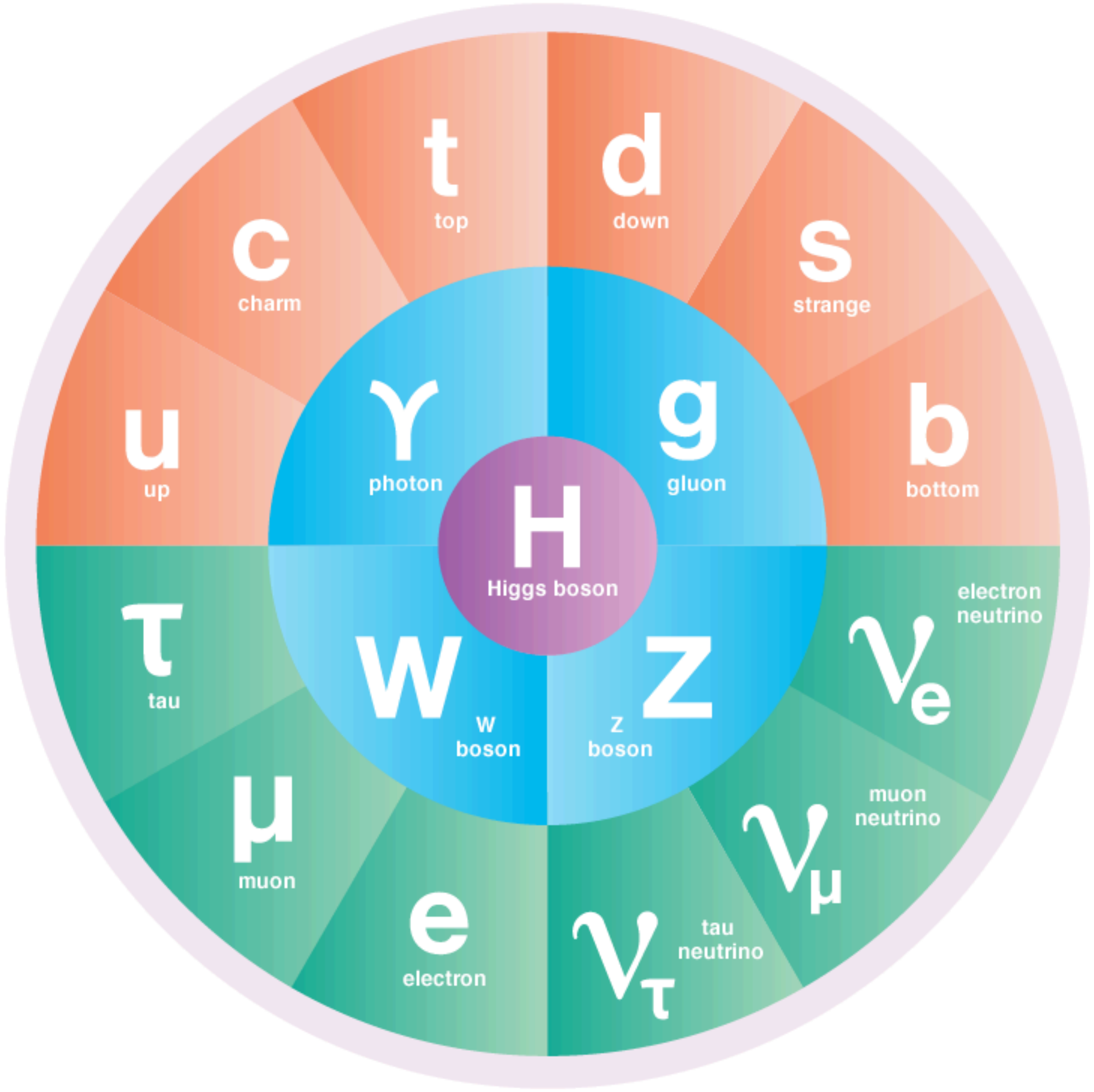
ML for Fundamental Physics: From the Smallest to the Largest Scales

IAS Geneva 2023

David Shih
October 4, 2023

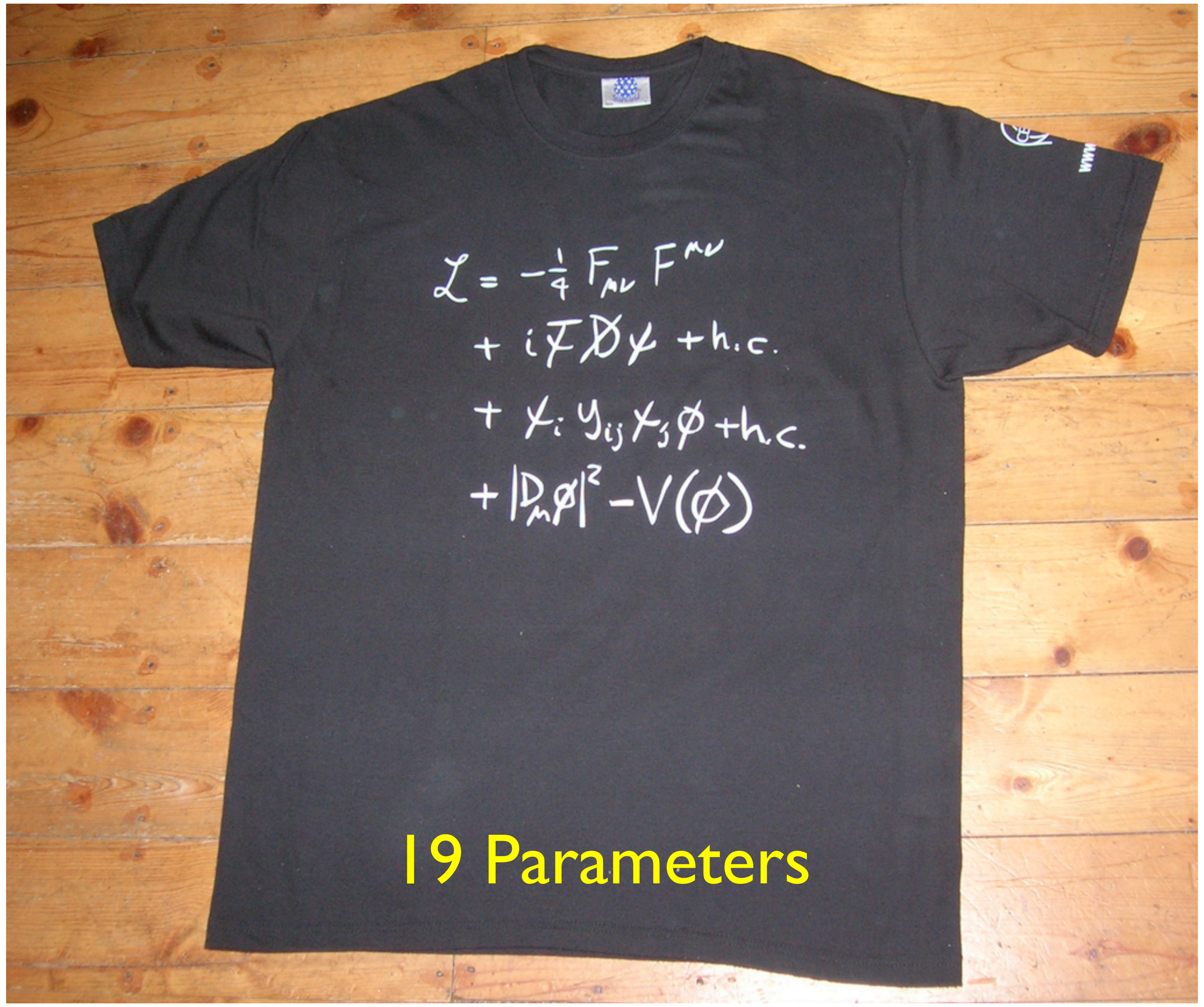
RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

The Standard Model of Particle Physics



3 Forces

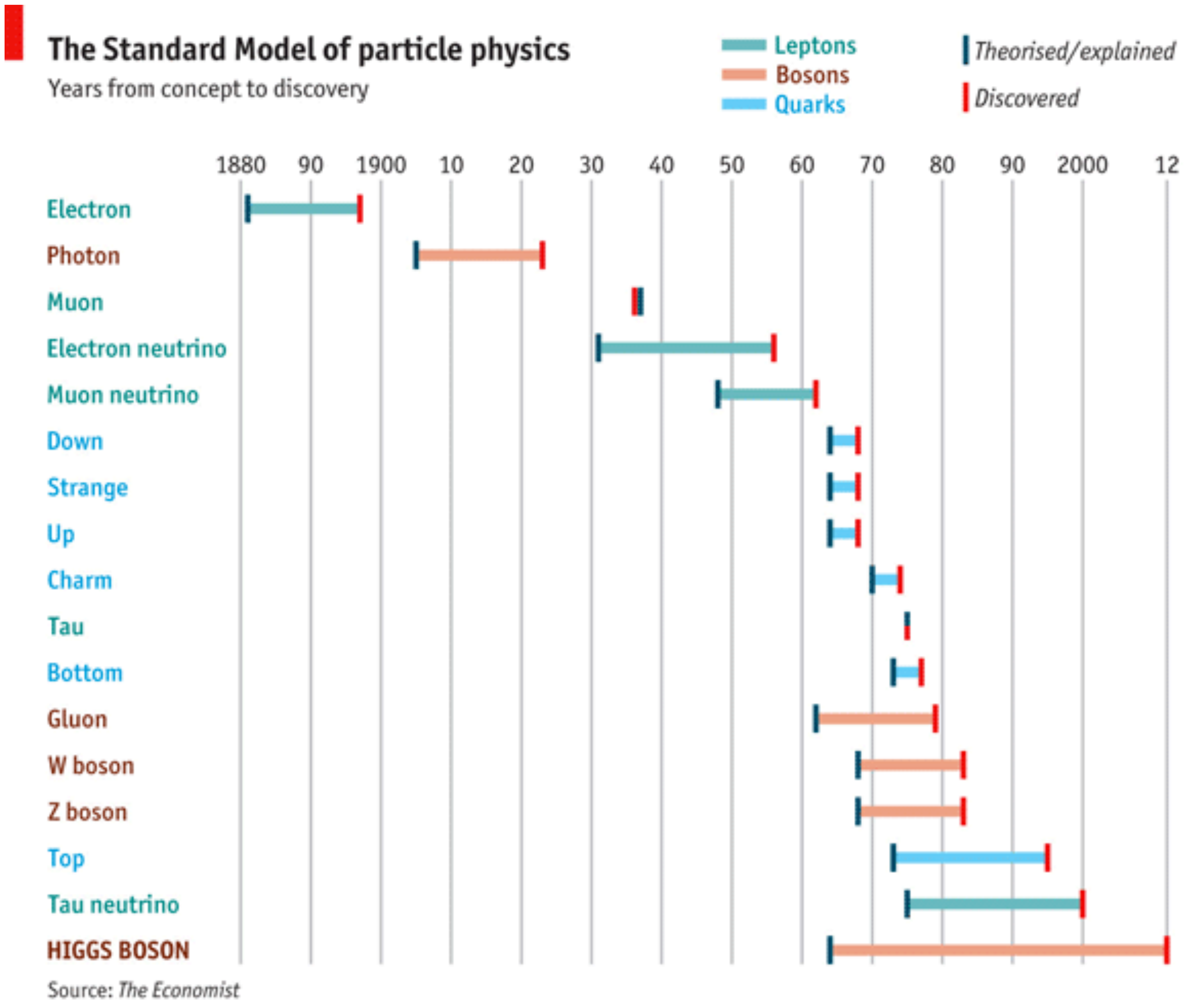
17 Particles



19 Parameters

The Standard Model describes all known fundamental matter and its interactions in the Universe.

The Standard Model of Particle Physics



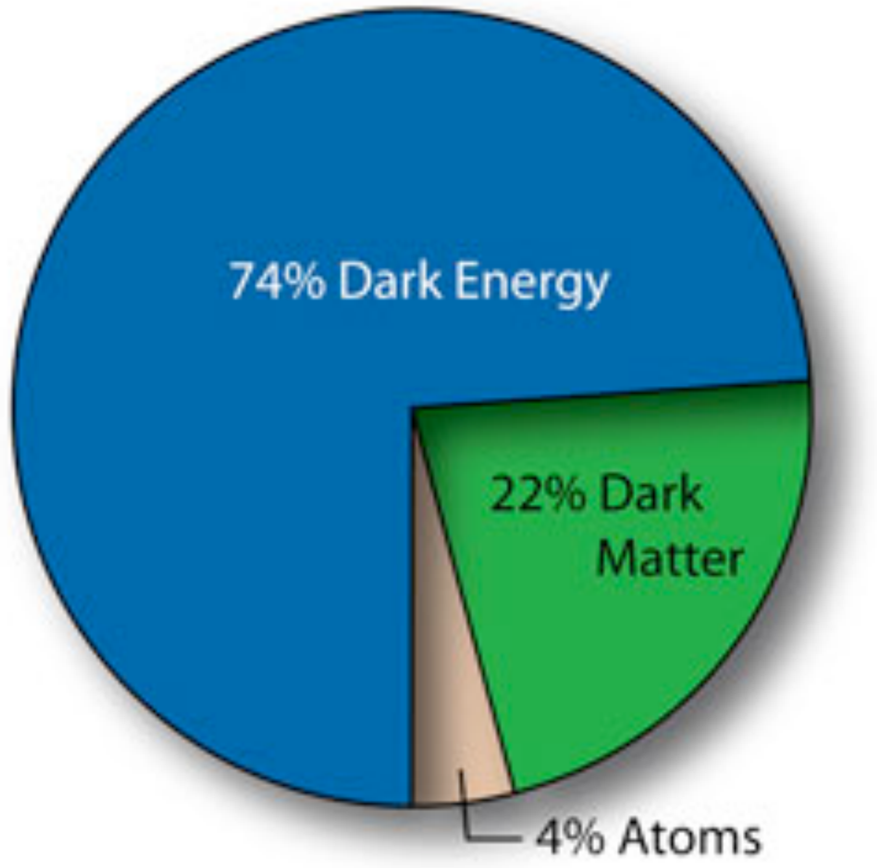
The Standard Model was largely established in the ‘60s, ’70s and ‘80s.

With the discovery of the Higgs boson by the LHC in 2012, it is finally complete.

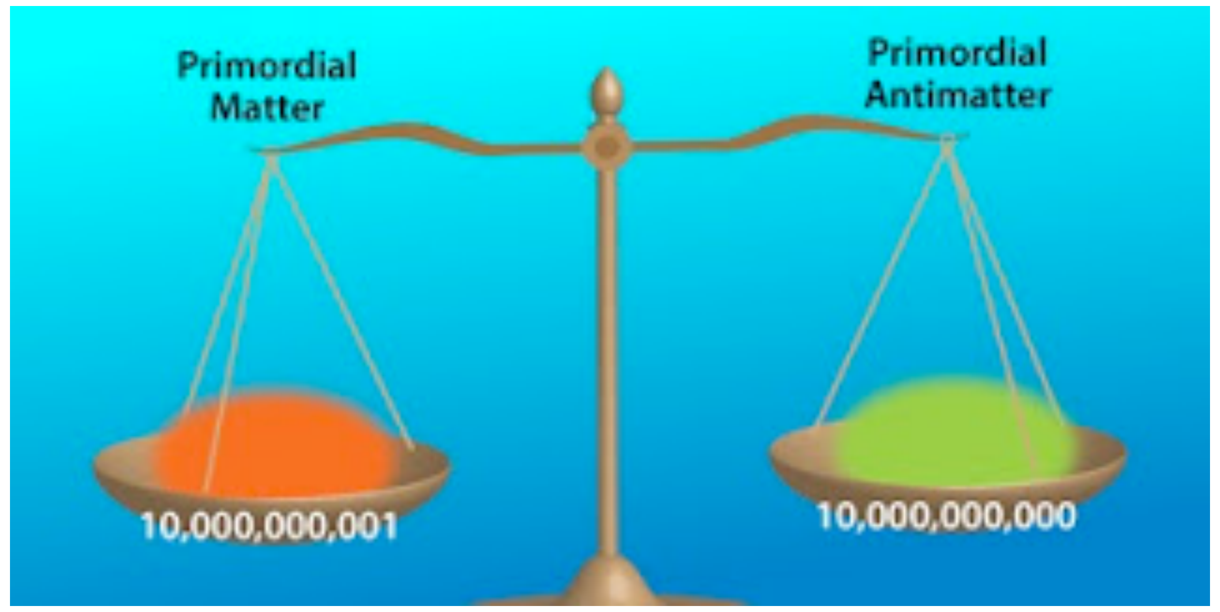
What “new physics” lies beyond the Standard Model?

Beyond the Standard Model

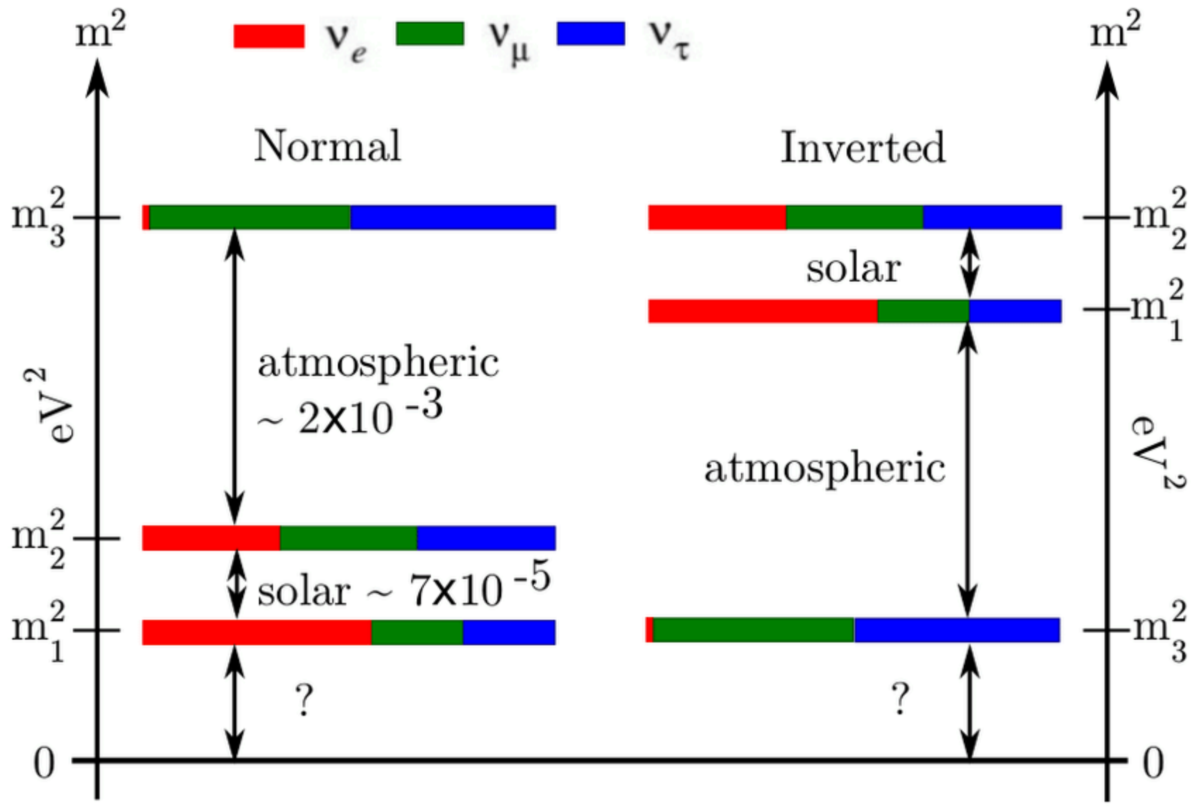
dark matter



matter/anti-matter asymmetry



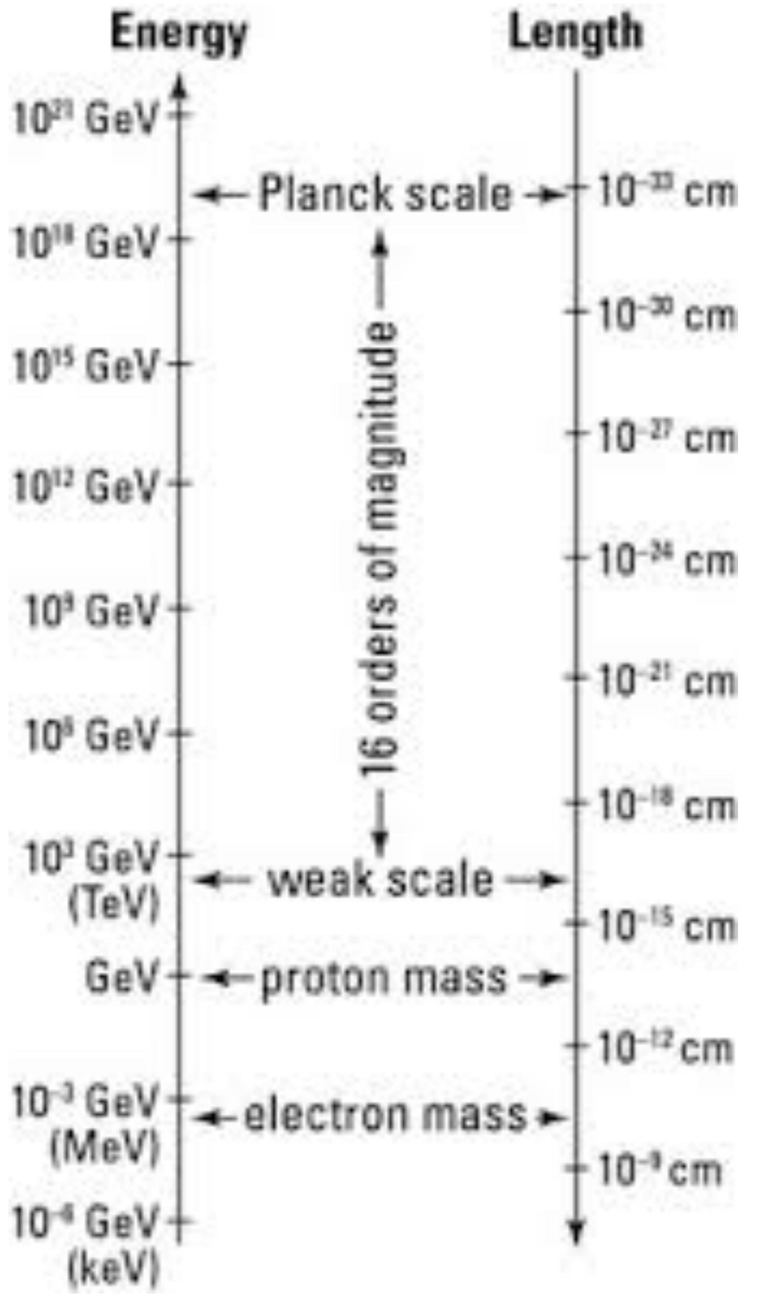
neutrino masses



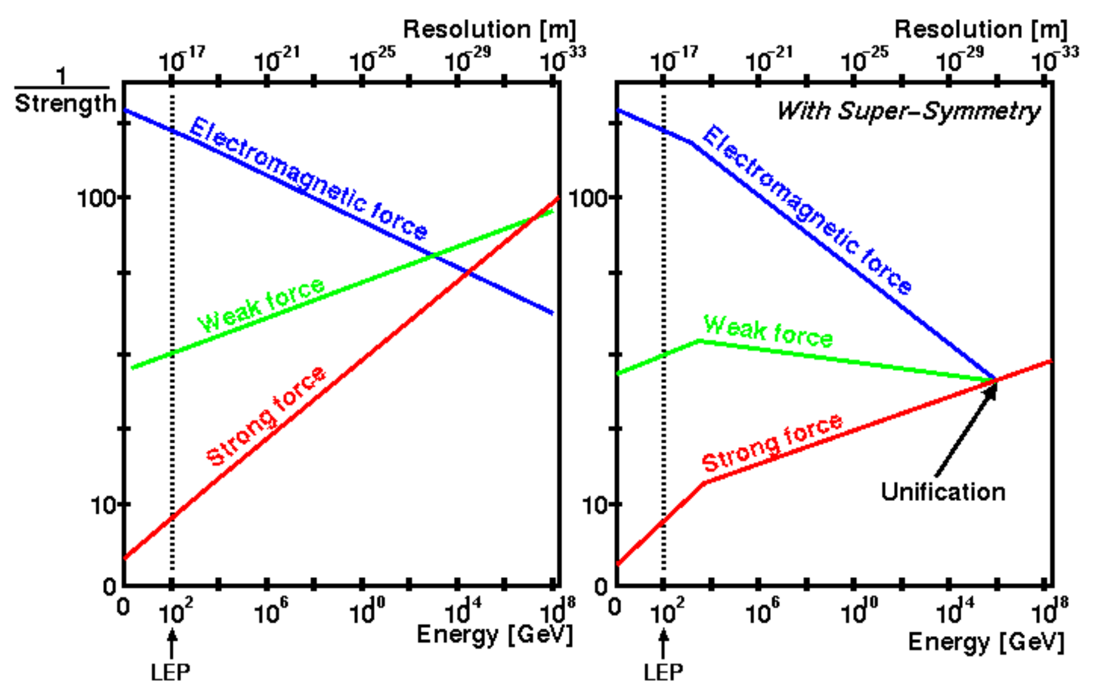
We know there must be “new physics” beyond the Standard Model…

