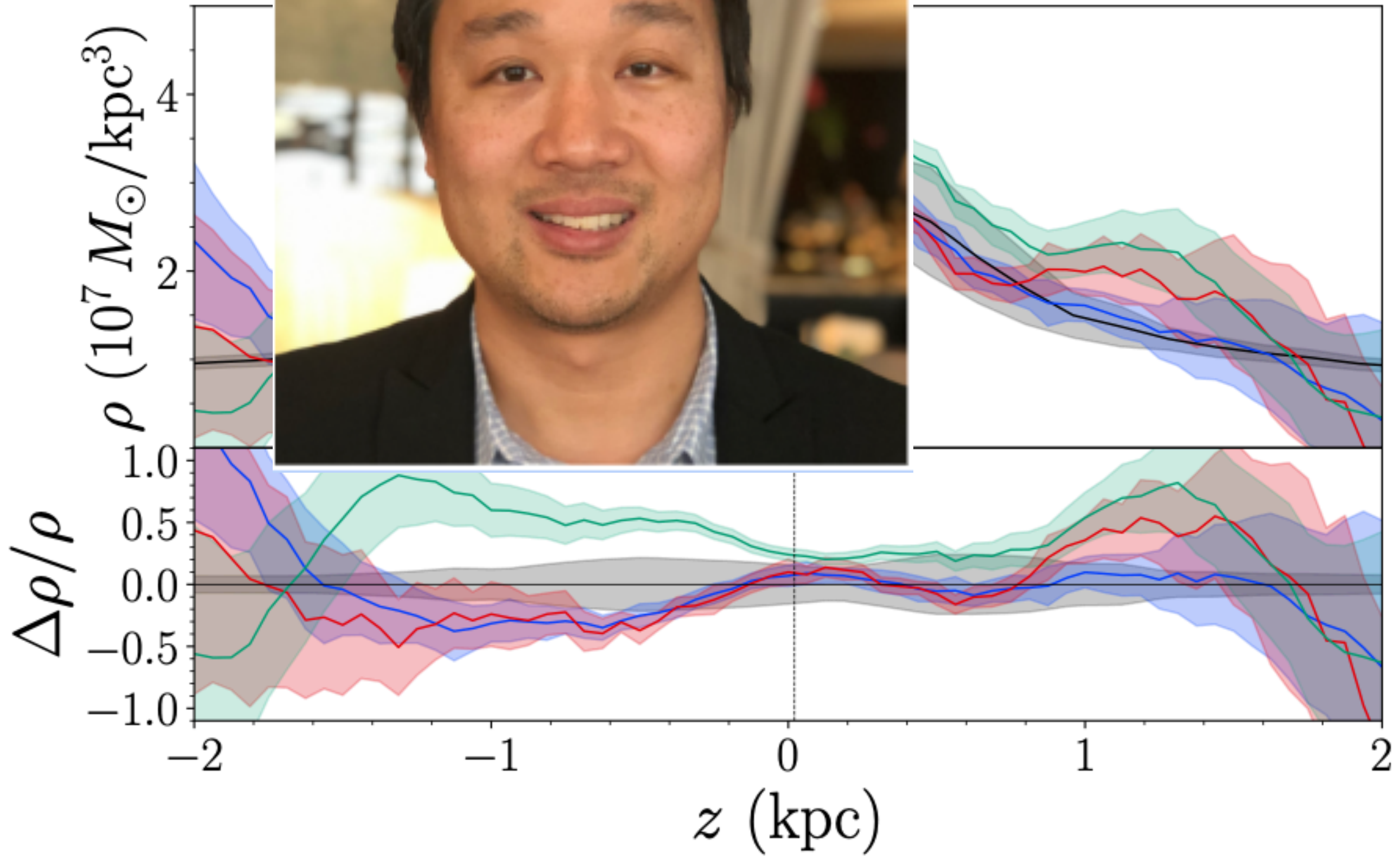


Measuring Galactic Dark Matter through Unsupervised Machine Learning

Matthew R. Buckley, Sung Hak Lim, Eric Putney, and David Shih
Department of Ph... ...nomy, Ru... ...Piscataway, NJ...



