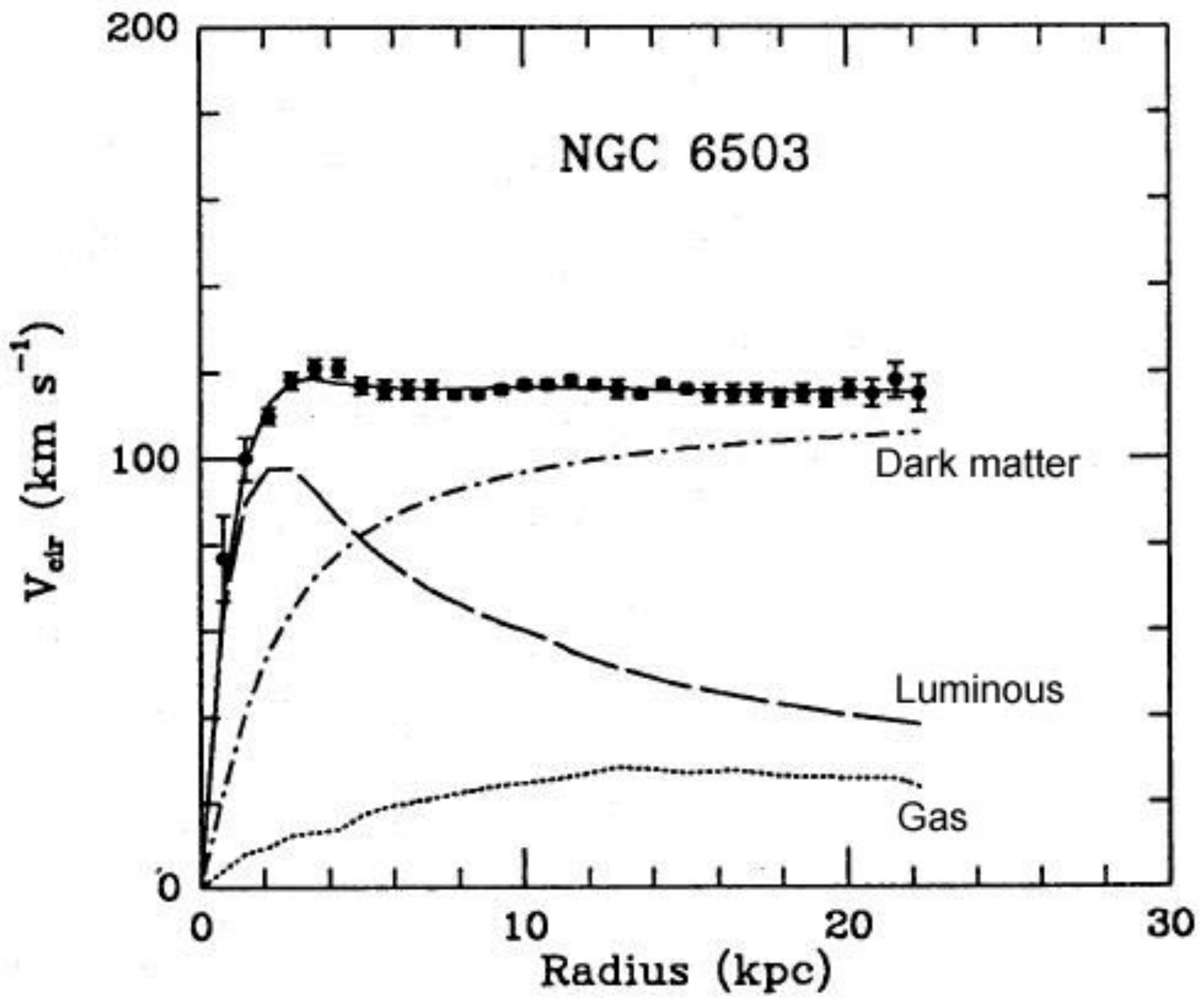
Dark Matter Models, Astrophysical Data, and Machine Learning

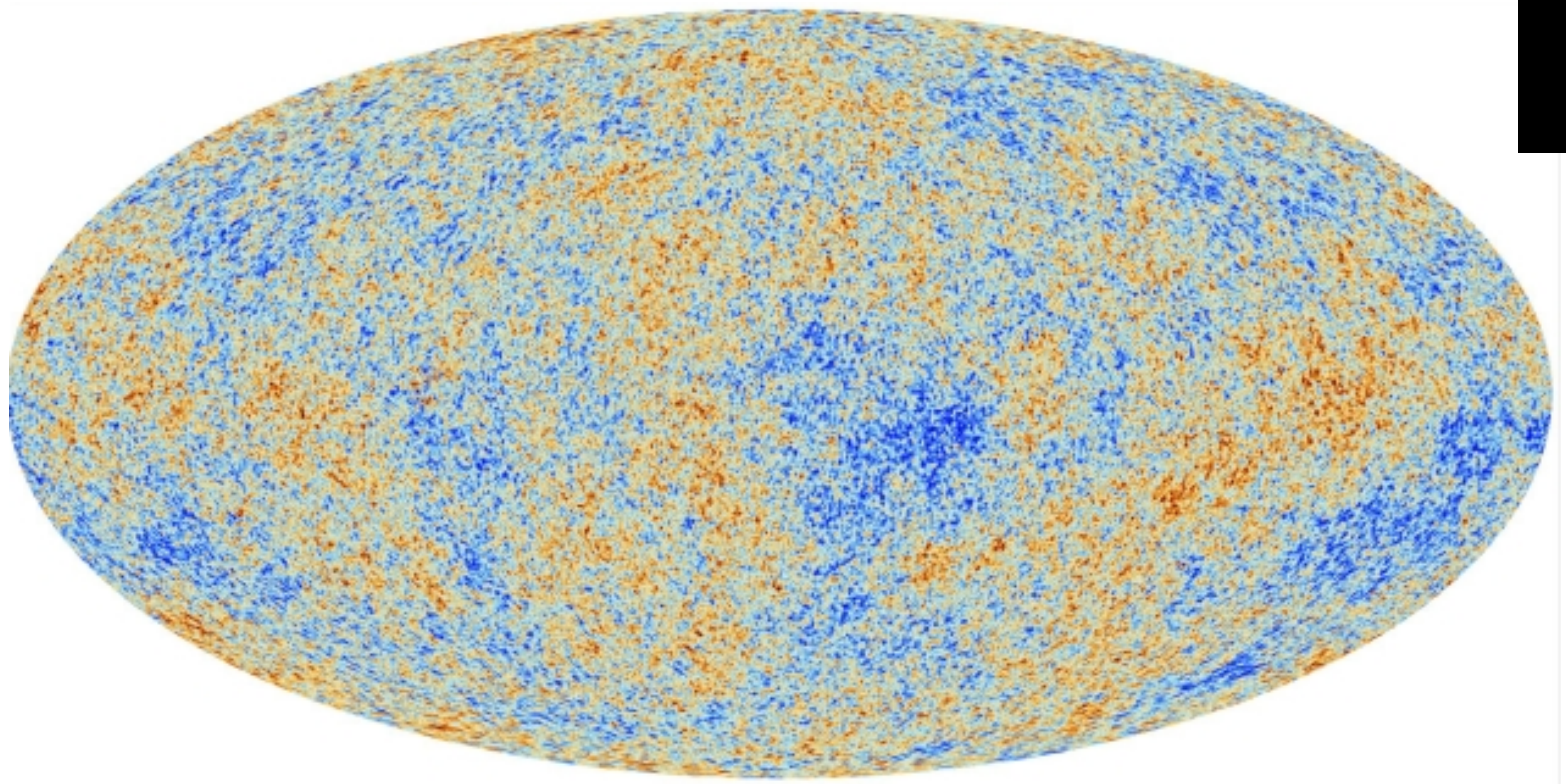
Matthew R Buckley
Rutgers University

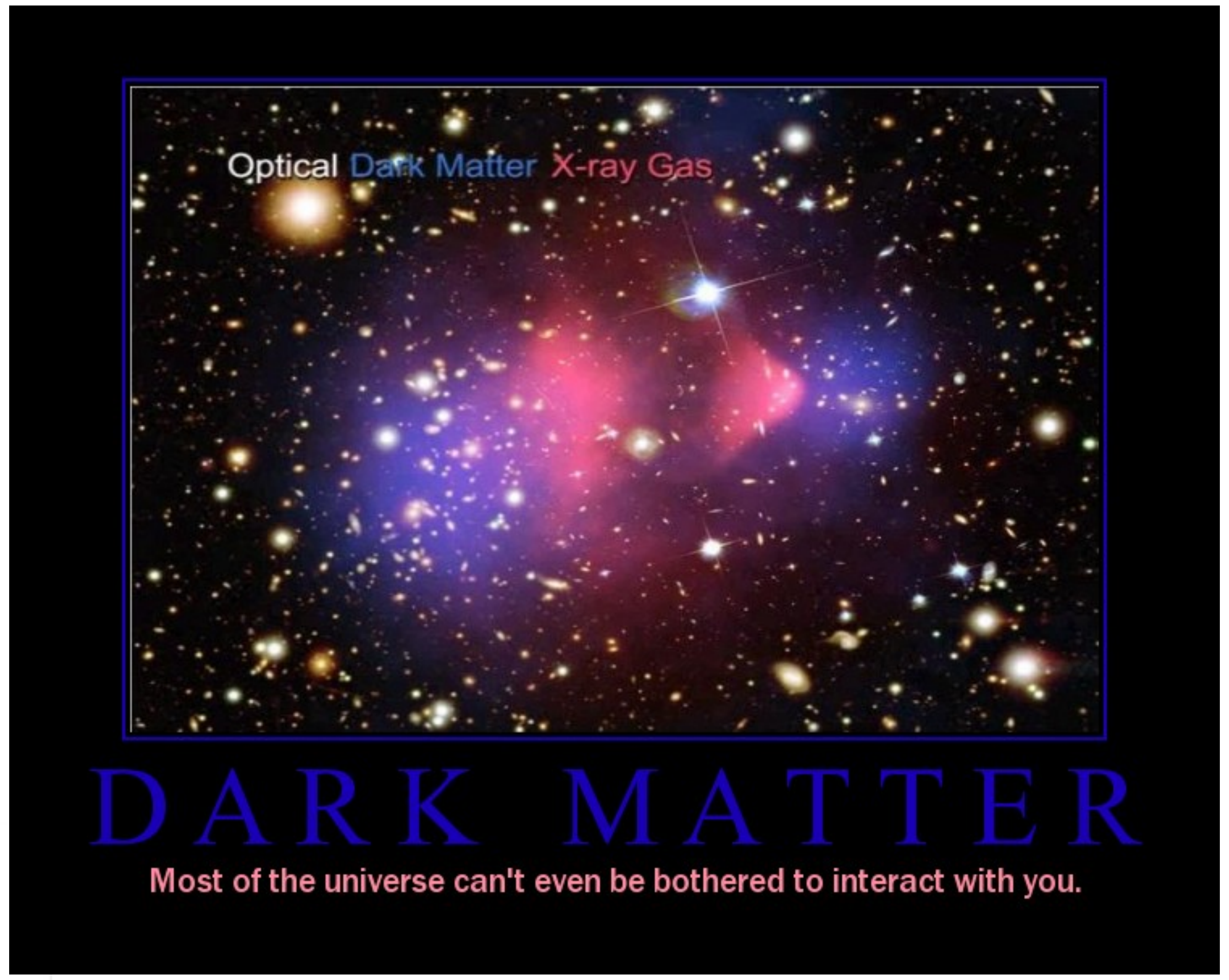
The Problem of Dark Matter

- …or, how does a particle theorist end up in an astronomy group?
  - 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.





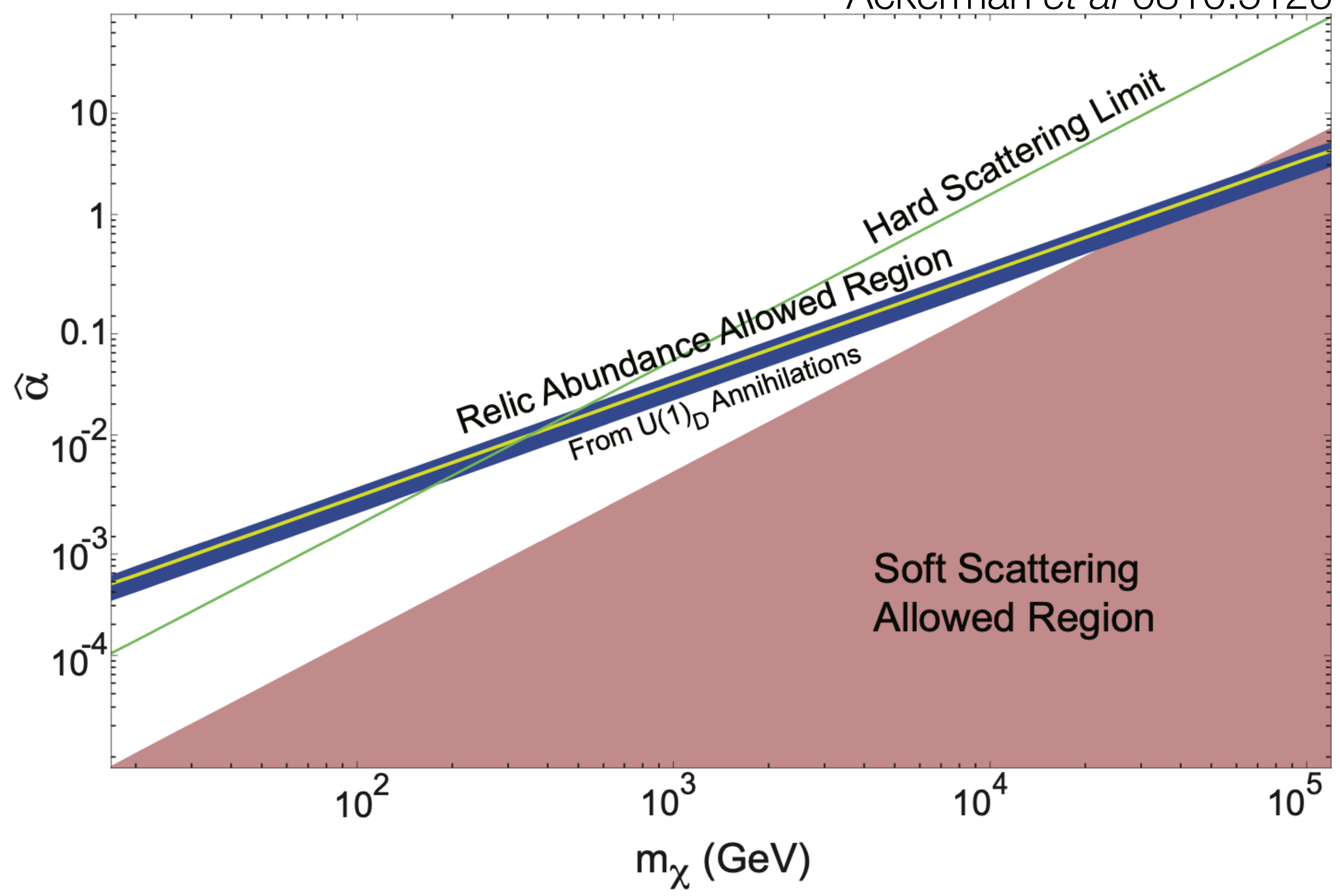
Why Should Dark Matter be Boring?

- Motivated in part by astrophysical anomalies, particle theorists started to branch out from SUSY/technicolor-inspired WIMPs, axions, sterile neutrinos
  - Complicated dark sectors imply complicated dynamics.

> Additionally, since the $U(1)_D$ effectively makes the dark halo a plasma (albeit a very cold, tenuous one), there may be other effects on structure formation that constrain this model [38]. We have estimated that the timescale for the Weibel instability in our model is short compared to relevant timescales for galactic dynamics. If this instability has a dramatic effect when subhalos collide during the assembly of a galactic halo, our $U(1)_D$ could be excluded for the entire range of interesting parameters. Further work is required to before we reliably understand the quantitative effects of such instabilities on galactic dynamics.