Beyond the Standard Model

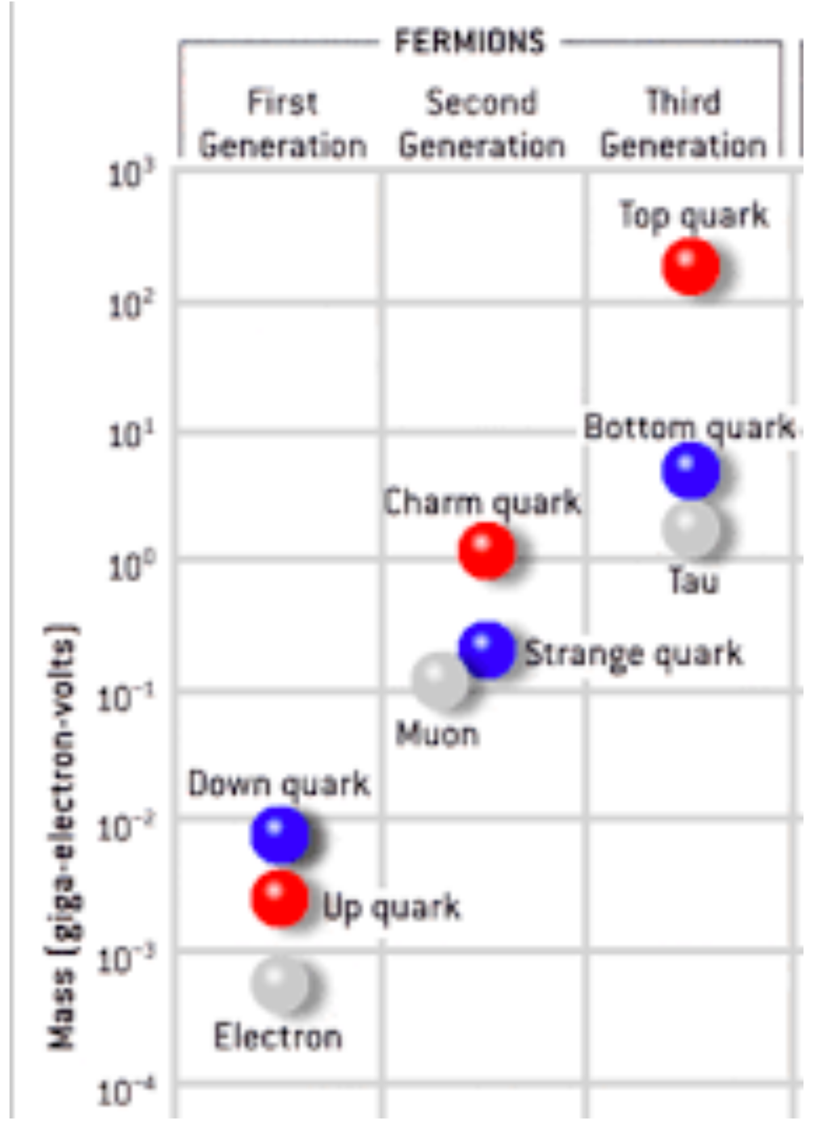
hierarchy problem



grand unification



flavor puzzle



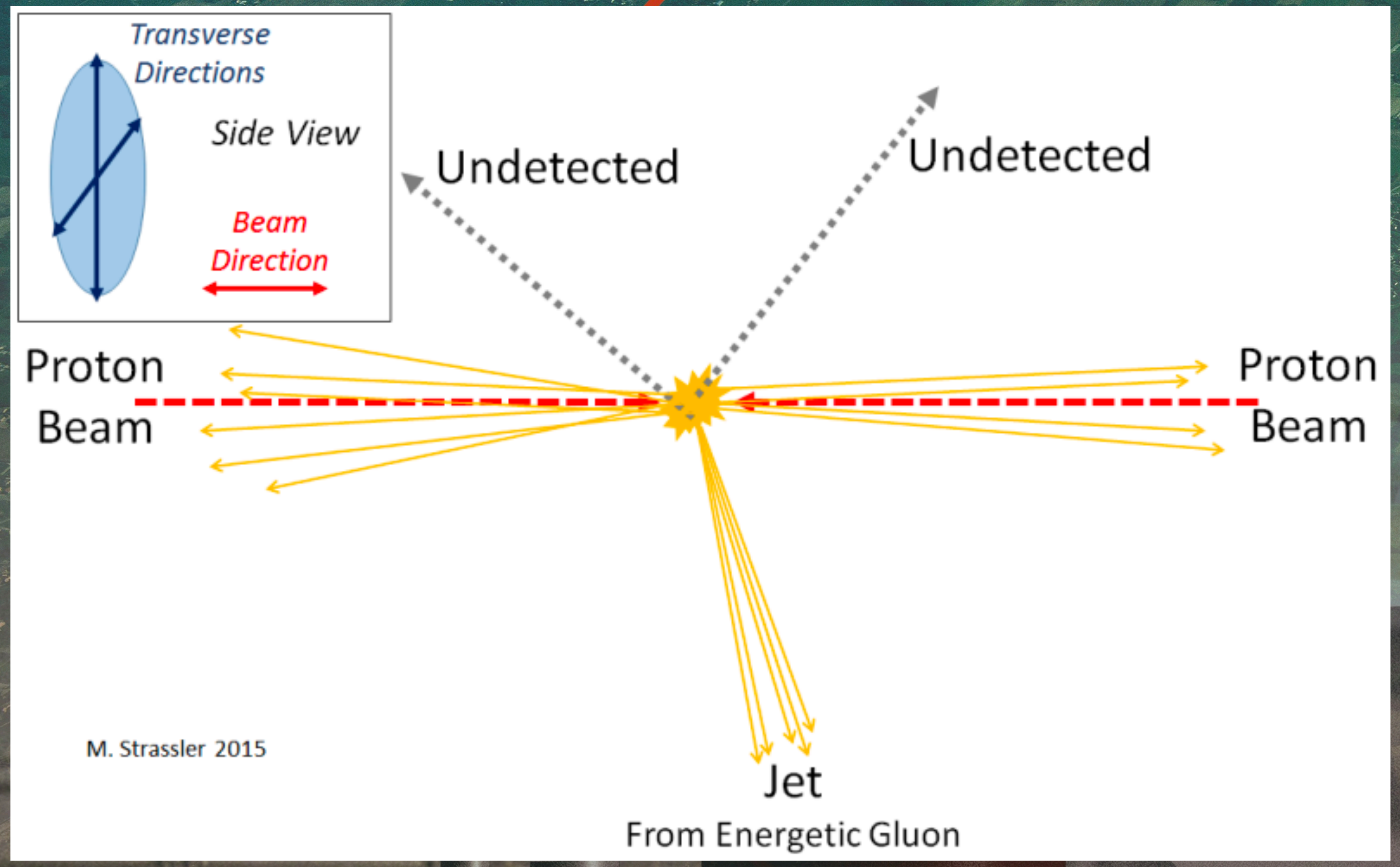
strong CP problem

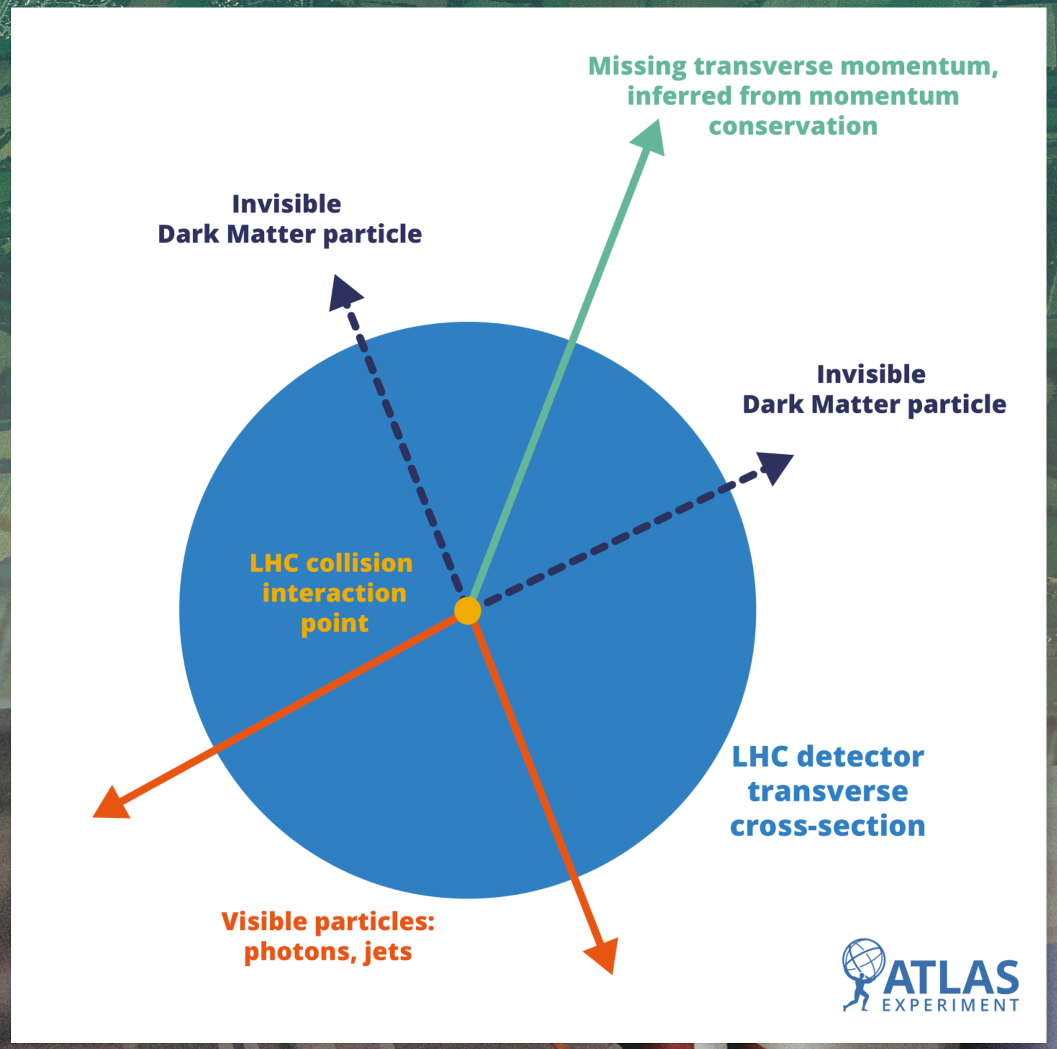
$$\mathcal{L} \supset \theta \, \frac{\alpha_s}{8\pi} G_{\mu\nu}\tilde{G}^{\mu\nu}$$

$$\theta \lesssim 10^{-10}$$

We know there must be "new physics" beyond the Standard Model…

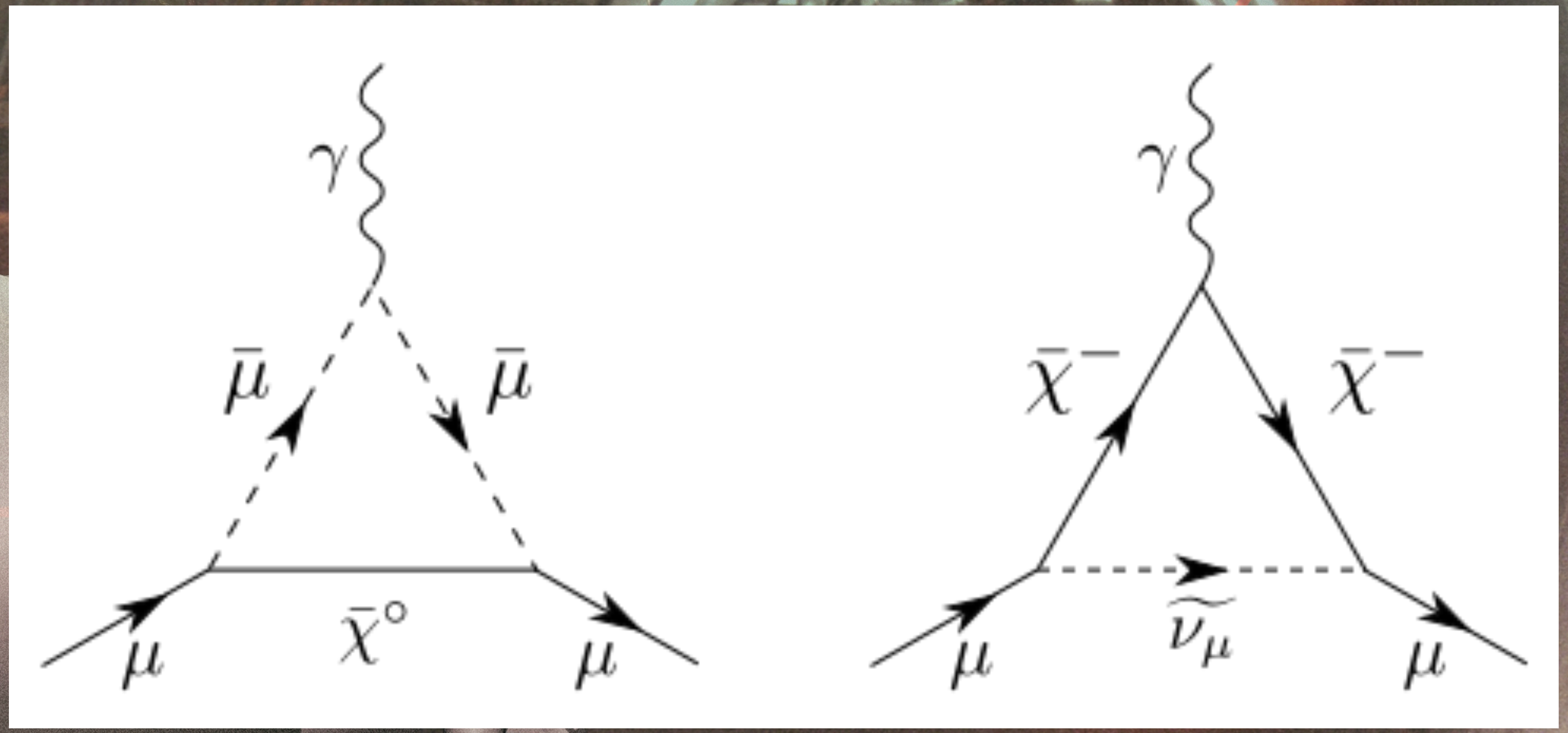
Searching for new physics at the smallest scales: Colliders





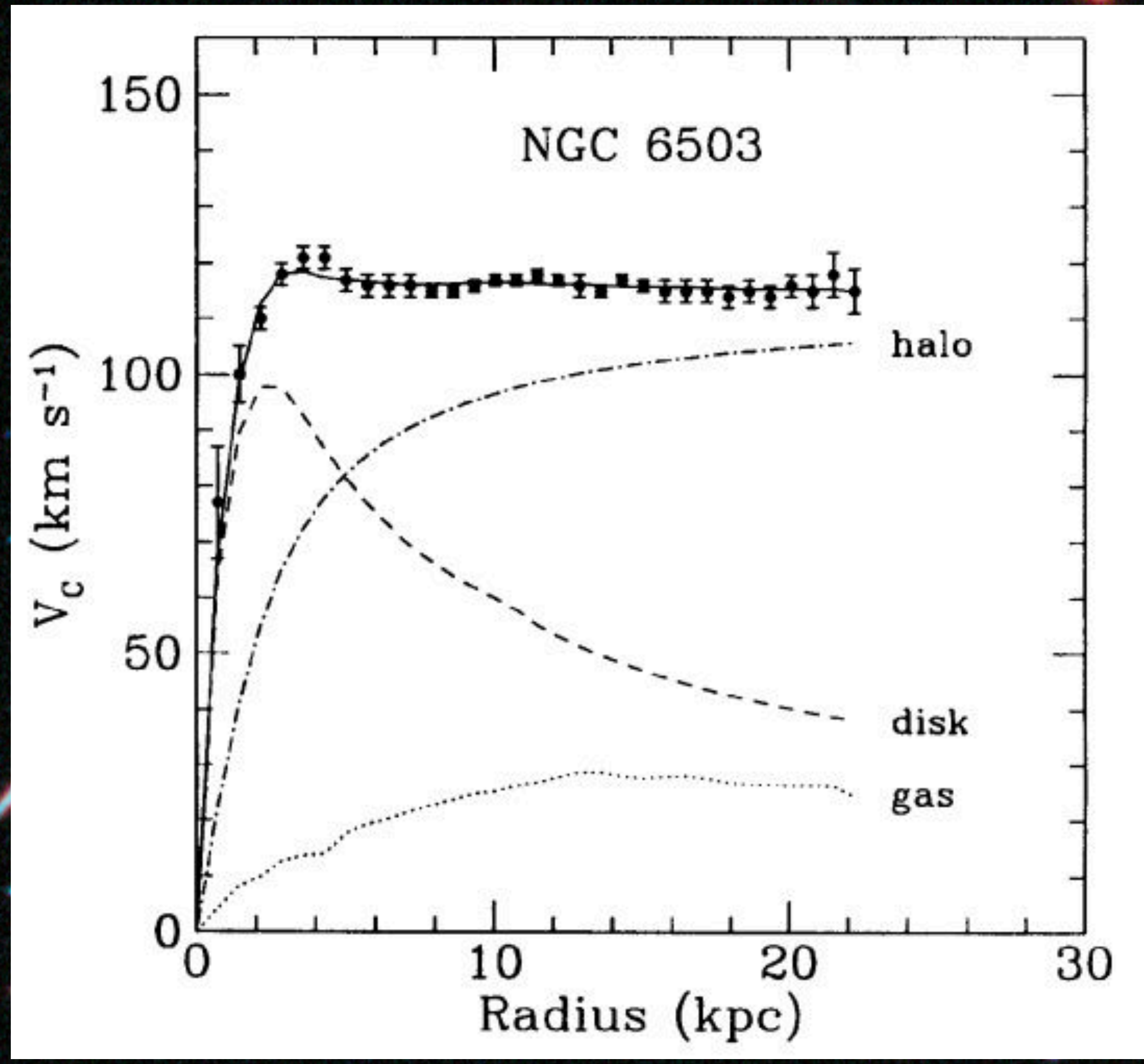
Directly produce the new physics

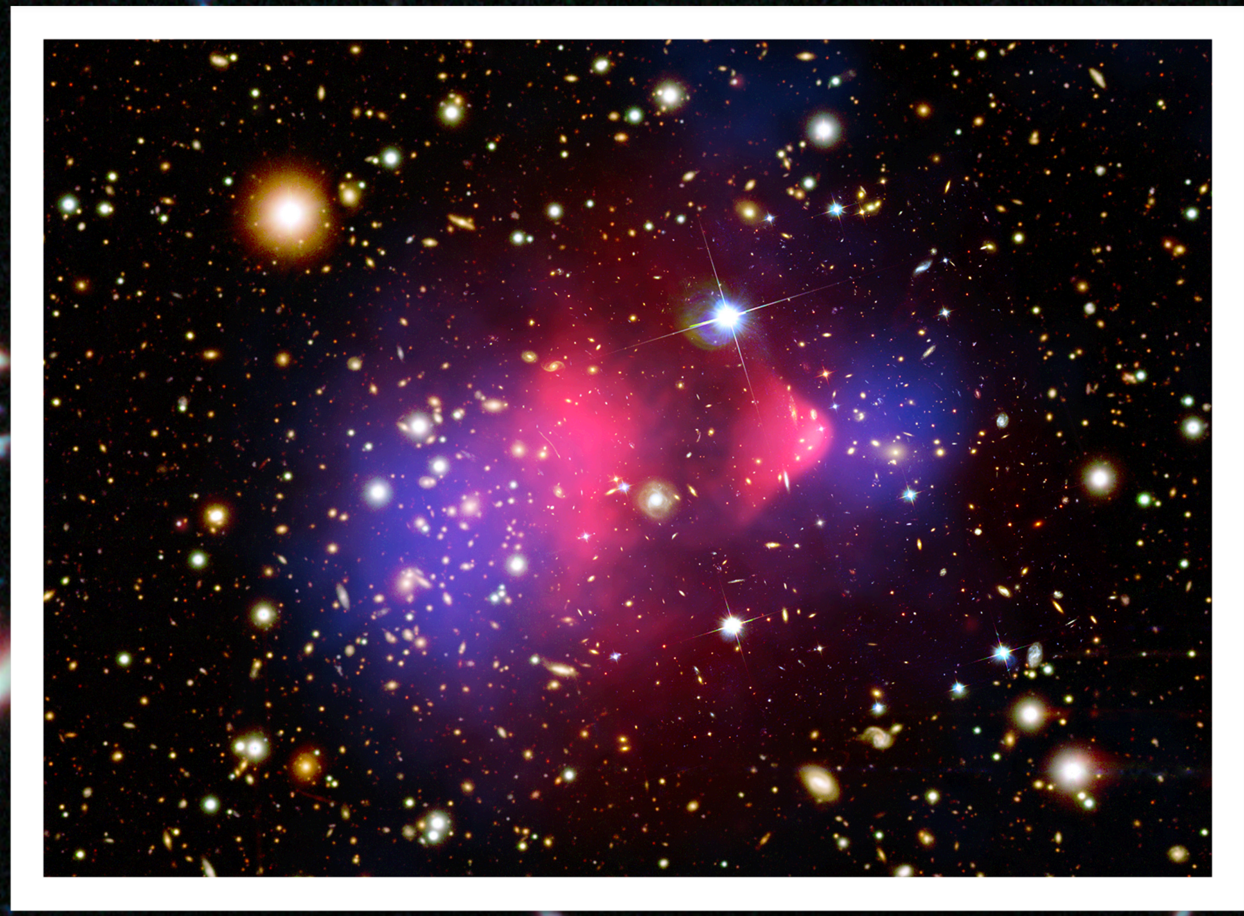
Indirect effects of new physics (precision tests of the Standard Model)



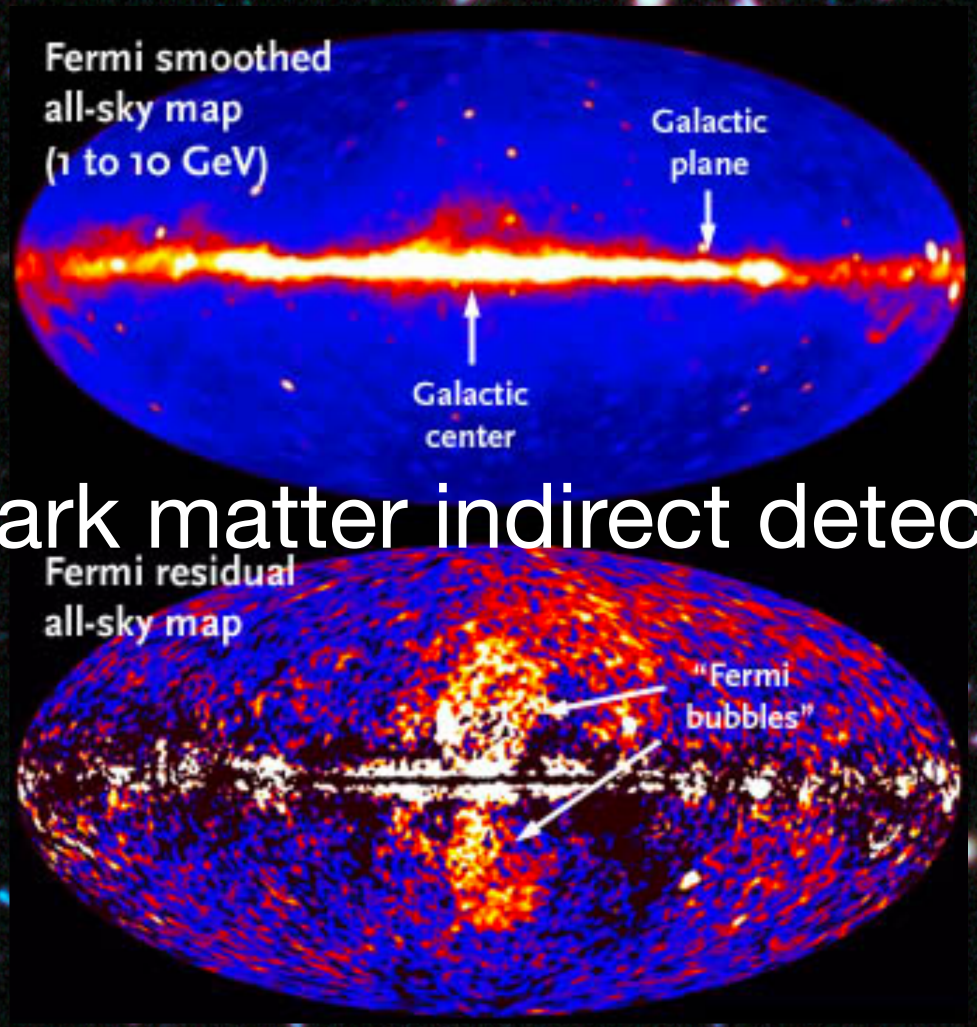
Searching for new physics at the largest scales: Astro/Cosmo