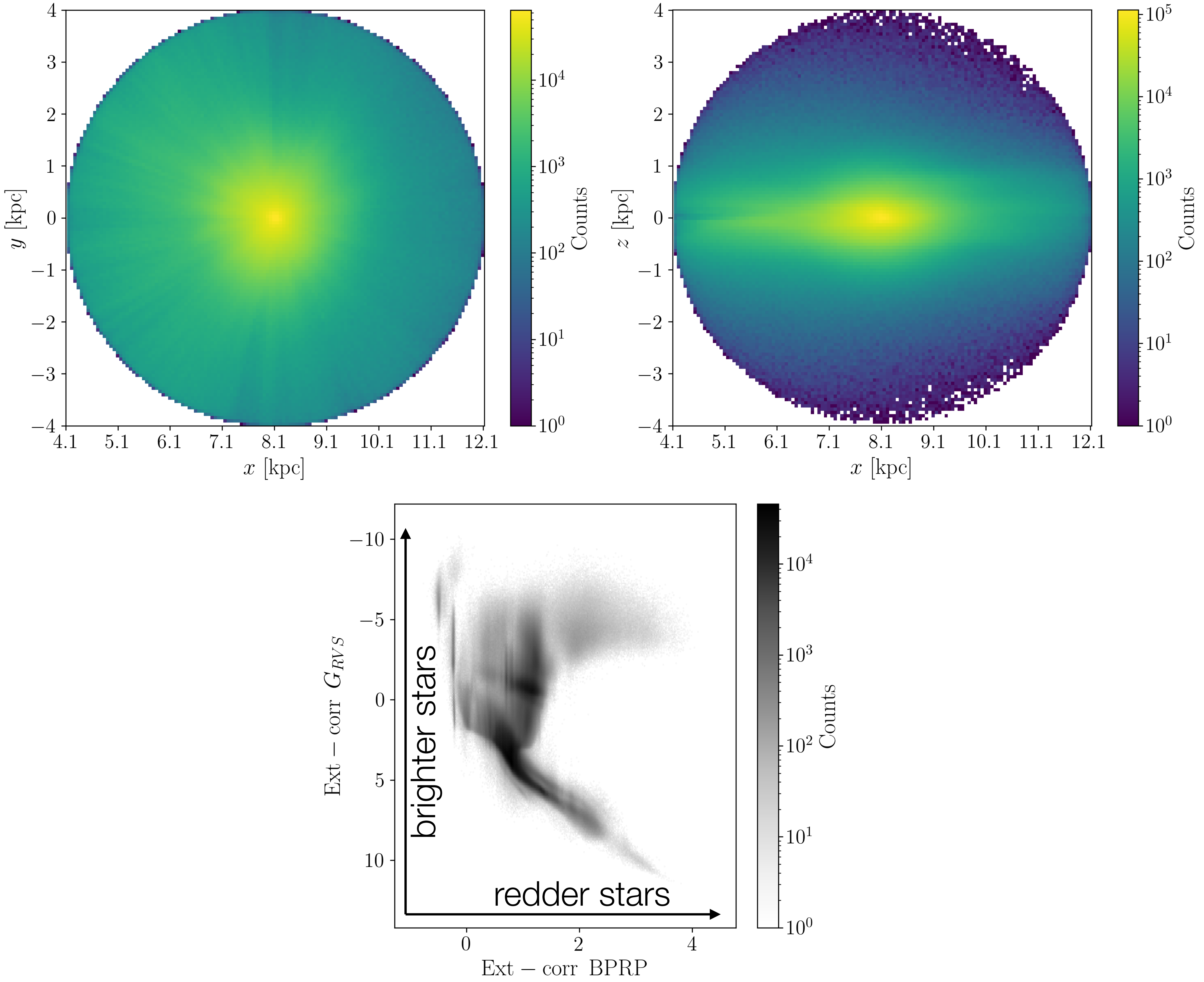




Buckley *et al* 2205.01129

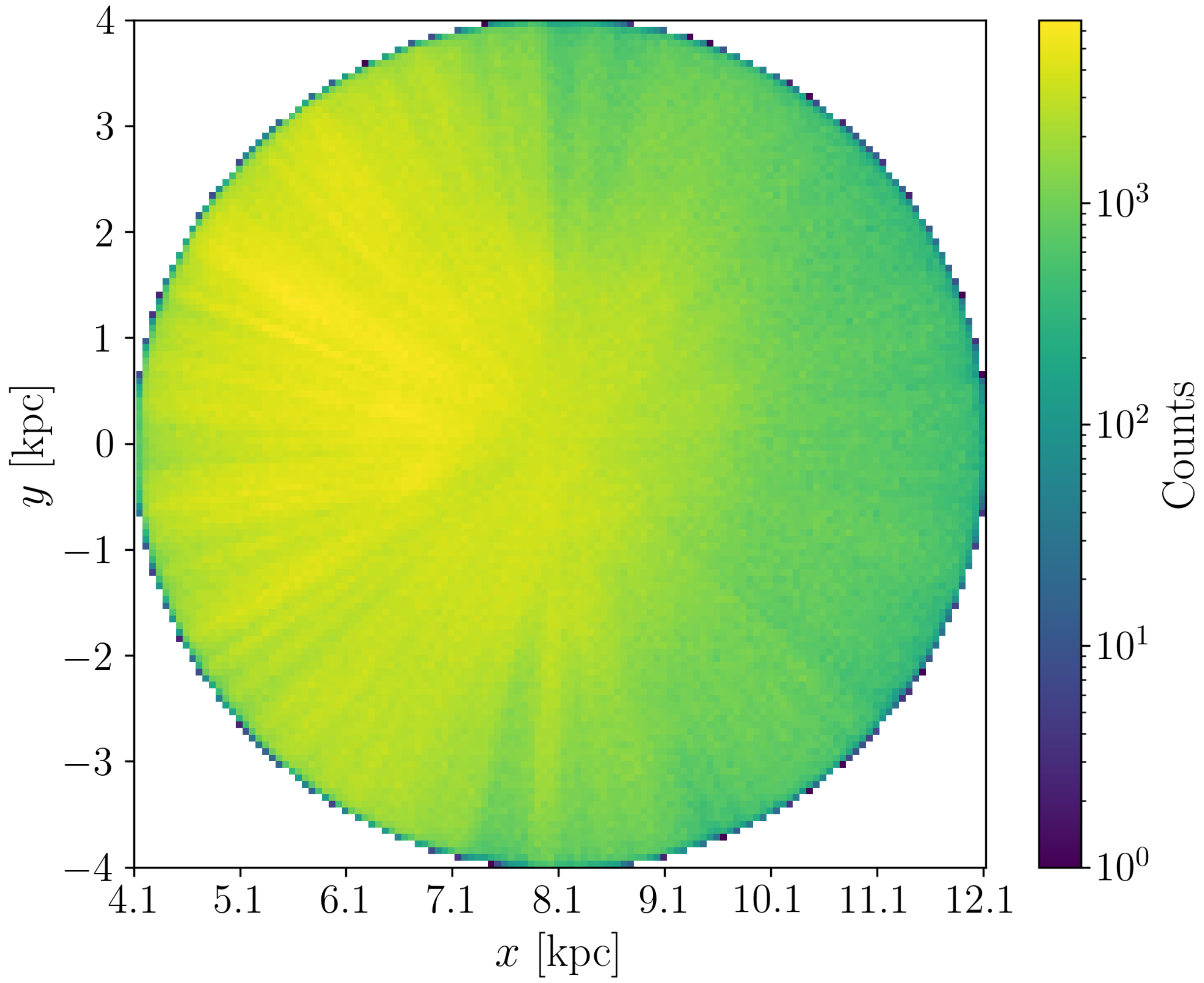
Dark Matter Density from Gaia

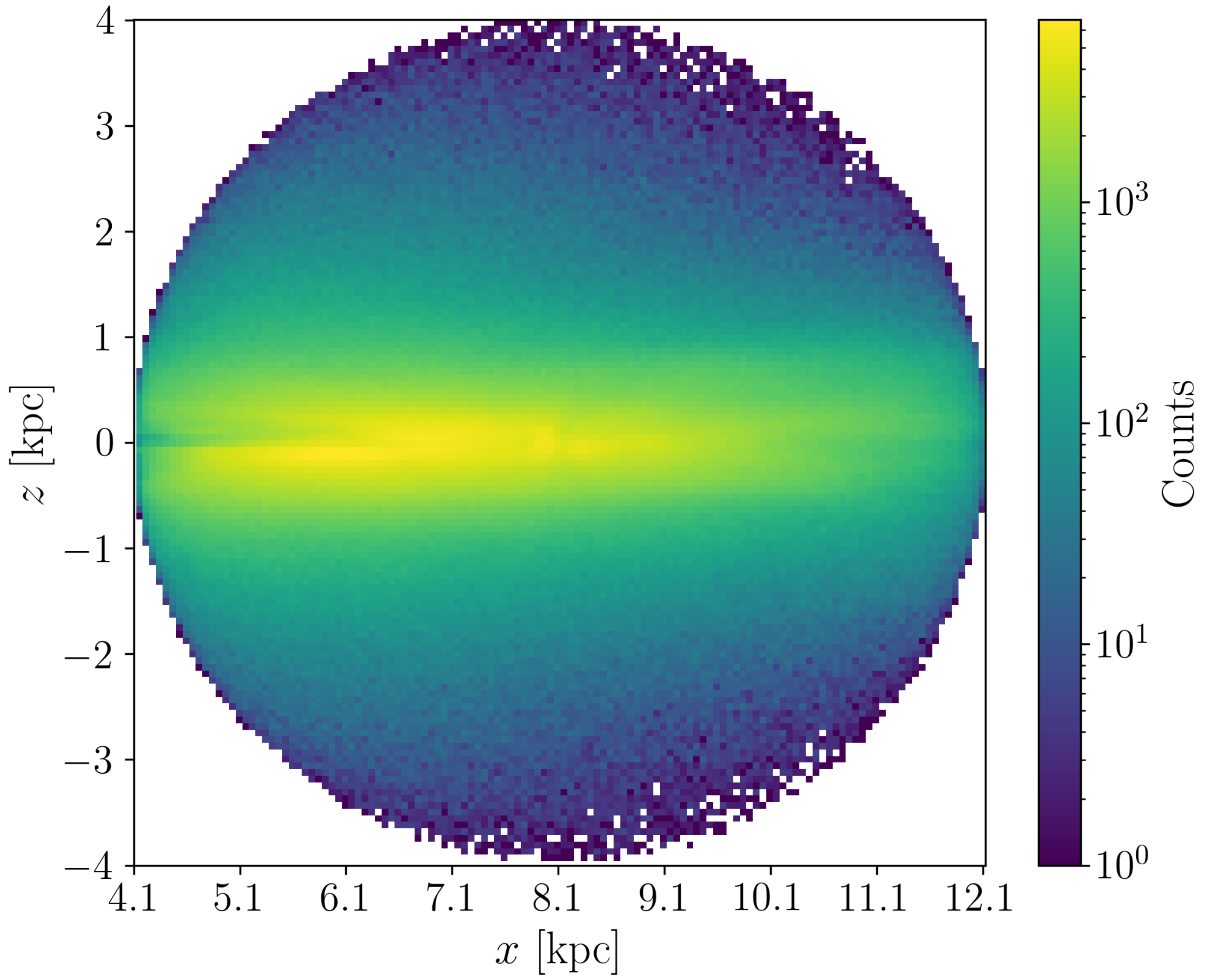
- Can we do this with real Gaia data?
- Real data is complicated:
  - Observations are not complete, and this completeness varies as a function of distance
  - And with which kinematic parameters are measured, and/or stellar properties
- The goal: get low-error measurements off of the Galactic disk,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