Theorist for ¯_(ツ)_/¯

Ackerman *et al* 0810.5126

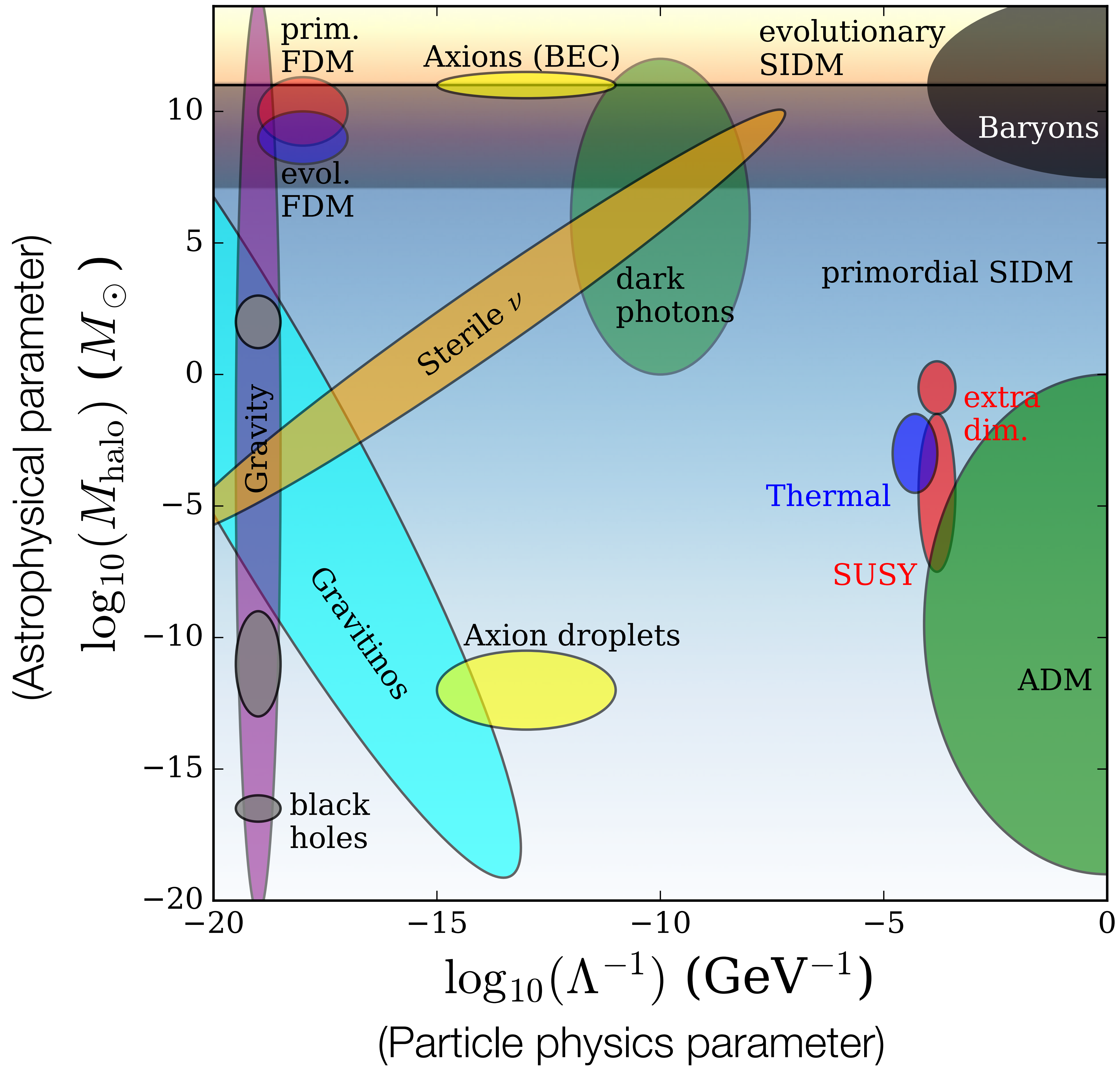


- Model building has to be in conversation with the data. What do these non-trivial interactions *do* to observable quantities?

A Language Barrier

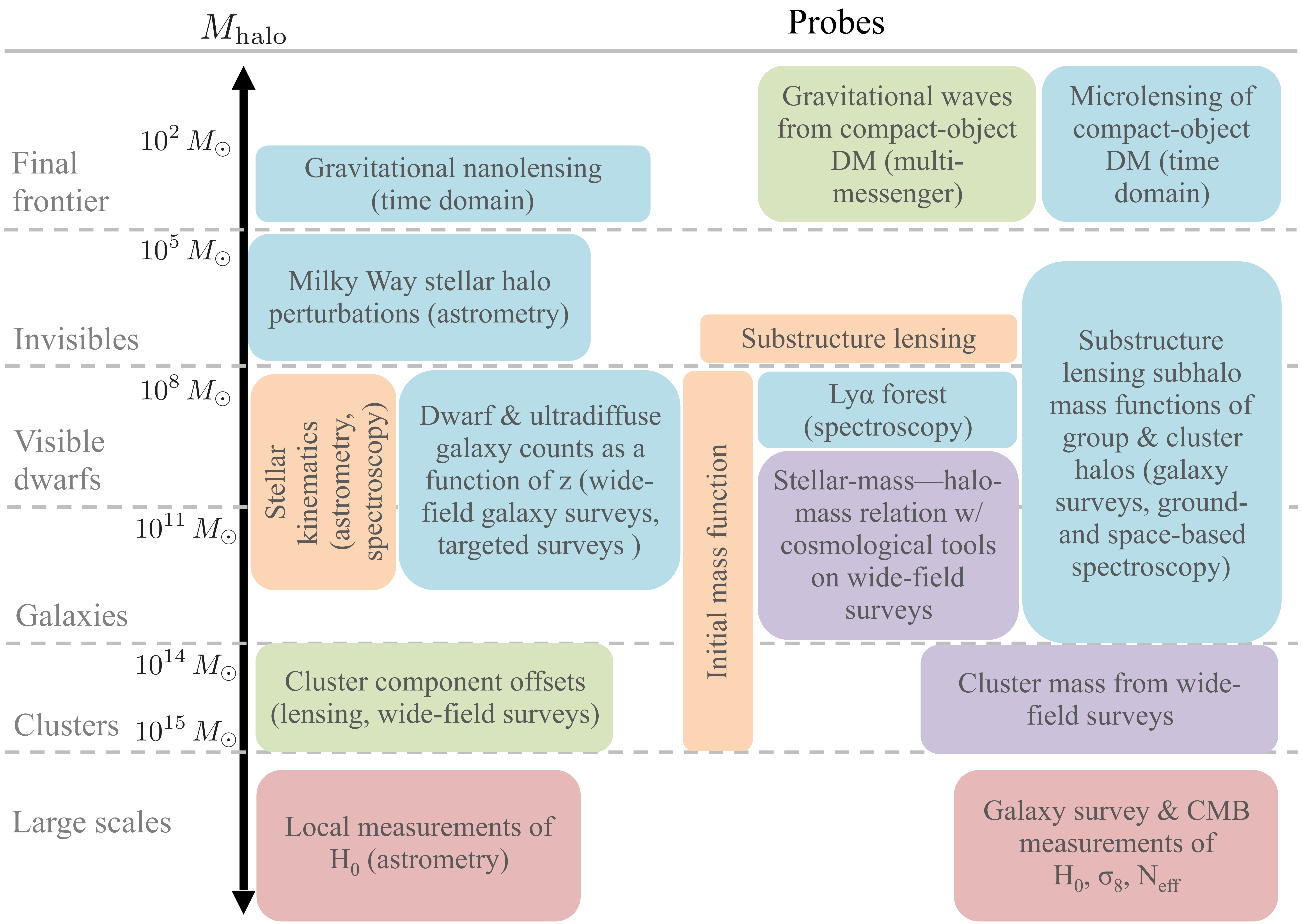
- Particle Theorists and Astrophysicists think about dark matter in very different ways.
- Non-trivial physics result in modifications in dark matter halos at specific scales.
  - Scales that might not be interesting in a “vanilla” CDM model.
- We are not going to figure out the model of dark matter from pure theory alone. We need data to point the way.

Buckley and Peter 1712.06615



Opportunities in Data

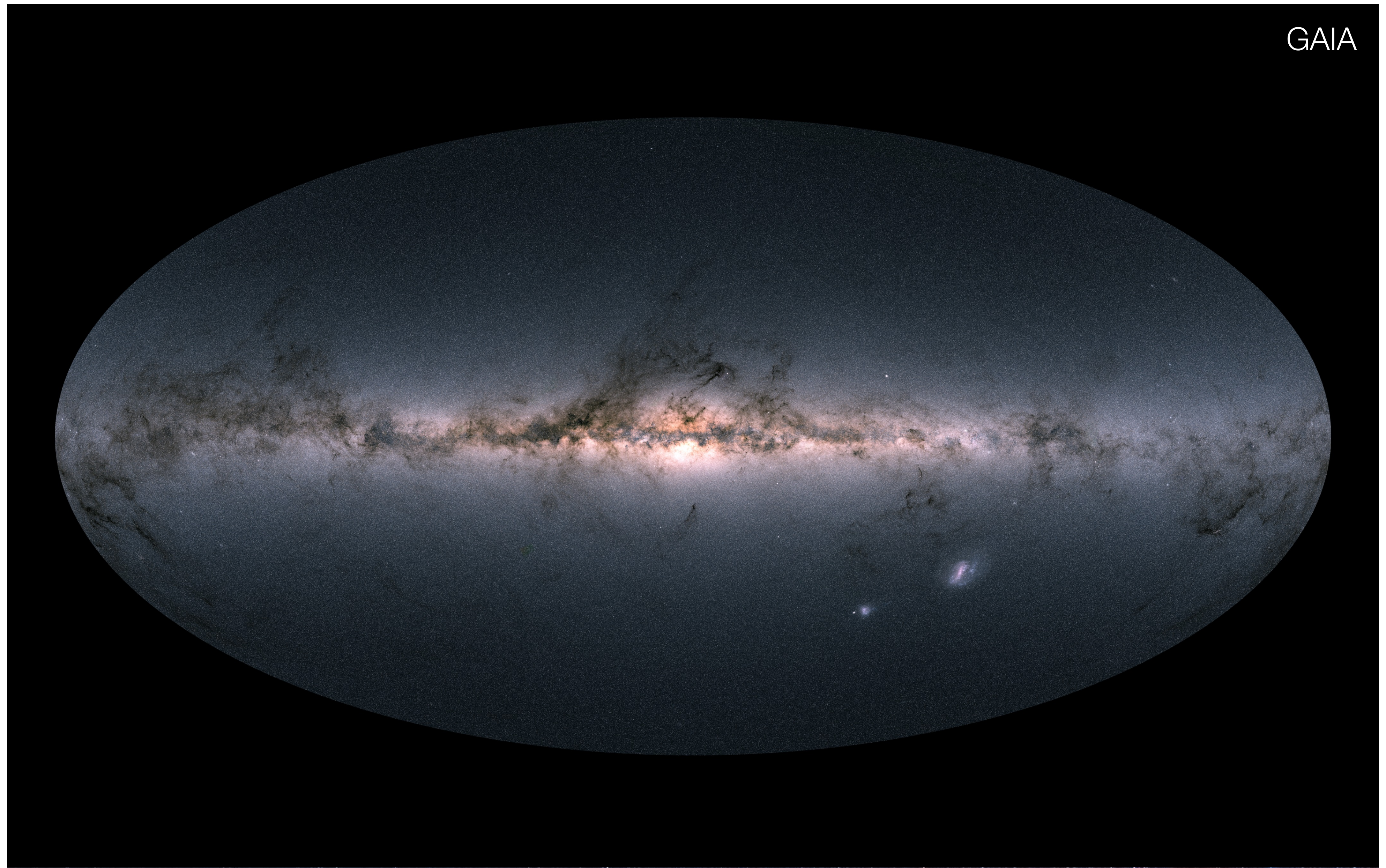
Buckley and Peter 1712.06615

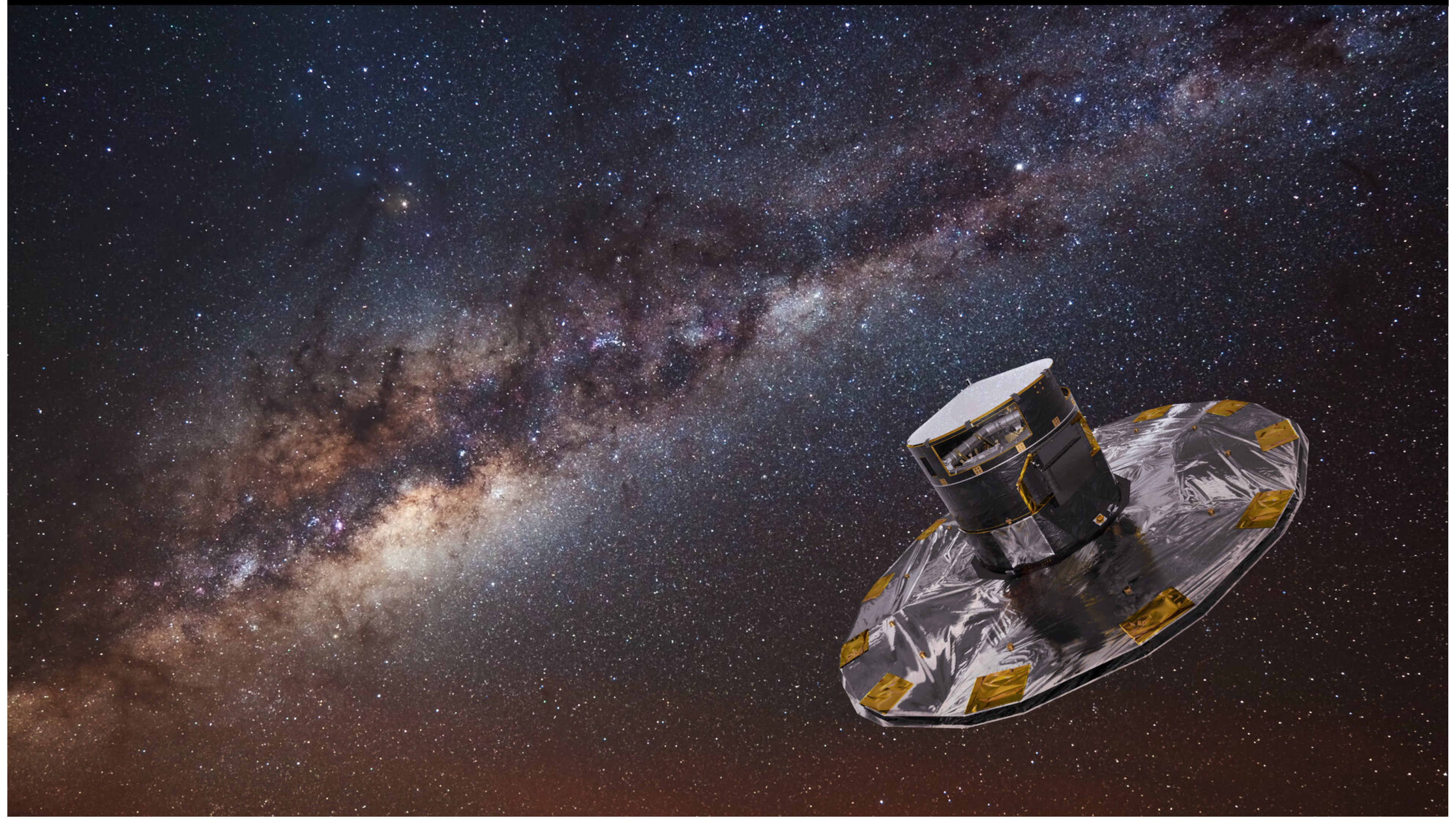


The *Gaia* Dataset

- 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.
- I’m interested in Gaia data as processed locations of stars within 4/5/6D kinematic space — not as individual images/spectra



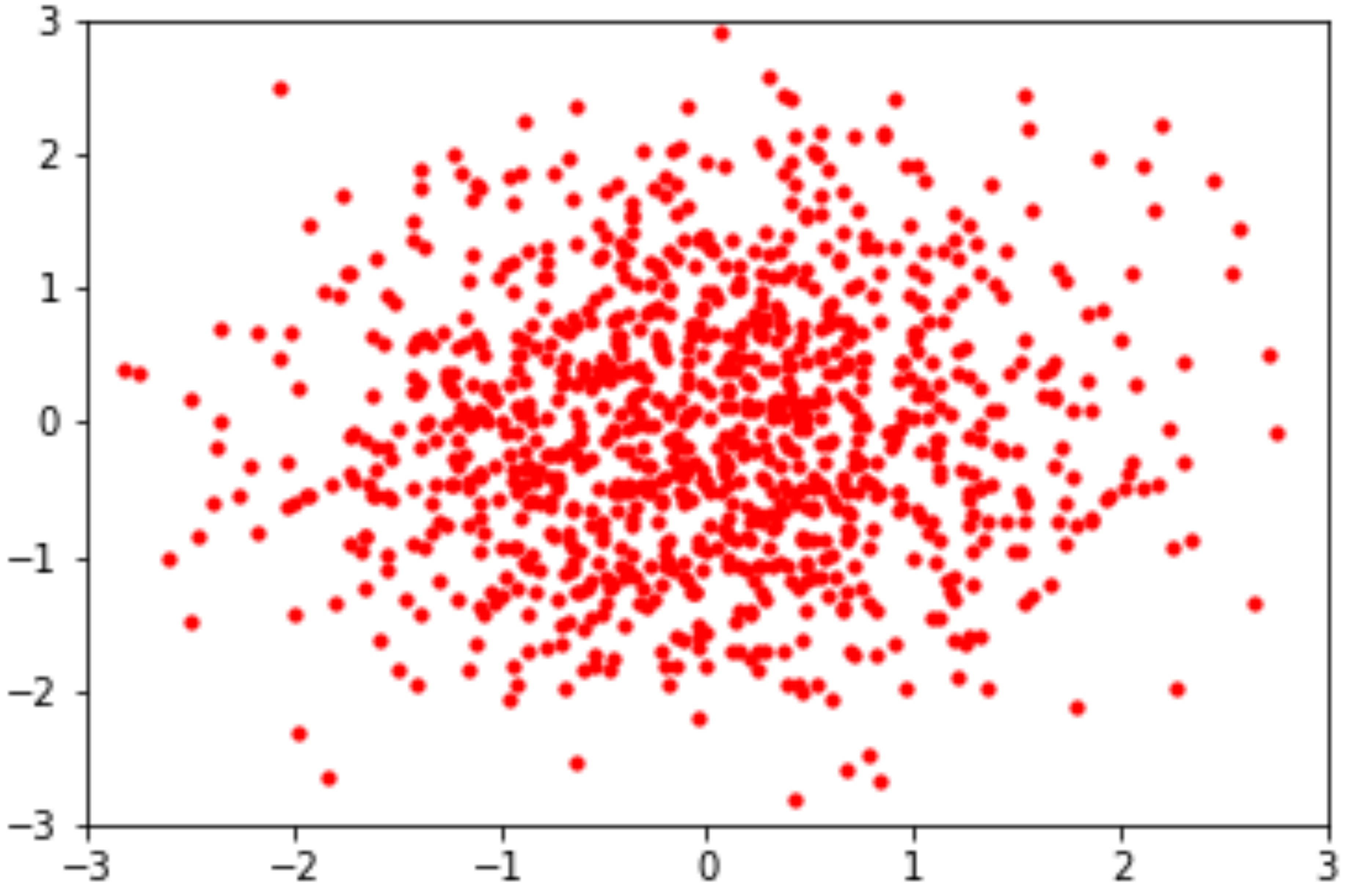


Machine Learning

- …kind of a hot topic these days
- Useful as a tool for physics:
  - Provides ways to analyze large, high-dimensional datasets
- Of particular interest: *normalizing flows* (see 1908.09257 for review)
  - Learn the transformation from a simple base distribution (Gaussians) to the (unknown) probability distribution of data.
  - Loss function is effectively the entropy of the dataset:

$$\mathcal{L} = -\sum_i \ln f(x_i)$$

- Makes the *phase space distribution* (and its derivatives) a tractable experimental “observable”