Astrophysical probes of dark matter

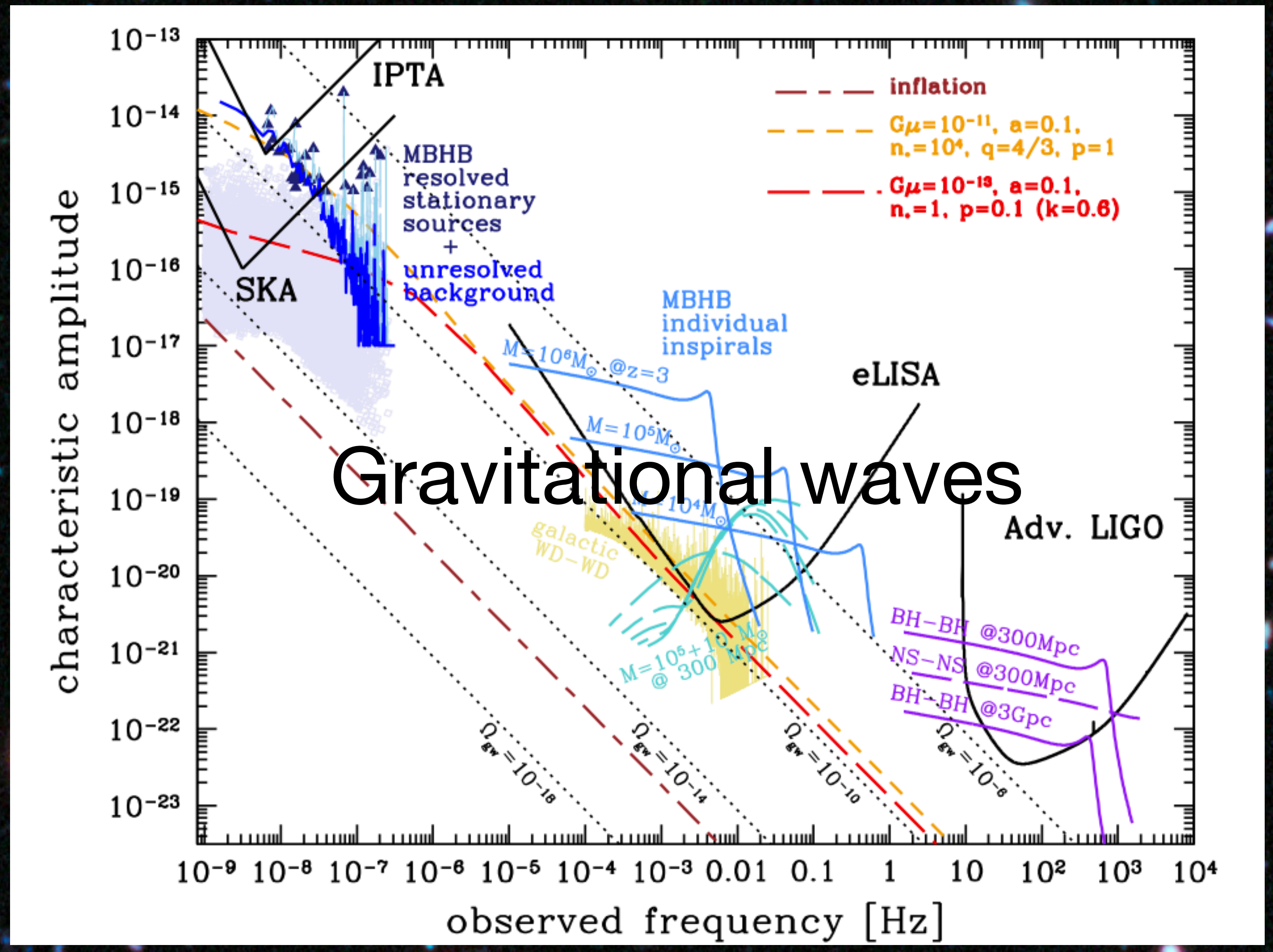




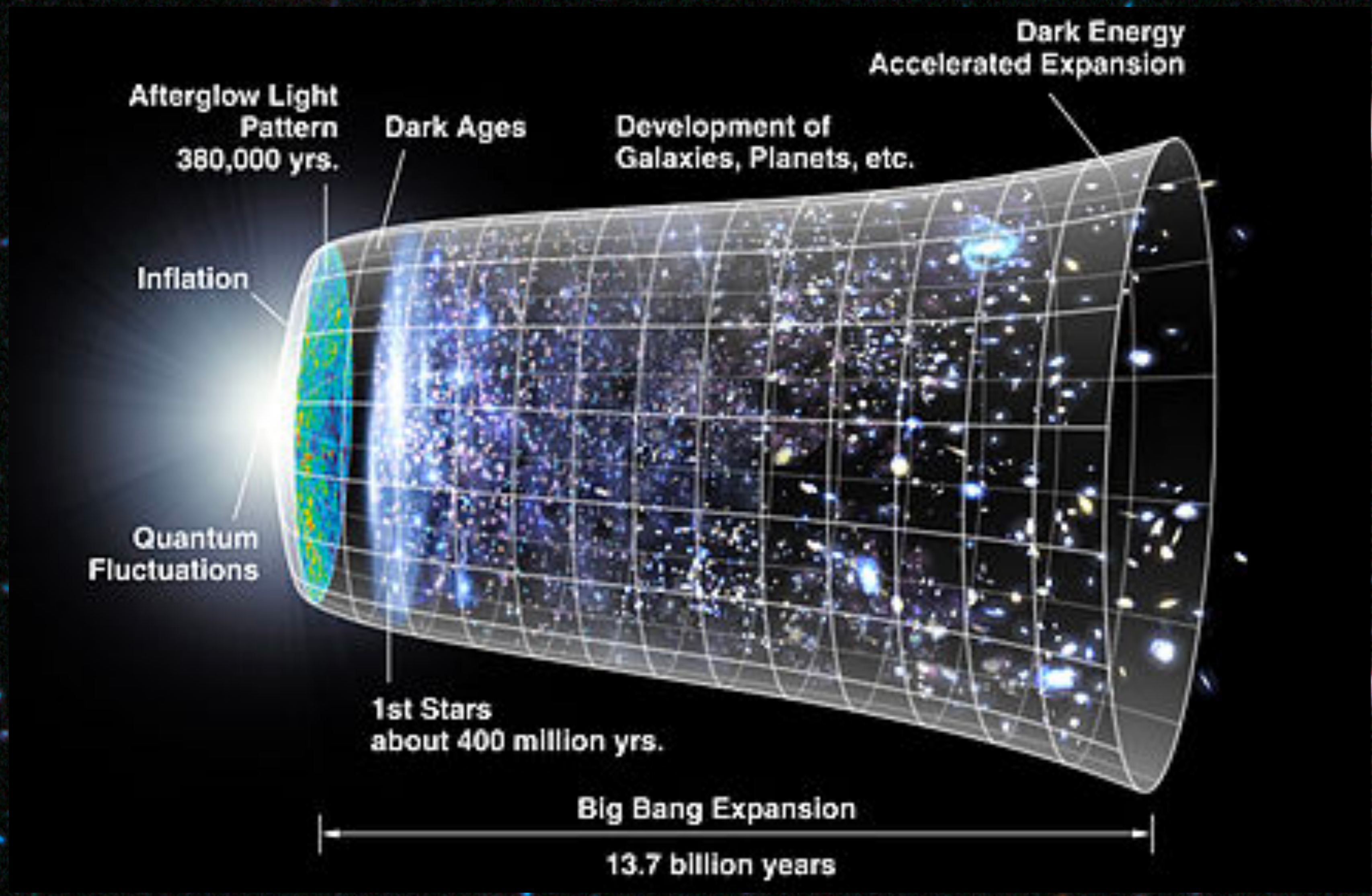
Dark matter indirect detection

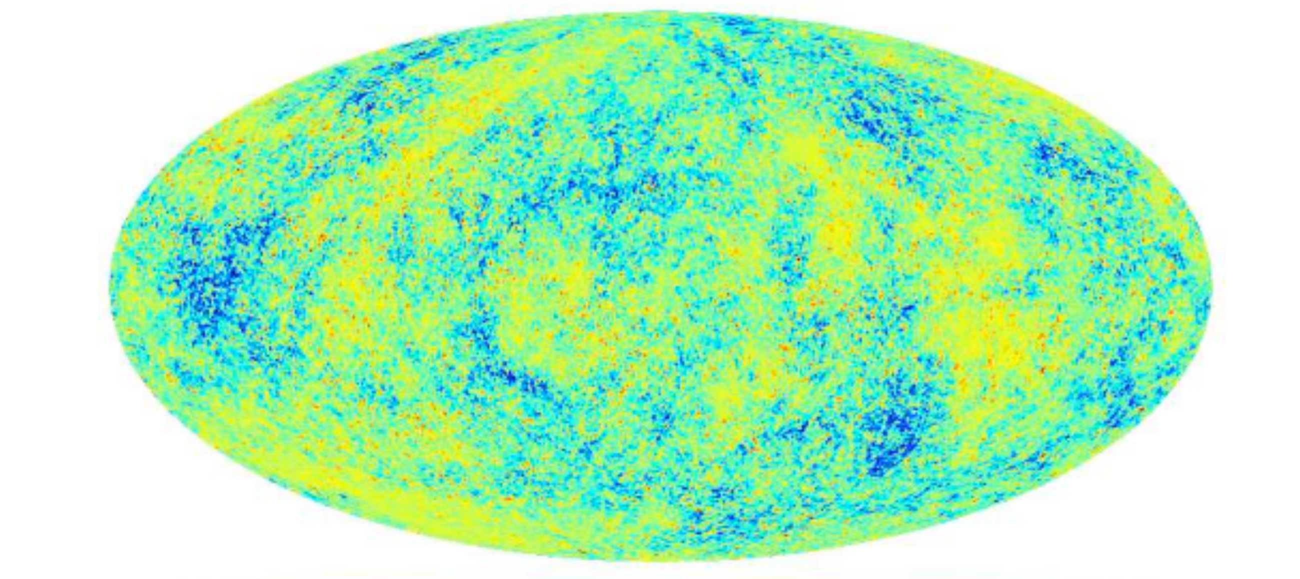


Gravitational waves



Early universe cosmology





Era of Big Data in HEP/Astro/Cosmo

- **LHC: 2010+, 10^{15} events, 10^2 PB (and growing)**
- Euclid: 2021+, 10^{10} objects, 10^2 PB
- Rubin (LSST): 2024(exp), 10^{10} objects, 10^2 PB
- Roman: 2027(exp), 10^9 objects, 10^1 PB
- SKA: 2030(exp), **~1-10 EB**

The Big Data era,
already familiar to HEP,
is coming for Astro/Cosmo

Modern ML methods will be essential to get the most out of these rich datasets

ML — a powerful new tool

Data → **Modern Machine Learning** → **Physics**

Modern machine learning is a **powerful new tool** which allows us to see **farther into the data** than ever before.

- ***Enabling*** new analyses that were previously impossible
- ***Enhancing*** sensitivity and precision
- ***Accelerating*** simulation and inference
- ***Unifying*** solutions to problems across different datasets and domains

ML for HEP

Data

Modern
Machine Learning

New physics searches

Triggering

Reconstruction/ Identification

Fast simulation

Measurement

Theory

…

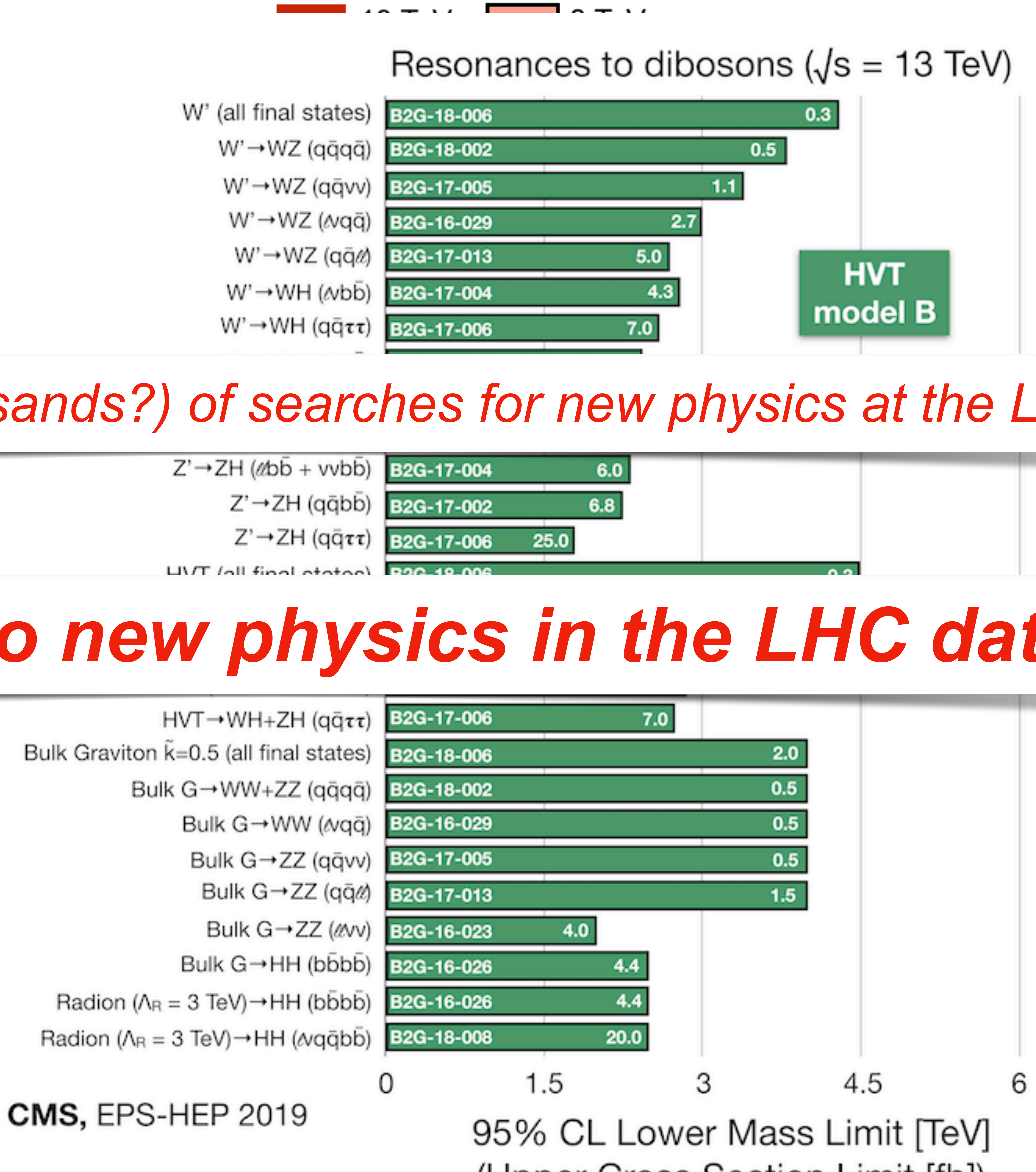
Not possible to survey everything in this talk!

Instead, will highlight selected examples

ML for New Physics Searches

By now, hundreds (thousands?) of searches for new physics at the LHC.

Is there really no new physics in the LHC data?



ML for New Physics Searches

All but a few of these LHC searches are optimized for ***specific models***

There could be vast, untapped discovery potential with ***ML-powered model-agnostic searches***

ML for New Physics Searches

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

Gregor Kasieczka (ed),[1] Benjamin Nachman (ed),[2,3] David Shih (ed),[4] Oz Amram,[5] Anders Andreassen,[6] Kees Benkendorfer,[2,7] Blaz Bortolato,[8] Gustaaf Brooijmans,[9] Florencia Canelli,[10] Jack H. Collins,[11] Biwei Dai,[12] Felipe F. De Freitas,[13] Barry M. Dillon,[8,14] Ioan-Mihail Dinu,[5] Zhongtian Dong,[15] Julien Donini,[16] Javier Duarte,[17] D. A. Faroughy[10] Julia Gonski,[9] Philip Harris,[18] Alan Kahn,[9] Jernej F. Kamenik,[8,19] Charanjit K. Khosa,[20,30] Patrick Komiske,[21] Luc Le Pottier,[2,22] Pablo Martín-Ramiro,[2,23] Andrej Matevc,[8,19] Eric Metodiev,[21] Vinicius Mikuni,[10] Inês Ochoa,[24] Sang Eon Park,[18] Maurizio Pierini,[25] Dylan Rankin,[18] Veronica Sanz,[20,26] Nilai Sarda,[27] Uroš Seljak,[2,3,12] Aleks Smolkovic,[8] George Stein,[2,12] Cristina Mantilla Suarez,[5] Manuel Szewc,[28] Jesse Thaler,[21] Steven Tsan,[17] Silviu-Marian Udrescu,[18] Louis Vaslin,[16] Jean-Roch Vlimant,[29] Daniel Williams,[9] Mikaeel Yunus[18]

https://arxiv.org/abs/2101.08320

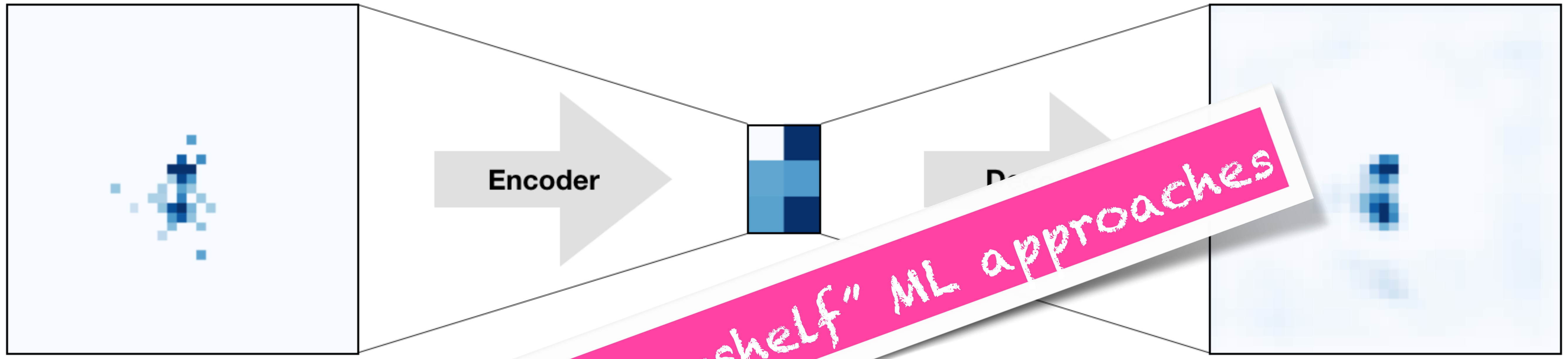
https://arxiv.org/abs/2105.14027

The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider

T. Aarrestad[a] M. van Beekveld[b] M. Bona[c] A. Boveia[e] S. Caron[d] J. Davies[c] A. De Simone[f,g] C. Doglioni[h] J. M. Duarte[i] A. Farbin[j] H. Gupta[k] L. Hendriks[d] L. Heinrich[a] J. Howarth[l] P. Jawahar[m,a] A. Jueid[n] J. Lastow[h] A. Leinweber[o] J. Mamuzic[p] E. Merényi[q] A. Morandini[r] P. Moskvitina[d] C. Nellist[d] J. Ngadiuba[s,t] B. Ostdiek[u,v] M. Pierini[a] B. Ravina[l] R. Ruiz de Austri[p] S. Sekmen[w] M. Touranakou[x,a] M. Vaškevičiūte[l] R. Vilalta[y] J.-R. Vlimant[t] R. Verheyen[z] M. White[o] E. Wulff[h] E. Wallin[h] K.A. Wozniak[α,a] Z. Zhang[d]

A lot of new ideas for model-agnostic searches!

ML for New Physics Searches



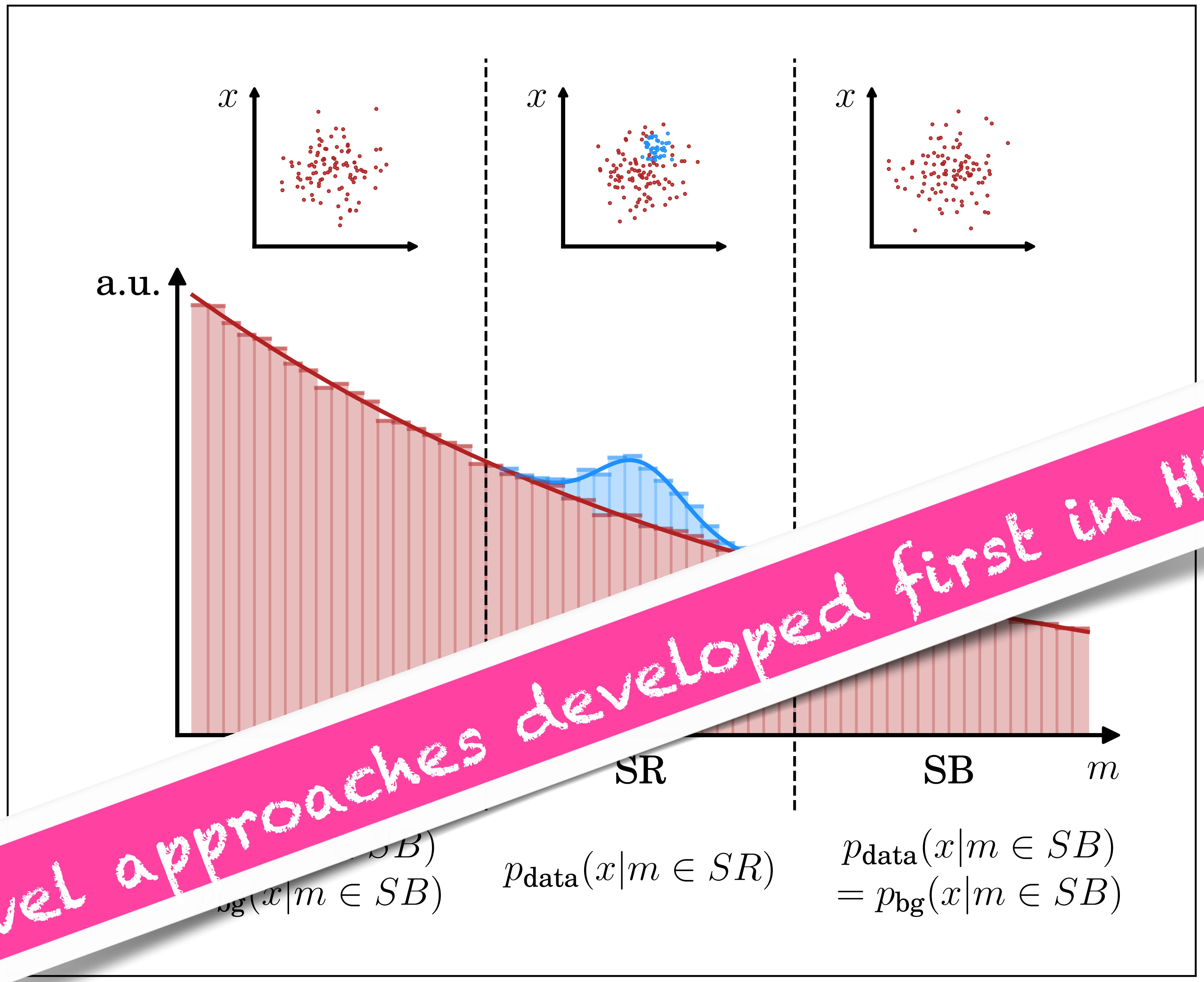
from 1808.08992

Mostly "off-the-shelf" ML approaches

Outlier detection

eg autoencoder based approaches:
Farina, Nakai & **DS** 1808.08992
Heimel et al 1808.08979
Cerri et al 1811.10276
....
Ostdiek 2109.01695
Mikuni, Nachman & **DS** 2111.06417
Dillon et al 2206.14225, 2301.04660
and many, many more!!

from 2109.00546

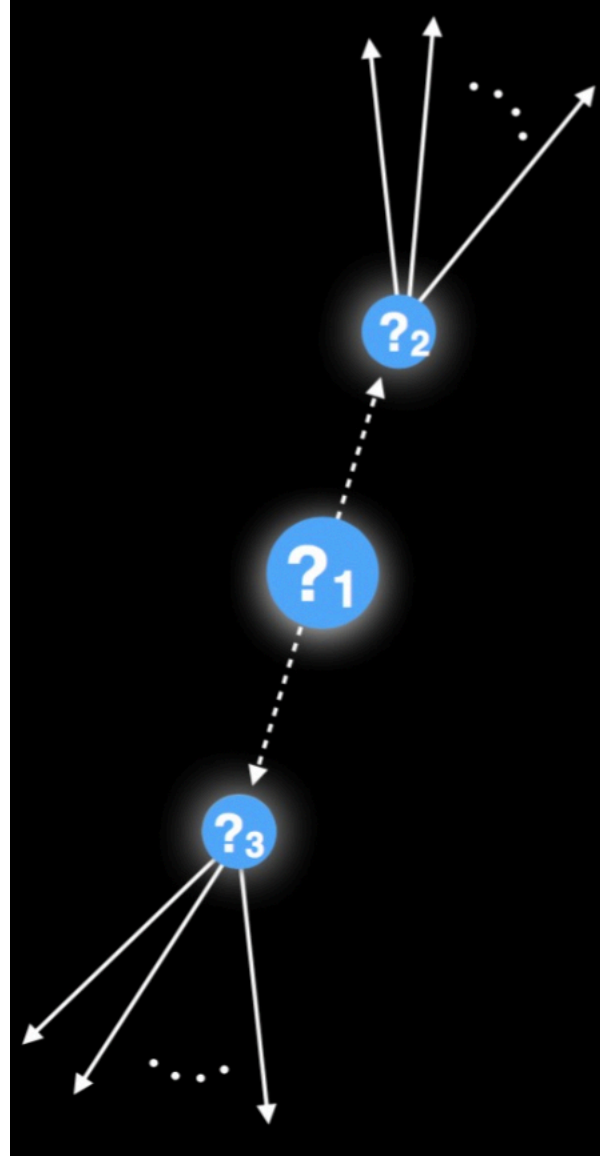


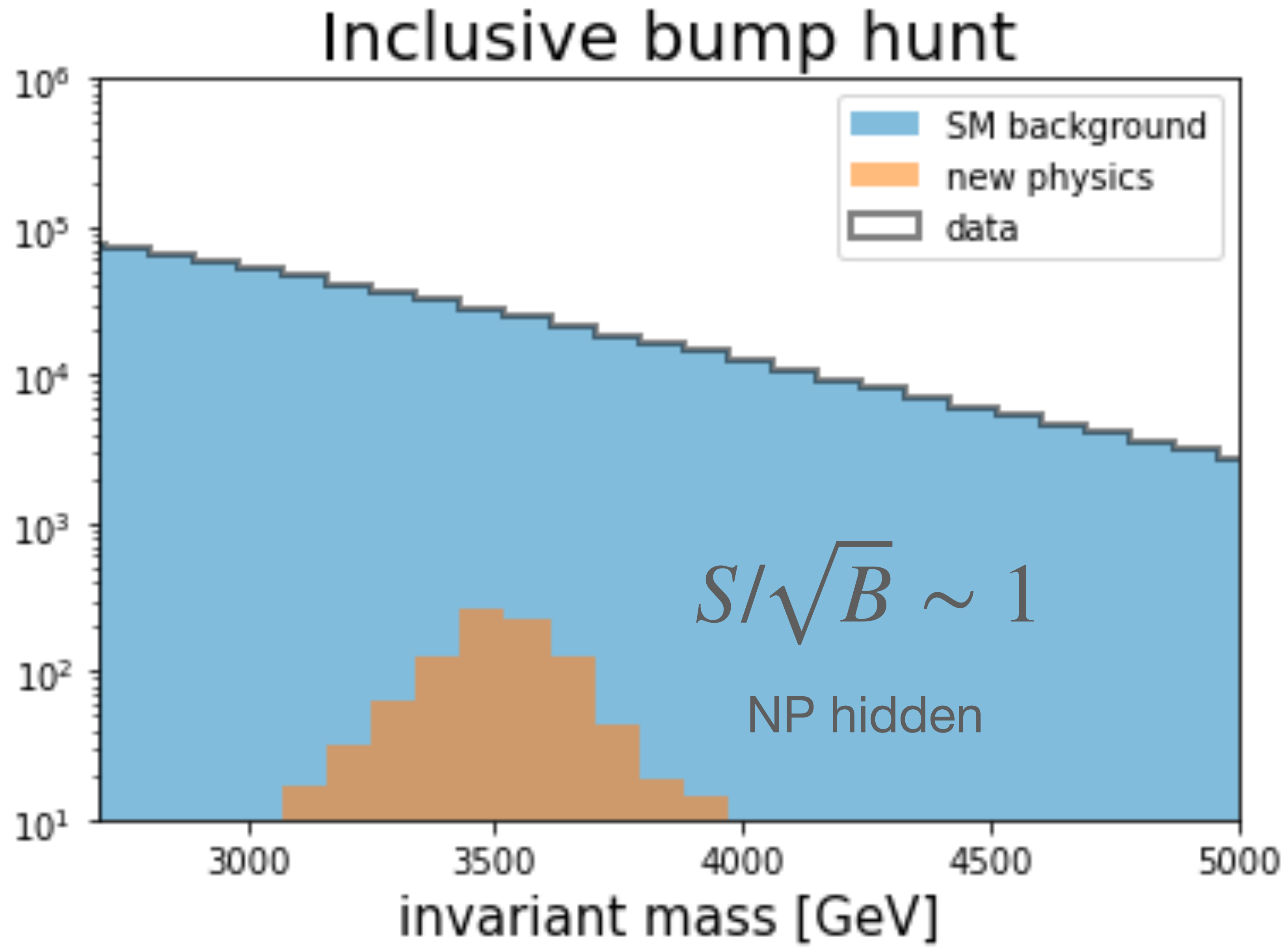
Novel approaches developed first in HEP

Overdensity detection

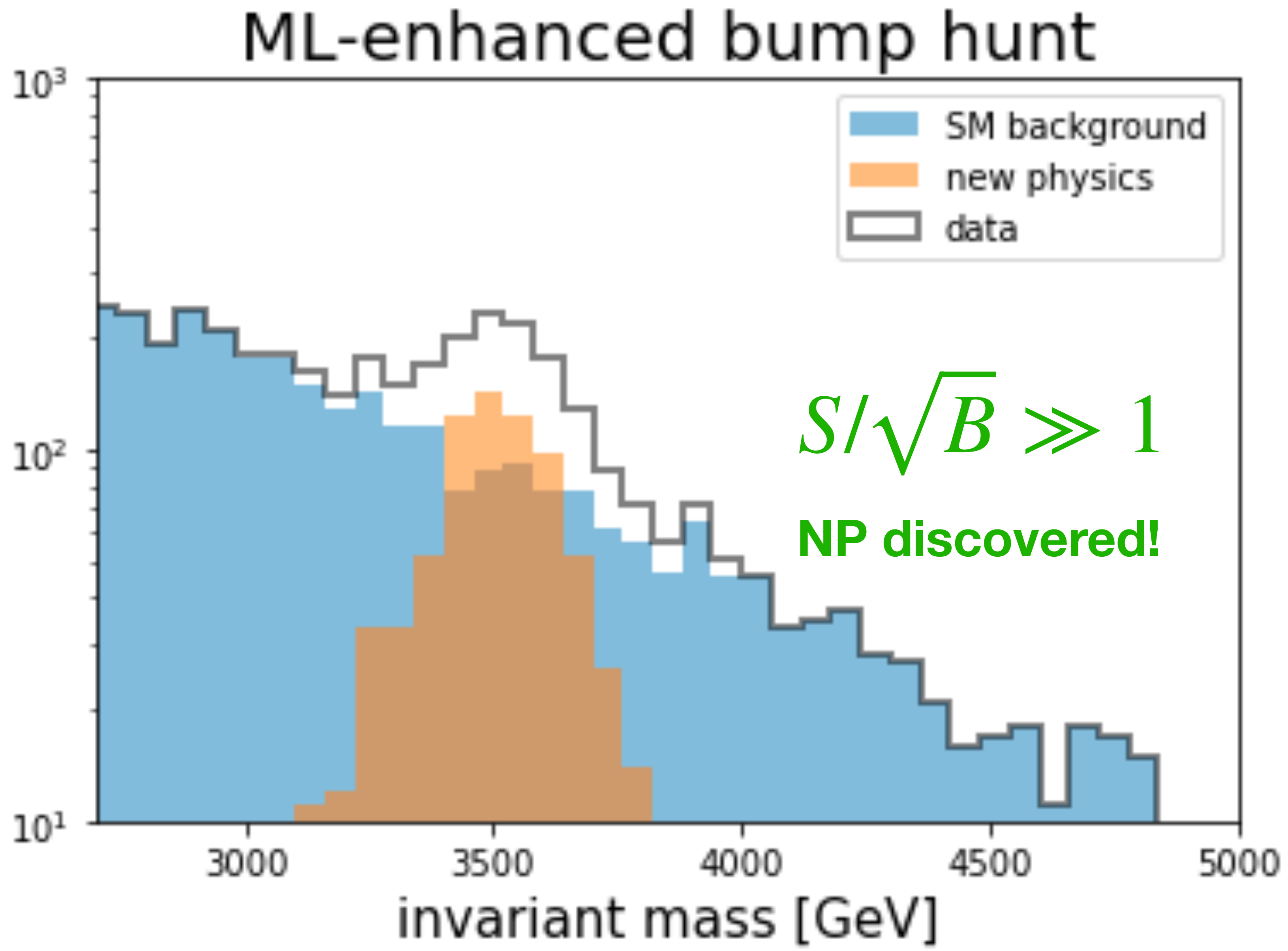
CWoLa Hunting [Collins, Howe & Nachman 1805.02664, 1902.02634]
ANODE [Nachman & **DS** 2001.04990]
SALAD [Andreassen, Nachman & **DS** 2001.05001]
SA-CWoLa [Benkendorfer et al 2009.02205]
CATHODE, LaCATHODE [**DS+** Hallin et al 2109.00546, 2210.14924]
CURTAINS [Raine et al 2203.09470]
FETA [Golling et al 2212.11285]