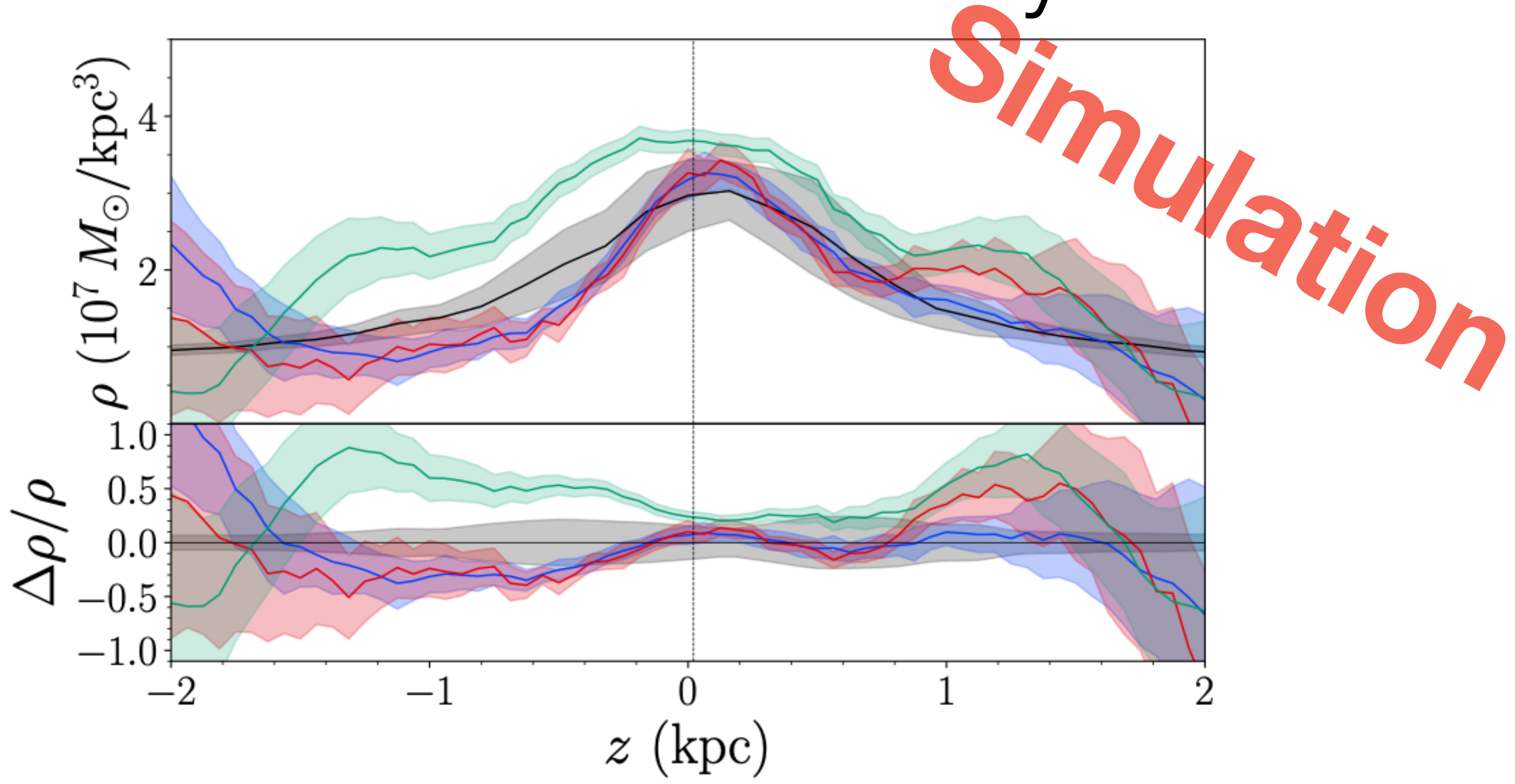


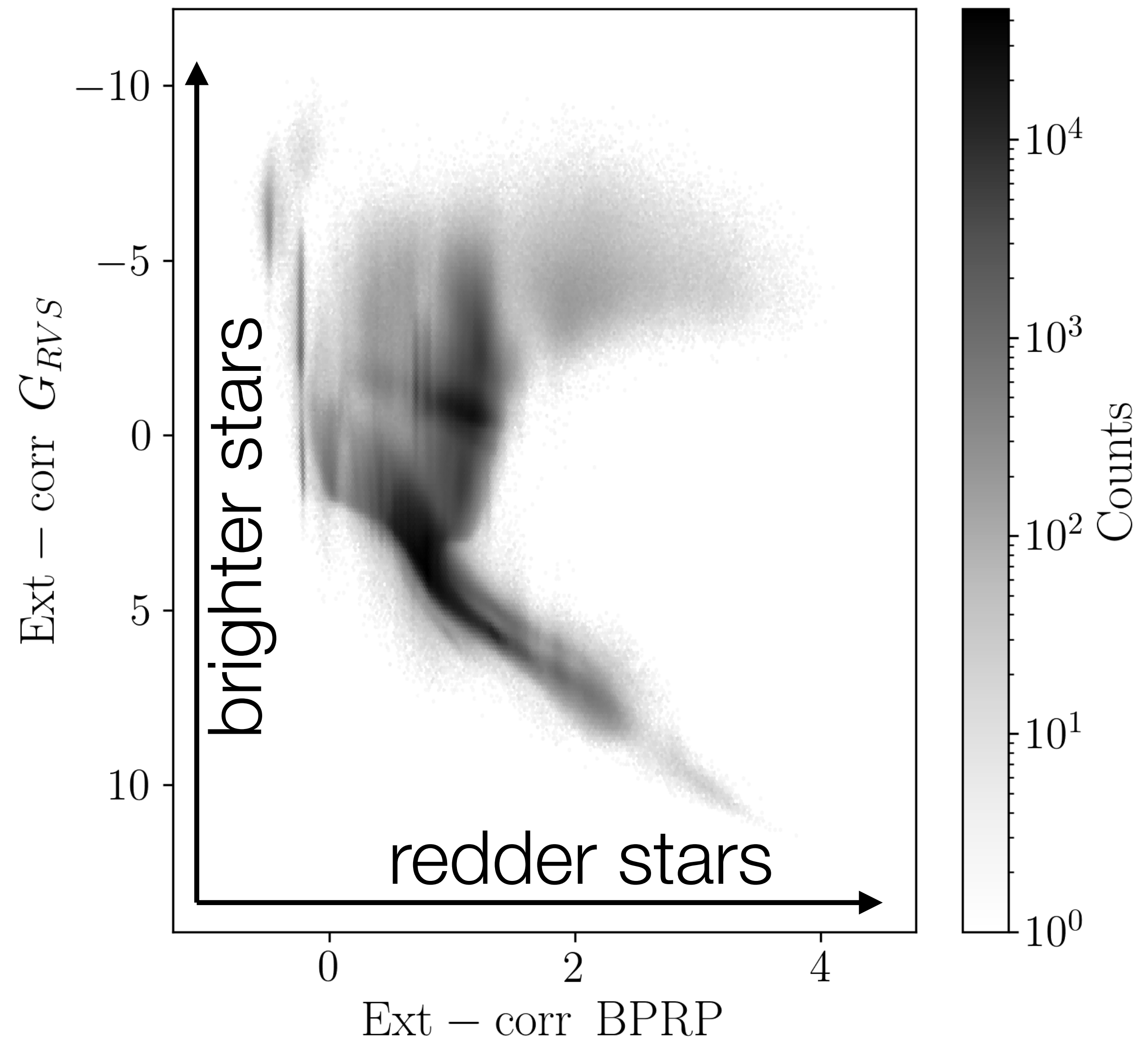
Dark Matter Density from Gaia

- 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.





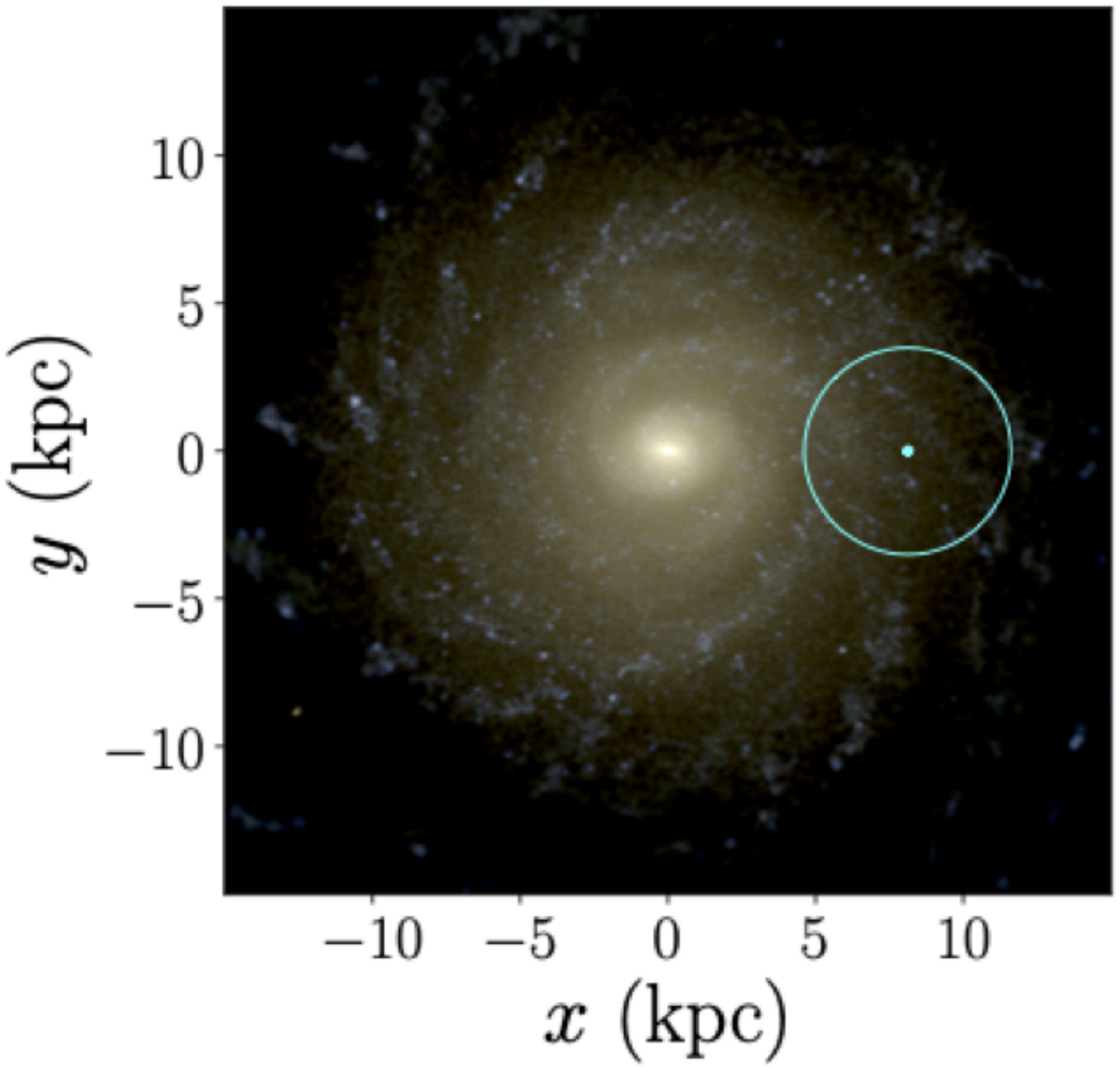




Upsampling Simulations

- 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!