Eric Jang

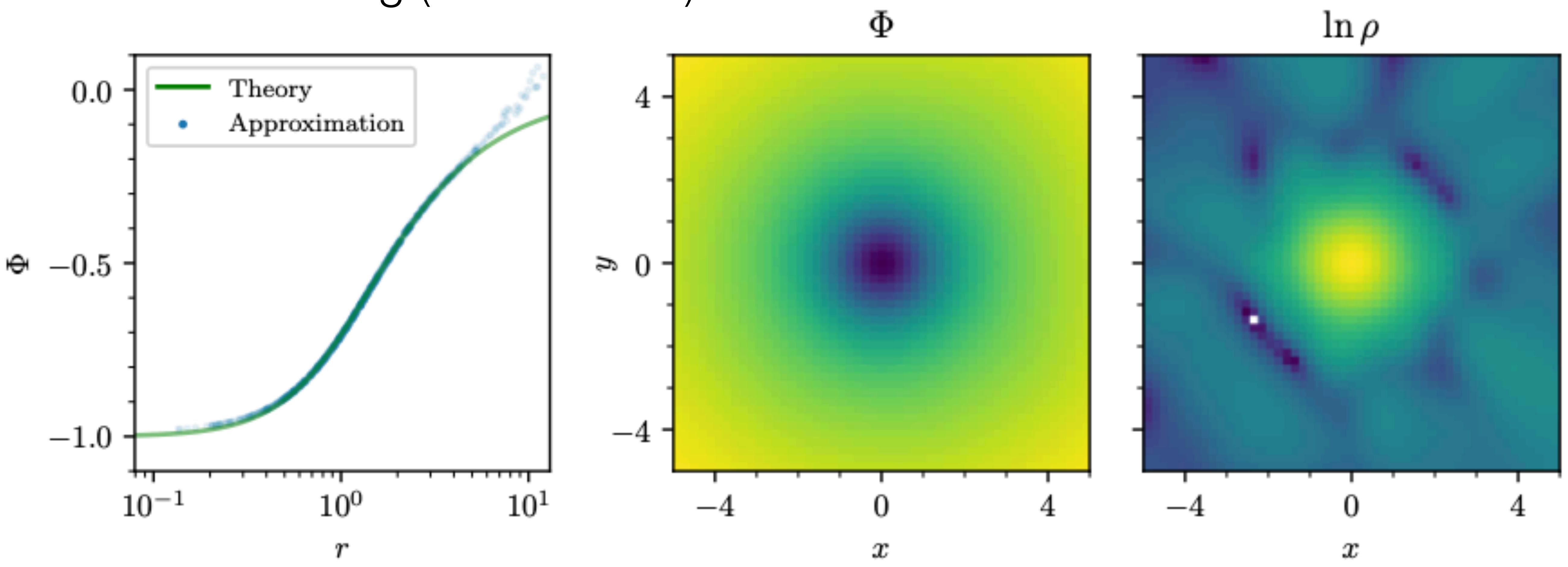
Dark Matter Density from Flows

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

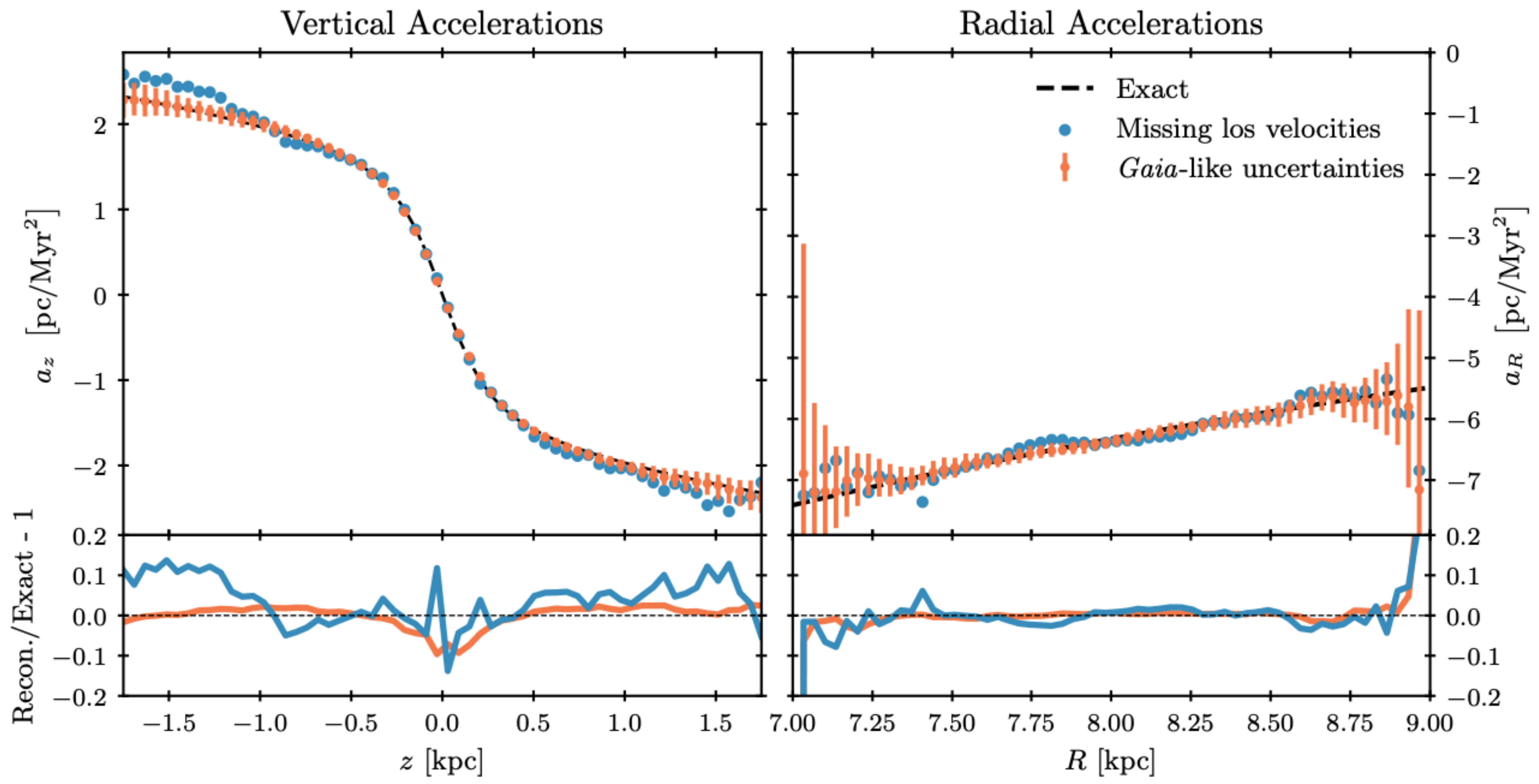
$$\frac{\partial f}{\partial t} + v_i \frac{\partial f}{\partial x_i} = \frac{\partial \Phi}{\partial x_i} \frac{\partial f}{\partial v_i}$$

- 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)

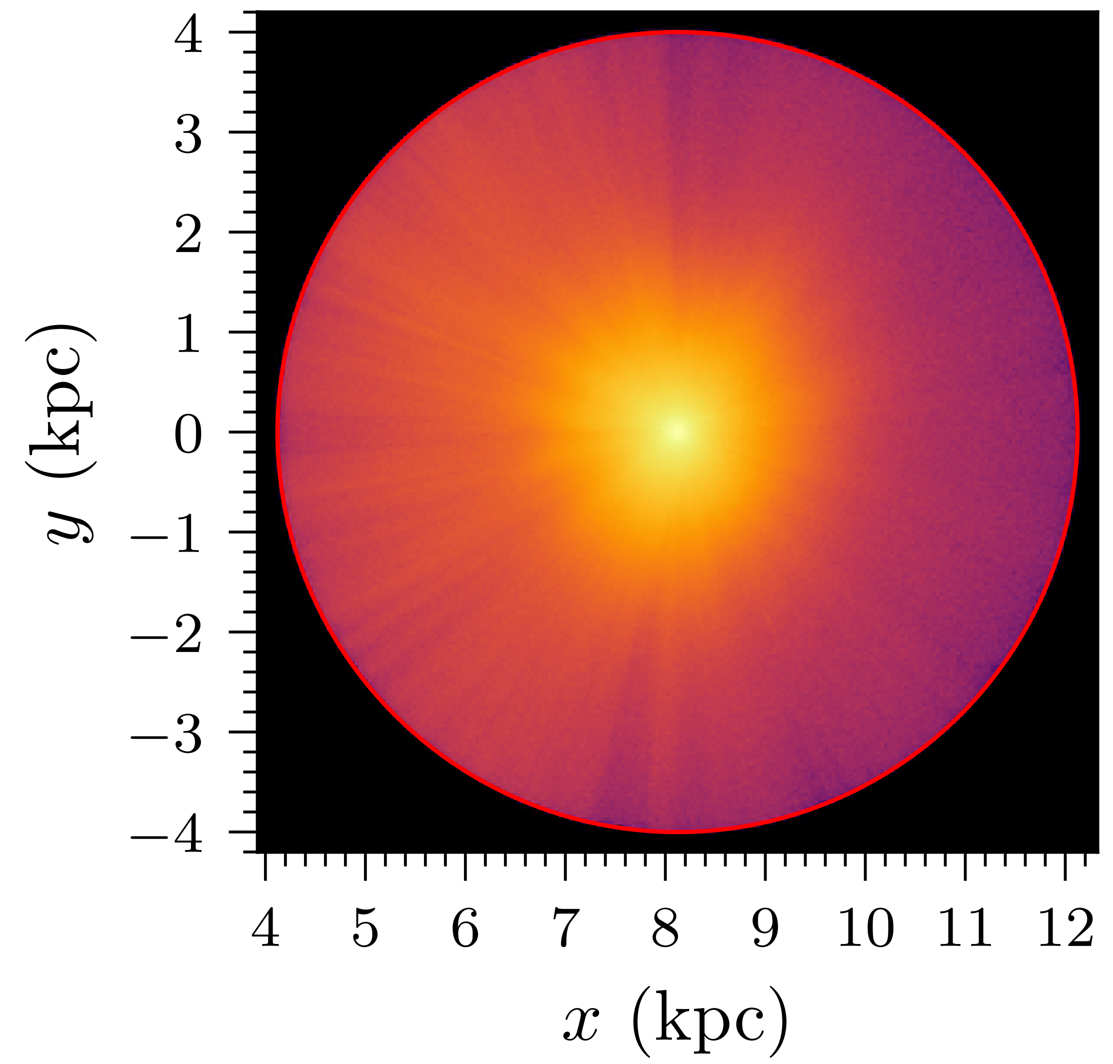


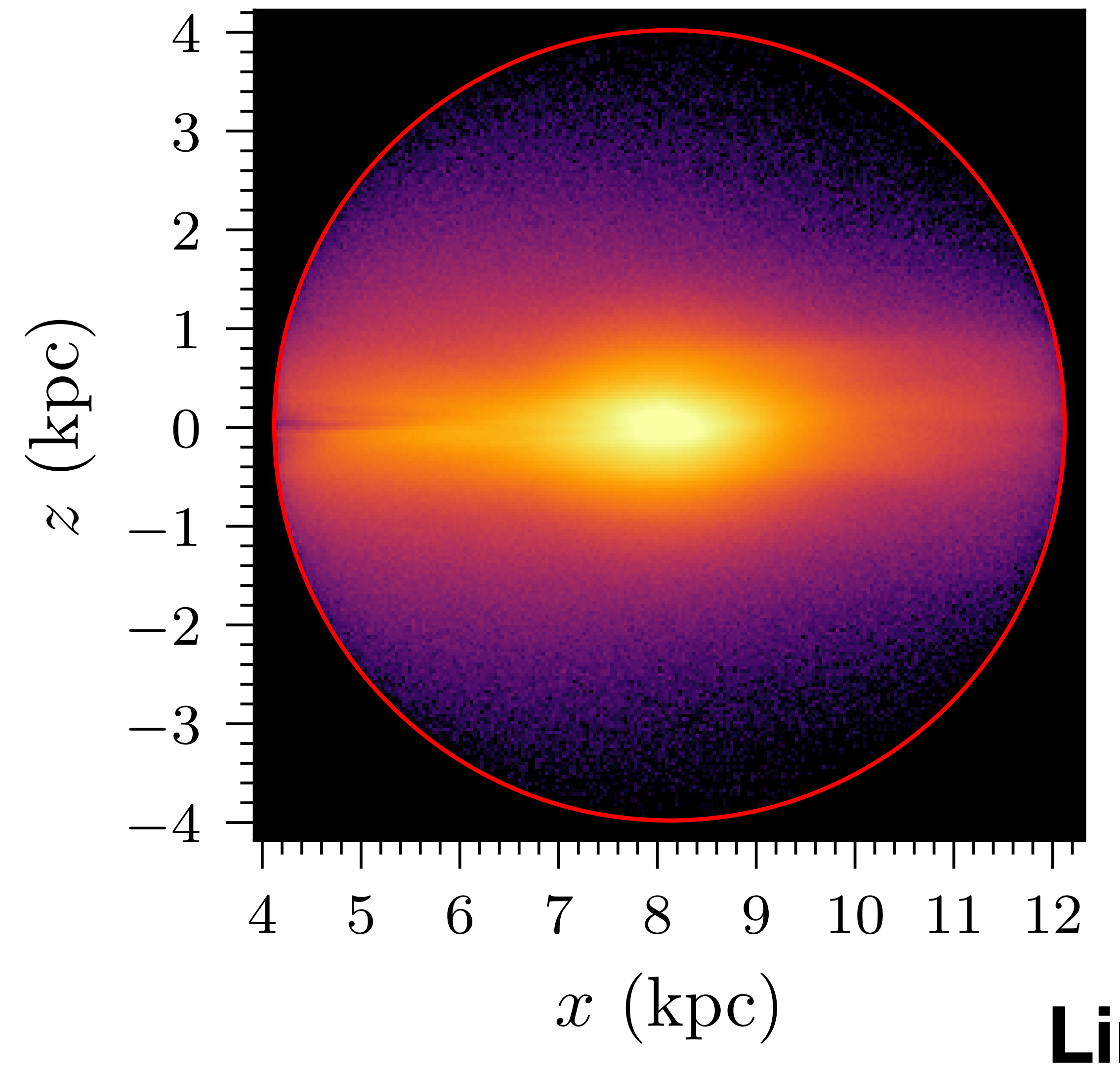
Real Data

- We tested with simulation (Buckley *et al* (2205.01129). Ask me for details
- But more interesting: real data
- Select 29,855,114 stars from the *Gaia* DR3 within 10 kpc of the Sun with full phase space information:

$$(\alpha, \delta, \varpi, \mu_\alpha^*, \mu_\delta, \mathrm{RVS})$$

  - Remove poorly measured stars. Retain error covariance matrix to allow for error propagation
  - 24,789,061 stars within 4 kpc of the Sun