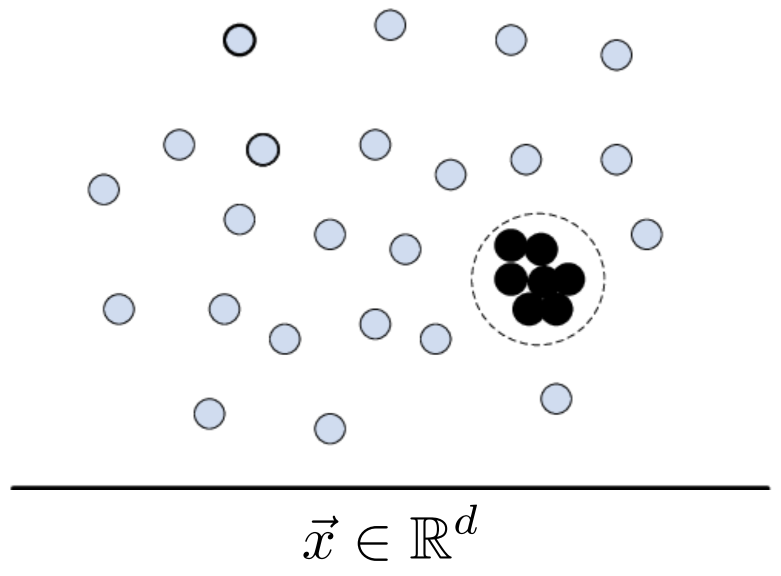
ML-enhanced bump hunts





Cut on anomaly score $R(x)$





x: ***additional*** features where NP could be localized

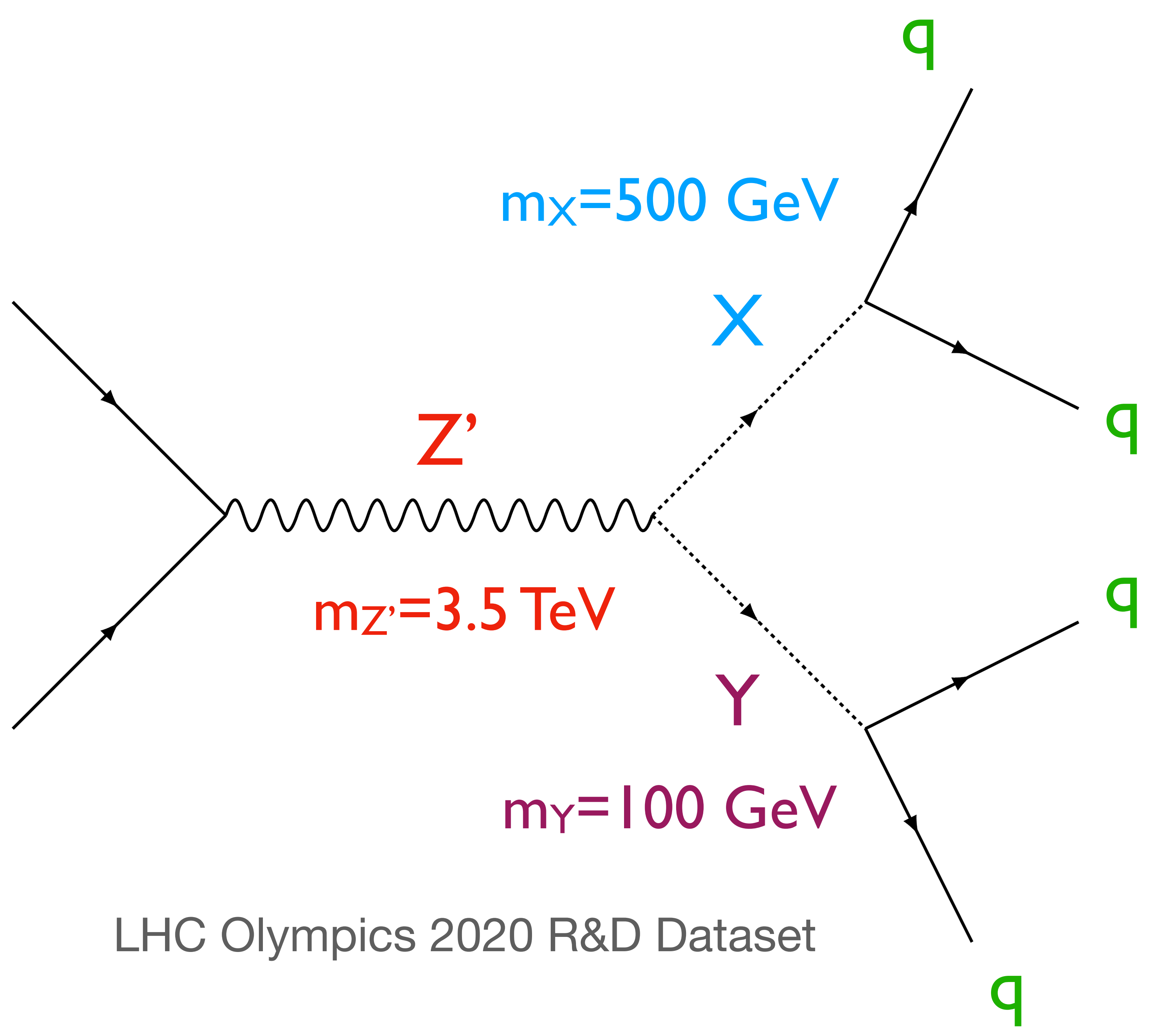
Learn model-agnostic **anomaly score $R(x)$** from data

provably optimal

$$R(x) = \frac{p_{data}(x)}{p_{bg}(x)}$$

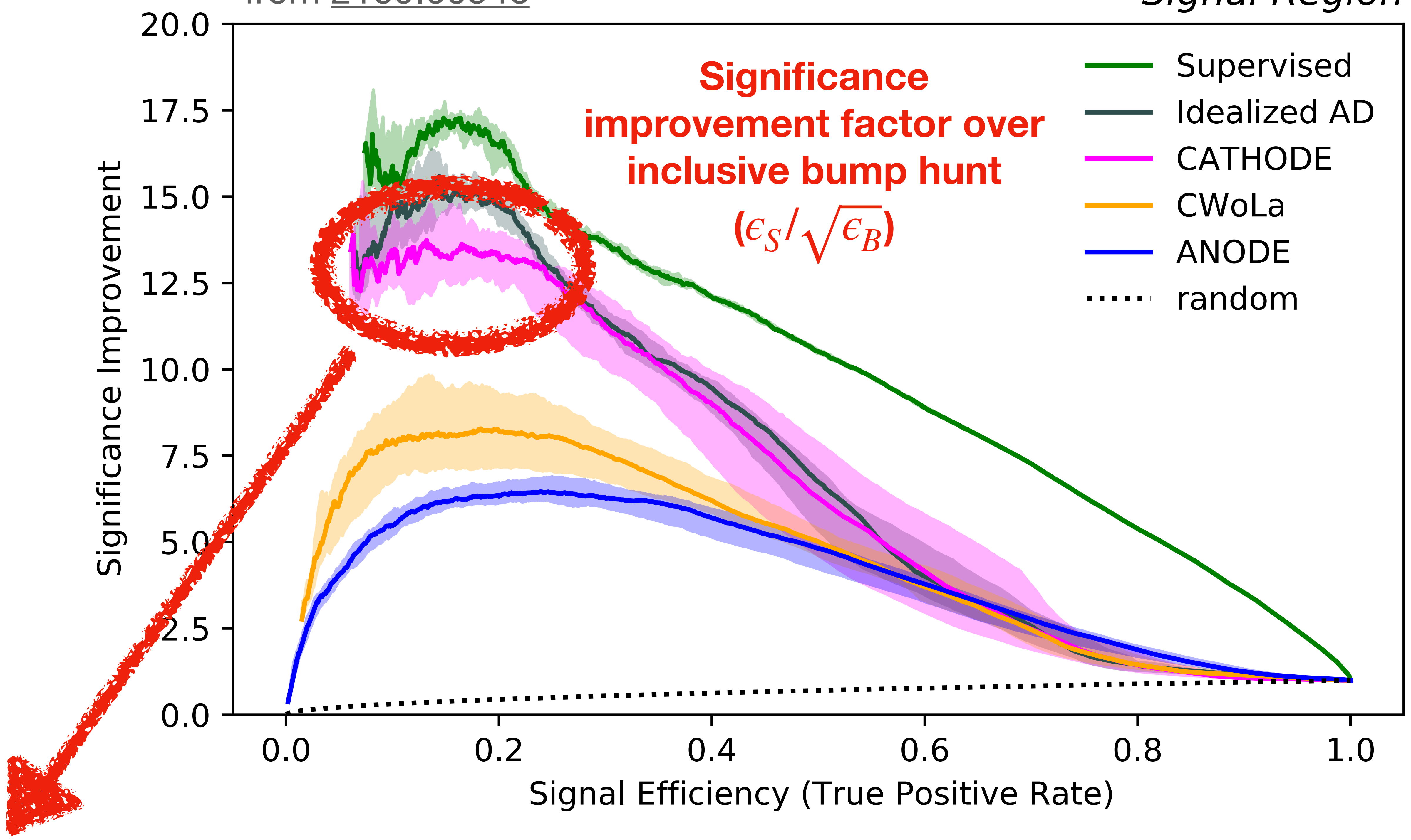
ML-enhanced bump hunts

New methods can achieve impressive performance gains over the inclusive bump hunt.



LHC Olympics 2020 R&D Dataset

from 2109.00546



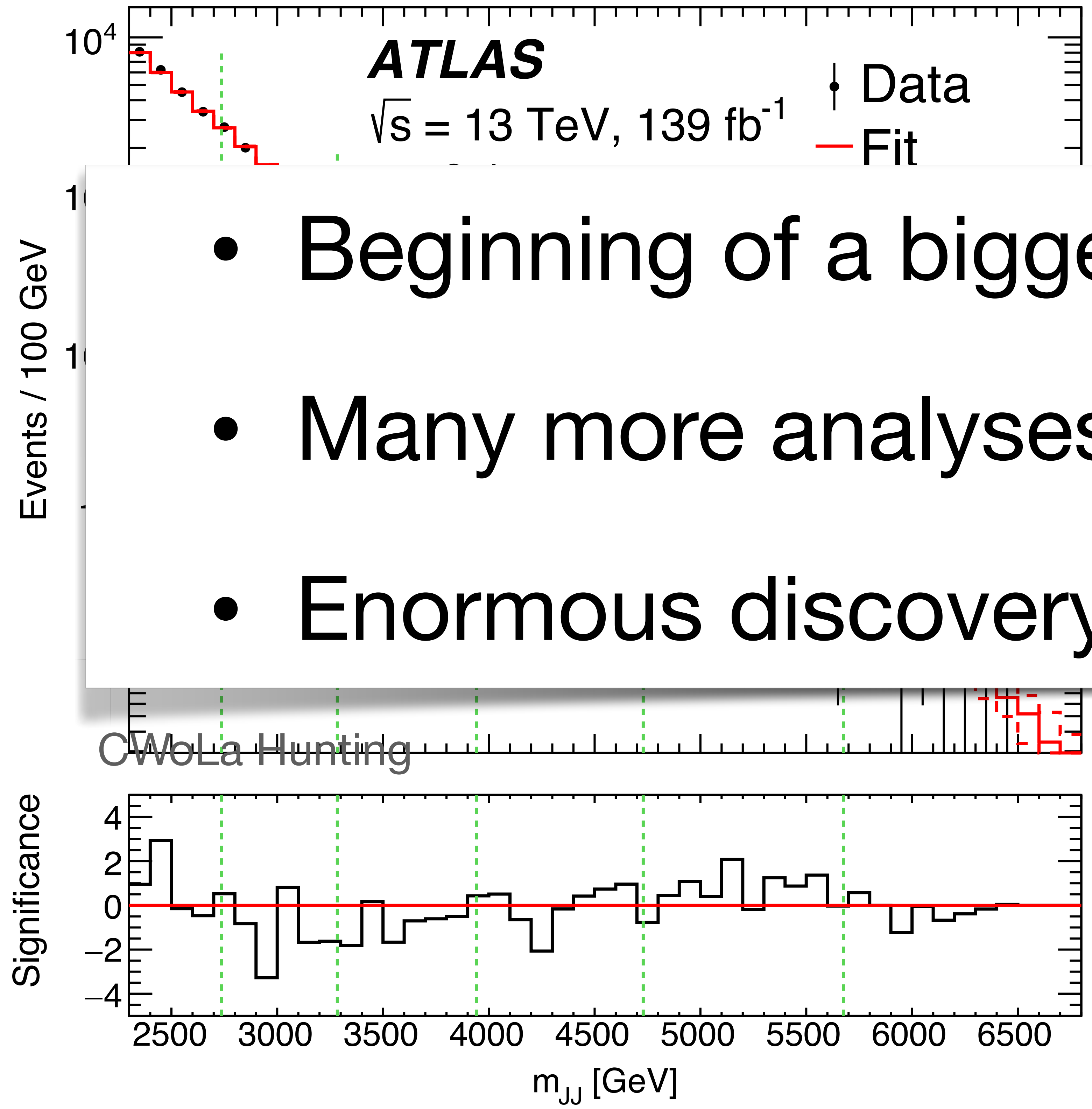
On this signal, ~2σ inclusive dijet ==> up to **~30σ with CATHODE method** [**DS+** Hallin et al 2109.00546]

New physics could be hiding in the data right now!

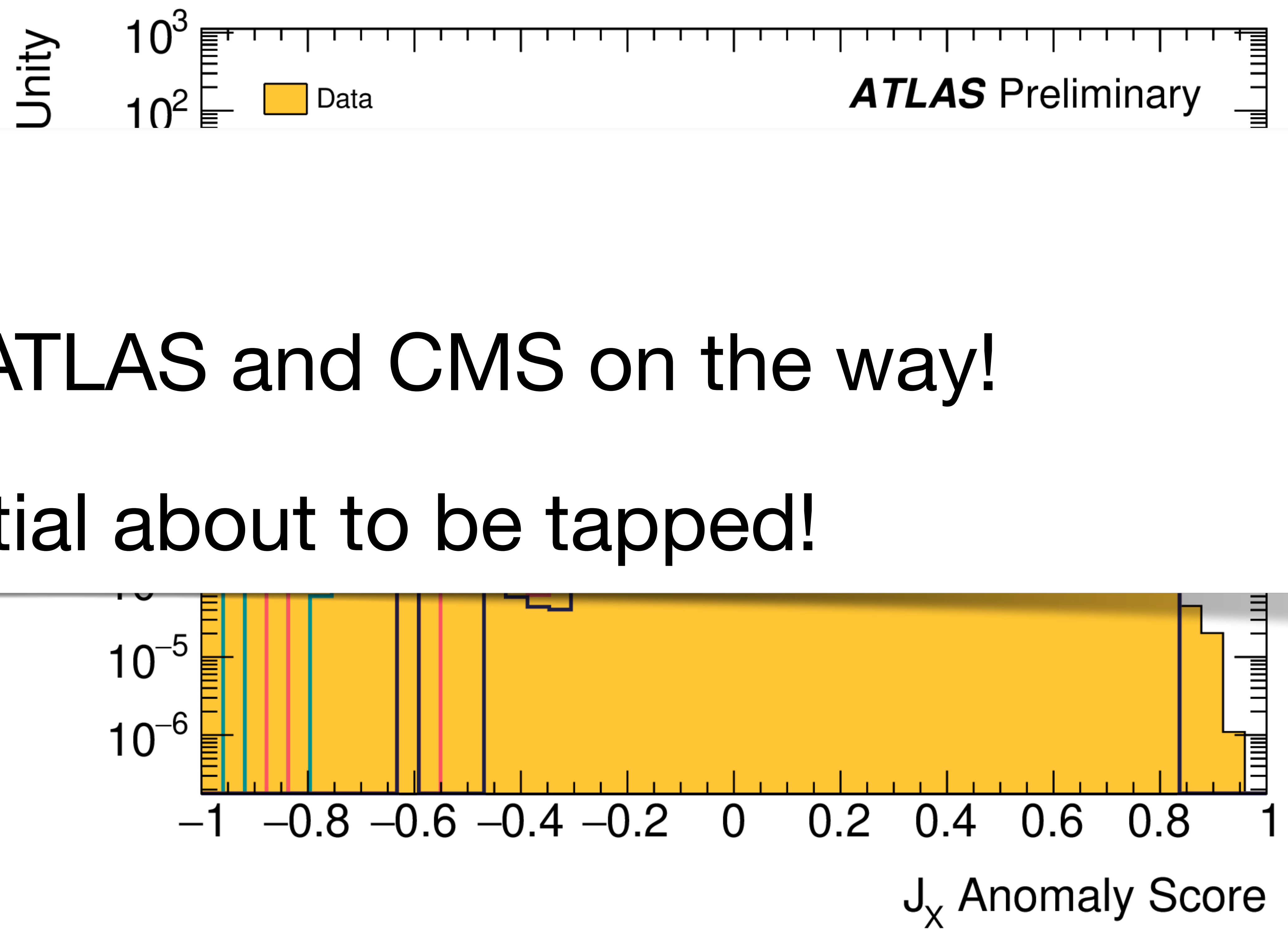
From proof-of-concept to reality

Proofs-of-concept are becoming actual LHC searches!

CWoLa Hunting
ATLAS, PRL **125** 131801 (2020)



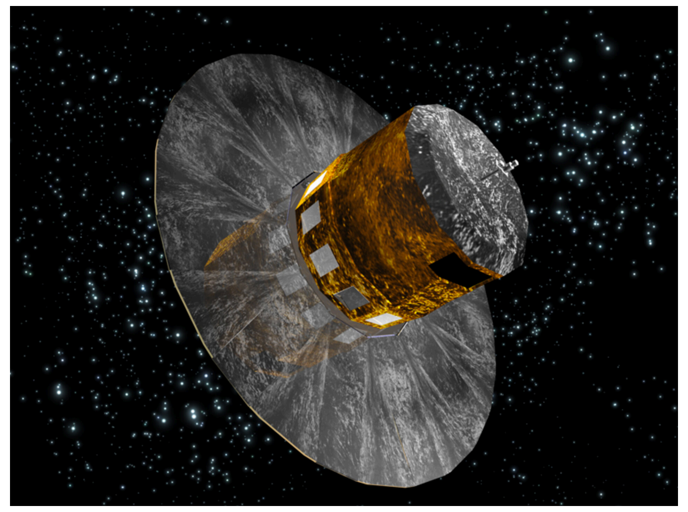
RNN VAE
ATLAS-CONF-2022-045



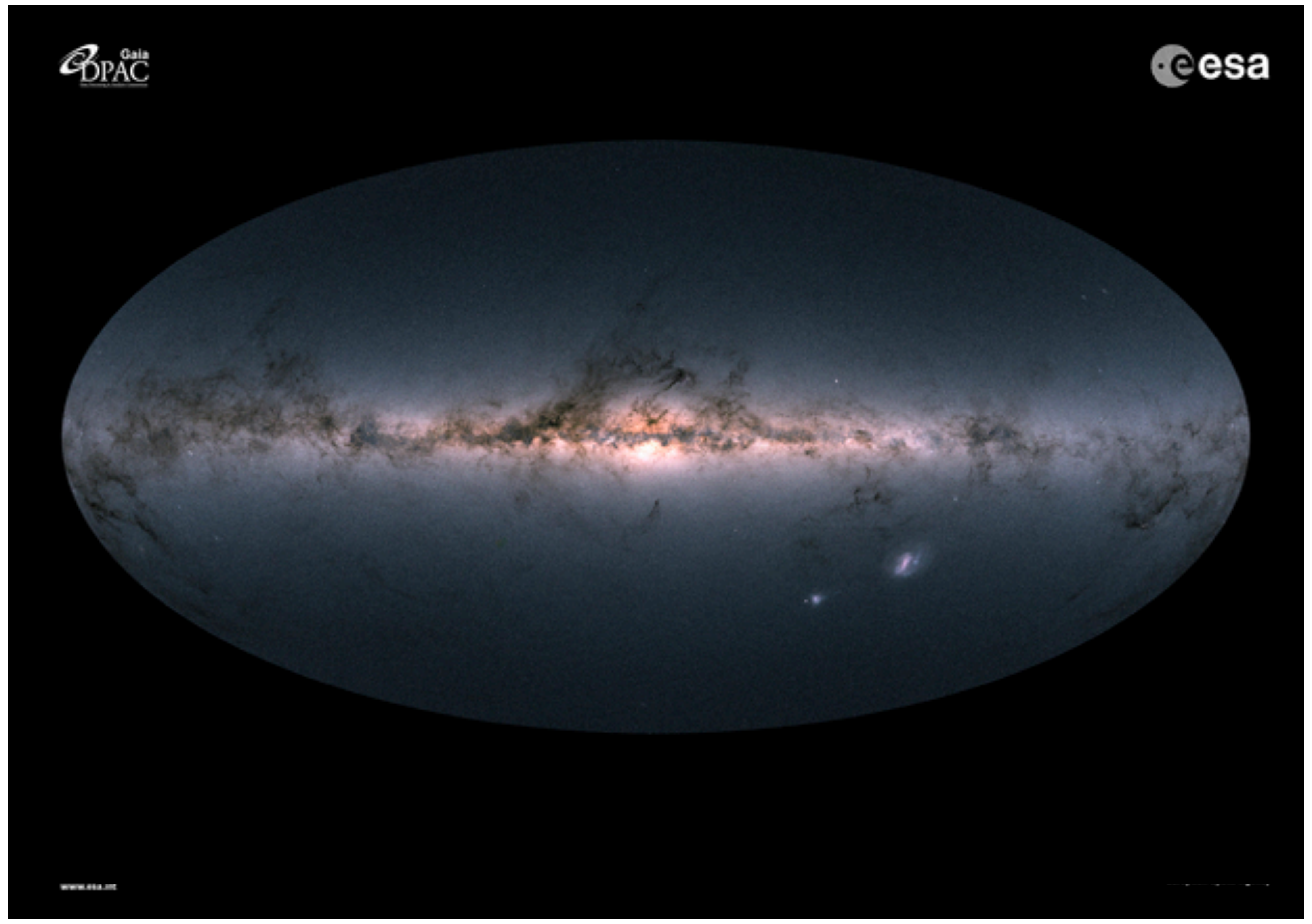
- Beginning of a bigger wave
- Many more analyses from ATLAS and CMS on the way!
- Enormous discovery potential about to be tapped!

From LHC → Astro

Searching for Stellar Streams in Gaia



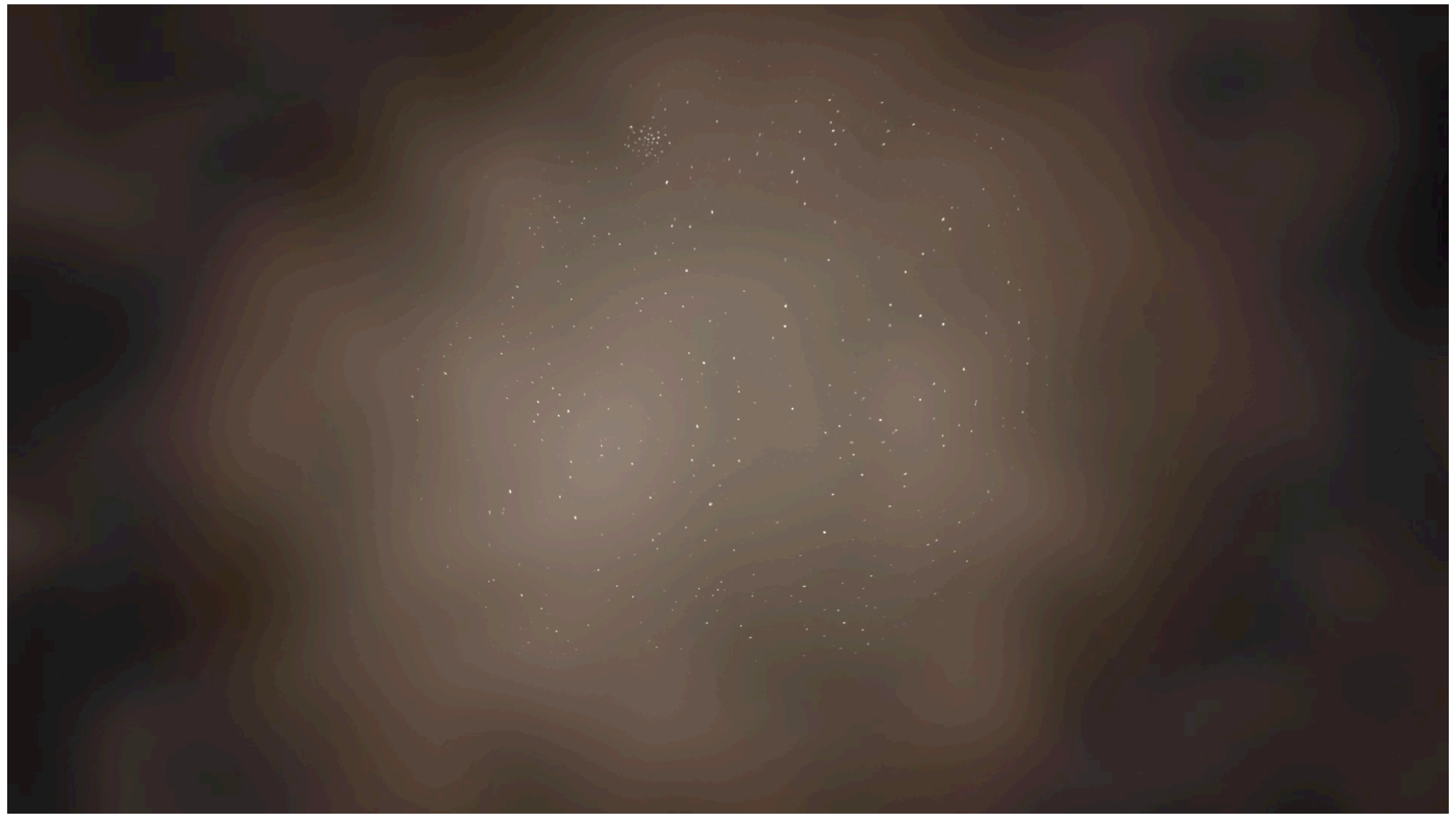
- We realized the same ML-enhanced bump hunt methods developed for LHC could be applied to **Gaia data** to search for **stellar streams**
  - *An example of power of ML to cut across domains!*



Gaia satellite:

- Launched in 2013; ongoing
- Angular positions, proper motions, color and magnitude of over **1 billion stars** in our Galaxy
- Distances and radial velocities for a smaller subset of nearby stars

Stellar Streams



credit: Gabriel Pérez Díaz

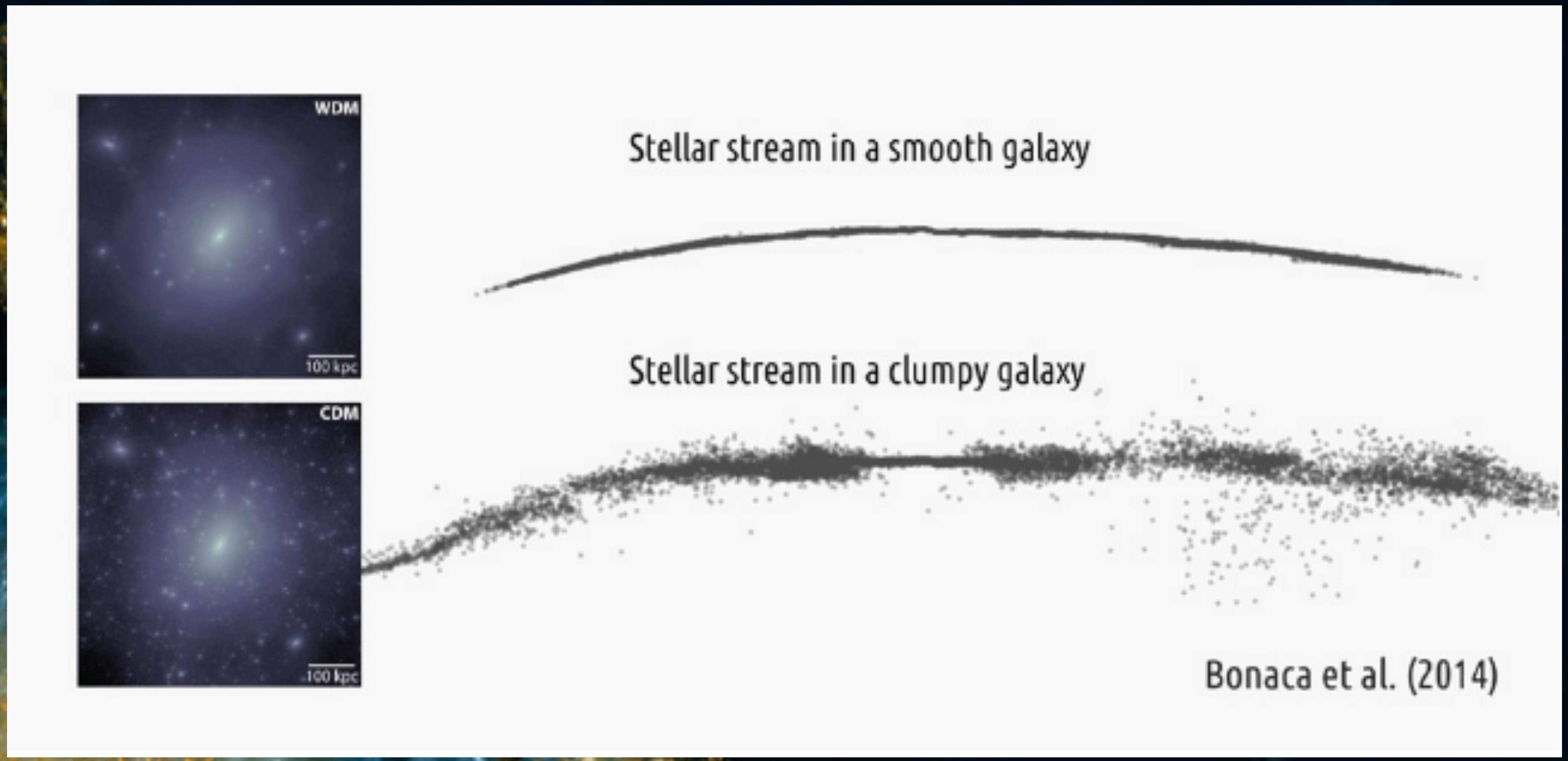
Stellar streams are the very old remnants of tidally disrupted globular clusters and dwarf galaxies.