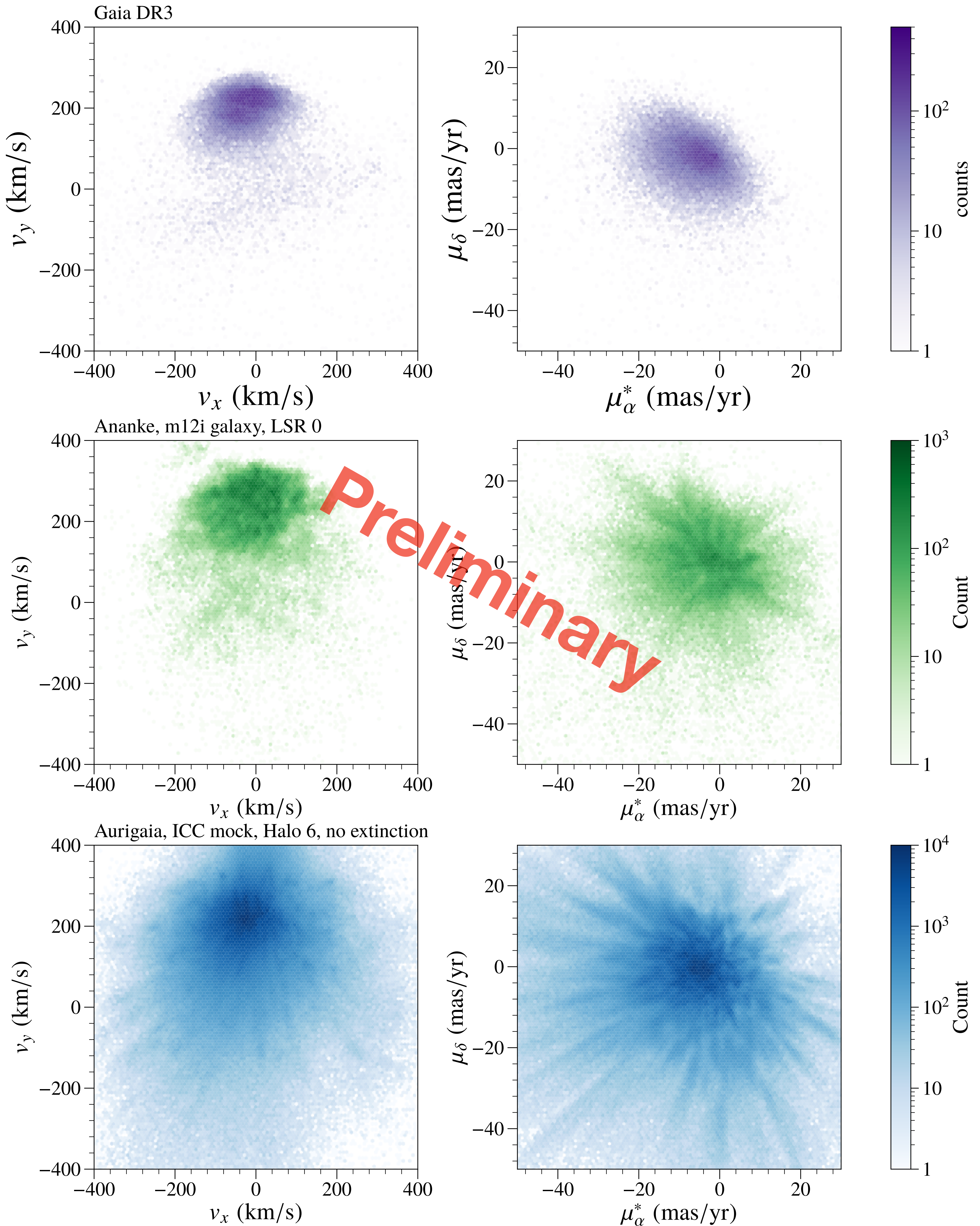
Galaxy h277 (N-Body Shop)



Lim *et al* (in prep)

Upsampling Simulations

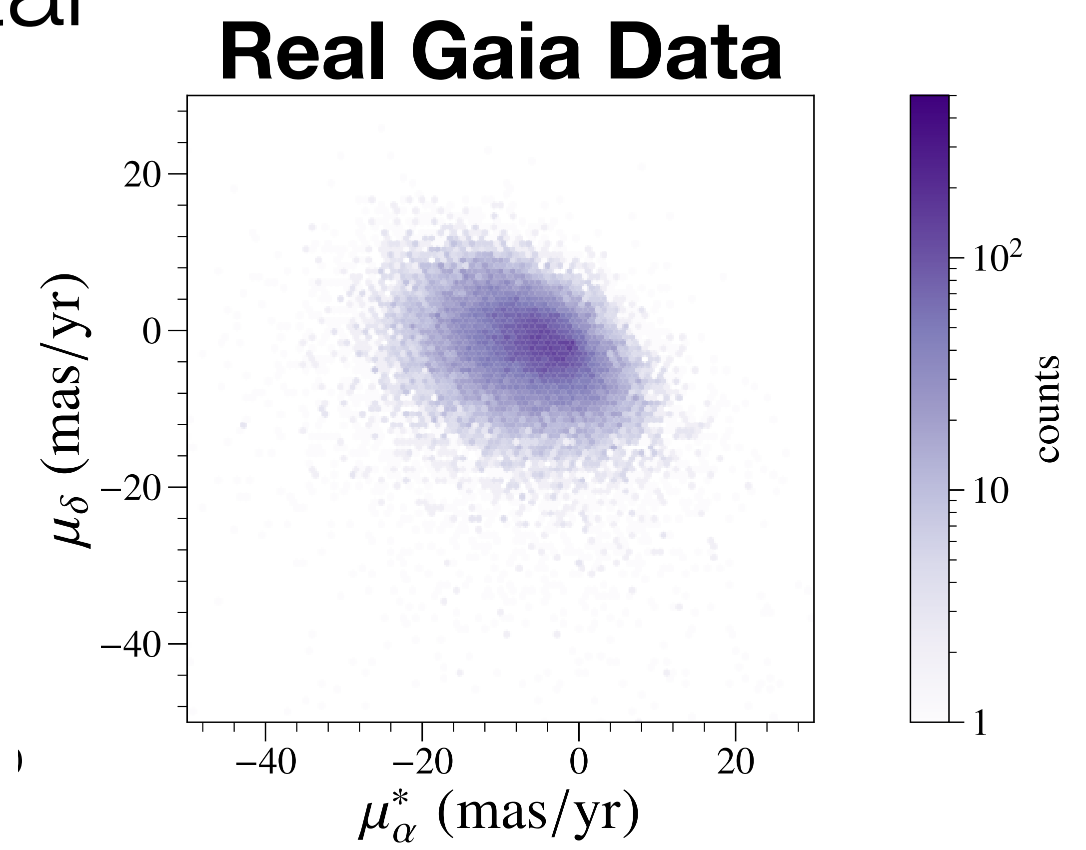
- Tools exist that can create "theorist-level" simulation for LHC machine learning.
- Much trickier for astrophysics. Can either:
  - Create by-hand analytic smooth models of the Galaxy or,
  - Use *N*-body hydrodynamical simulations
- But in the latter case, there complications:
  - Every galaxy is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