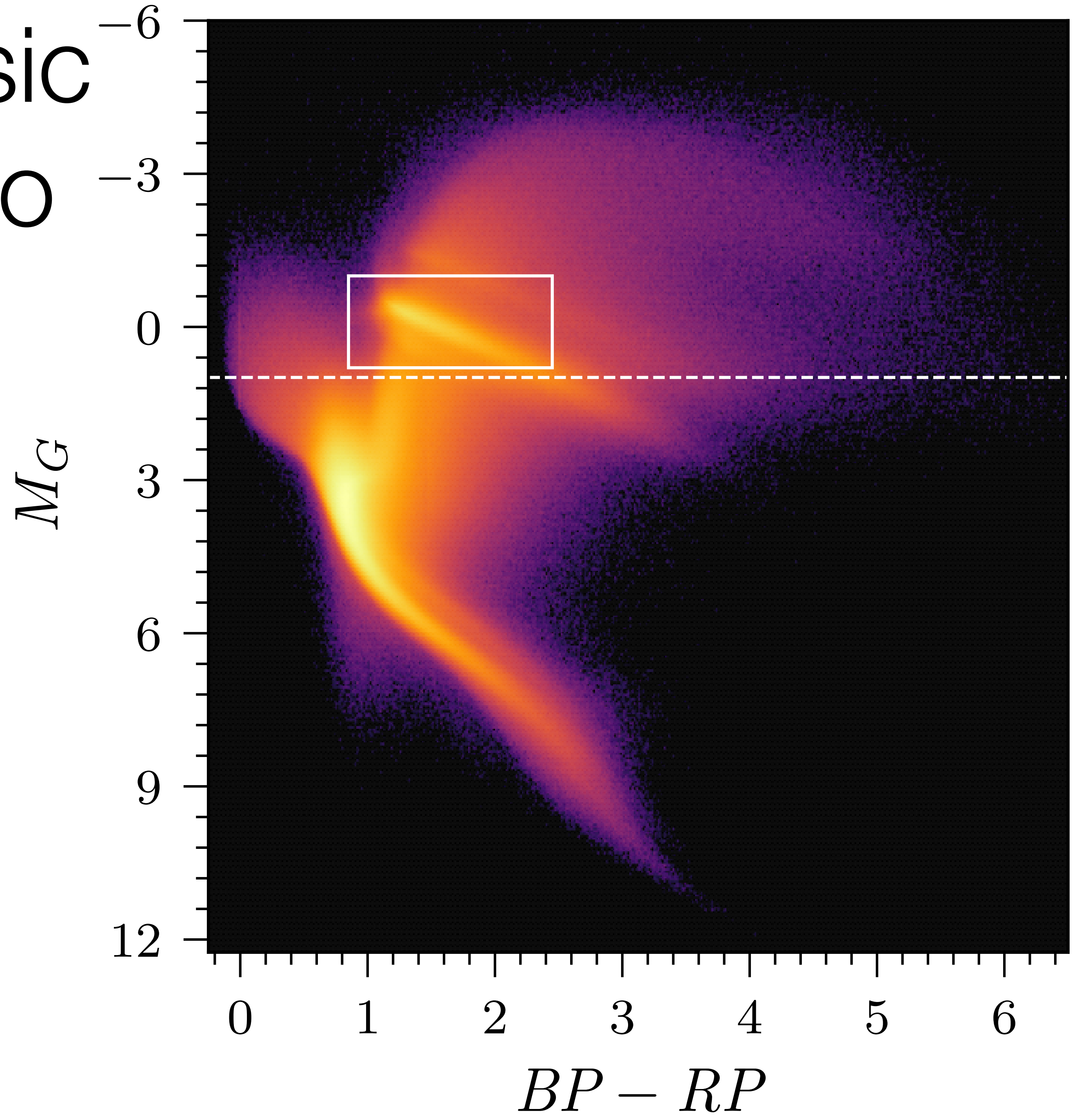


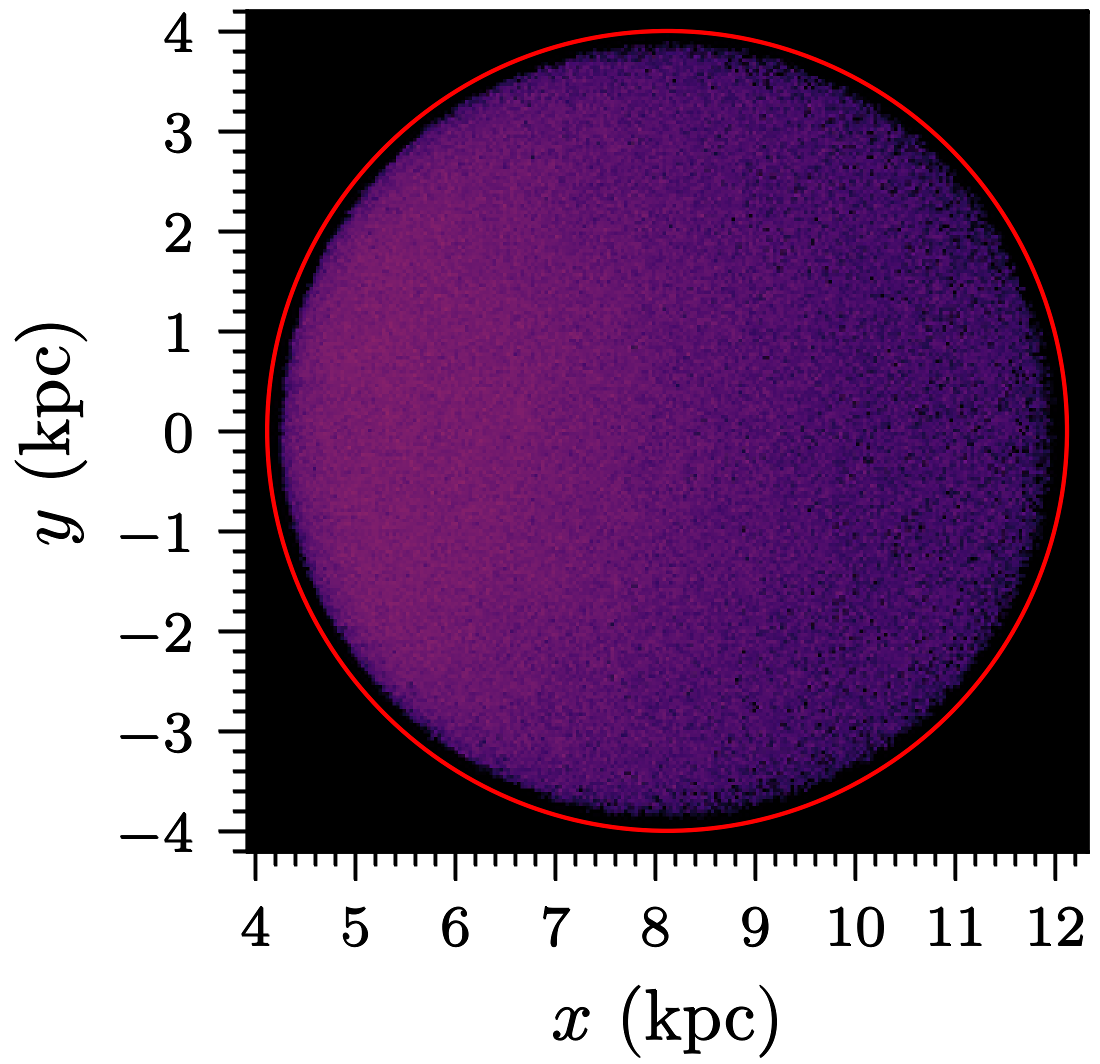


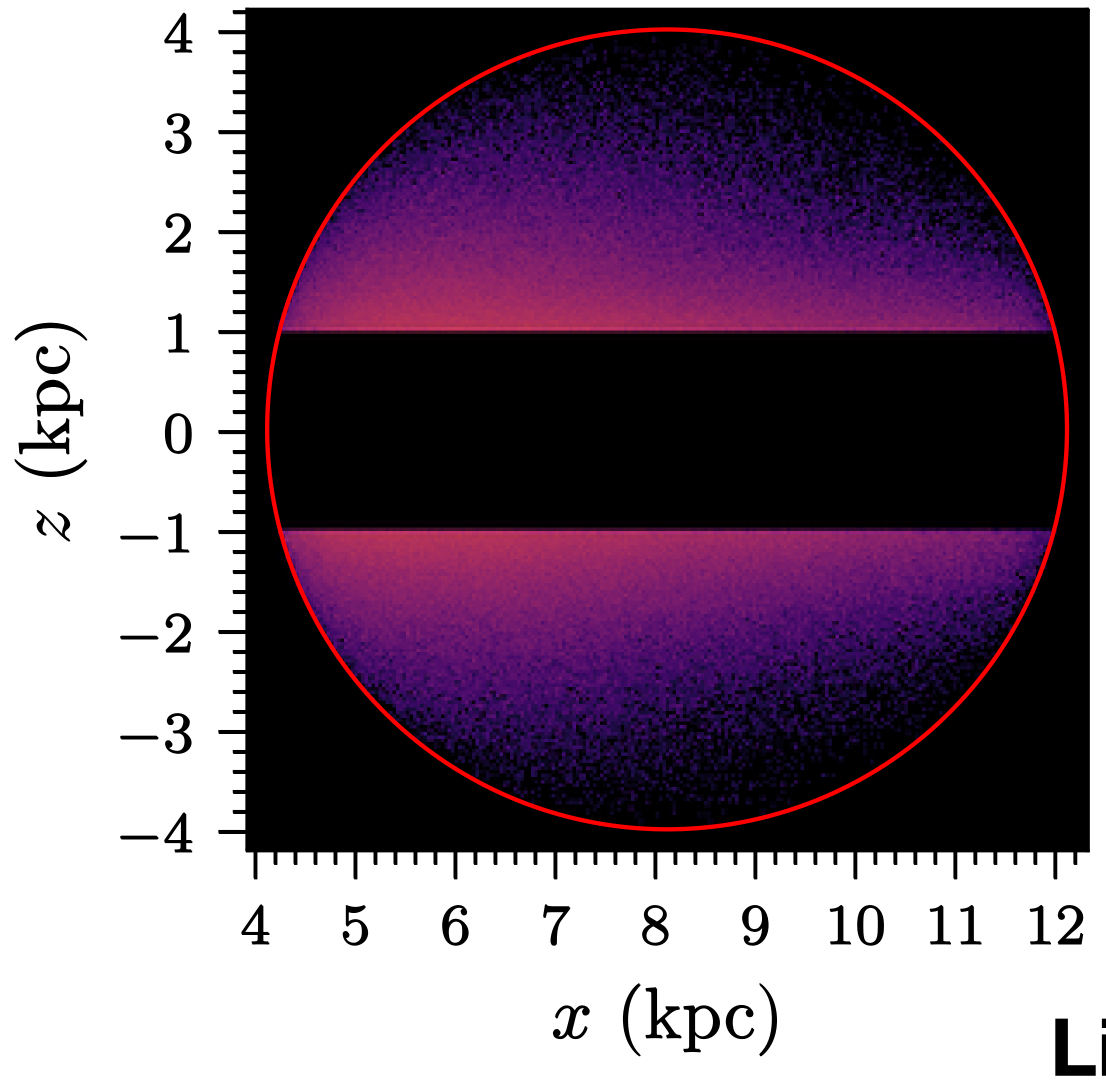
Lim *et al* 2305.13358

Real Data

- Our sample is biased: distant stars that are bright are included, but not nearby dim stars.
- *Gaia* spectrometer complete down to observed magnitude $G_{\rm RVS} < 14$
- Require stars have intrinsic magnitude low enough so that they could be seen everywhere in our 4 kpc volume



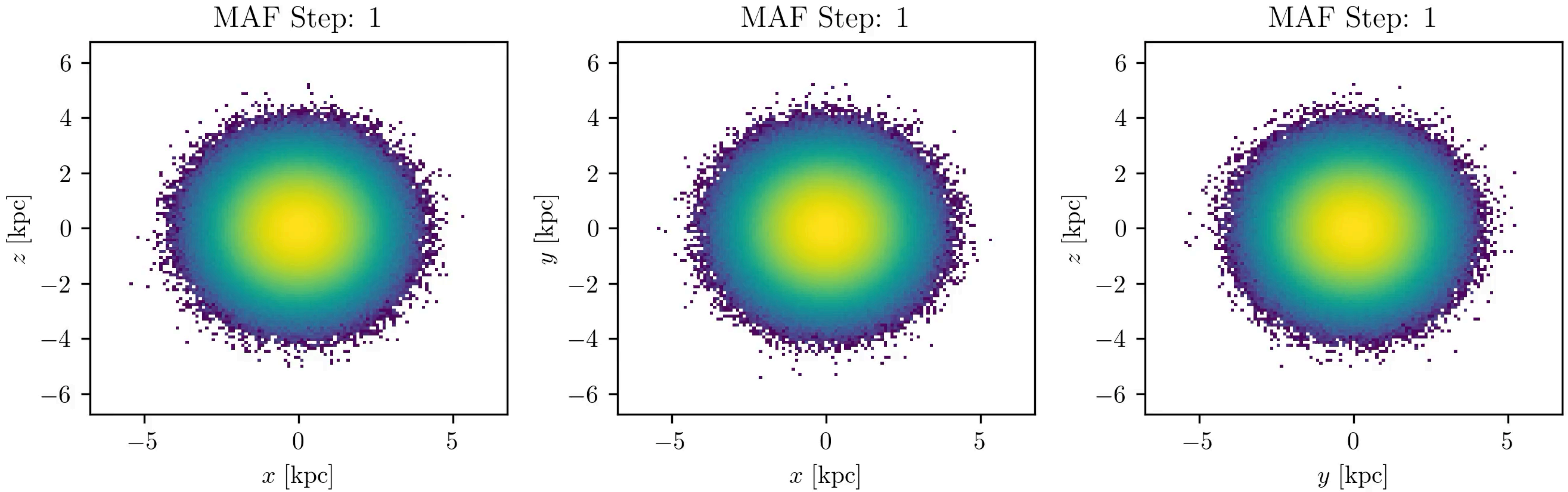




Lim *et al* 2305.13358

Normalizing Flows

- Train our flows on the real data
  - For error propagation we reperturb data within errors and redo entire analysis
  - To estimate statistical errors we “bootstrap” (sample *with* replacement) new datasets and redo entire analysis

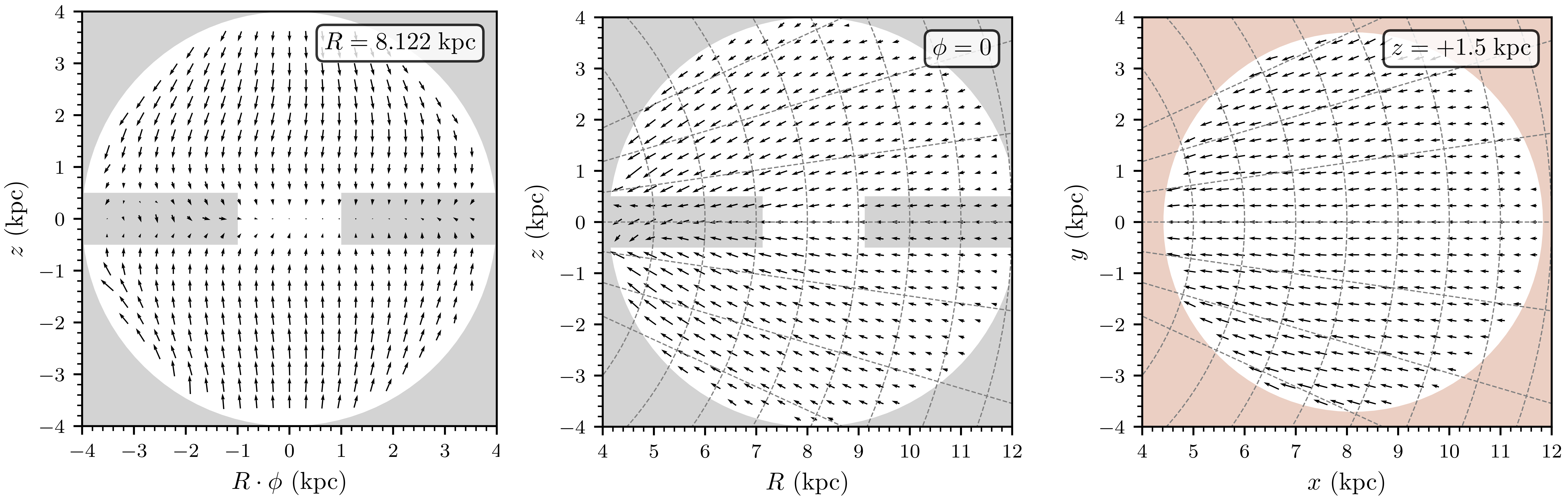


Accelerations

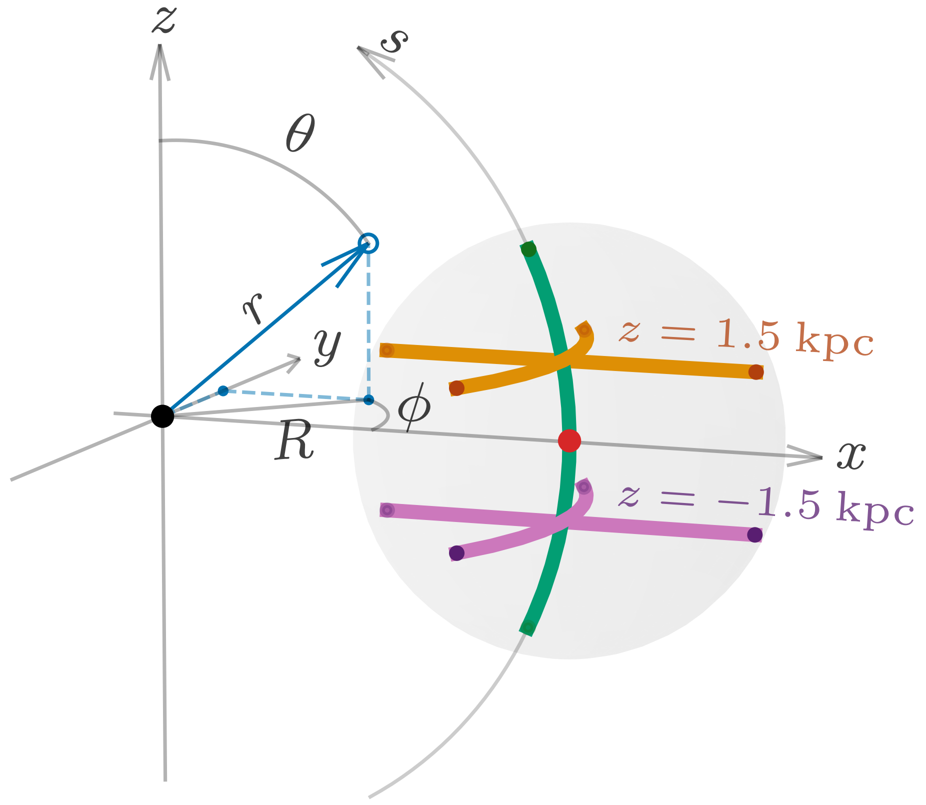
- At each x, sample 10K v from the MAF, numerically finding $\vec{a} = -\vec{\nabla}\Phi$ by minimizing the MSE $$\sum \left| v_i \frac{\partial f}{\partial x_i} - \frac{\partial \Phi}{\partial x_i} \frac{\partial f}{\partial v_i} \right|^2 = \sum \left| \frac{\partial f}{\partial t} \right|^2$$