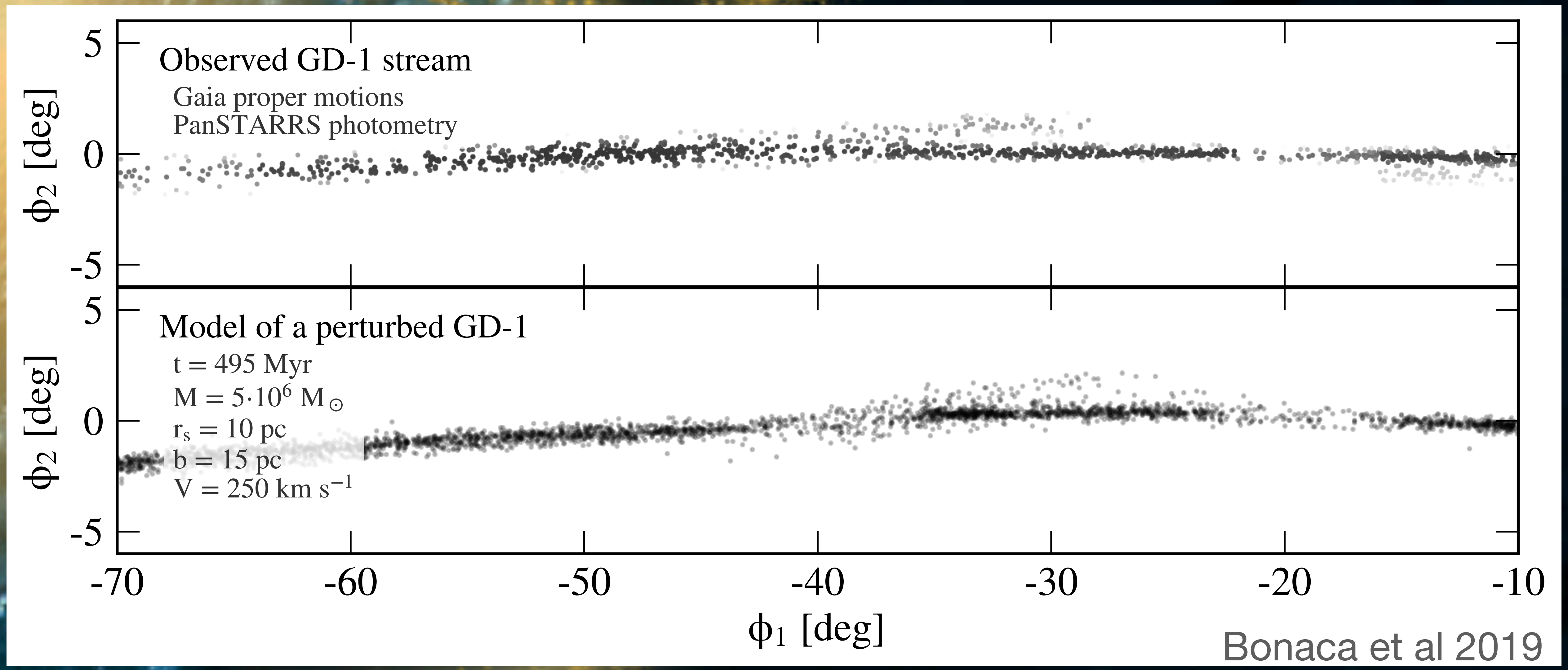
credit: S. Payne-Wardenaar / K. Malhan, MPIA

Collection of stars moving together along a common orbit — concentrated spatially and in velocity.

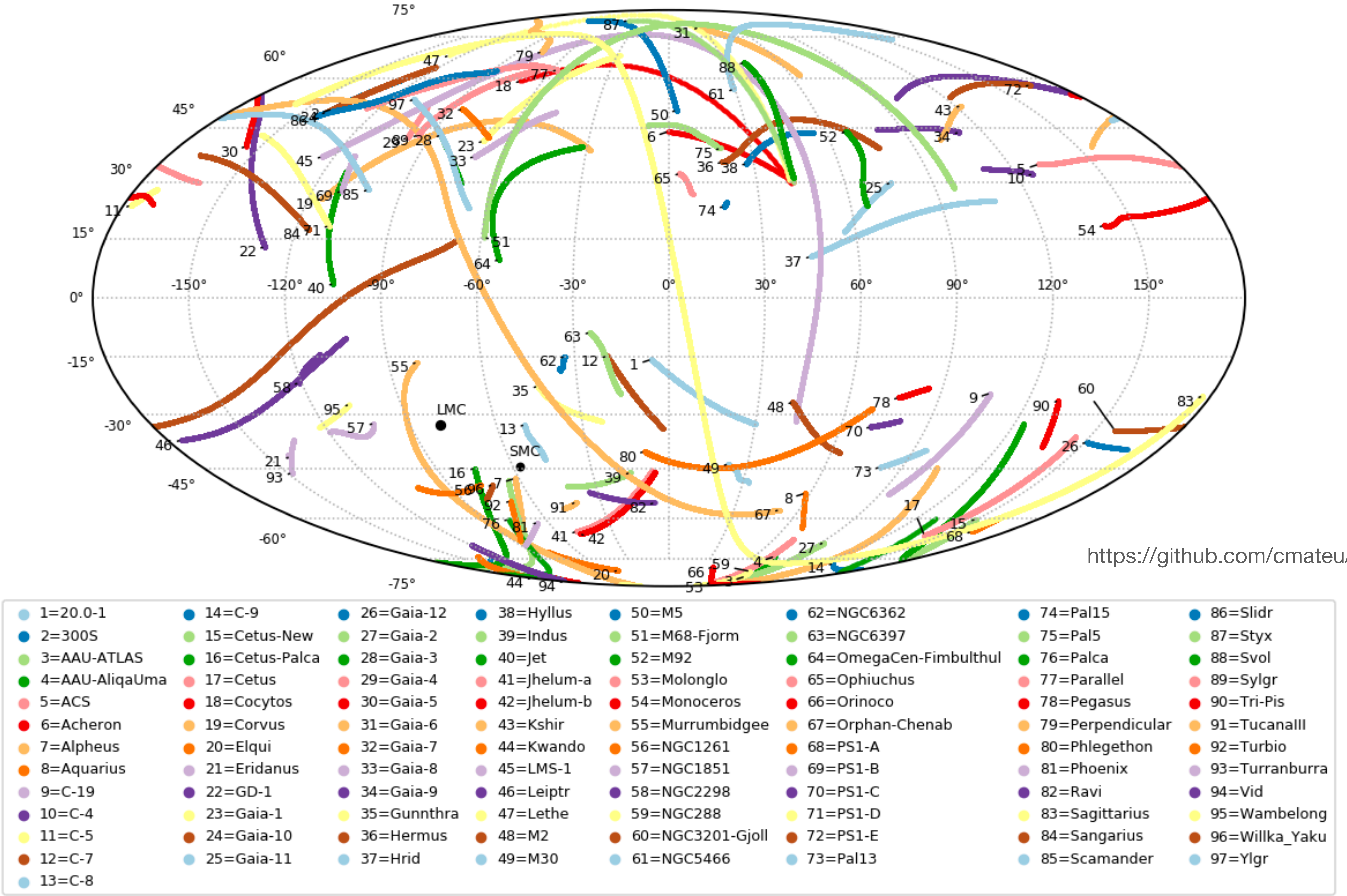
Stellar Streams

Stellar streams could be unique astrophysical probes into dark matter substructure





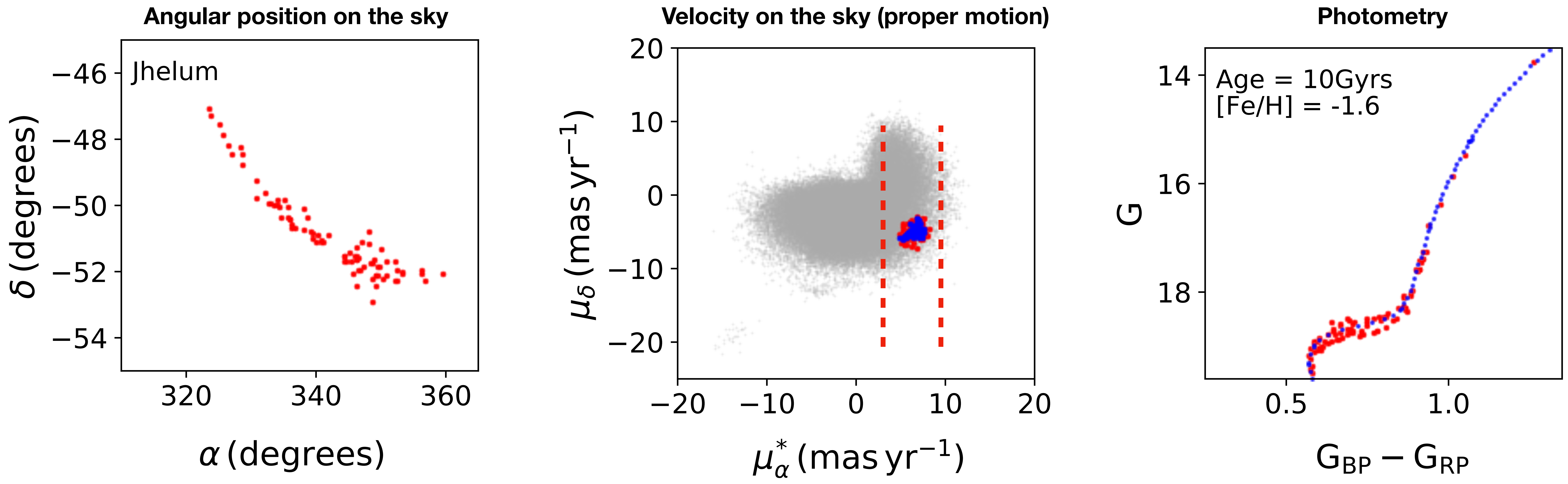
Known Stellar Streams of the Milky Way



https://github.com/cmateu/galstreams

Via Machinae

[DS, Buckley, Necib ‘23] [DS, Buckley, Necib, Tamanas ‘21]

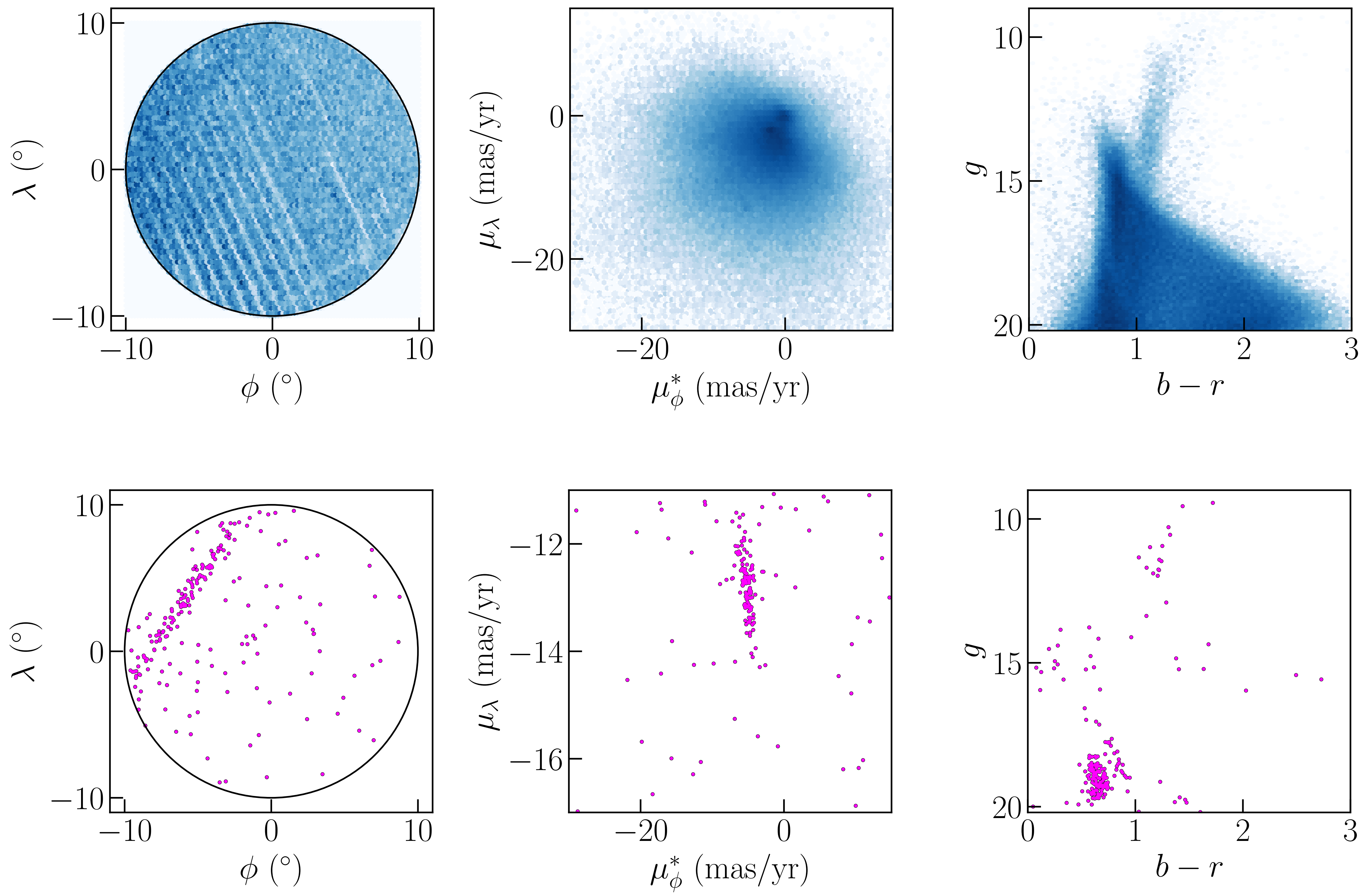


- Streams are local overdensities in multiple features — ideal for **enhanced bump hunt methods**!
- Choose either proper motion coordinate as resonant feature
- Learn anomaly score (using normalizing flows) with remaining five features

Core method — illustrated with GD-1 Stream

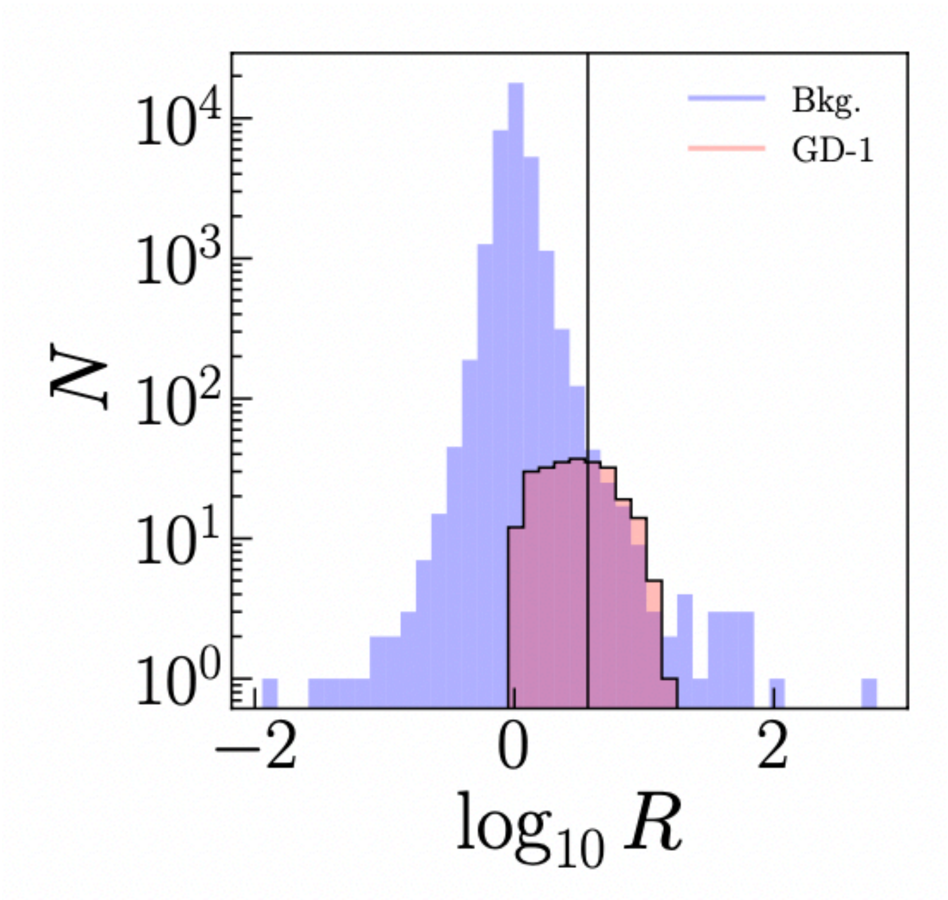
[DS, Buckley, Necib, Tamanas ’21]

Fully data driven, simulation independent!



All stars in a patch of the sky containing (part of) GD-1 (ra,dec)=(148.6,24.2)

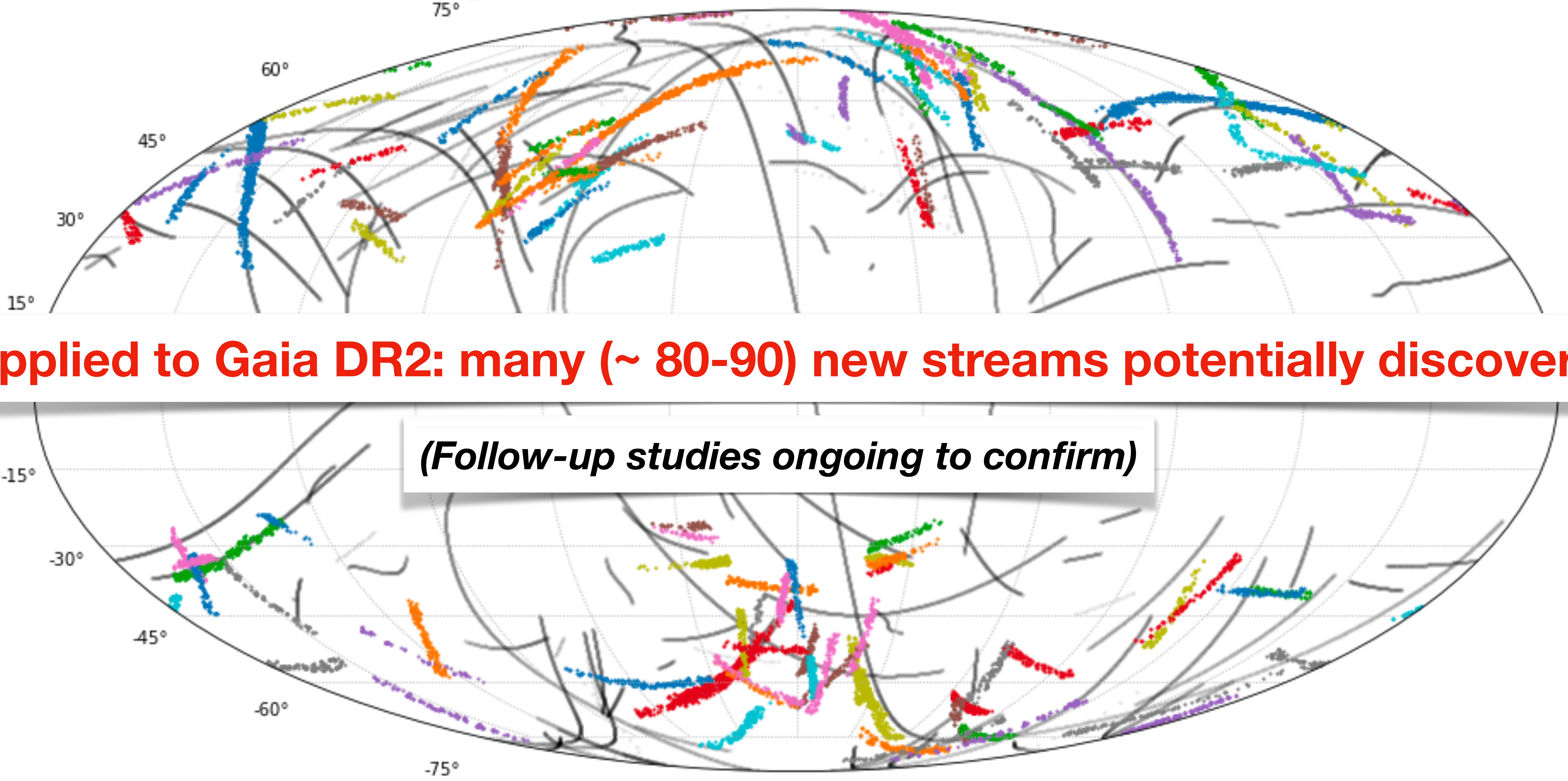
Stars in SR after cut on $R(x)$ obtained from ANODE



The method works!

New stream candidates from Gaia DR2

[DS, Buckley, Necib 2303.01529]



Applied to Gaia DR2: many (~ 80-90) new streams potentially discovered!

(Follow-up studies ongoing to confirm)

Direct phase space density estimation of stellar tracers from Gaia

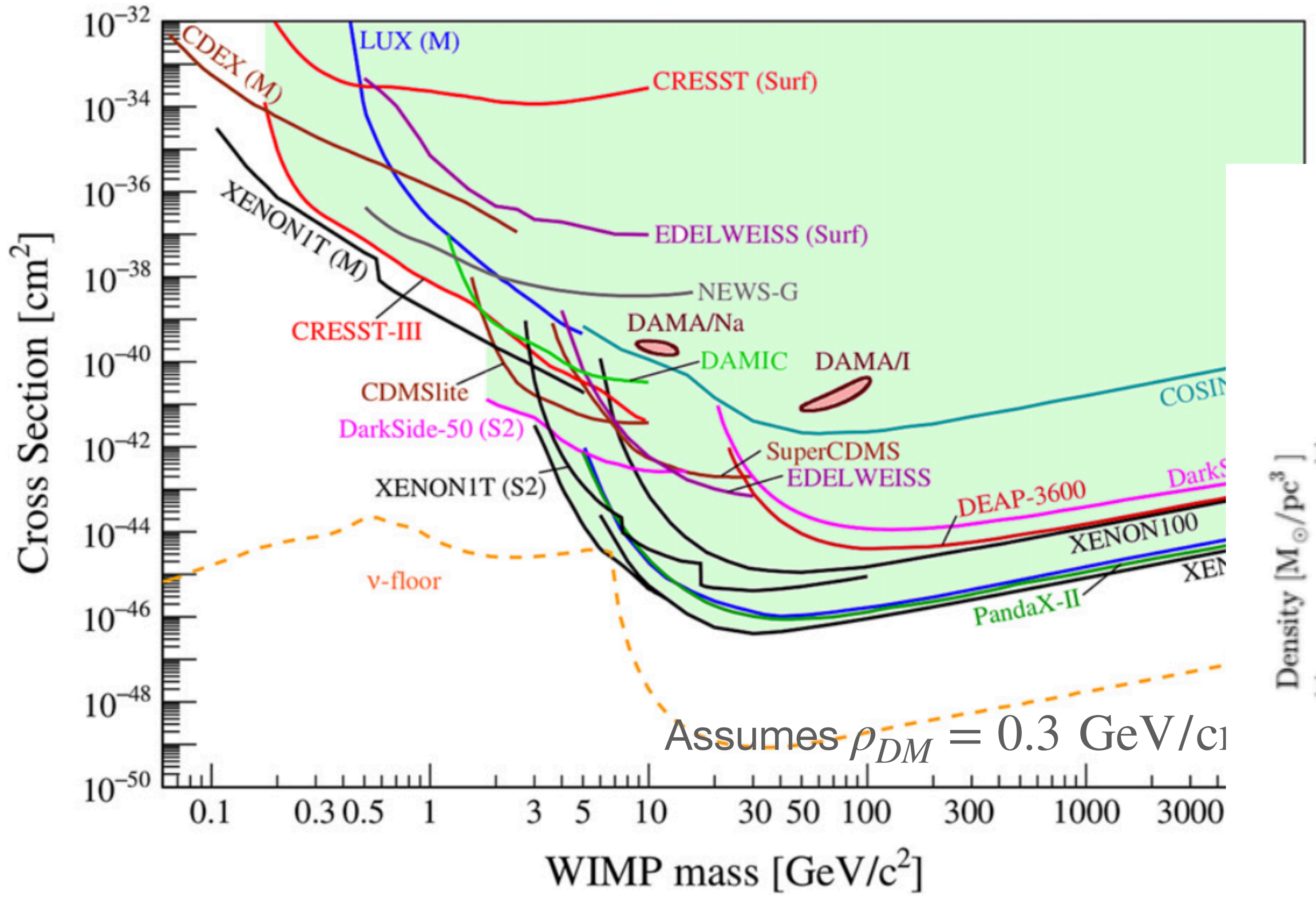
Buckley, Lim, Putney & **DS** 2205.01129, 2305.13358
Green et al 2011.04673, 2205.02244, Naik et al 2112.07657, An et al 2106.05981

- We realized that training normalizing flows on the Gaia dataset could have other interesting applications
- The full 6D phase space density $p(\vec{x}, \vec{v})$ of all the stars in the Galaxy (or at least all the nearby ones) carries a wealth of information about Galactic dynamics.
- In particular, we can directly infer the mass density $\rho(\vec{x})$ of the Galaxy from knowledge of $p(\vec{x}, \vec{v})$, and from that the mass density $\rho_{DM}(\vec{x})$ of the dark matter.