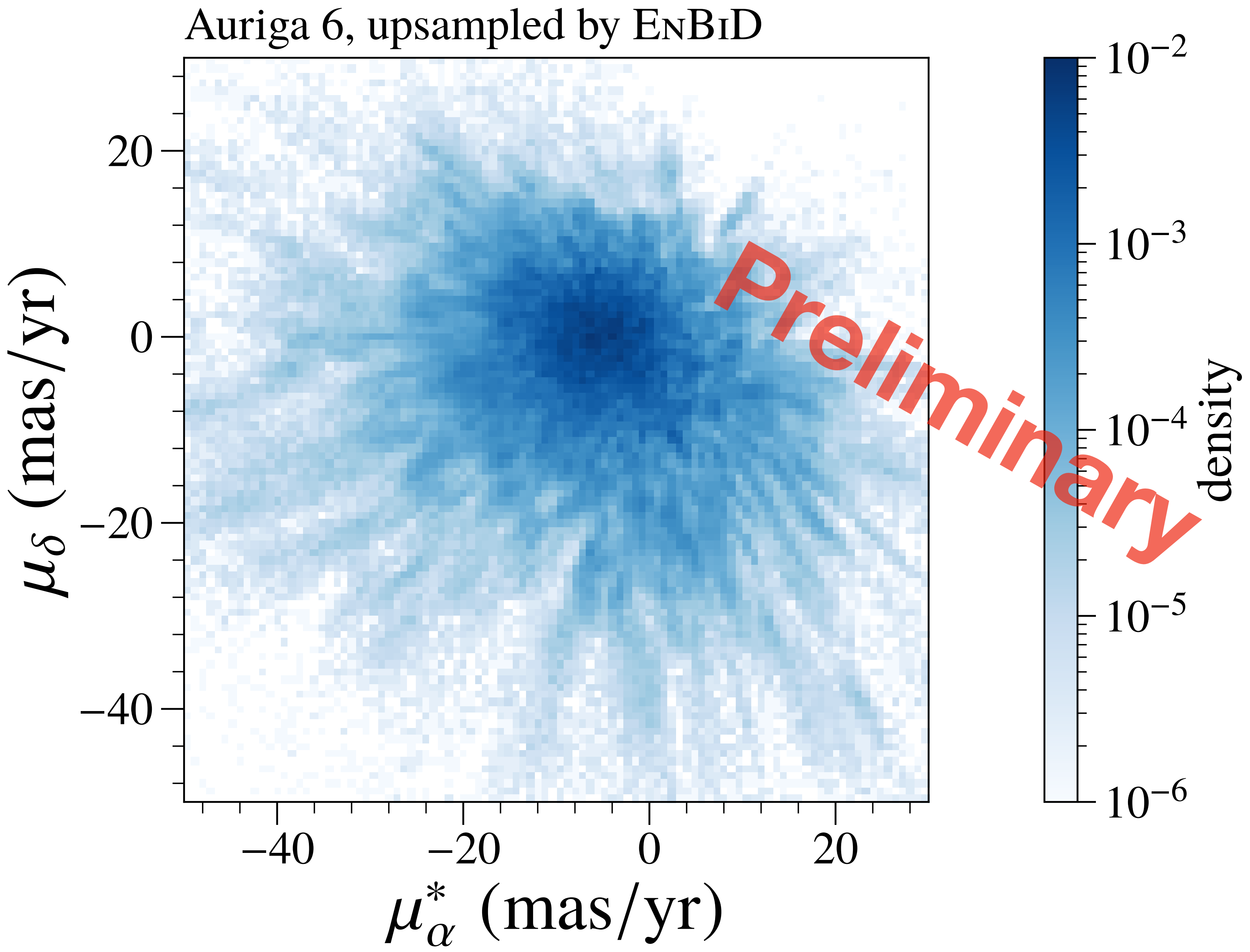


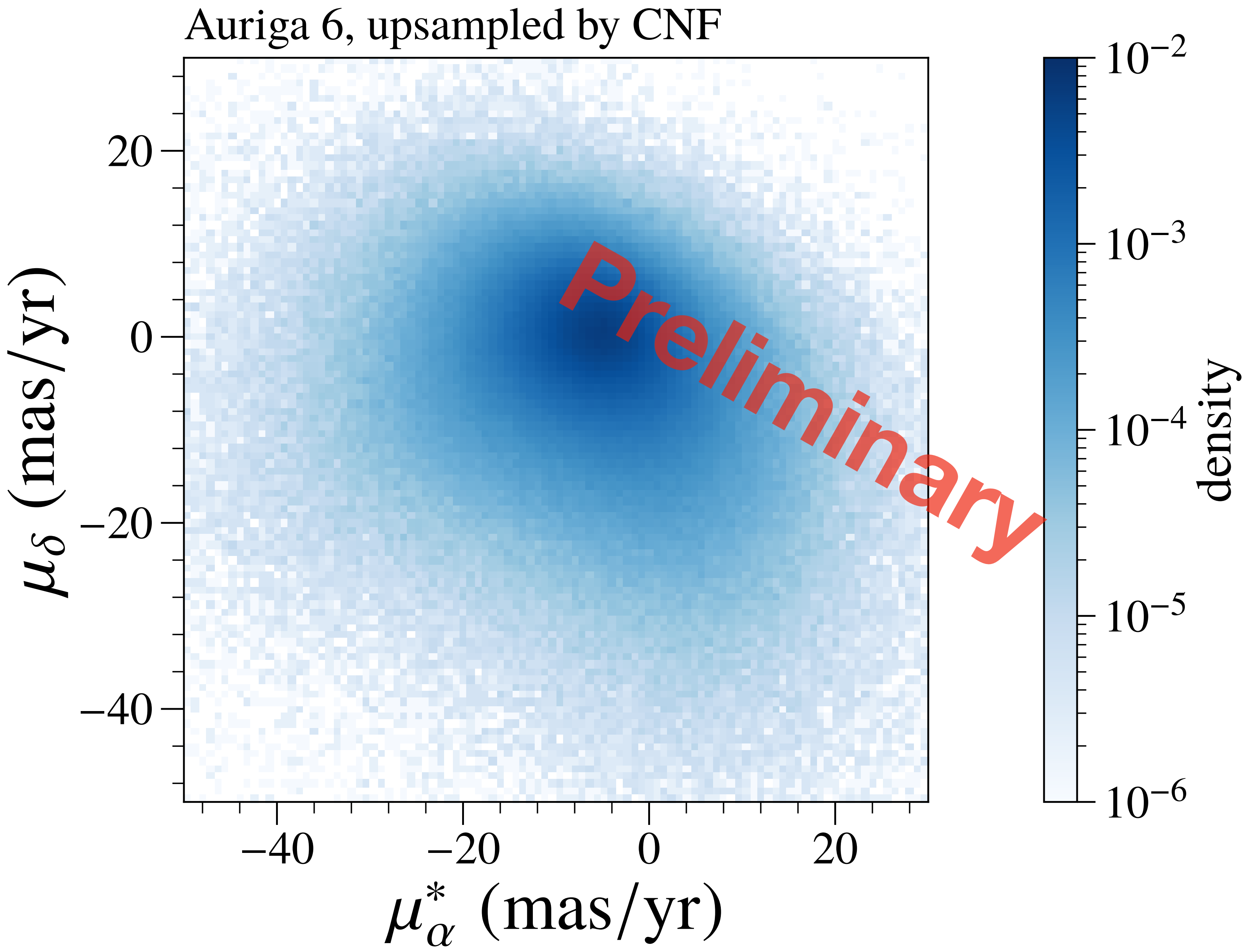
Lim *et al* (in prep)