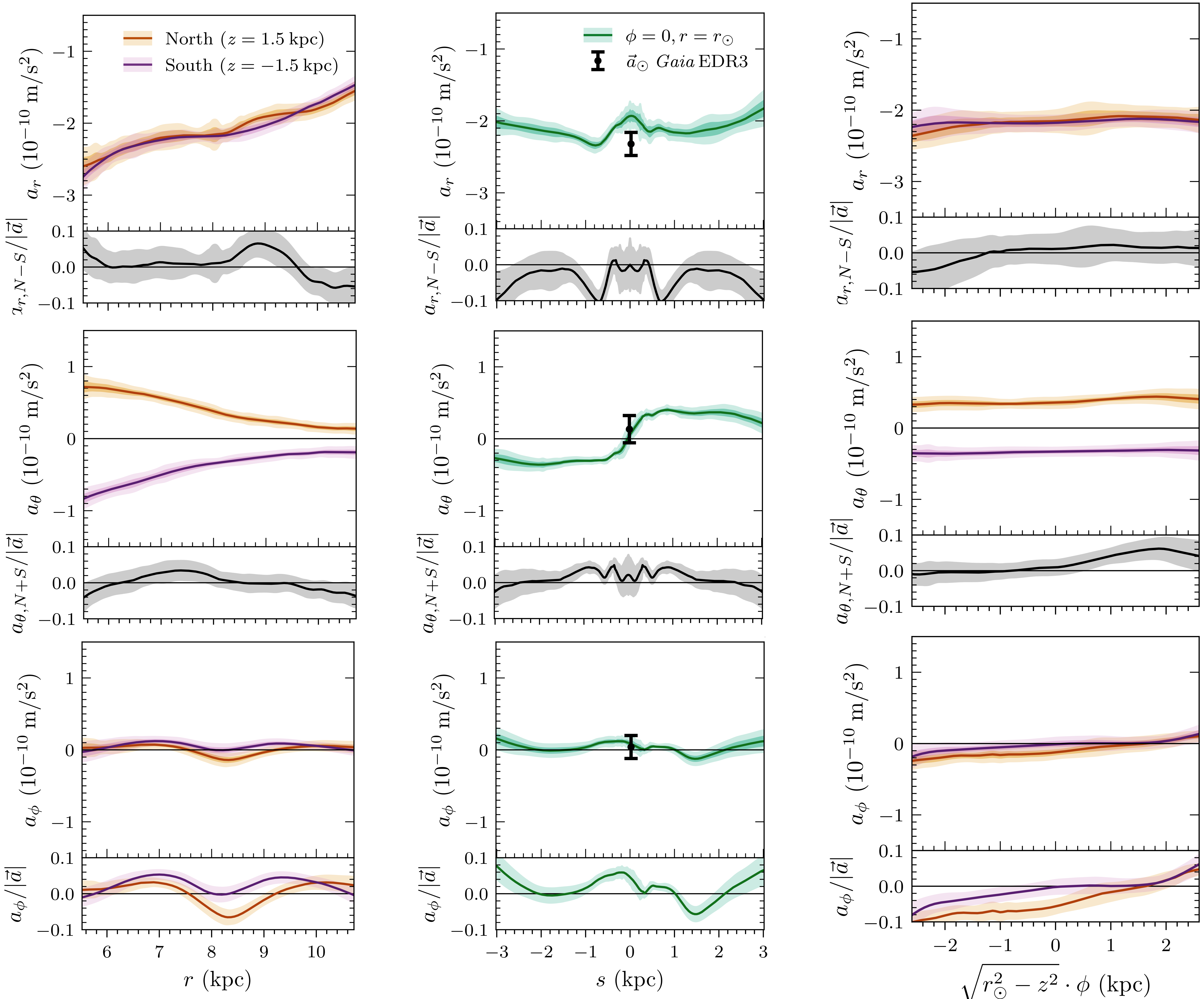
$$\boxed{\vec{a}_{\odot,\mathrm{MAF}} = (-1.94 \pm 0.22, 0.08 \pm 0.08, -0.06 \pm 0.08) \times 10^{-10}\ \mathrm{m/s^2}}$$

$$\vec{a}_{\odot,\mathrm{quasar}} = (-2.32 \pm 0.16, 0.04 \pm 0.16, -0.14 \pm 0.19) \times 10^{-10}\ \mathrm{m/s^2}$$



Accelerations





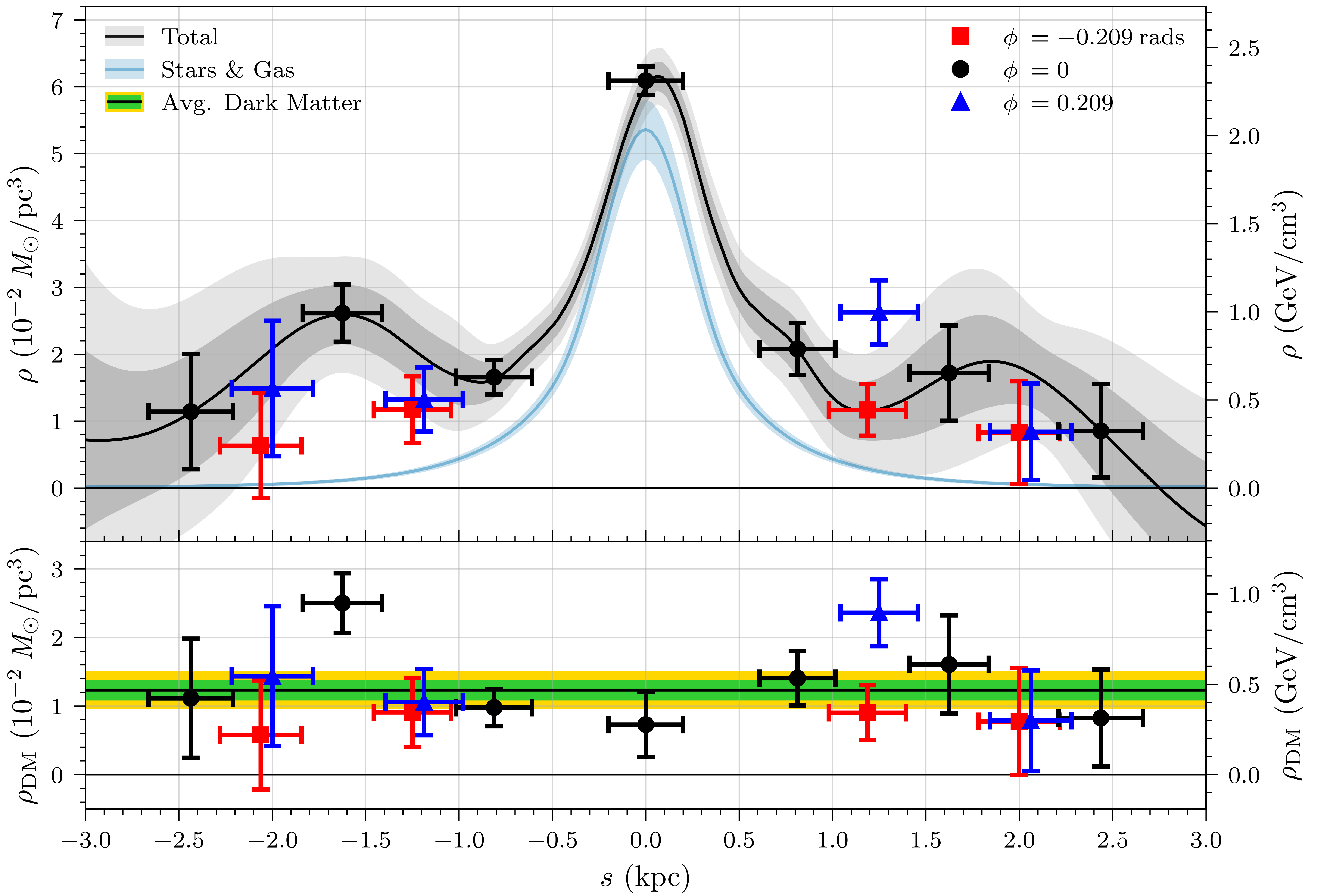
Lim *et al* 2305.13358

Mass Density

- Densities obtained from kernel averaged

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

  - 0.5 kpc width, 0.2 kpc height
  - Nearby points have correlated densities.
- Most computationally intensive part of our analysis. No 3D scans (yet)
- Baryonic profile from McKee et al (2015)



Dark Matter Density

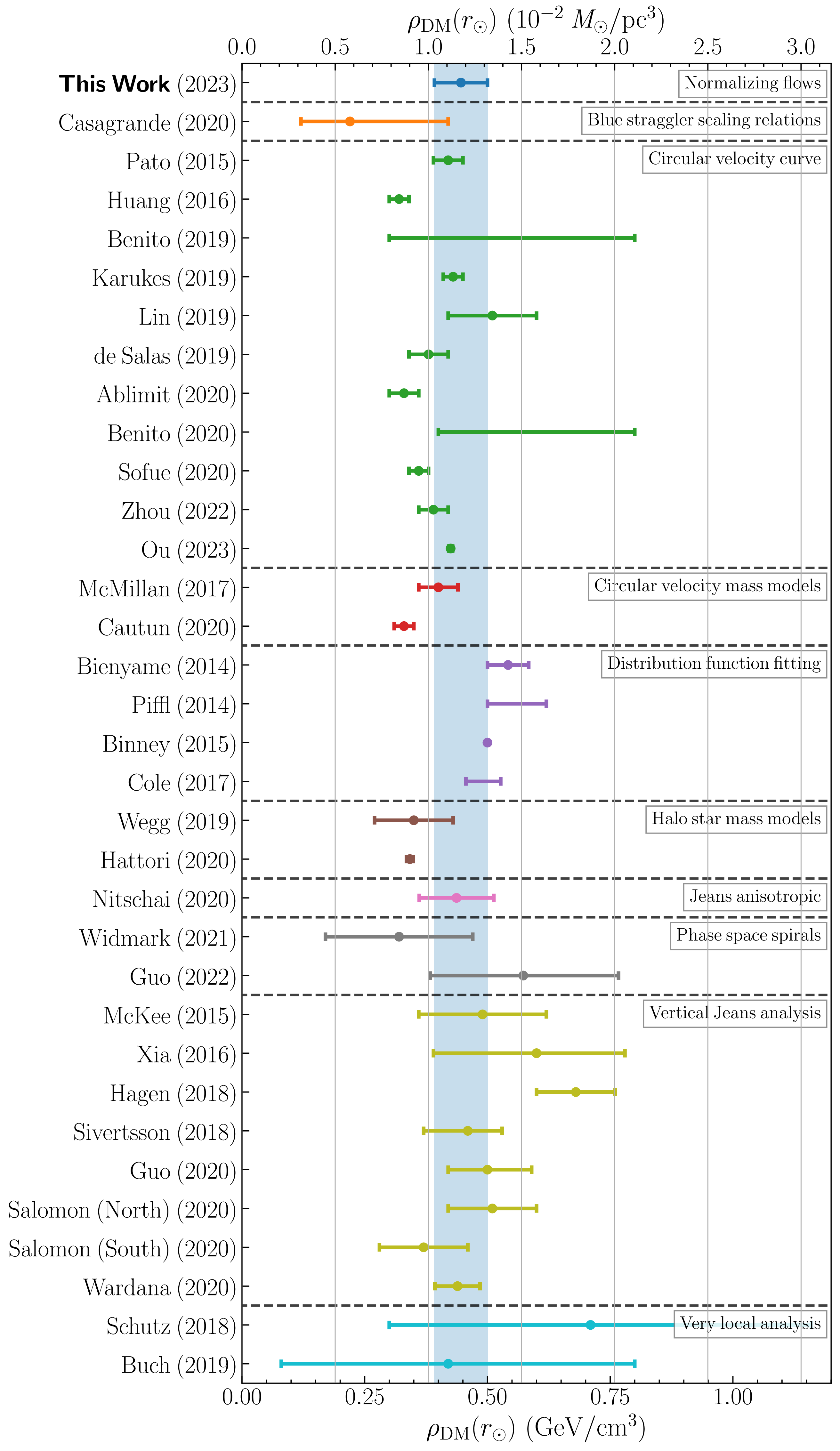
- At the Solar location, we find

$$\rho_{\rm DM} = 0.32 \pm 0.18\ {\rm GeV/cm^3}$$

- Assuming spherical symmetry

$$\rho_{\rm DM}(r = r_\odot) = 0.47 \pm 0.05\ {\rm GeV/cm^3}$$

- Broadly consistent with previous measurements, with competitive realistic errors.
    - Baryonic errors subdominant everywhere off of the disk.