Local dark matter density

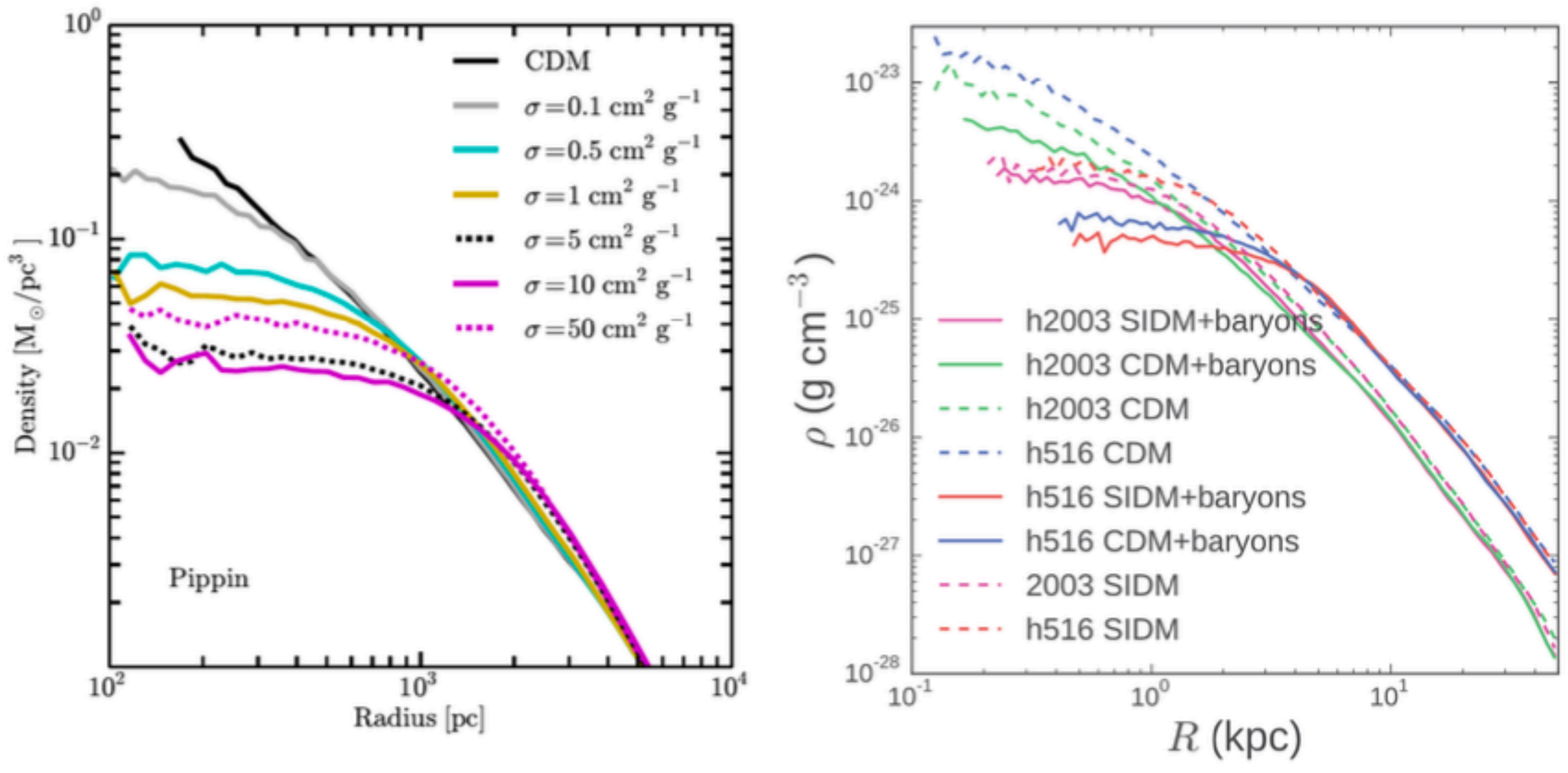
Knowing the local dark matter density $\rho_{DM}(x)$ is very important for many reasons:

from 2104.07634



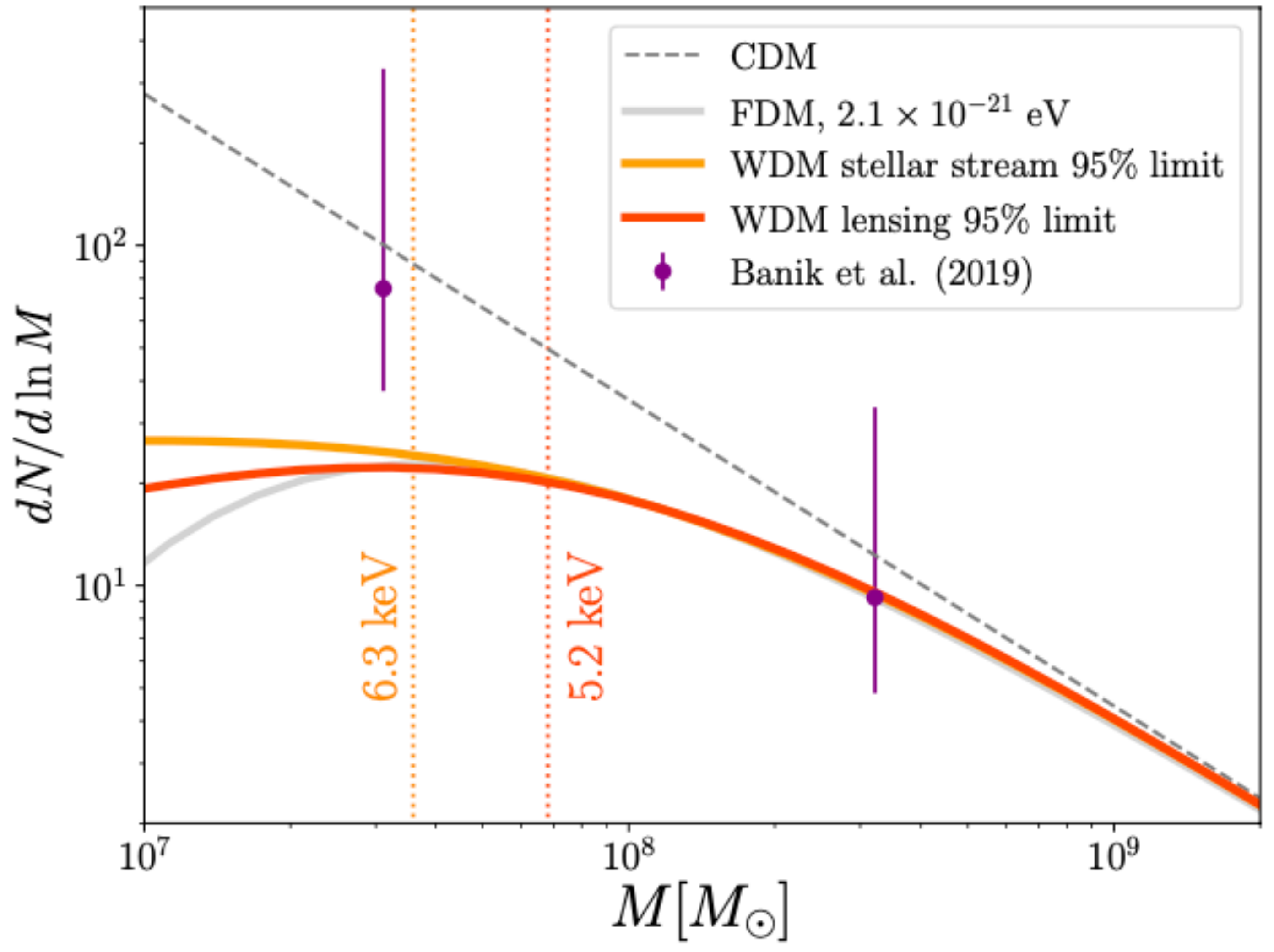
Input to direct detection experiments

from 1812.00044



Make contact with models (NFW, Einasto, etc) and simulations, learn more about Galaxy formation and nature of dark matter

from 2001.05503



Could potentially resolve the presence of dark matter substructure

Idea: mass density from phase space density

Buckley, Lim, Putney & **DS** 2205.01129, 2305.13358
Green et al 2011.04673, 2205.02244, Naik et al 2112.07657, An et al 2106.05981

- Liouville theorem: phase space density is conserved
- Stars are well-approximated as collisionless, only interacting through long-ranged gravitational force
- So they must obey the **collisionless Boltzmann equation**:

$$\left[\cancel{\frac{\partial}{\partial t}} + \vec{v} \cdot \frac{\partial}{\partial \vec{x}} + \vec{a}(\vec{x}) \cdot \frac{\partial}{\partial \vec{v}}\right] p(\vec{x}, \vec{v}; t) = 0$$

Dynamical equilibrium
(expected to be approximately valid)

Accelerations: $\vec{a}(\vec{x}) = -\nabla \Phi(\vec{x})$

$\Phi(\vec{x})$: gravitational potential of the Galaxy (DM+stars+gas…)

Idea: mass density from phase space density

Buckley, Lim, Putney & **DS** 2205.01129, 2305.13358
Green et al 2011.04673, 2205.02244, Naik et al 2112.07657, An et al 2106.05981

$$\left[\vec{v}\cdot\frac{\partial}{\partial\vec{x}}+\vec{a}(\vec{x})\cdot\frac{\partial}{\partial\vec{v}}\right]p(\vec{x},\vec{v})=0$$

- Just from knowledge of $p(\vec{x},\vec{v})$ and its derivatives we can determine the accelerations $\vec{a} = -\nabla\Phi$
- Taking another derivative gives us the mass density of the Galaxy!

$$4\pi G\rho = \nabla^2\Phi = \nabla\cdot\vec{a}$$

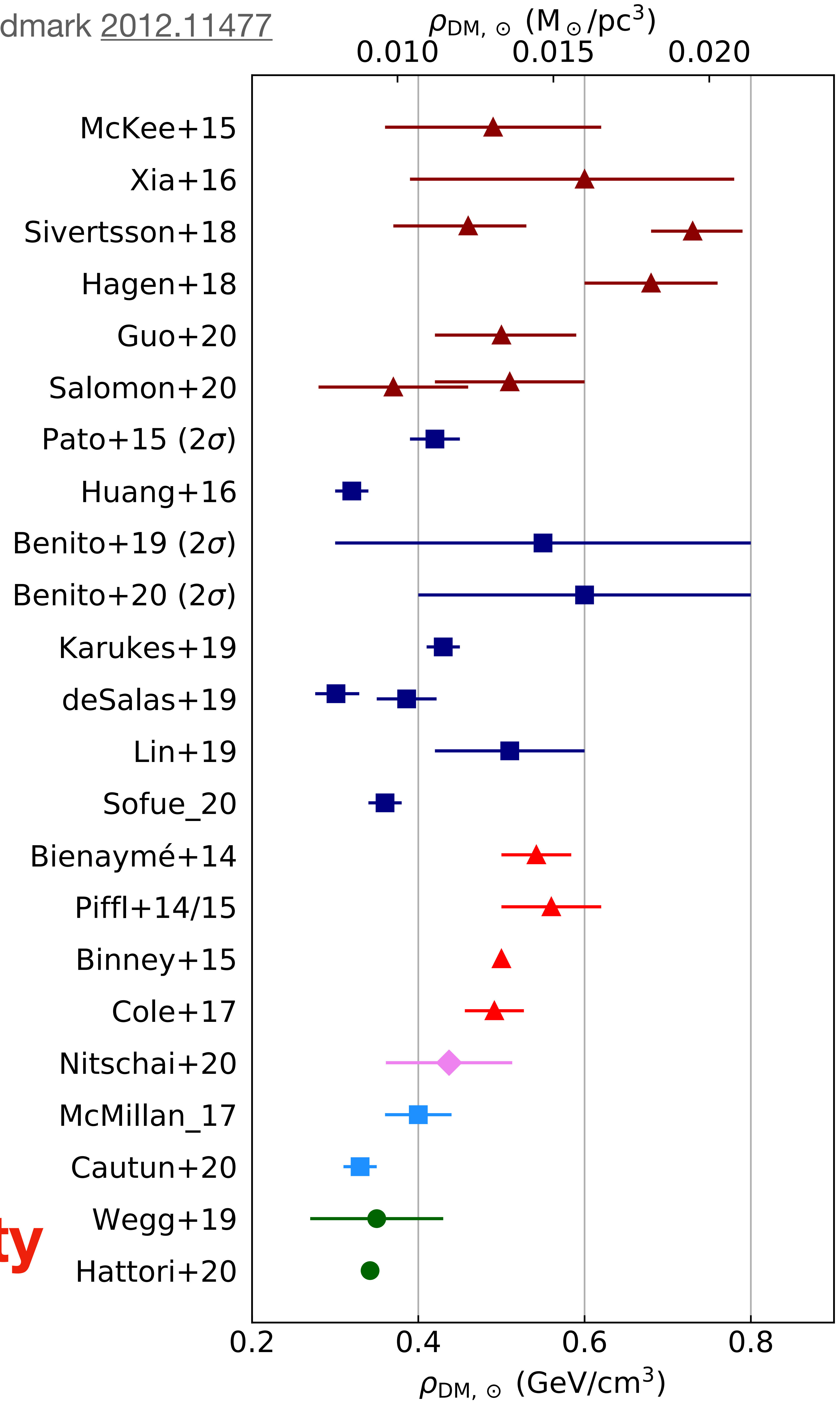
From de Salas & Widmark 2012.11477

Comparison with previous approaches

- Existing measurements typically use **Jean's equation** (second moment of Boltzmann equation) or **rotation curves**
- **They make many assumptions** (axisymmetry, reflection symmetry, simple parametric models…) and **bin the data**
- Results can seem precise but might not be accurate (biased)

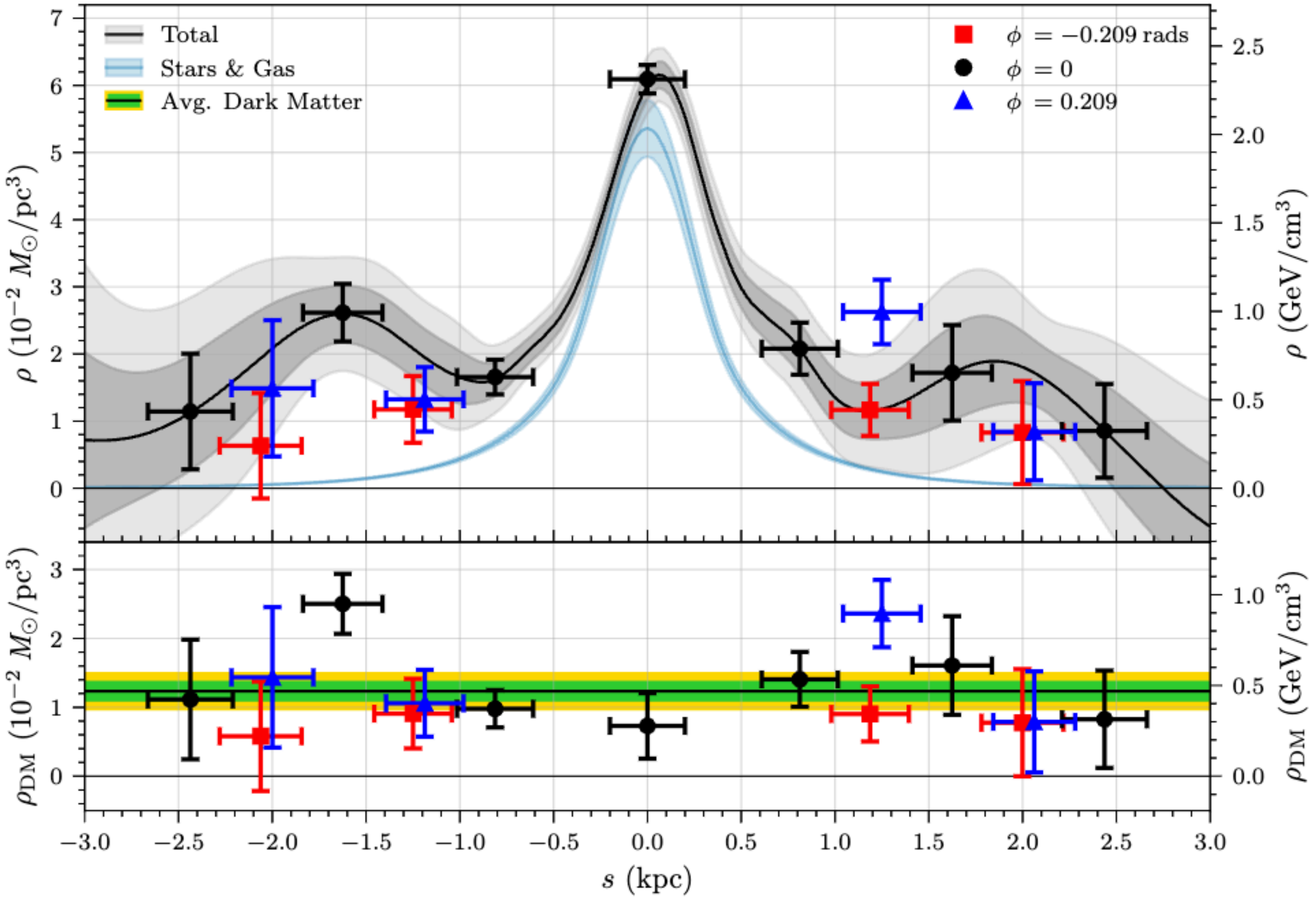
Our approach using normalizing flows is model-free, does not assume symmetries, and is unbinned

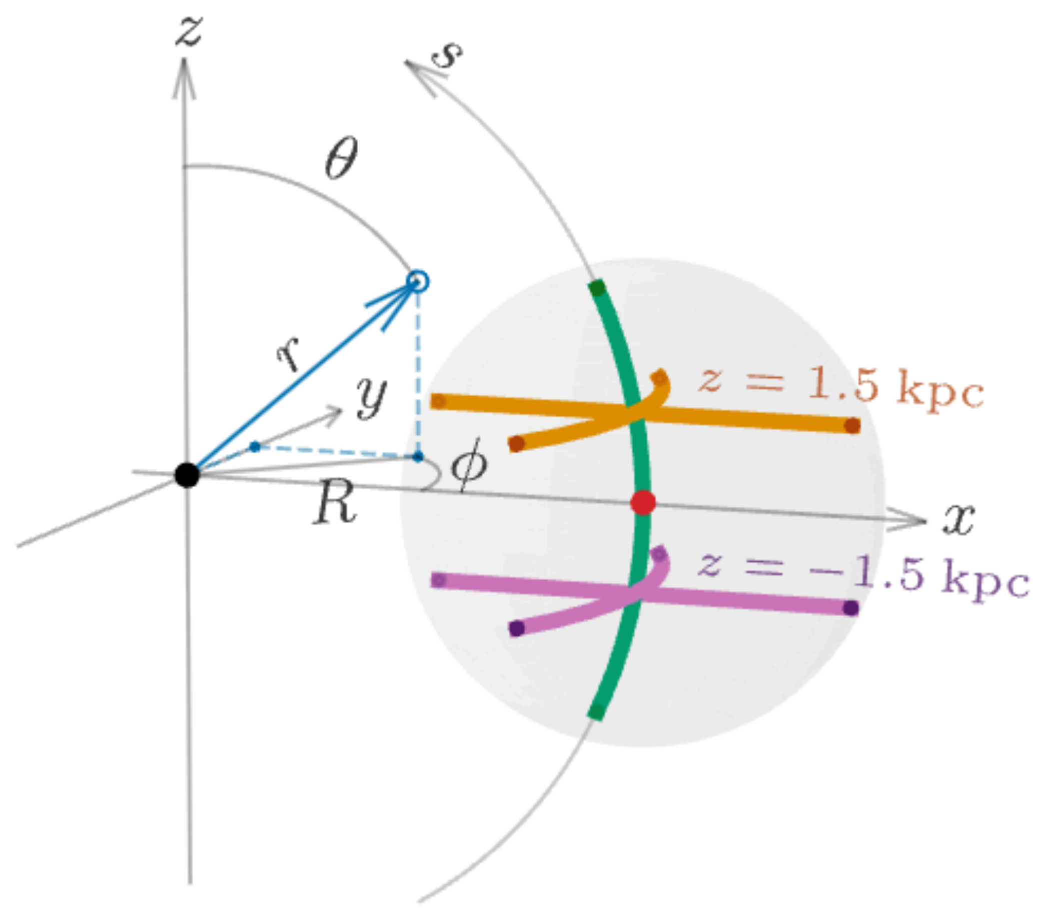
First ever fully 3d measurement of dark matter density in the solar neighborhood



Results: mass density

Lim, Putney, Buckley & **DS** 2305.13358





Density	$(10^{-2}\ M_\odot/\mathrm{pc}^3)$	$(\mathrm{GeV/cm}^3)$	χ^2_ν
$\rho_\odot$	6.17 ± 0.20	2.34 ± 0.08	
$\rho_{b,\odot}$	5.34 ± 0.42	2.03 ± 0.16	
$\rho_{\mathrm{DM},\odot}$	0.83 ± 0.47	0.32 ± 0.18	
$\bar{\rho}_{\mathrm{DM}}(r = r_\odot)$	1.18 ± 0.14	0.47 ± 0.05	1.38

Result is consistent with nonzero, spherically symmetric DM density!

Error bars include:

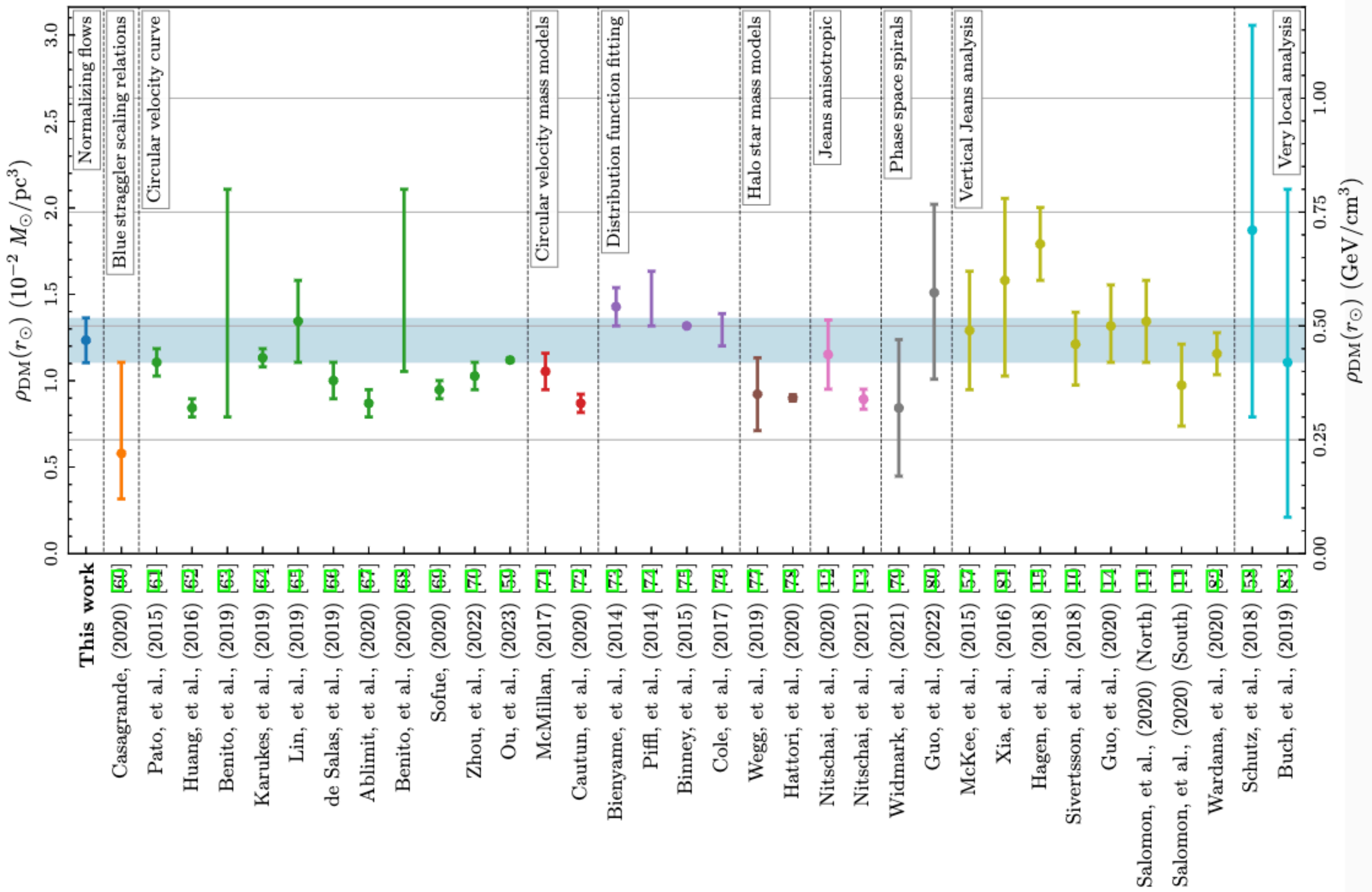
- MAF training variance
- Gaia measurement error
- Finite training statistics

Results: mass density

Lim, Putney, Buckley & **DS** 2305.13358

Our result:

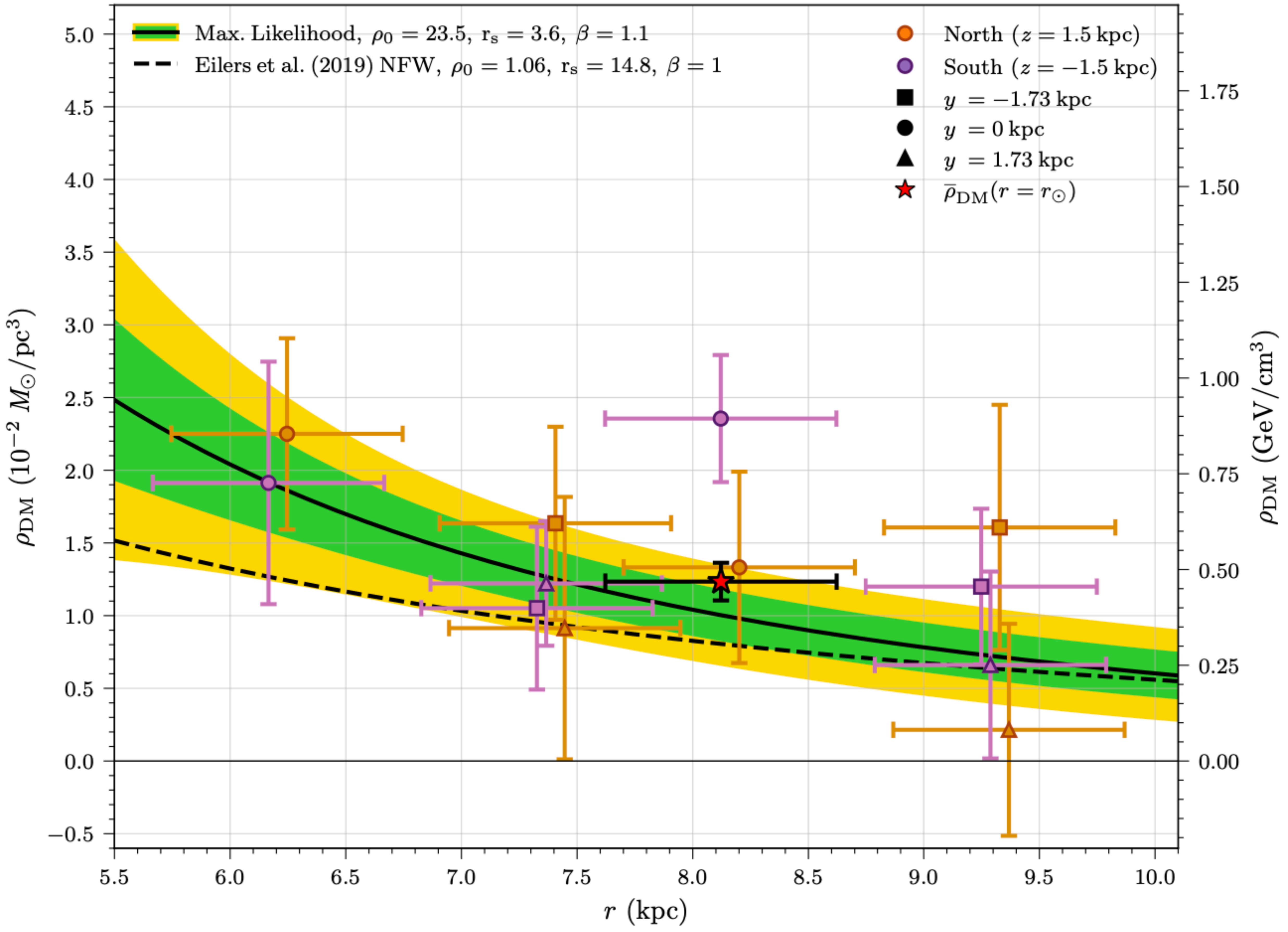
$$\rho_{DM}(r_\odot) = 0.47 \pm 0.05 \ \mathrm{GeV/cm^3}$$