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







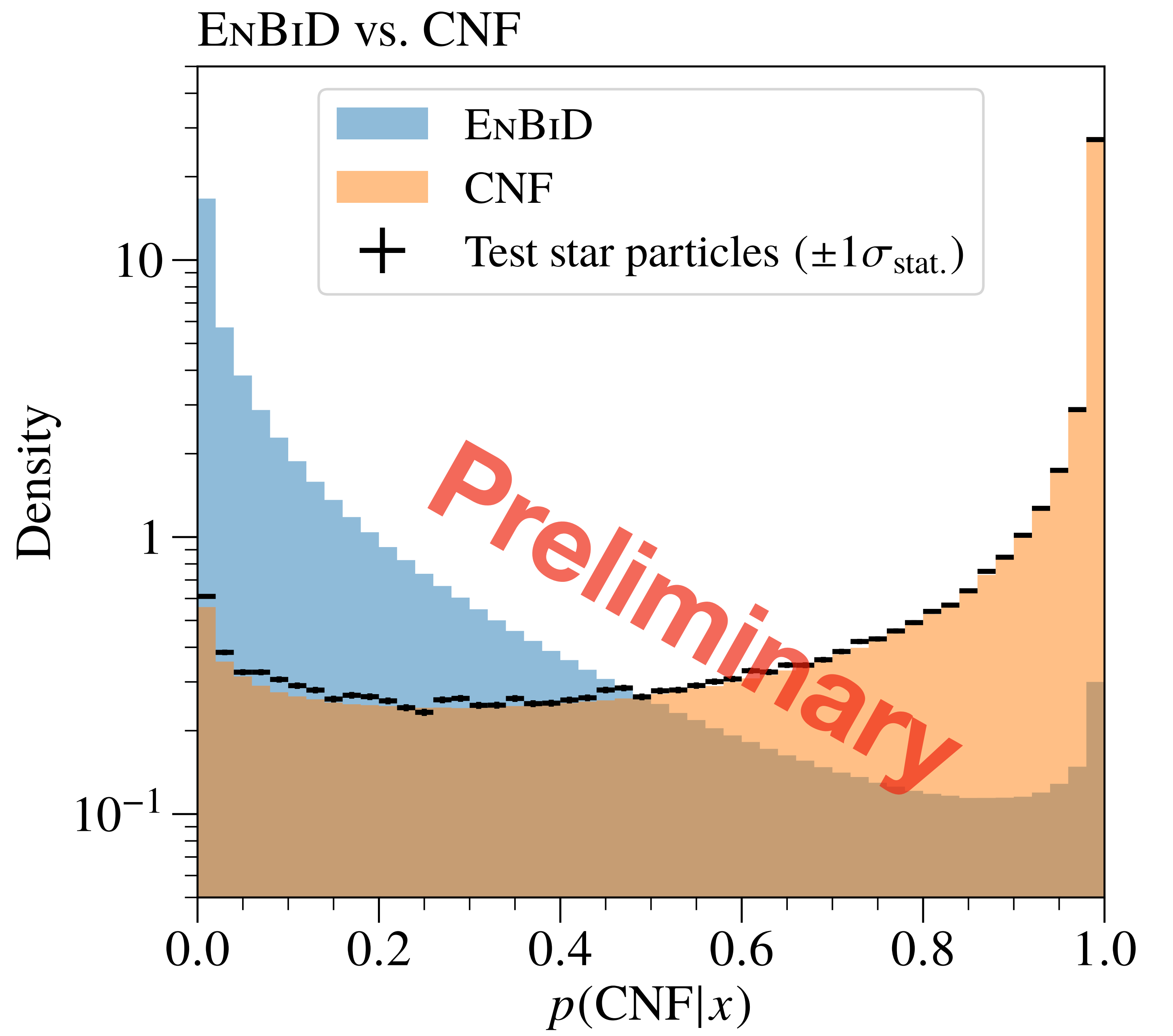
Lim *et al* (in prep)

Upsampling Simulations

- David's favorite metric (with a twist):
- 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

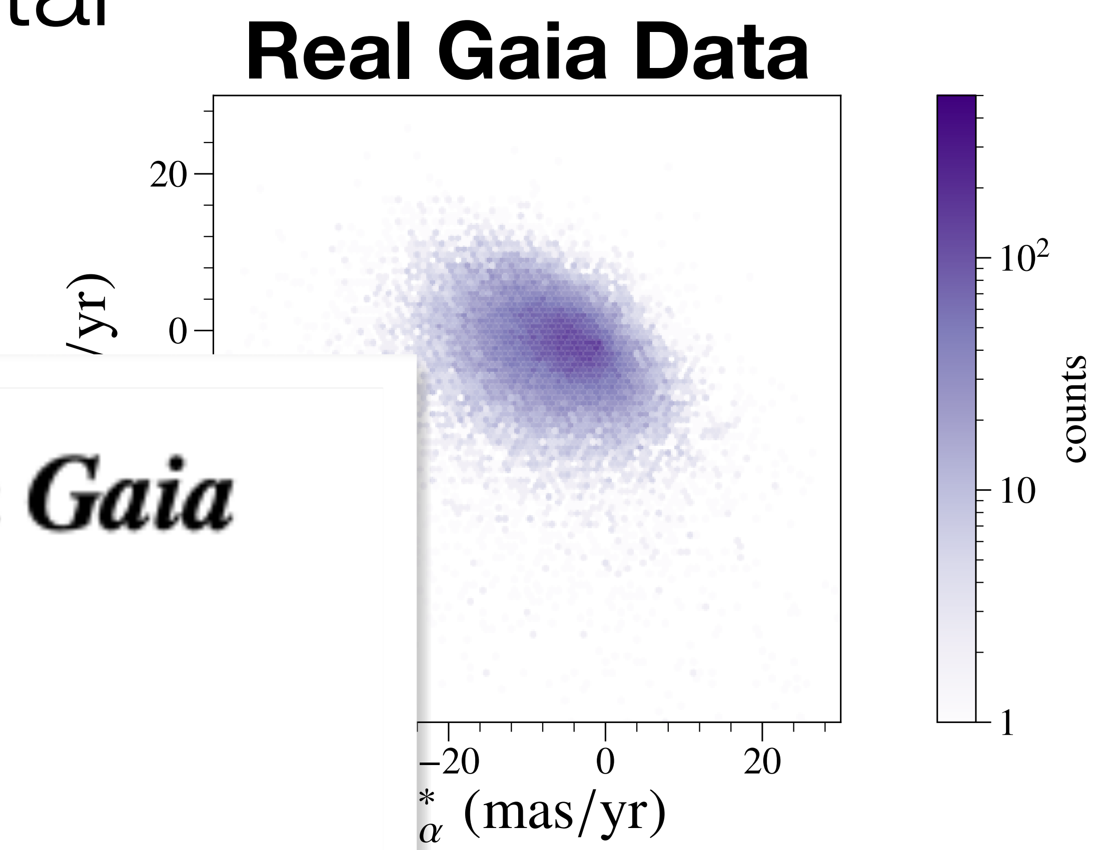
Preliminary

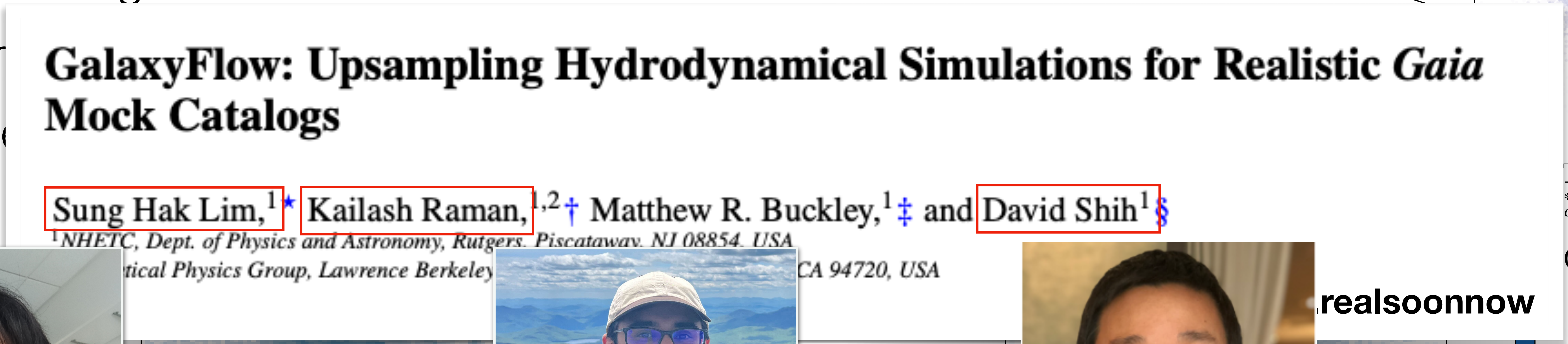


Lim *et al* (in prep)