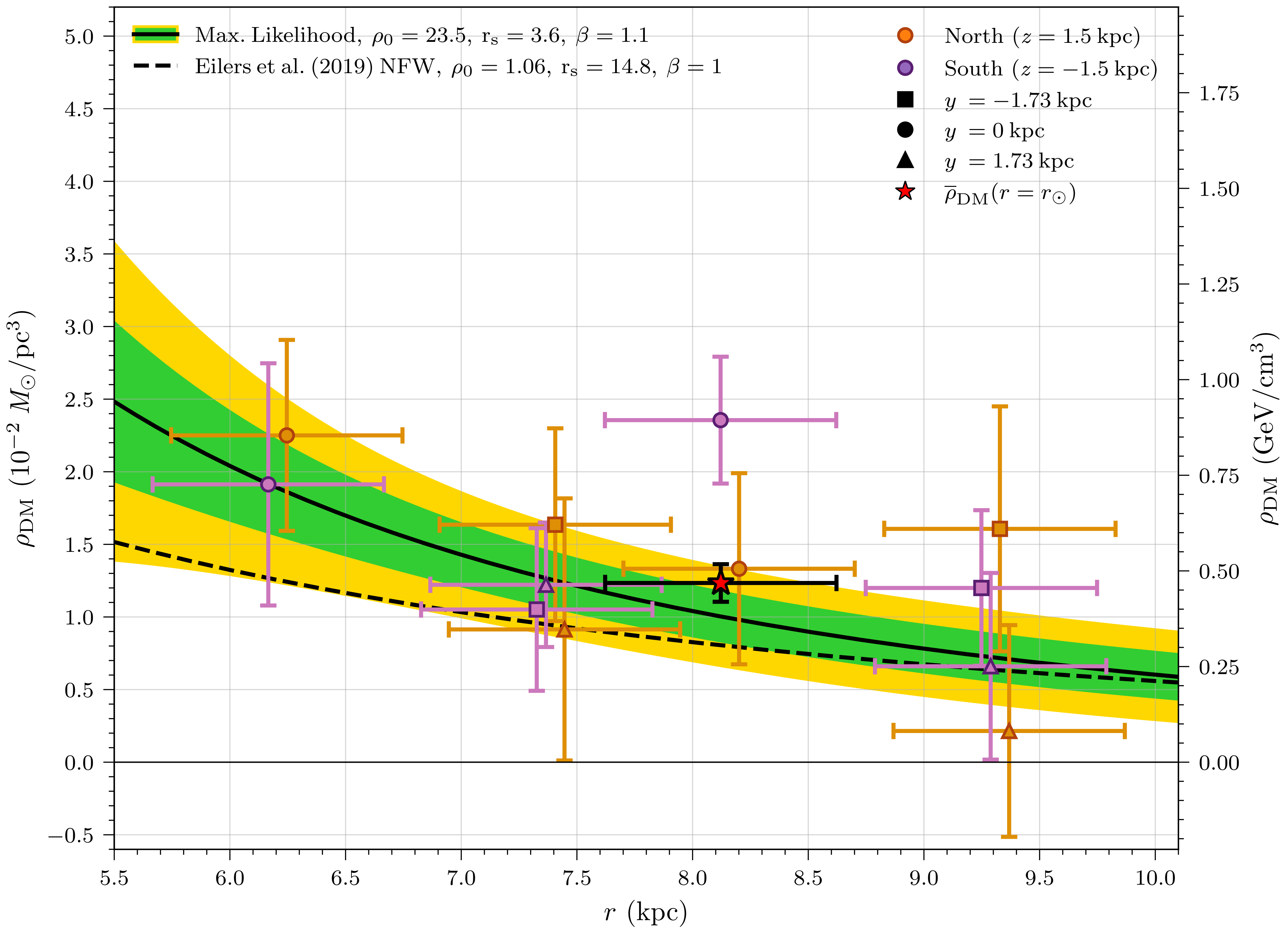


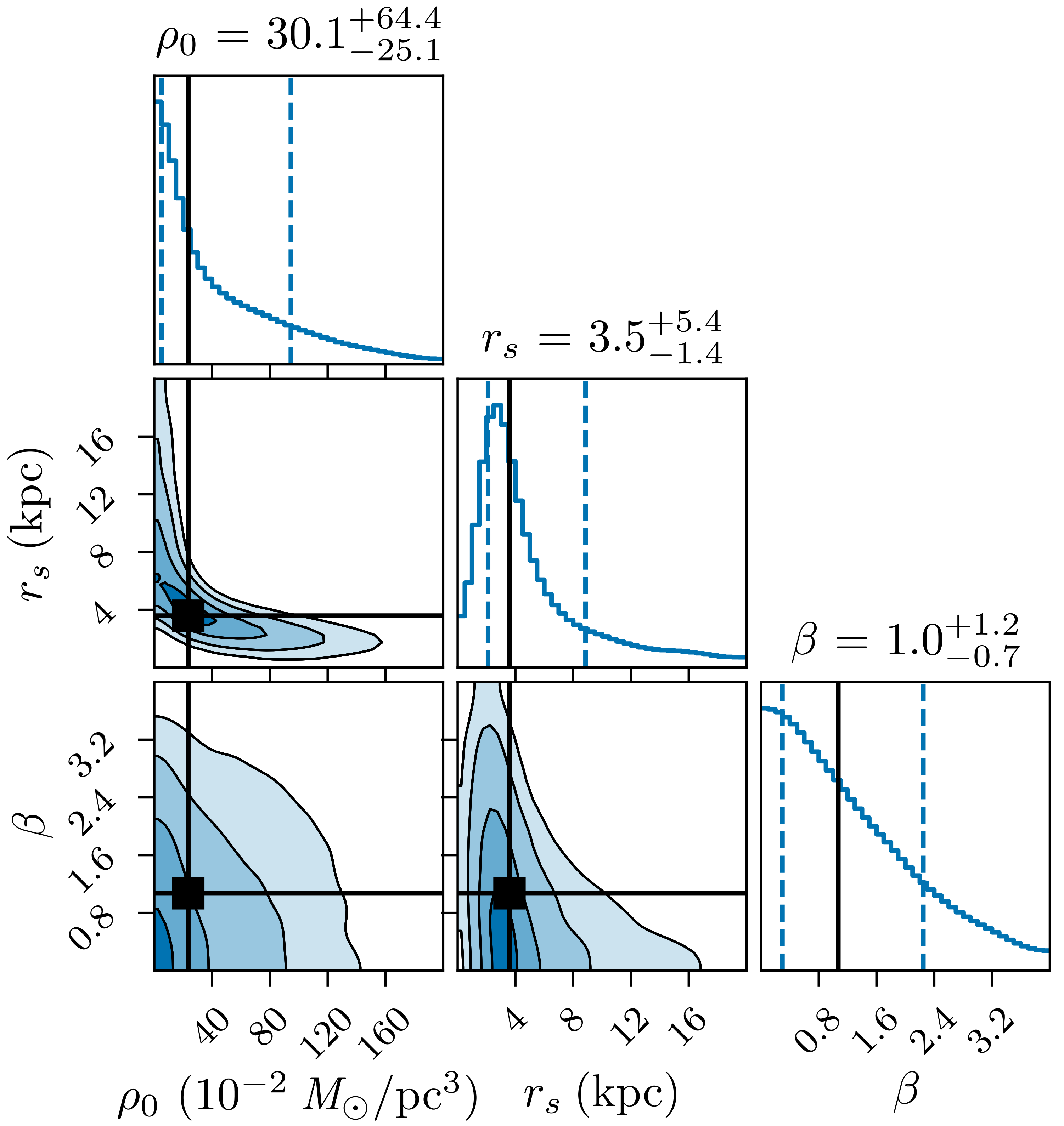
Lim *et al* 2305.13358

Dark Matter Profile

- Fit radial profile to a generalized NFW.
  - Data doesn't extend to the GC. Yet.

$$\rho_{\rm DM}(r) = \frac{\rho_0}{(r/r_s)^\beta (1 + r/r_s)^{3-\beta}}$$



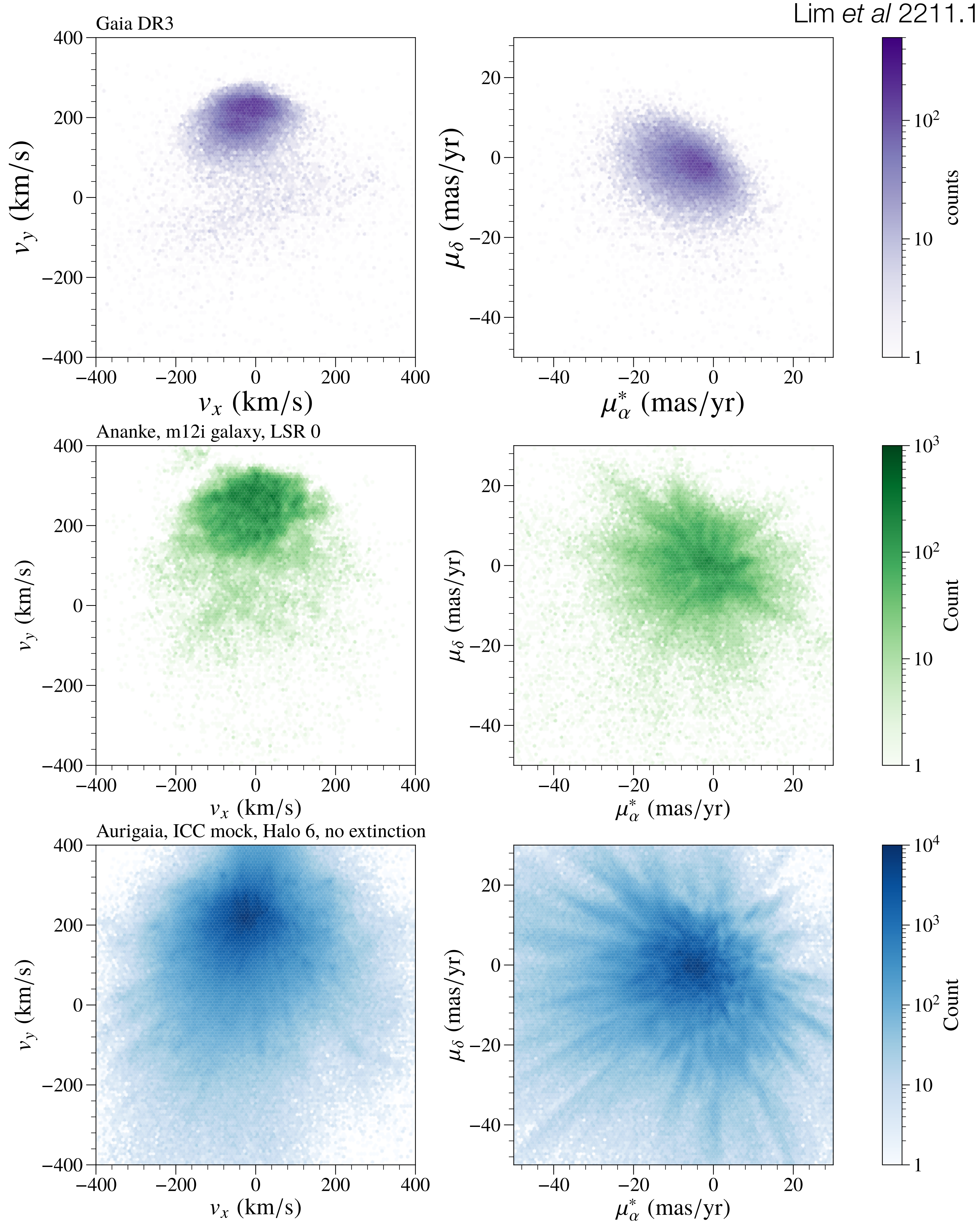


Lim *et al* 2305.13358

Upsampling Simulations

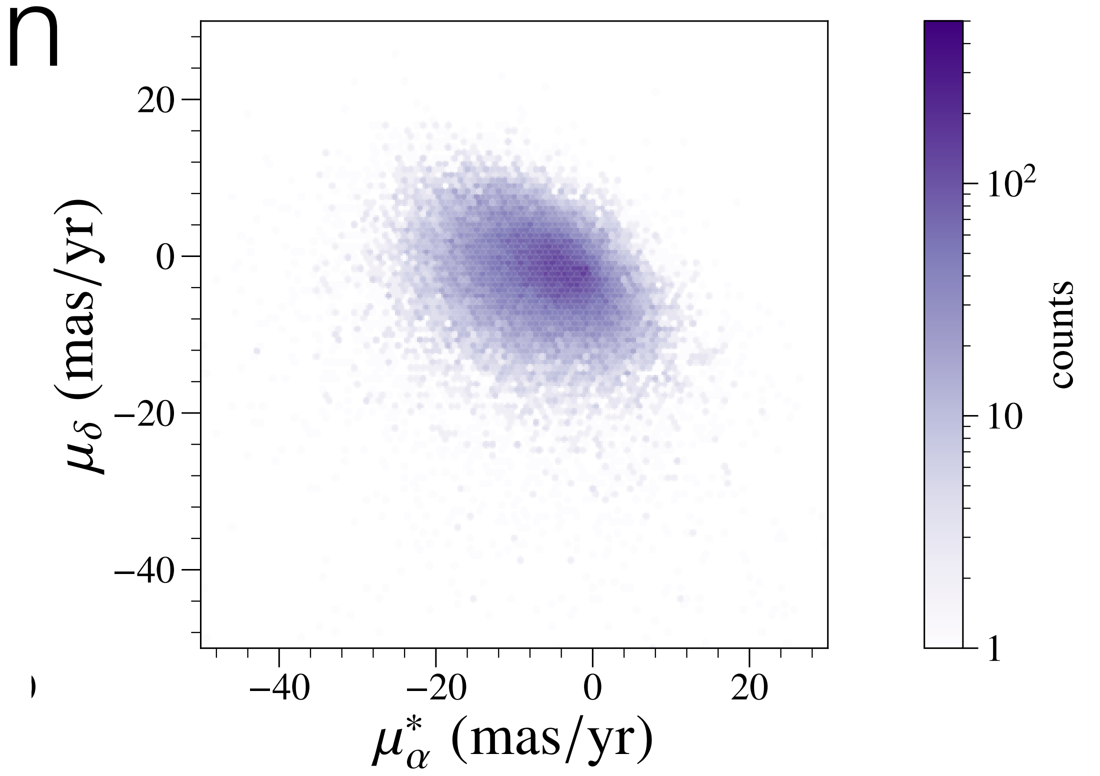
- How to create synthetic galaxy observations? Can either:
  - Create by-hand analytic smooth models of the Galaxy or,
  - Use *N*-body hydrodynamical simulations
- But in the latter case, there complications:
  - Every galaxy is unique.
  - Simulations work on the level of tens of millions of "star particles," not hundreds of billions of *stars*.
- Upsampling required!
  - But existing upsamplers are "clumpy"