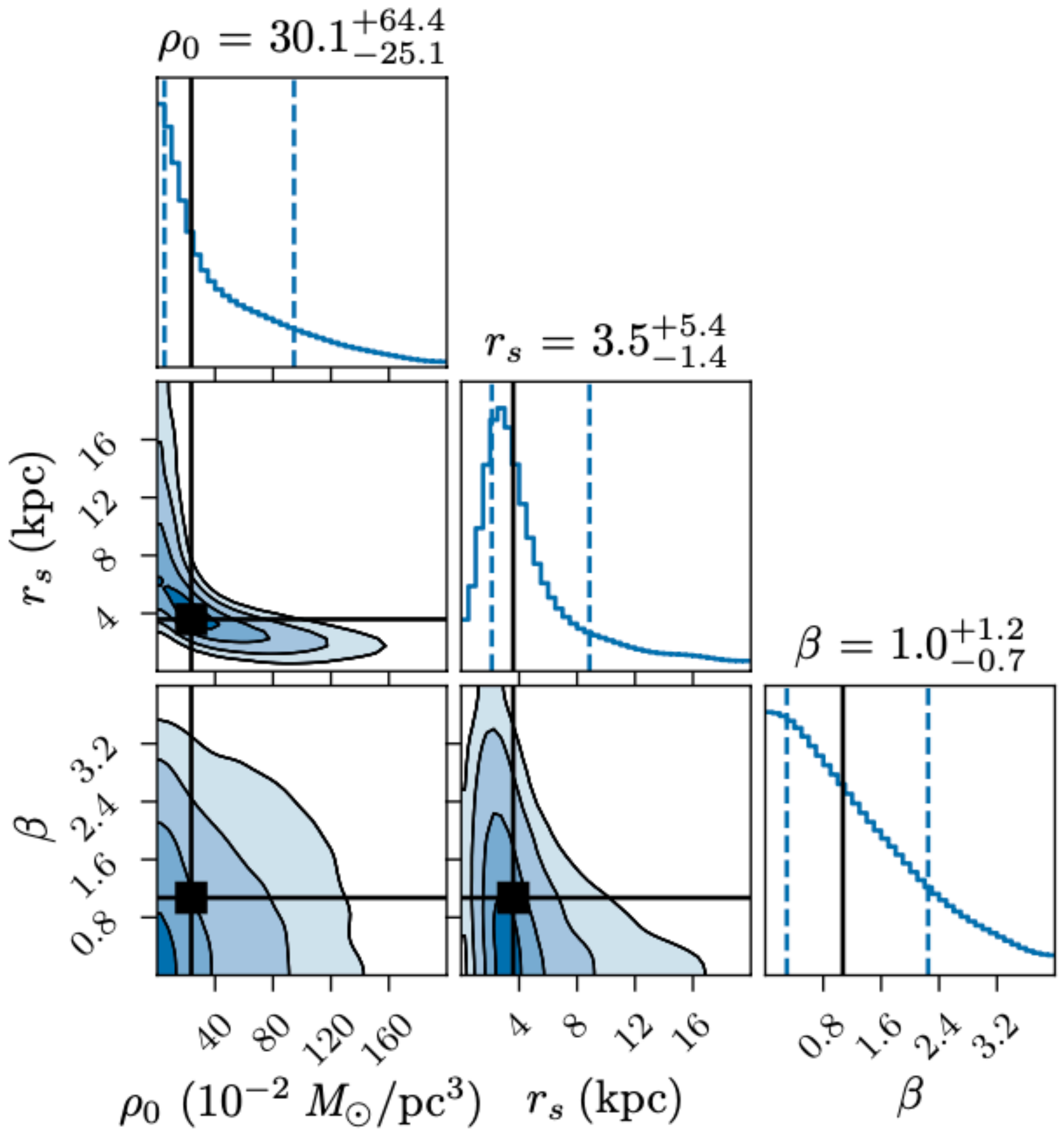


Excellent agreement with previous measurements, with hopefully more realistic error bars

Results: mass density

Lim, Putney, Buckley & **DS** 2305.13358





Radial profile broadly consistent with recent NFW fits

Summary/Outlook

- Modern machine learning is a powerful new tool revolutionizing fundamental physics with Big Data.
- There has been an explosion of new methods and proofs-of-concept.
- Many new methods are beginning to be applied to real data in the HEP and Astro/Cosmo domains.
- These are exciting times! New discoveries await!

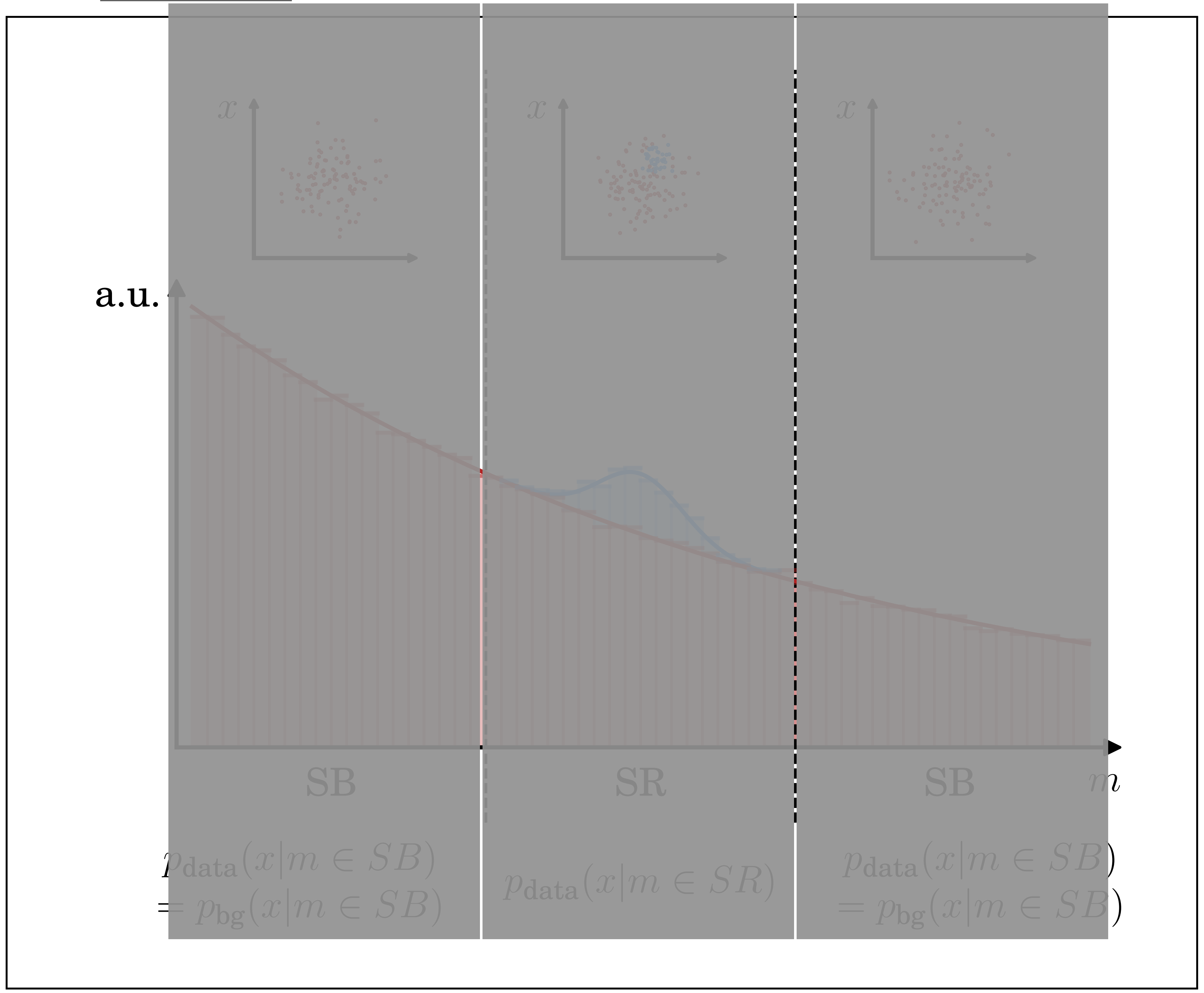
Thanks for your attention!

Backup

Example: Classifying Anomalies THrough Outer Density Estimation (CATHODE)

DS+ Hallin et al 2109.00546, 2210.14924

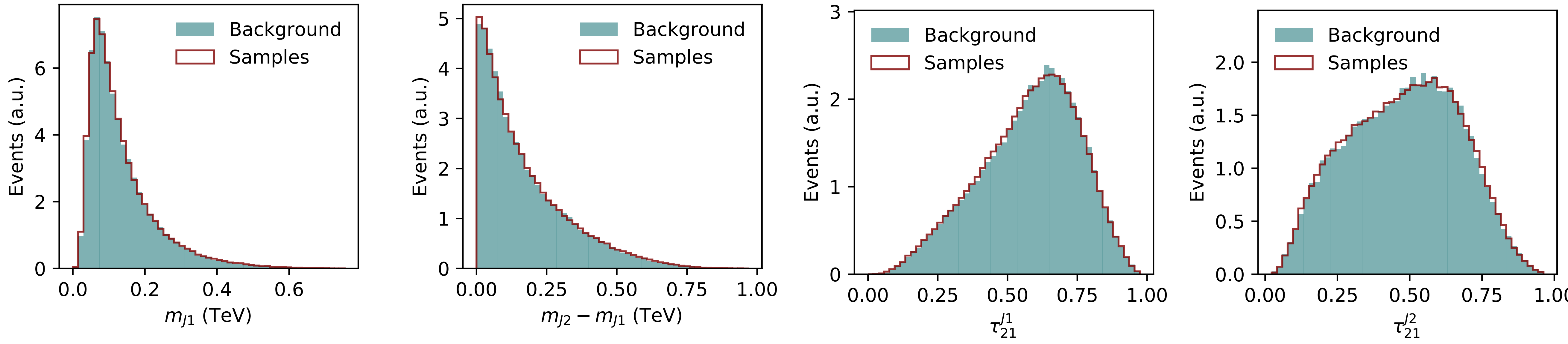
from 2109.00546



1. Train generative model (eg normalizing flow) on ***sidebands*** to learn background model

$$p_{data}(x\,|\,m \in SB) = p_{bg}(x\,|\,m \in SB)$$

2. Sample from this model for $m \in SR$ to obtain synthetic background events in the ***signal region***.



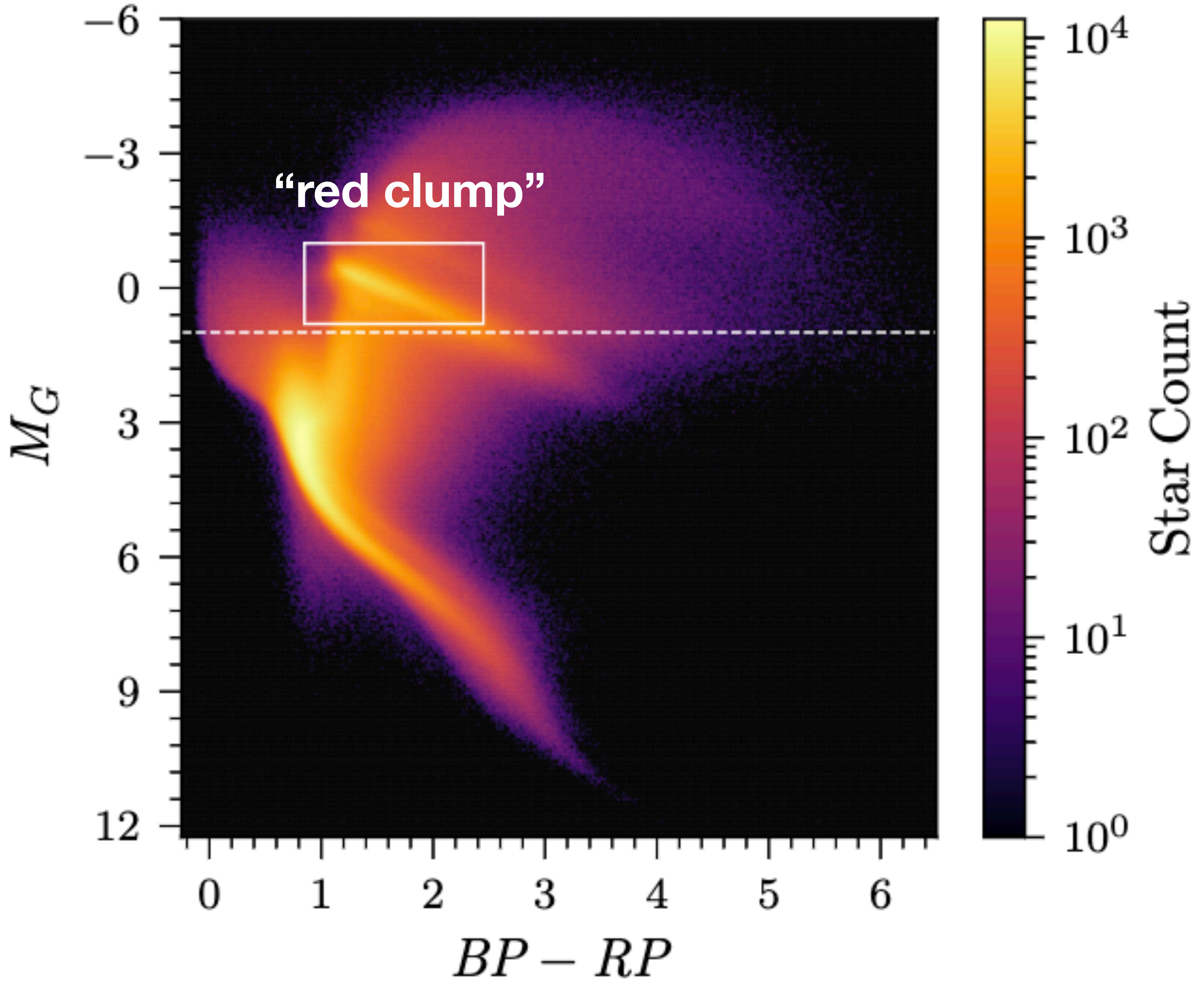
3. Train binary classifier on data vs synthetic background to learn

$$R(x) = \frac{p_{data}(x)}{p_{bg}(x)}$$

From proof-of-concept to real data

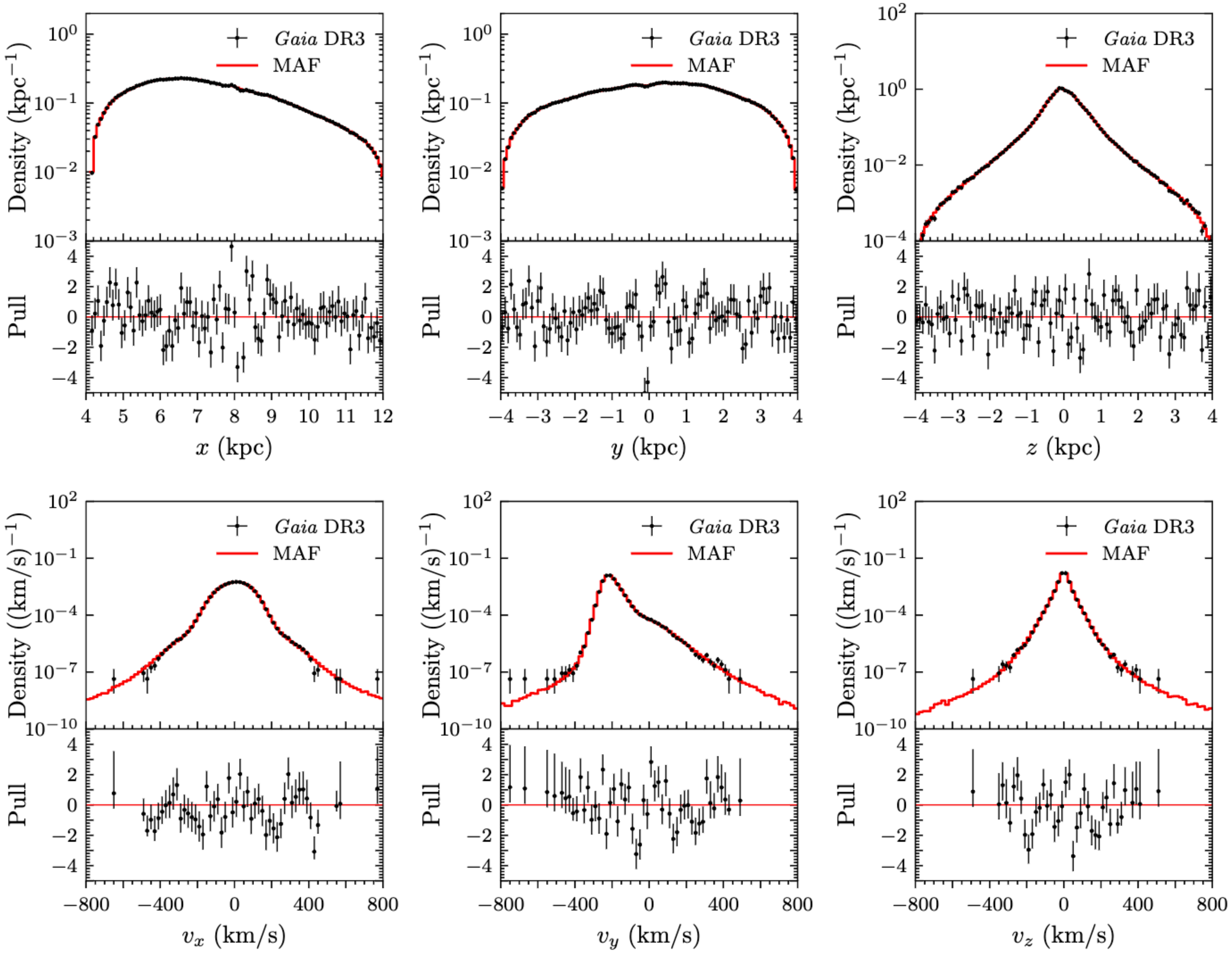
Buckley, Lim, Putney & **DS** 2205.01129, 2305.13358

- After validating our method with a realistic hydrodynamical cosmological simulation, we applied it to Gaia DR3.
- Selected stars in Gaia DR3 within 4 kpc with
  - full 6d features
  - brightness cut to ensure completeness
- dominated by **“red clump” stars** which are supposed to be a good equilibrium tracer population => **5.8M stars**



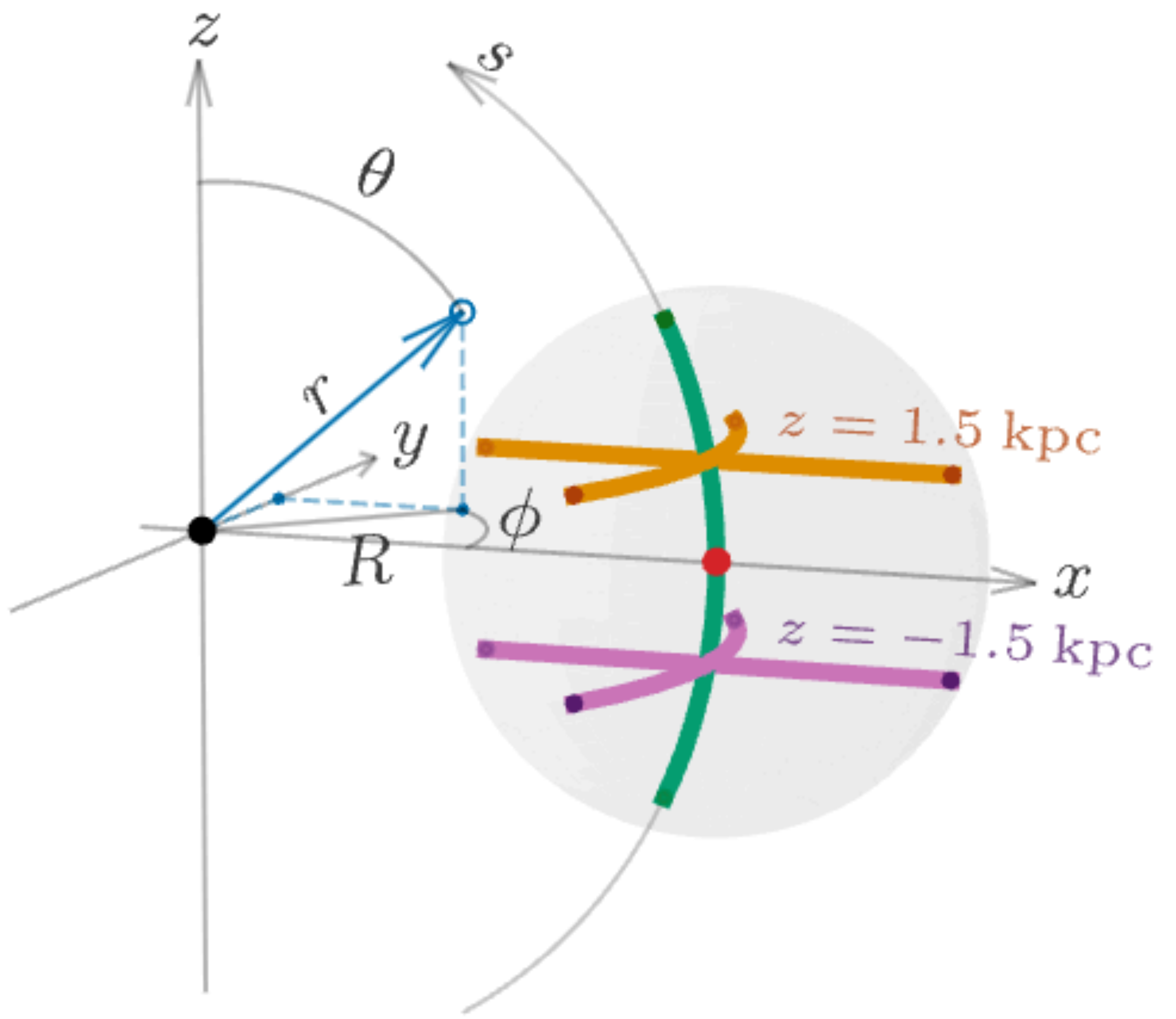
Results: density estimation

Lim, Putney, Buckley & **DS** 2305.13358



Results: accelerations

Lim, Putney, Buckley & **DS** 2305.13358



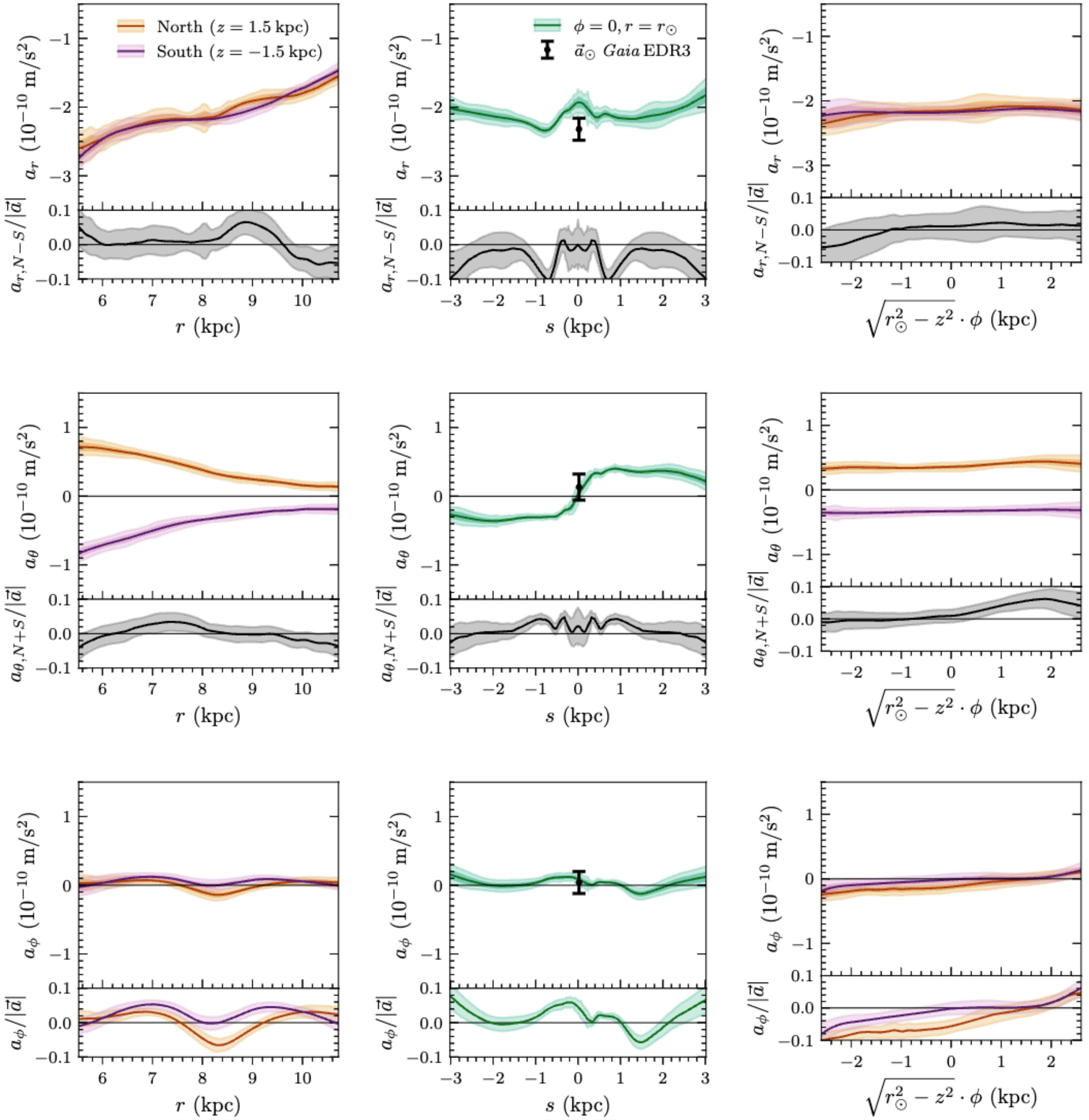
Symmetries to ~10% level:

- north-south
- azimuthal (phi)

=> Expected from dynamical equilibrium

	Gaia EDR3 [56]	**This work**
a_x $(10^{-10}\mathrm{m/s^2})$	-2.32 ± 0.16	-1.94 ± 0.22
a_y $(10^{-10}\mathrm{m/s^2})$	0.04 ± 0.16	0.08 ± 0.08
a_z $(10^{-10}\mathrm{m/s^2})$	-0.14 ± 0.19	-0.06 ± 0.08
$\lvert\vec{a}\rvert$ $(10^{-10}\mathrm{m/s^2})$	2.32 ± 0.16	1.94 ± 0.22

TABLE I: Galactic acceleration at the Solar location $\vec{a}_\odot$ in Cartesian coordinates, calculated by averaging the solution to the Boltzmann equation within a 100 pc sphere centered on the Sun. We list for comparison the acceleration at the Solar location obtained from *Gaia* DR3 quasar measurements [56].