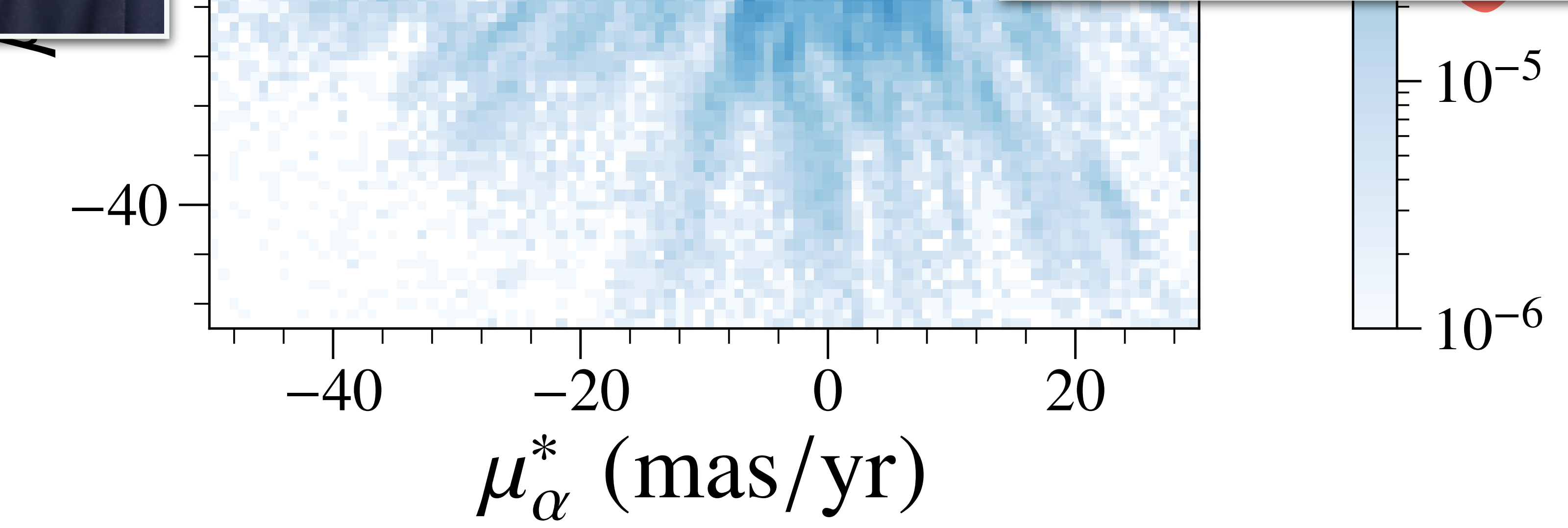
Upsampling Simulations

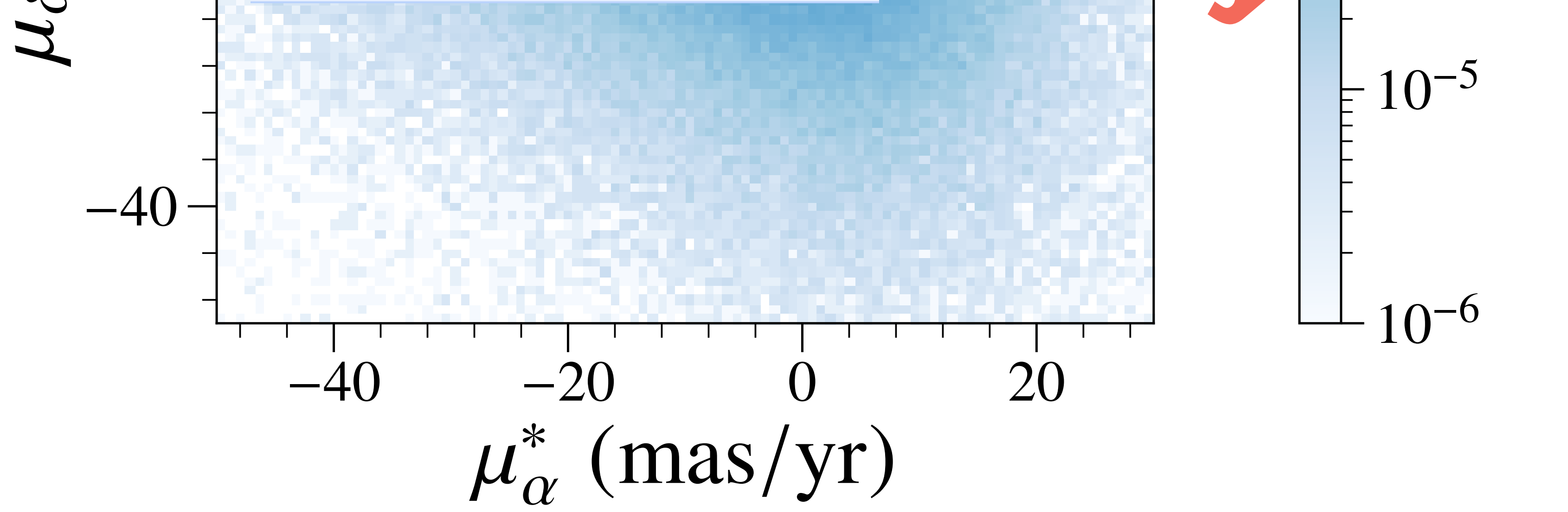
- Use normalizing flows (CNFs) to learn the density distribution of simulation star particles, then generate synthetic stars from the flow.
  - Demonstrating with stars near the “Sun”
  - Much sr
  - Confirm




GalaxyFlow: Upsampling Hydrodynamical Simulations for Realistic *Gaia* Mock Catalogs

Sung Hak Lim,[1]* Kailash Raman,[1,2]† Matthew R. Buckley,[1]‡ and David Shih[1]§
[1]NHETC, Dept. of Physics and Astronomy, Rutgers, Piscataway, NJ 08854, USA
tical Physics Group, Lawrence Berkeley CA 94720, USA