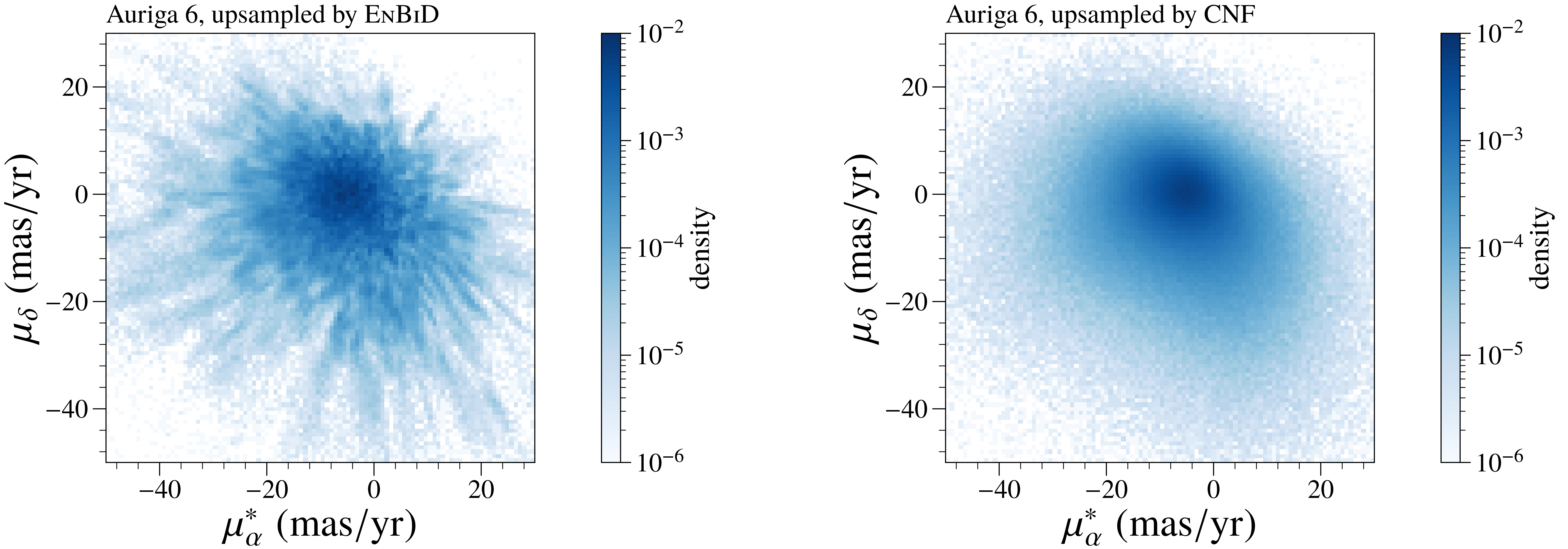
Lim *et al* 2211.11765



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

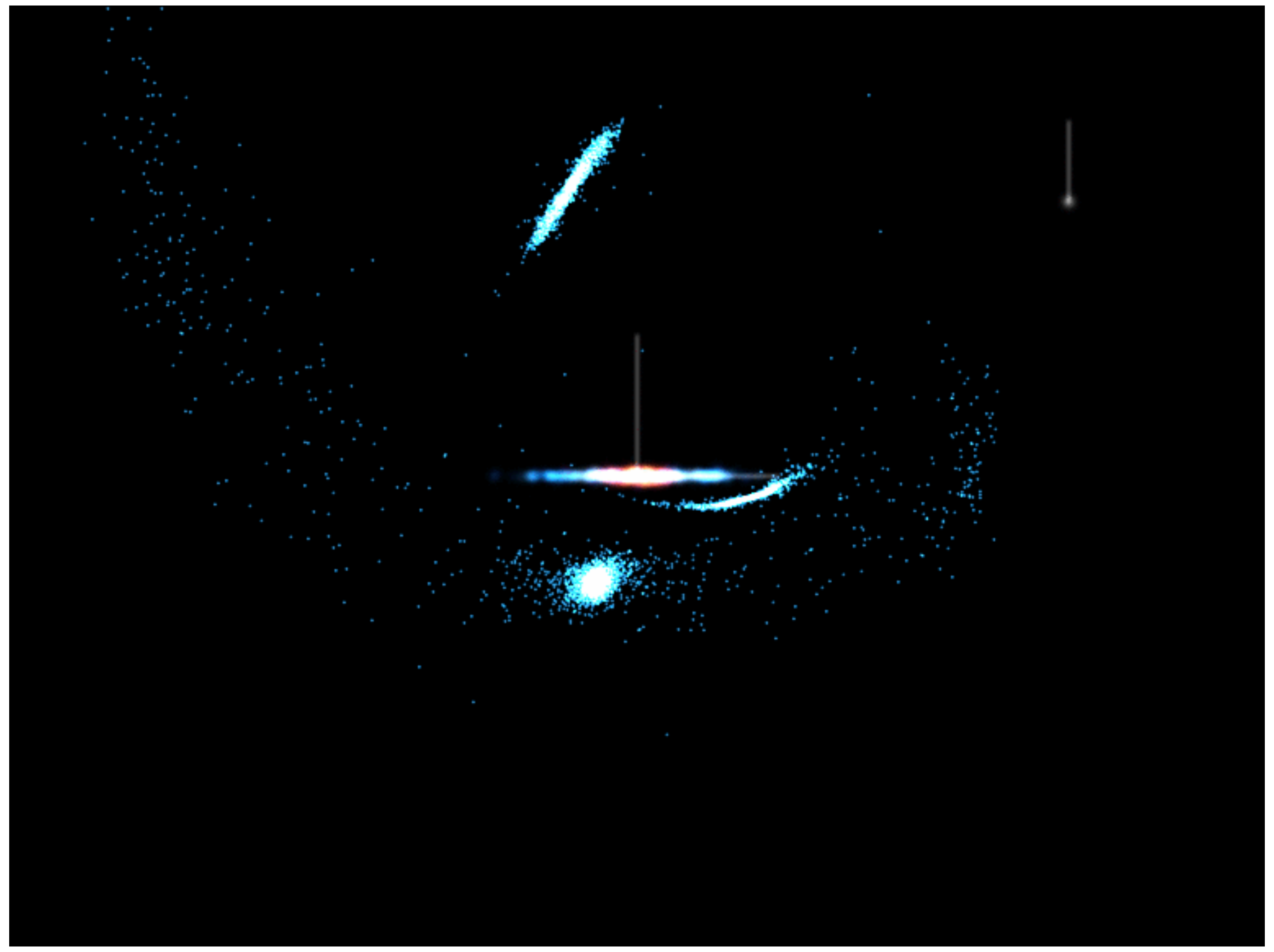


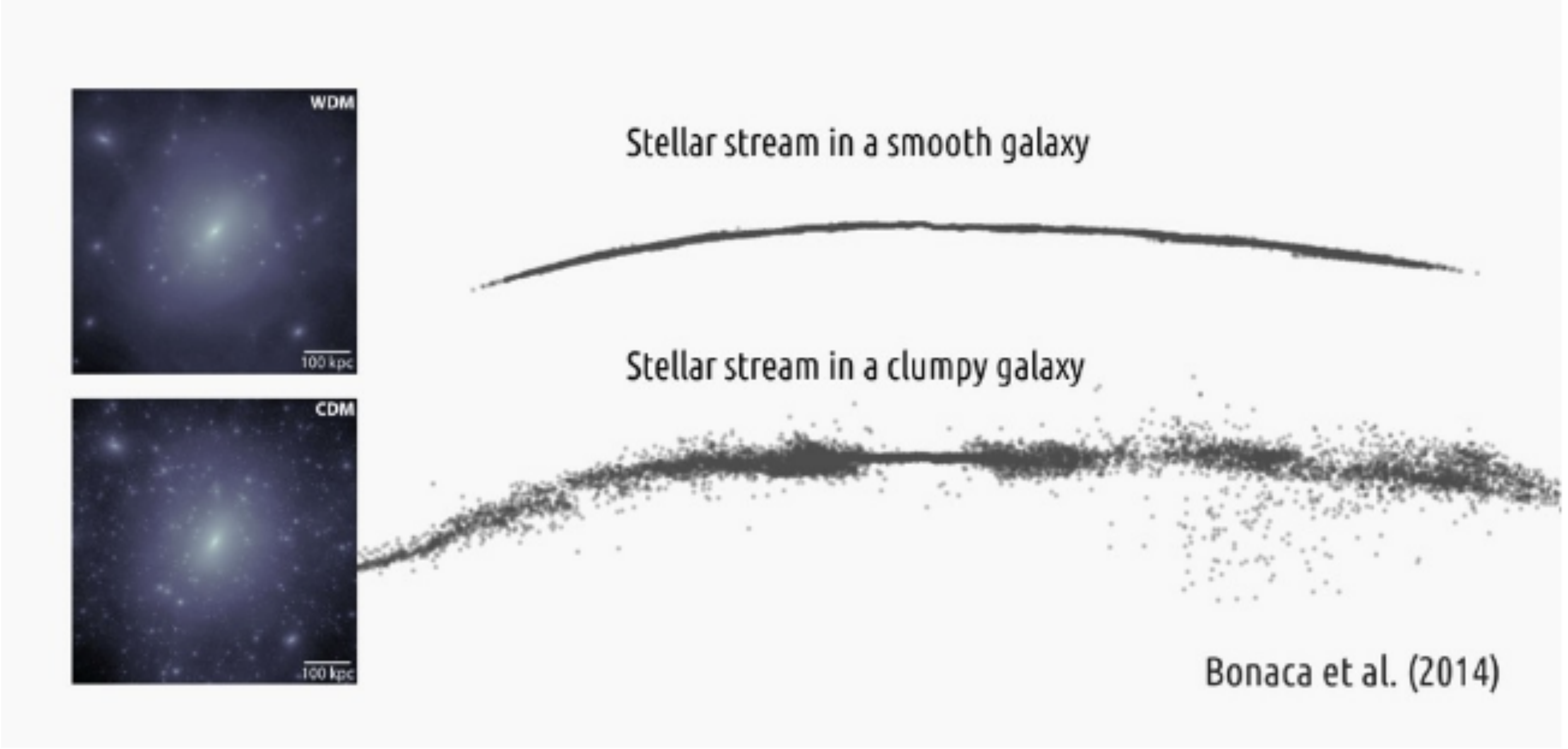


Lim *et al* 2211.11765

Searching for Substructure

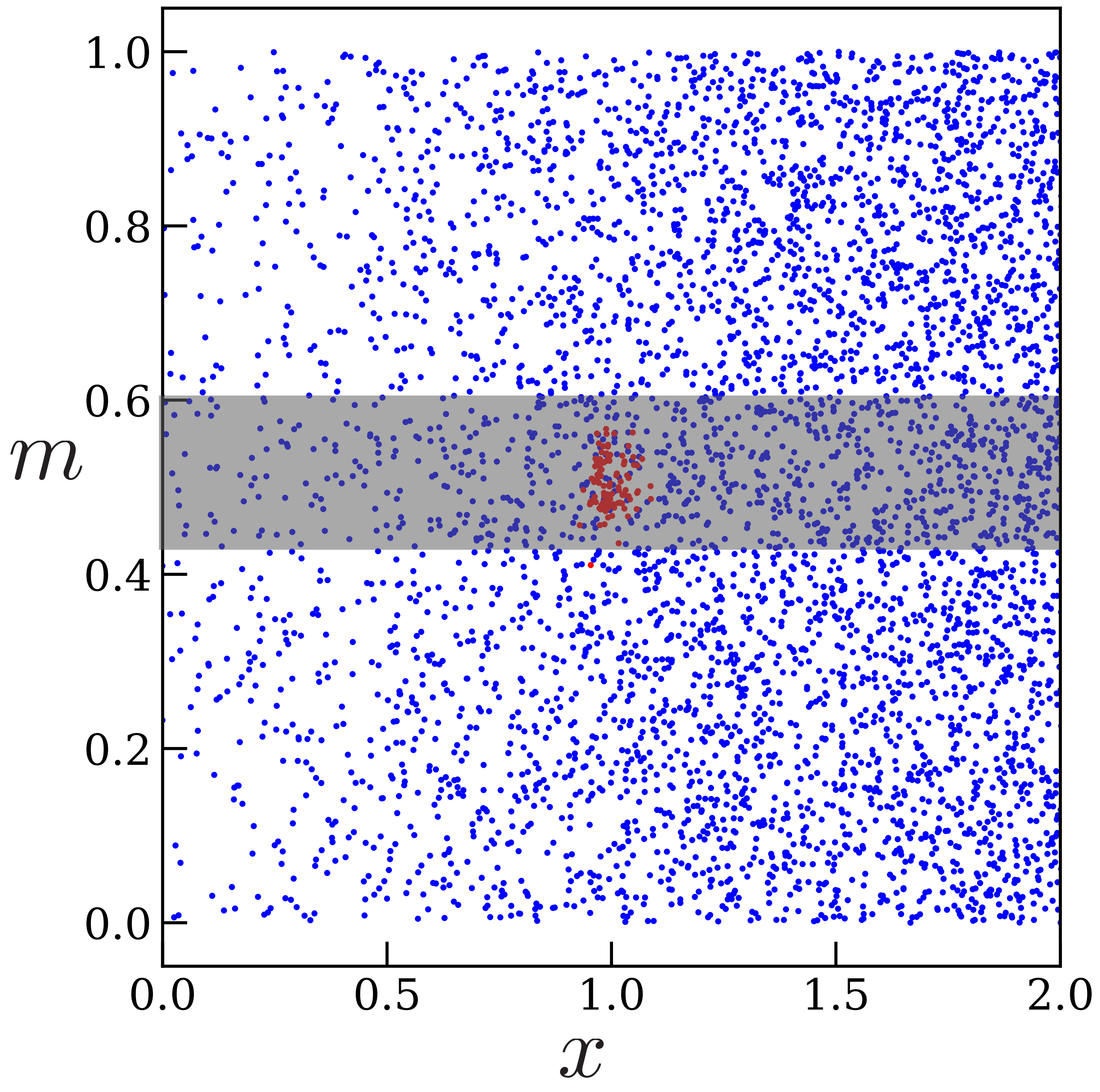
- Stellar substructure (streams, tidal debris, etc) provide both a merger history, probes for the overall Galactic potential, *and* probes of dark substructure.
- My original motivation: 1) find a bunch of stellar streams without assuming a potential, and 2) use them to constrain the dark matter.
    - Got *slightly* distracted by step 1)





Searching for Substructure

- 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.
- Use ANODE (Nachman & Shih 2001.04990) and CATHODE (Hallin *et al* 2109.00546), LHC-developed *anomaly detectors*.
- Learn phase space density in two ways: directly in a *search region*, and interpolated from a *control region*. Wherever the ratio is large, you have an anomaly

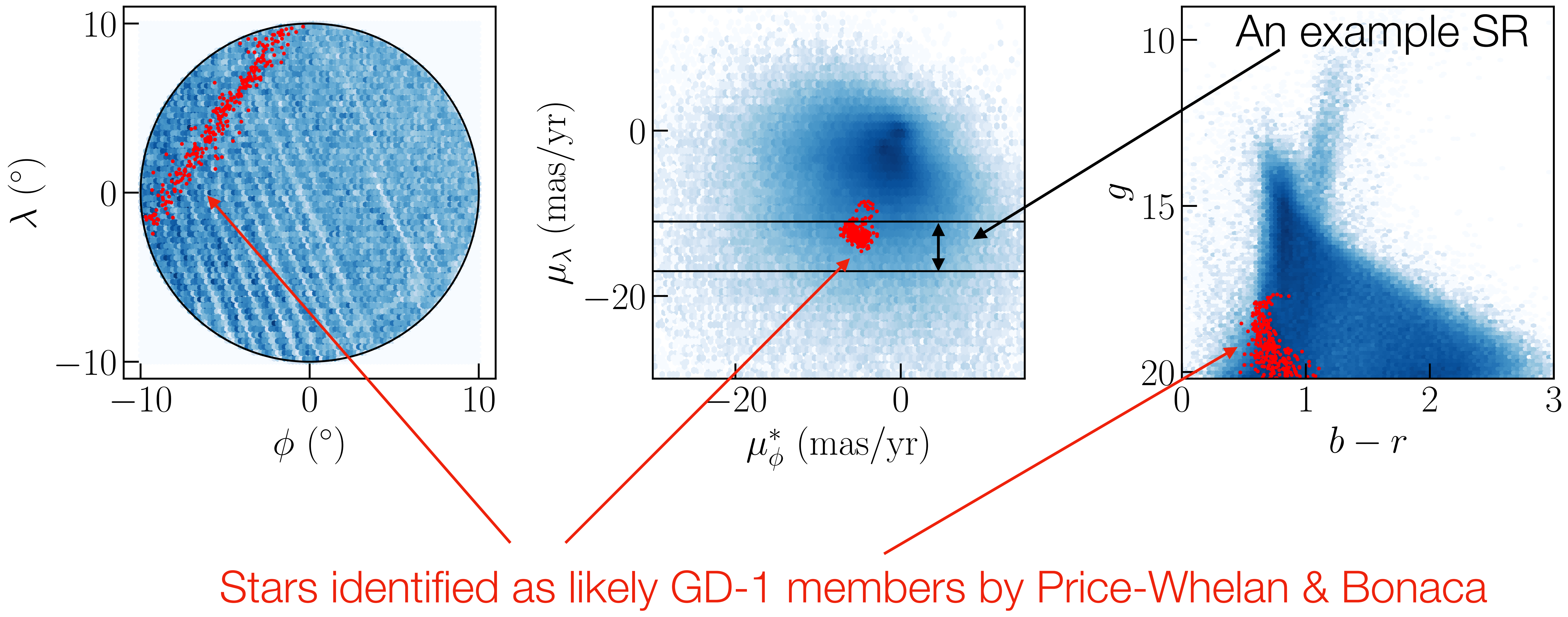


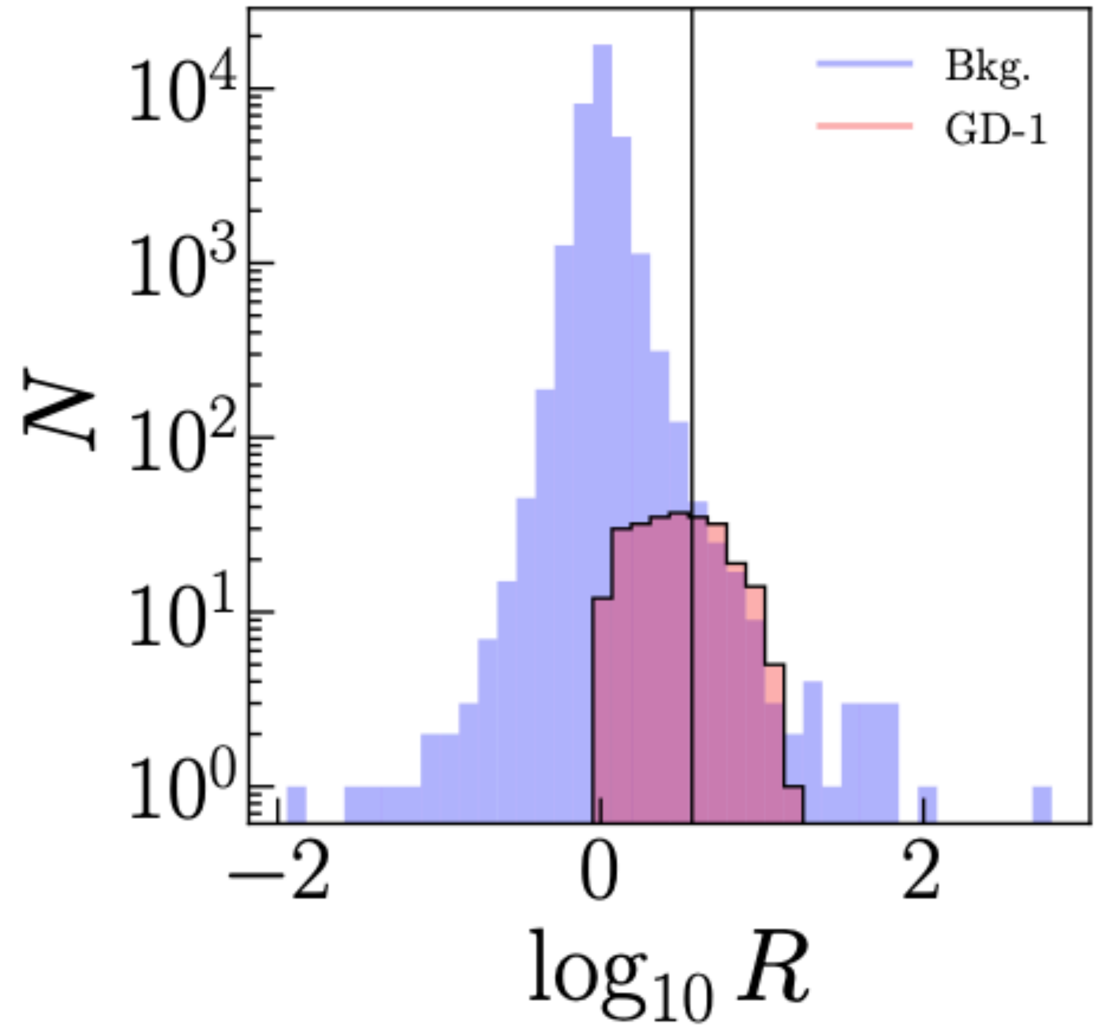
$$R(\vec{x}|m \in \mathrm{SR}) = \frac{P(\vec{x}|m \in \mathrm{SR})}{P_{\mathrm{CR}}(\vec{x}|m \in \mathrm{SR})}$$

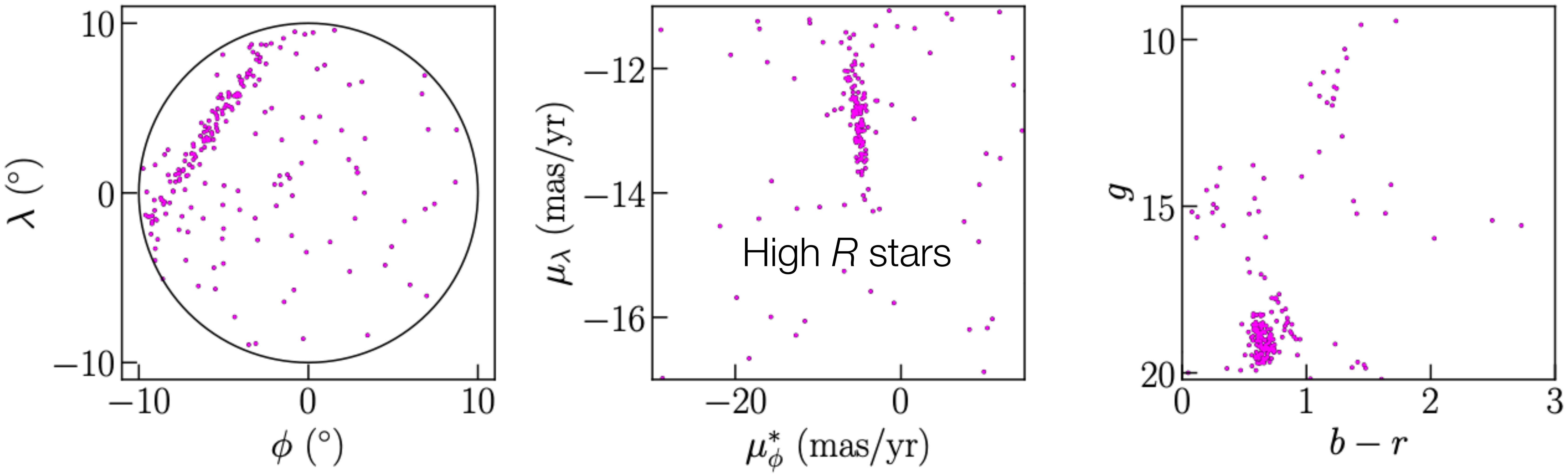
Shih *et al* (2104.12789)

Searching for Substructure

- First testing on well-known and distinct GD-1 stream.
  - Have some approximation of “truth-level" stream membership.

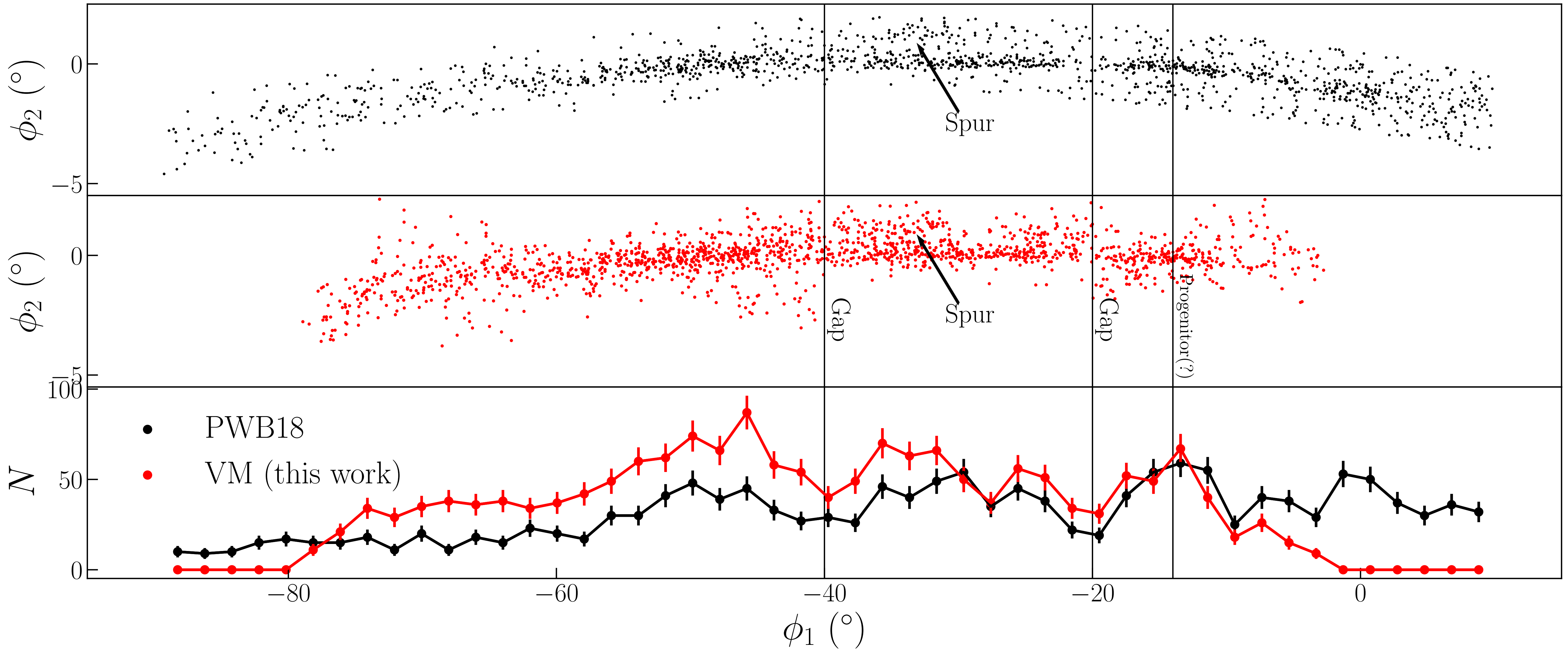




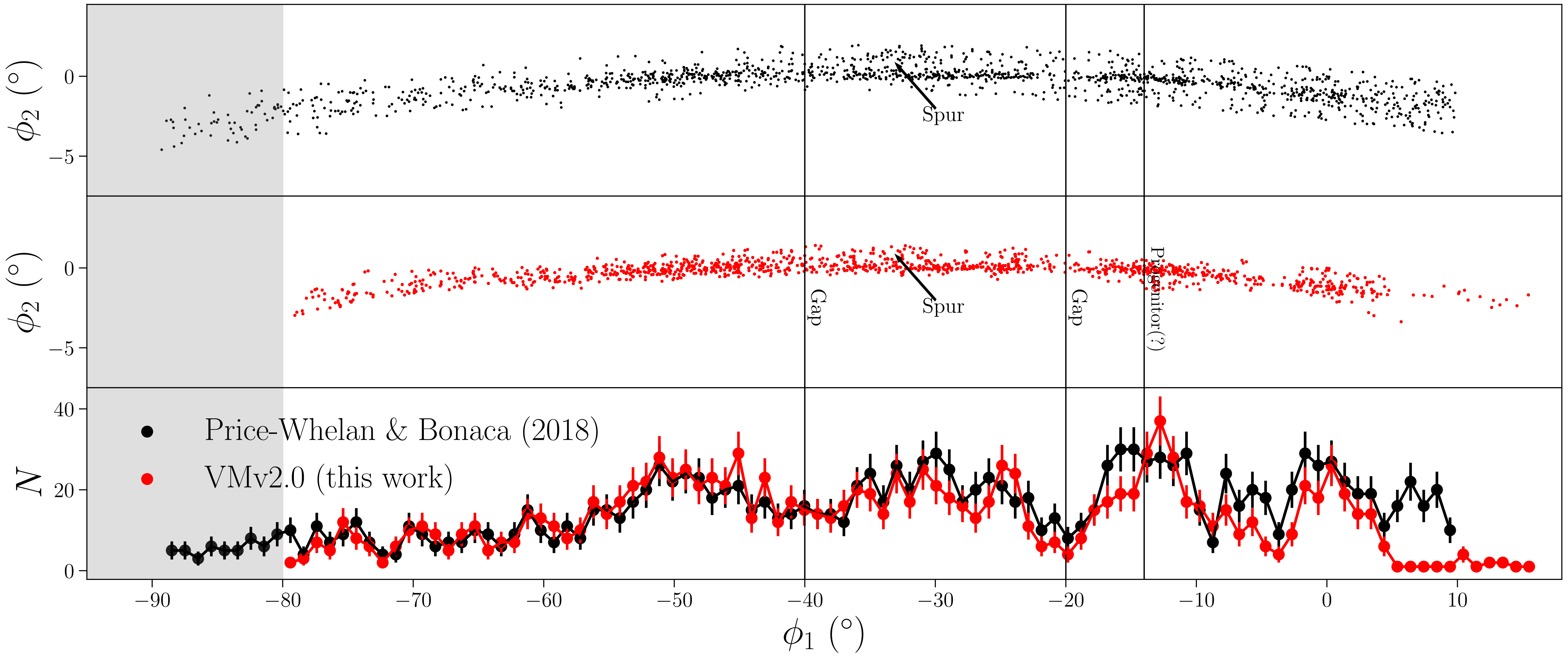


Shih *et al* (2104.12789)

Searching for Substructure



Shih, Buckley, Necib, and Tamanas 2104.12789

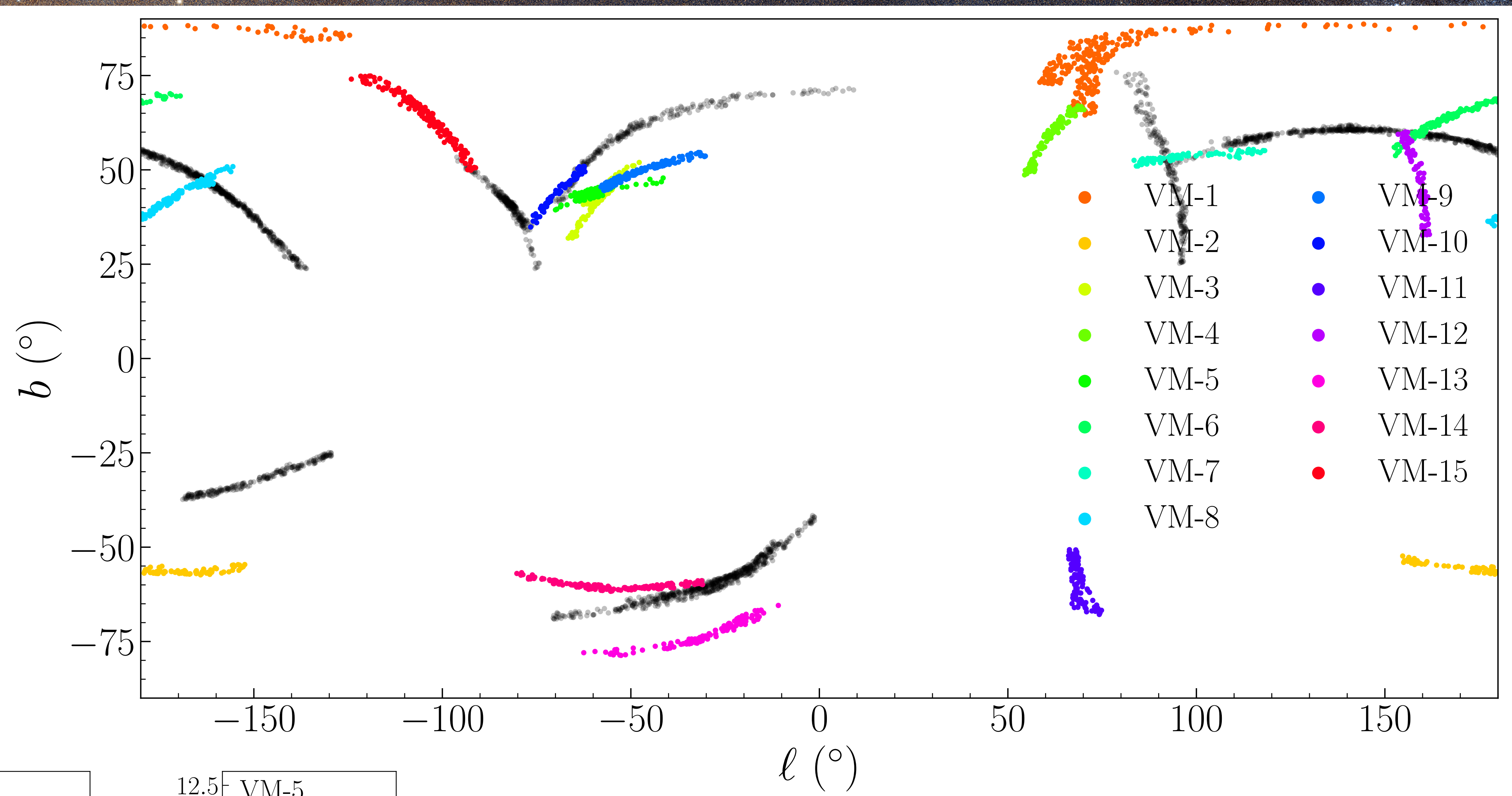


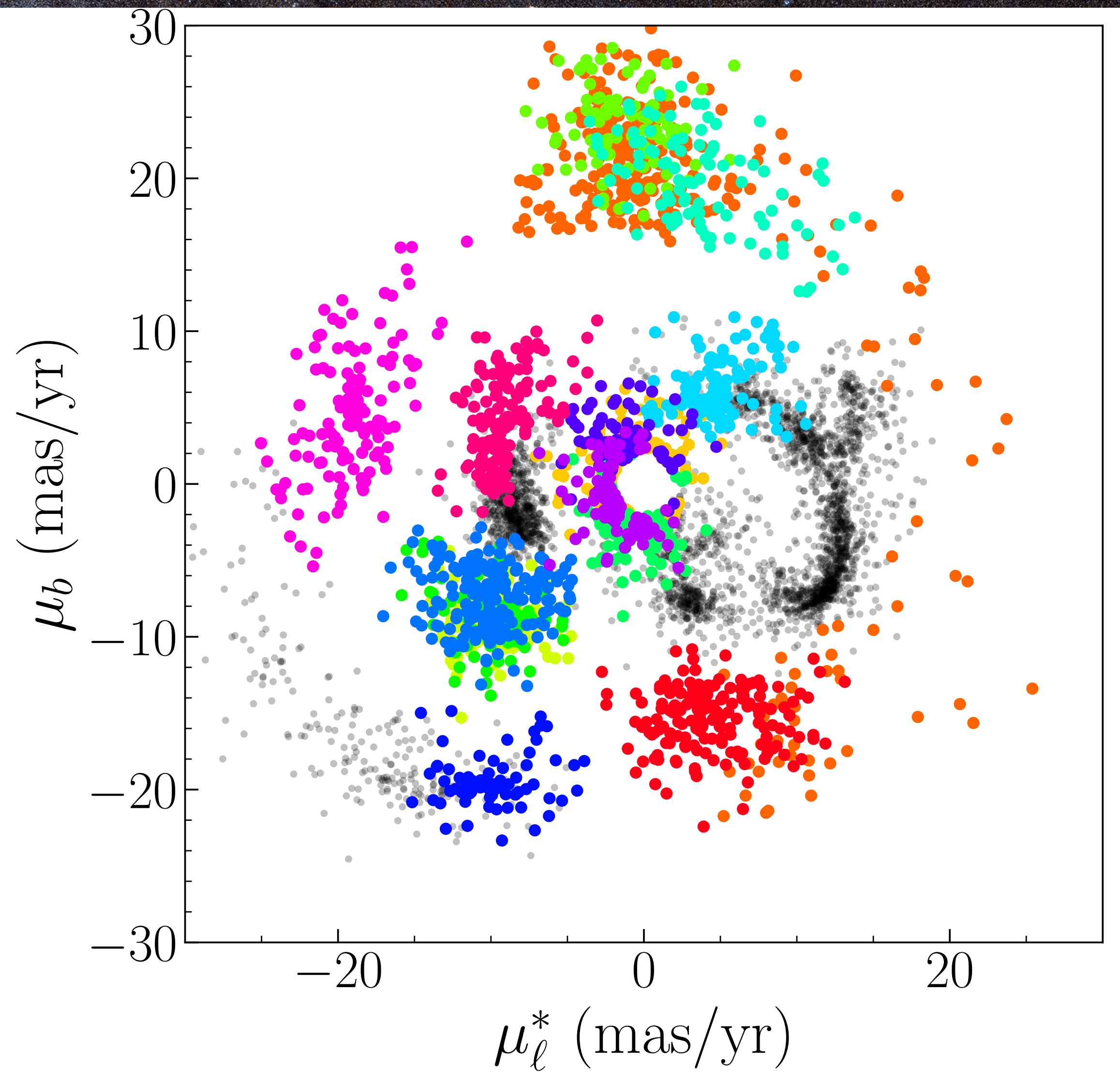
Shih, Buckley, and Necib 2303.01529

Shih et al (2104.12789)

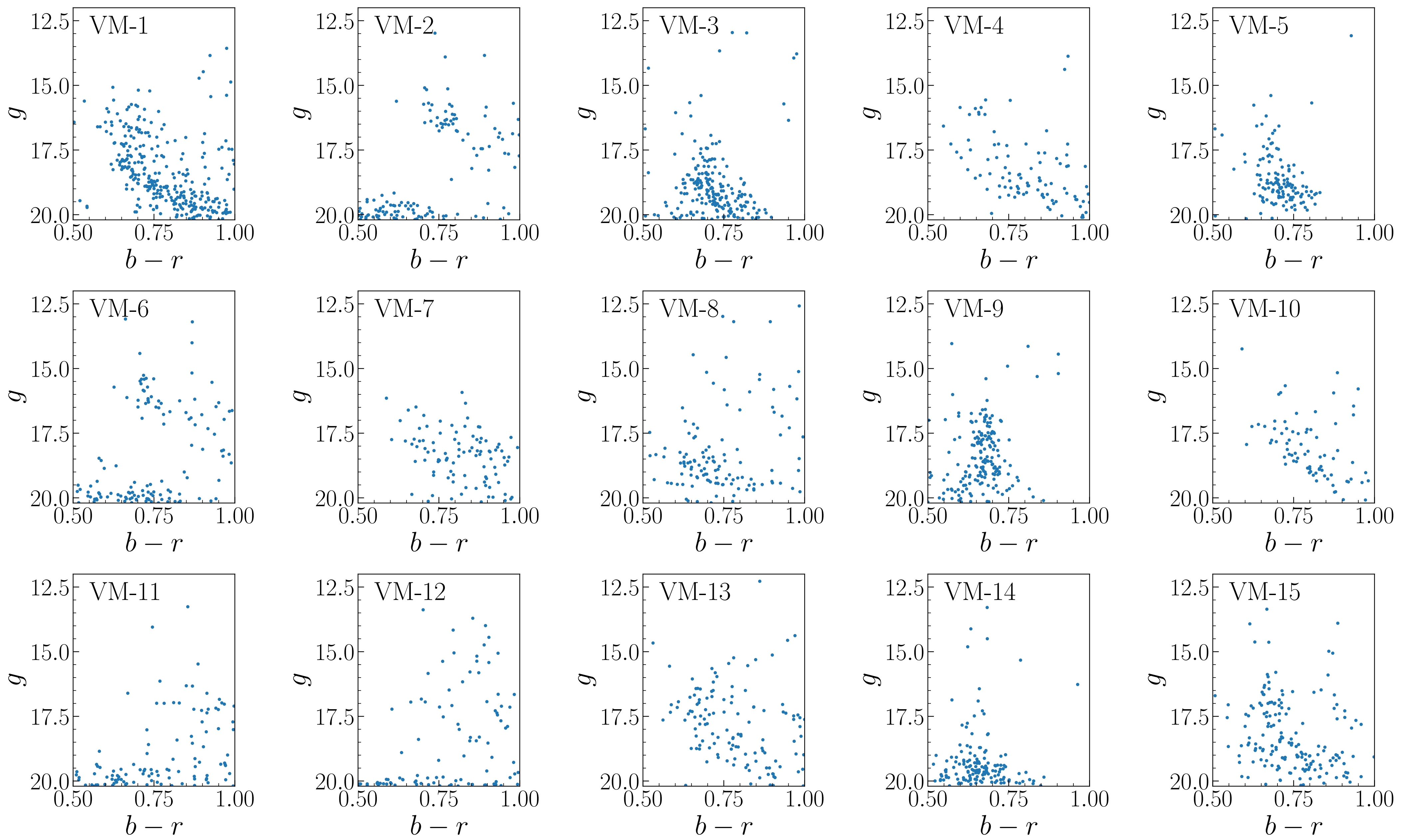
Searching for Substructure

- We identify 82 stream candidates in *Gaia* DR2 expect a false-positive rate of ~10%.
  - Here are the top 15.
  - Redoing this for DR3 now.





Shih *et al* (2104.12789)



- In addition to these new streams, we can now also characterize their phase-space densities, a new handle on their evolution.
  - We can now start returning to the original goal: use these streams as probes of the dark matter.

Conclusions

- Theorists can create vast array of viable particle physics models for dark matter compatible with existing data.
- We need more data to determine which is correct. And we (theorists) need to understand what our models would predict in the existing data.
- Machine learning gives new ways to understand large, high-dimensional datasets.
  - And new ways to connect simulations to observations
  - But raises new questions about model predictions
- As a theorist, my goal is to have a clear anomaly in the data, so that I can go back to playing model-building games.