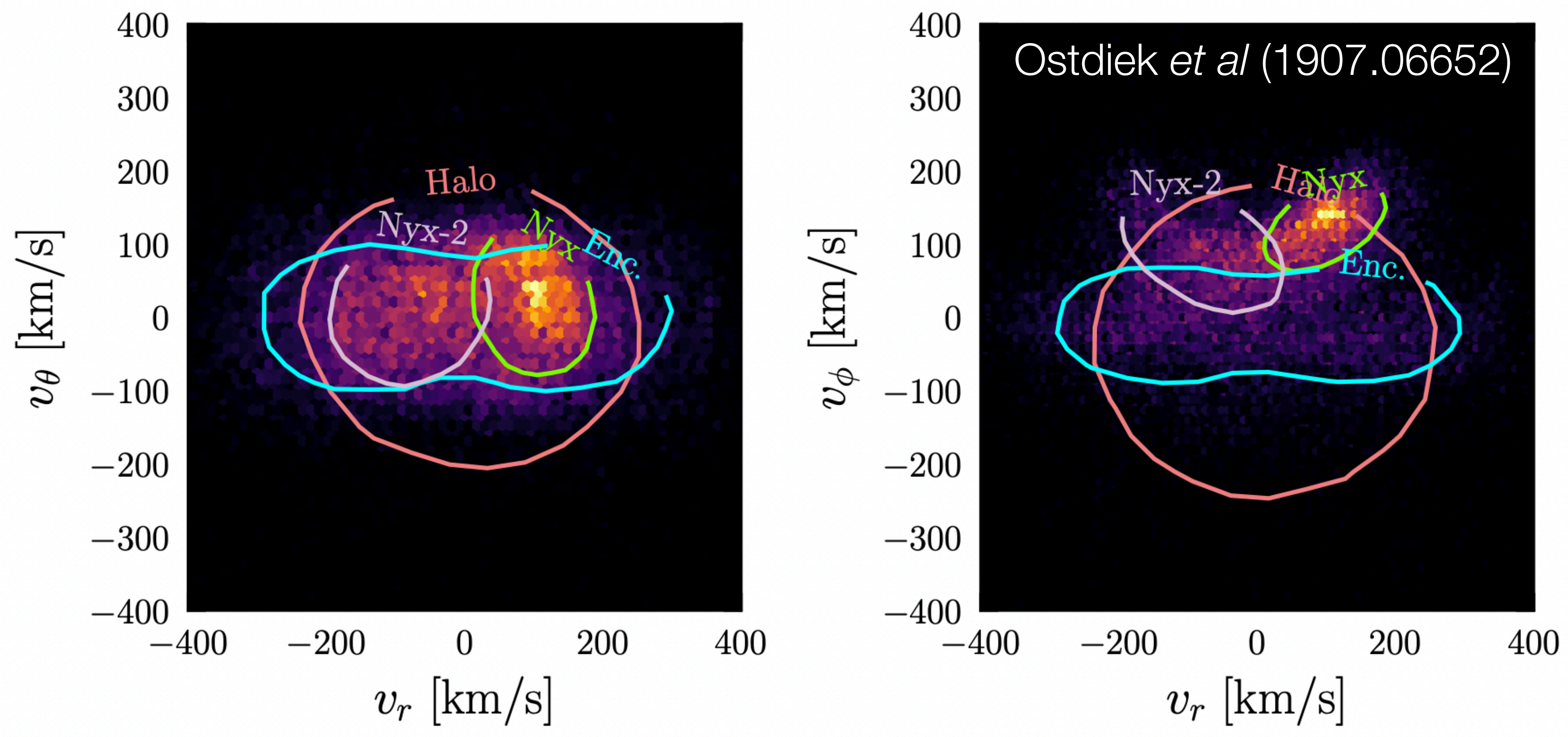


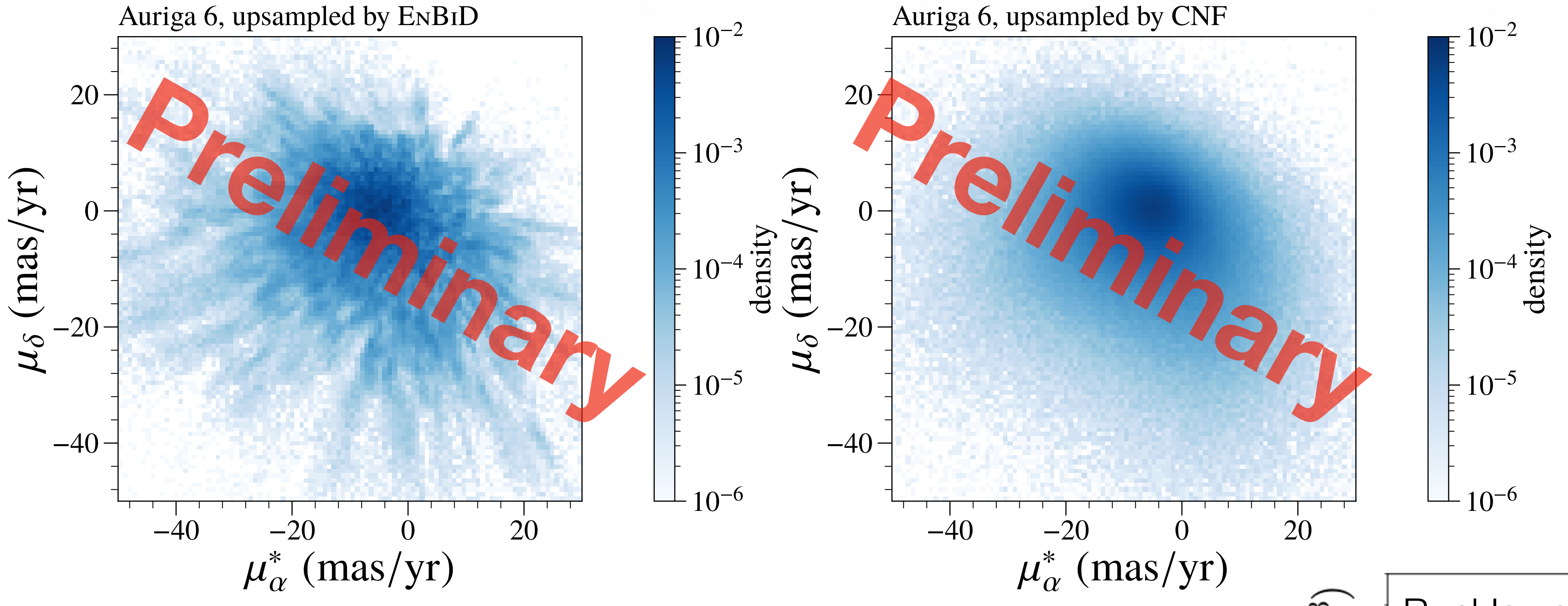


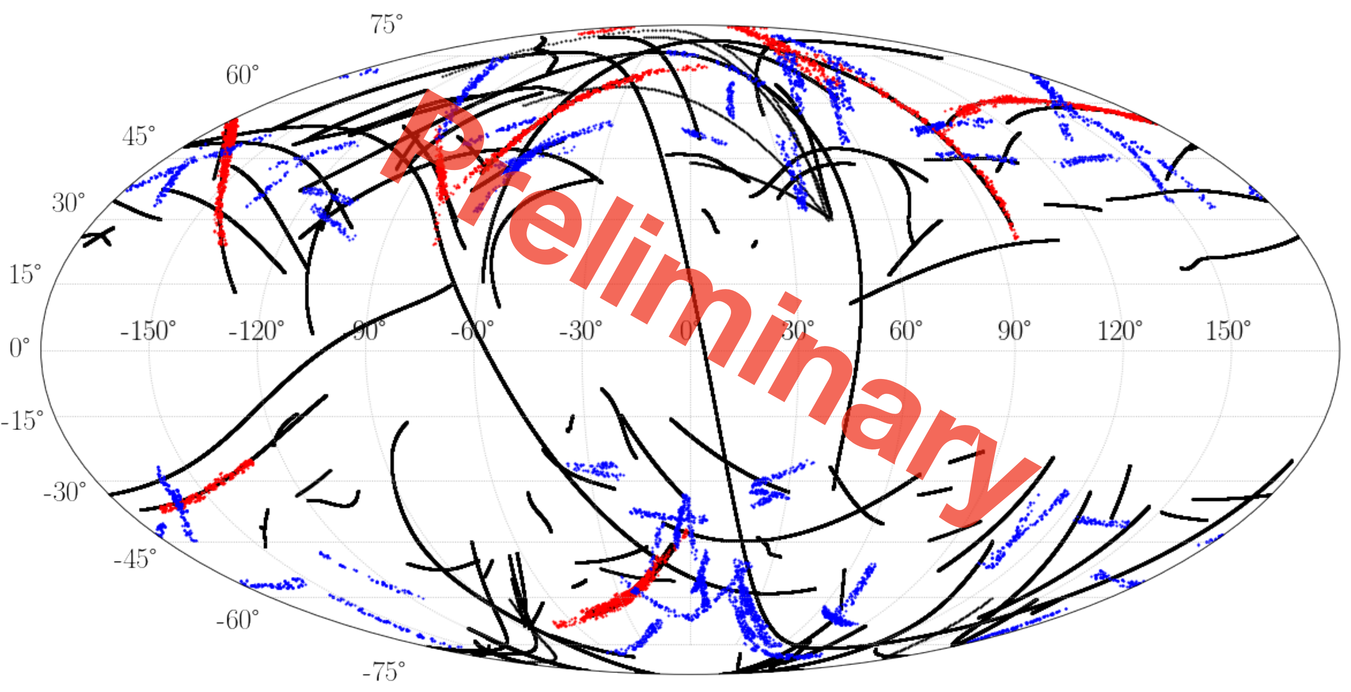
Lim *et al* (in prep)

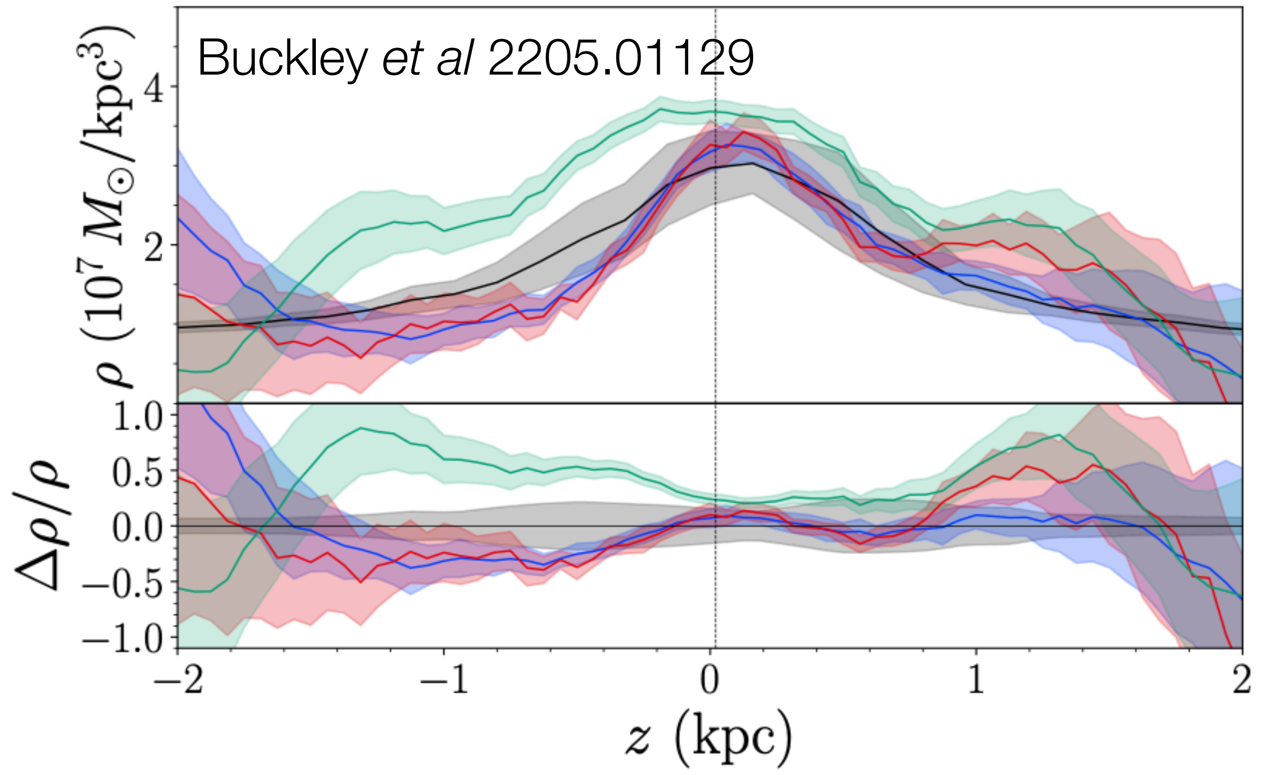
Conclusions

- 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!









ML Applications for Gaia

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