More on determining the accelerations

Buckley, Lim, Putney & **DS** 2205.01129, 2305.13358
Green et al 2011.04673, 2205.02244, Naik et al 2112.07657, An et al 2106.05981

$$\left[\vec{v}\cdot\frac{\partial}{\partial\vec{x}}+\vec{a}(\vec{x})\cdot\frac{\partial}{\partial\vec{v}}\right]p(\vec{x},\vec{v})=0$$

- How can we solve for 3 acceleration functions $\vec{a}(\vec{x})$ with just a single equation?
- $\vec{a}(\vec{x})$ doesn't depend on velocity! So this is actually an infinite number of equations for $\vec{a}(\vec{x})$, one for each choice of $\vec{v}$
- We choose to perform least-squares minimization over a sample of velocities to determine best-fit $\vec{a}(\vec{x})$

$$L(\vec{a}(\vec{x}))=\frac{1}{N}\sum_{\alpha=1}^{N}\left(\left[\vec{v}_\alpha\cdot\frac{\partial}{\partial\vec{x}}+\vec{a}(\vec{x})\cdot\frac{\partial}{\partial\vec{v}}\right]p(\vec{x},\vec{v}_\alpha)\right)^2$$

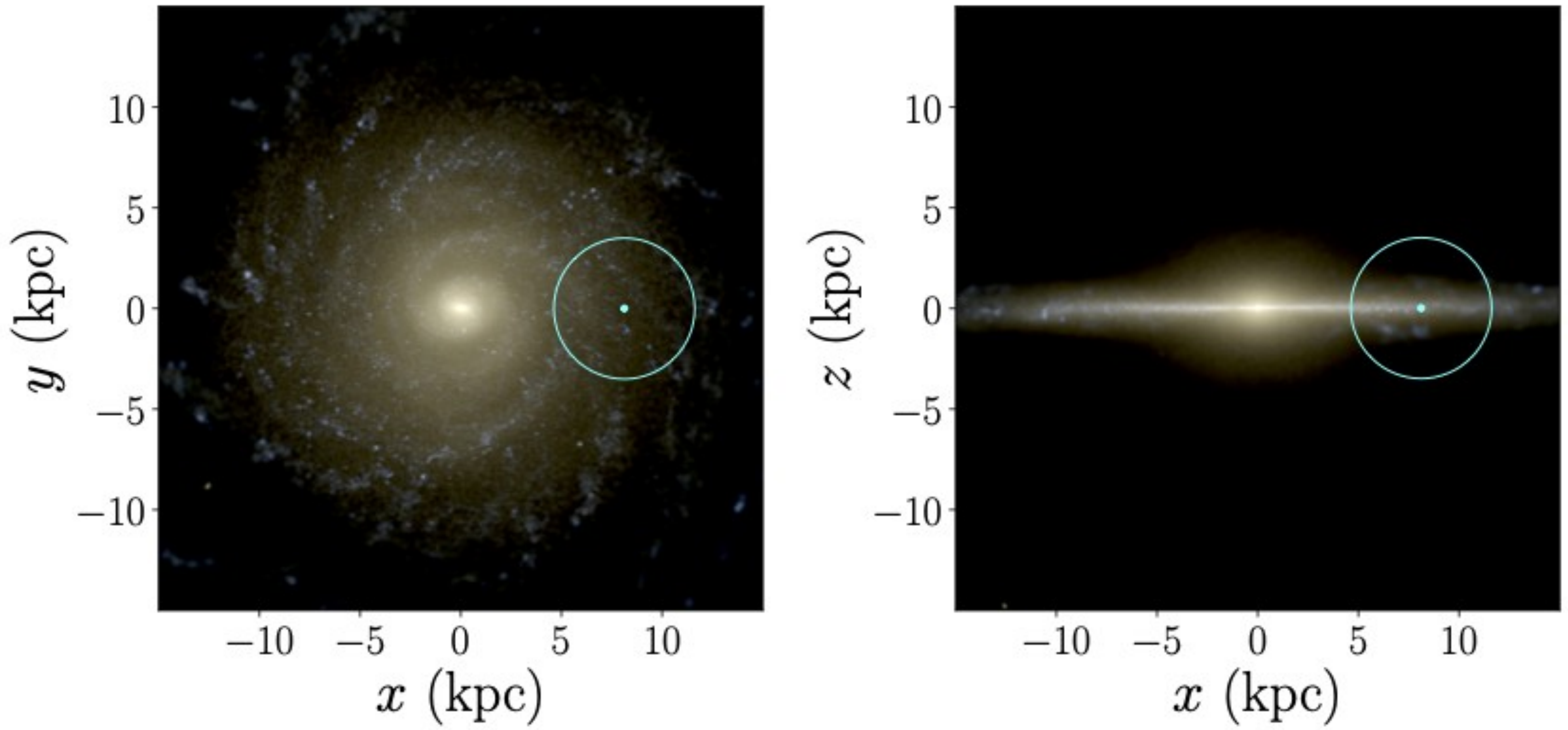
Our work: first to use Nbody+hydro simulation

Proof-of-concept

Buckley, Lim, Putney & **DS** 2205.01129

- Training data: state-of-the-art Nbody+hydro galaxy simulation from “Nbody shop” collaboration [https://b2share.eudat.eu/records/c9f232d8ac804785aad35004177a704e]

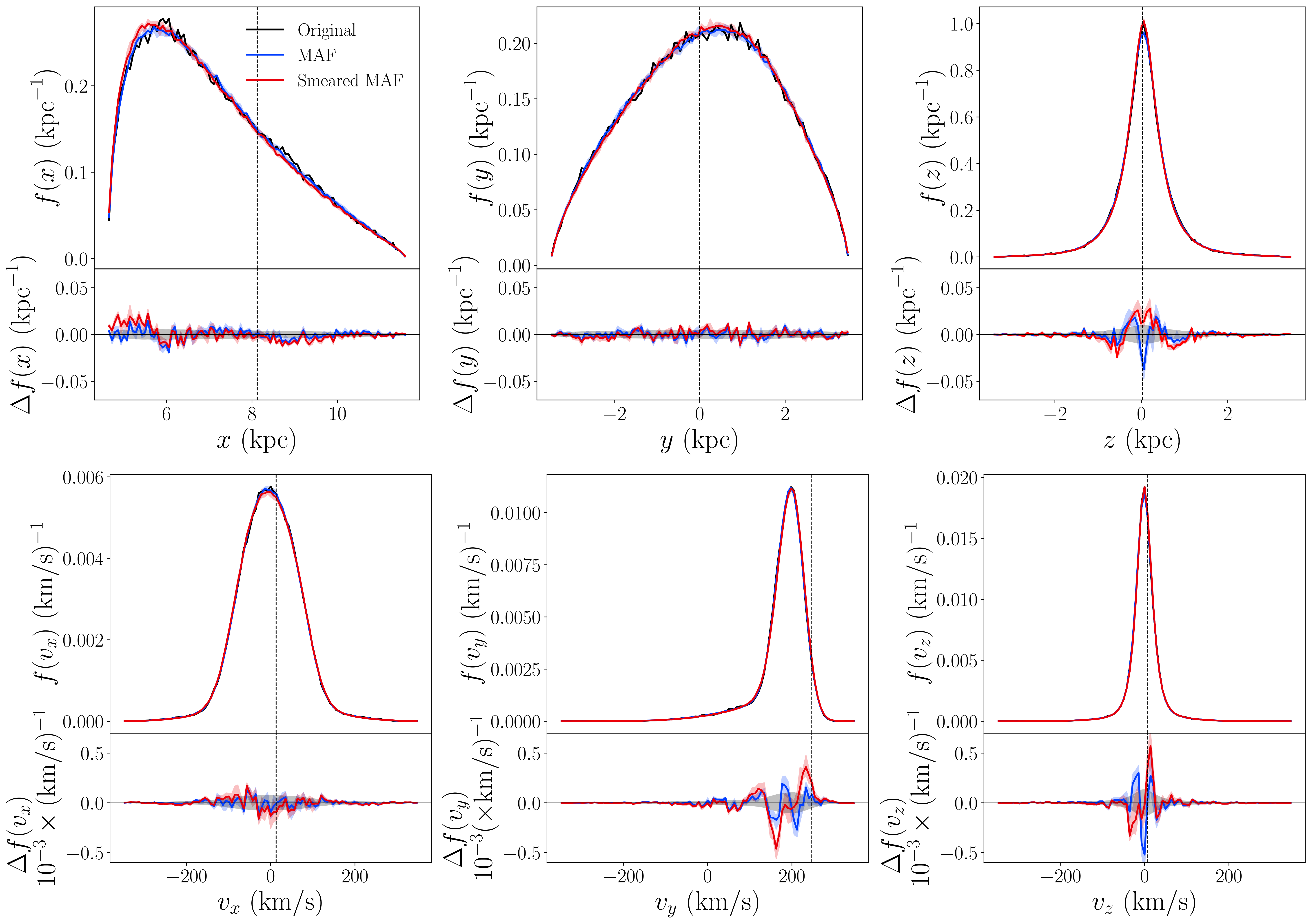
- Milky Way like Galaxy h277



- number of ~~stars~~ star particles
  153,174 (<< size of Gaia 6D dataset)
- observer’s location
  [8.122, 0., 0.0208] kpc
- observing radius = 3.5 kpc
- simulation resolution: 0.173 kpc
- Using only kinematic information:
  position and velocity

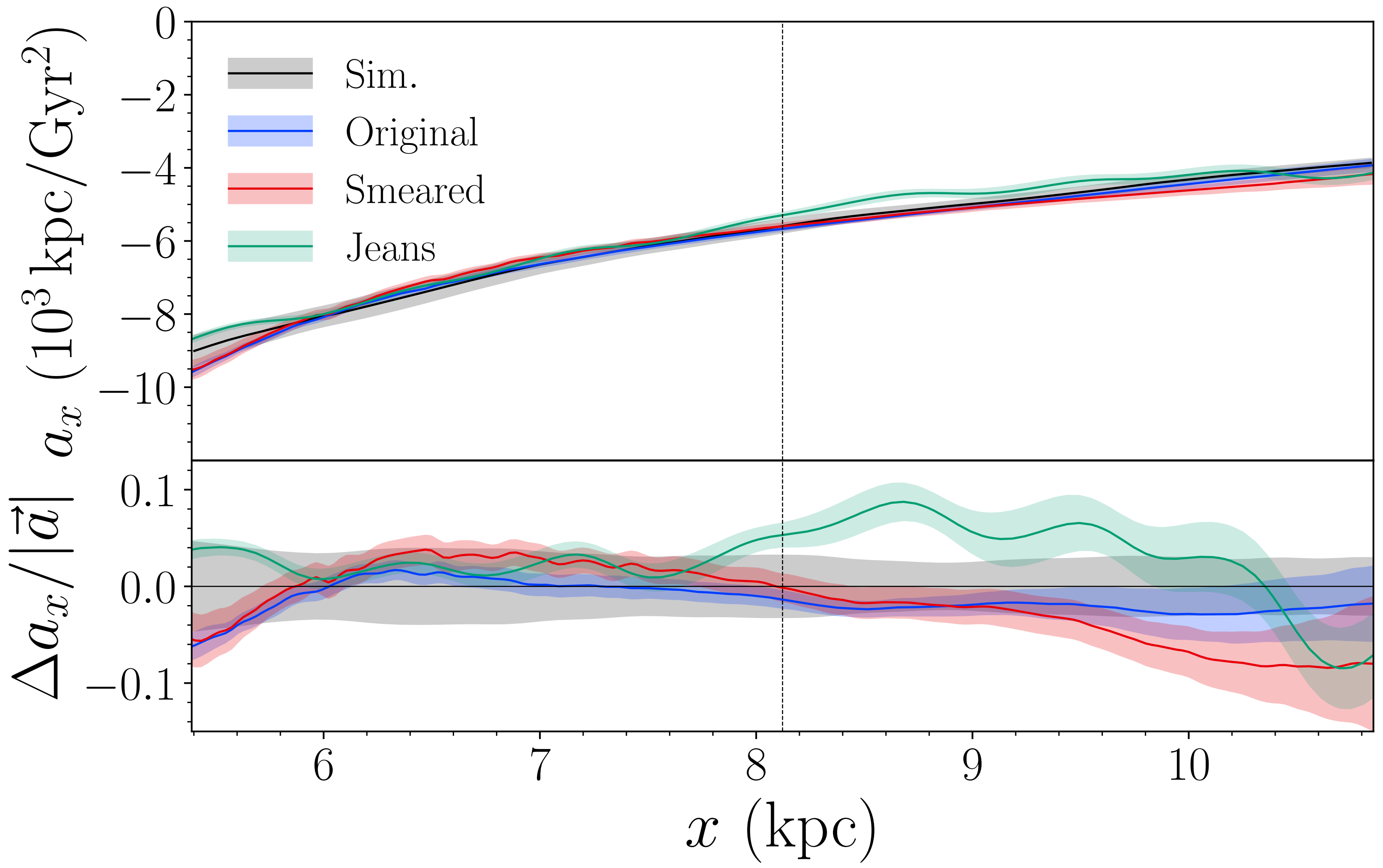
Results: density estimation

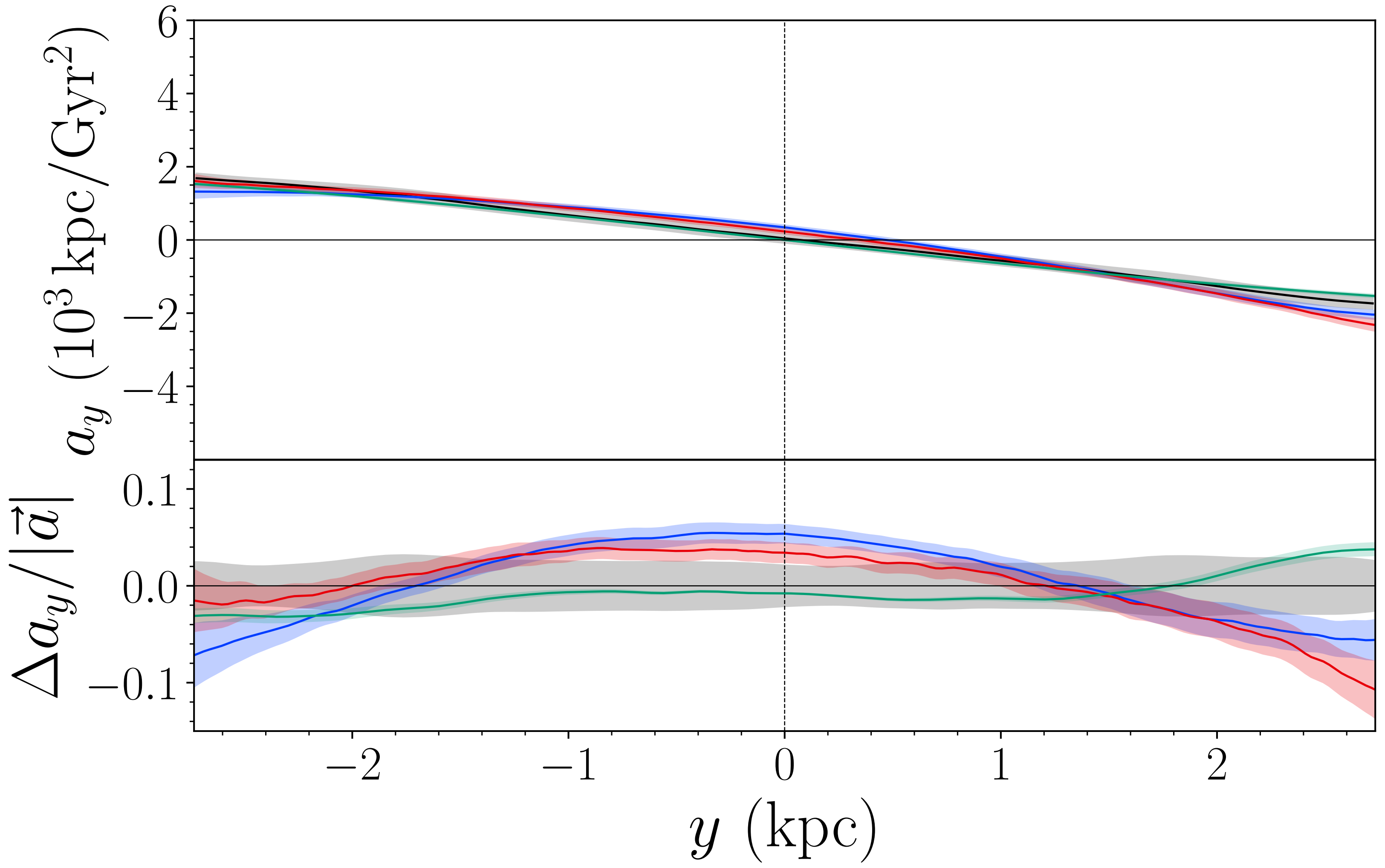
Buckley, Lim, Putney & **DS** 2205.01129

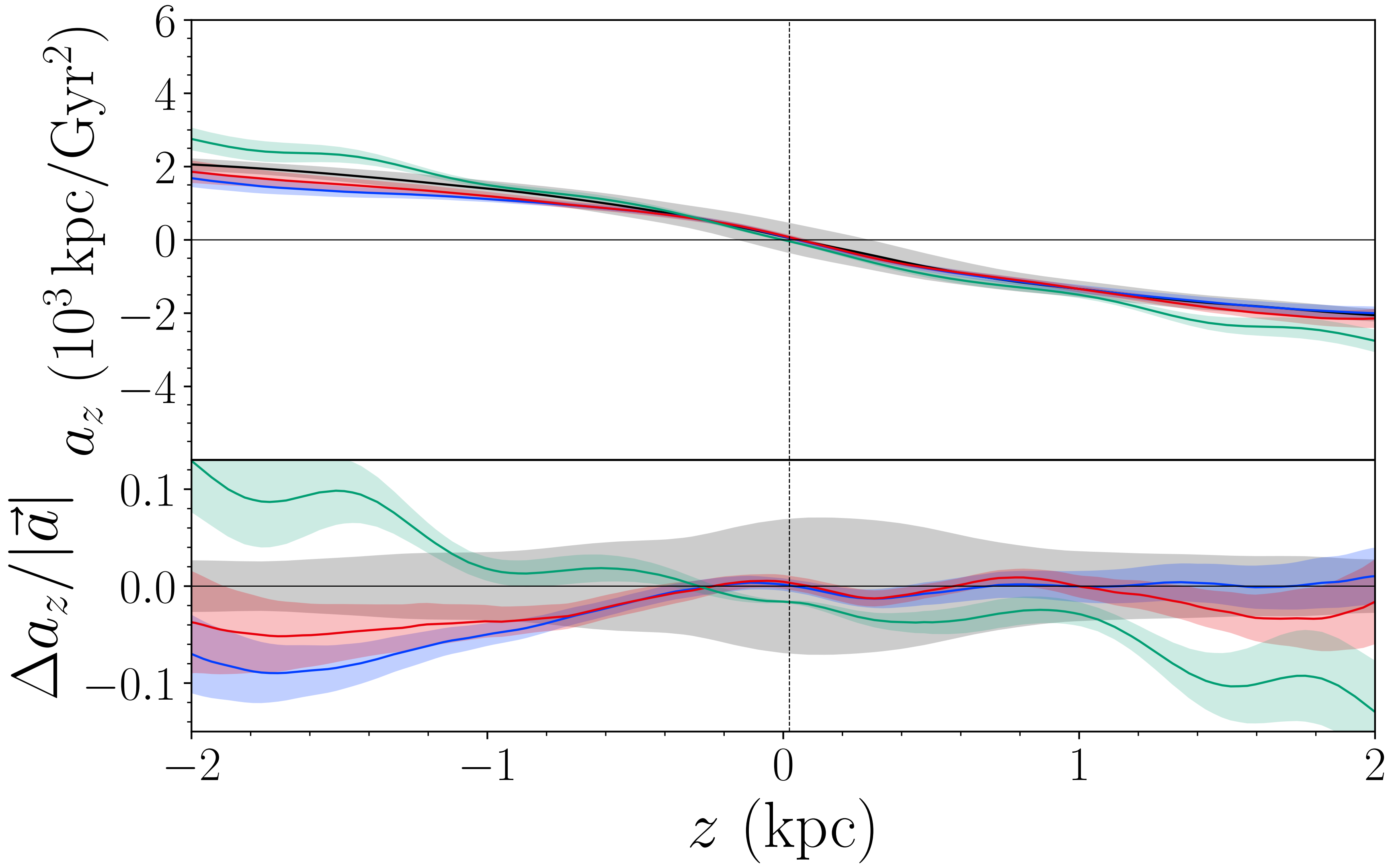


Results: accelerations

Buckley, Lim, Putney & **DS** 2205.01129







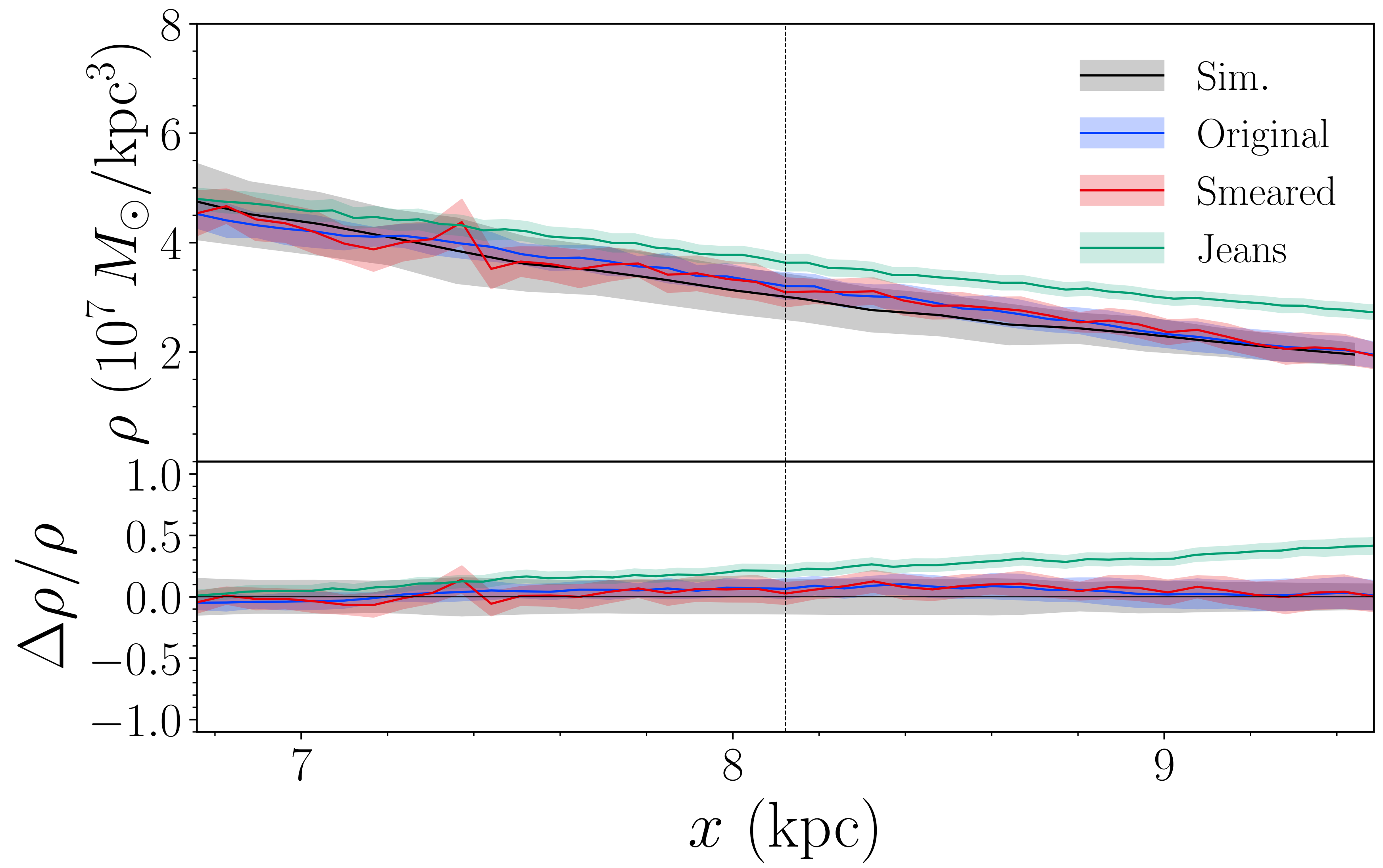
Accelerations to within 5% accuracy!

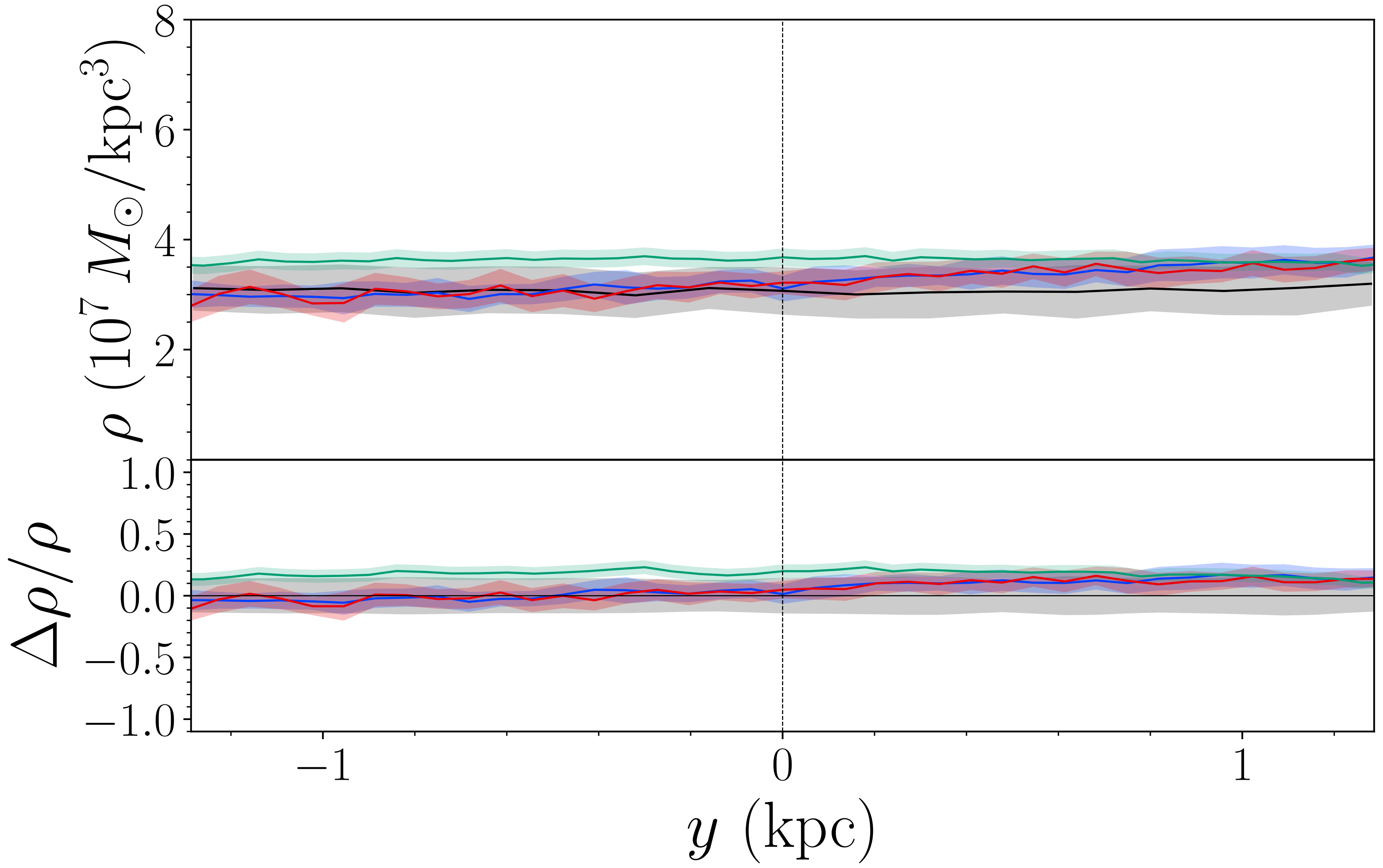
We estimated uncertainties from:

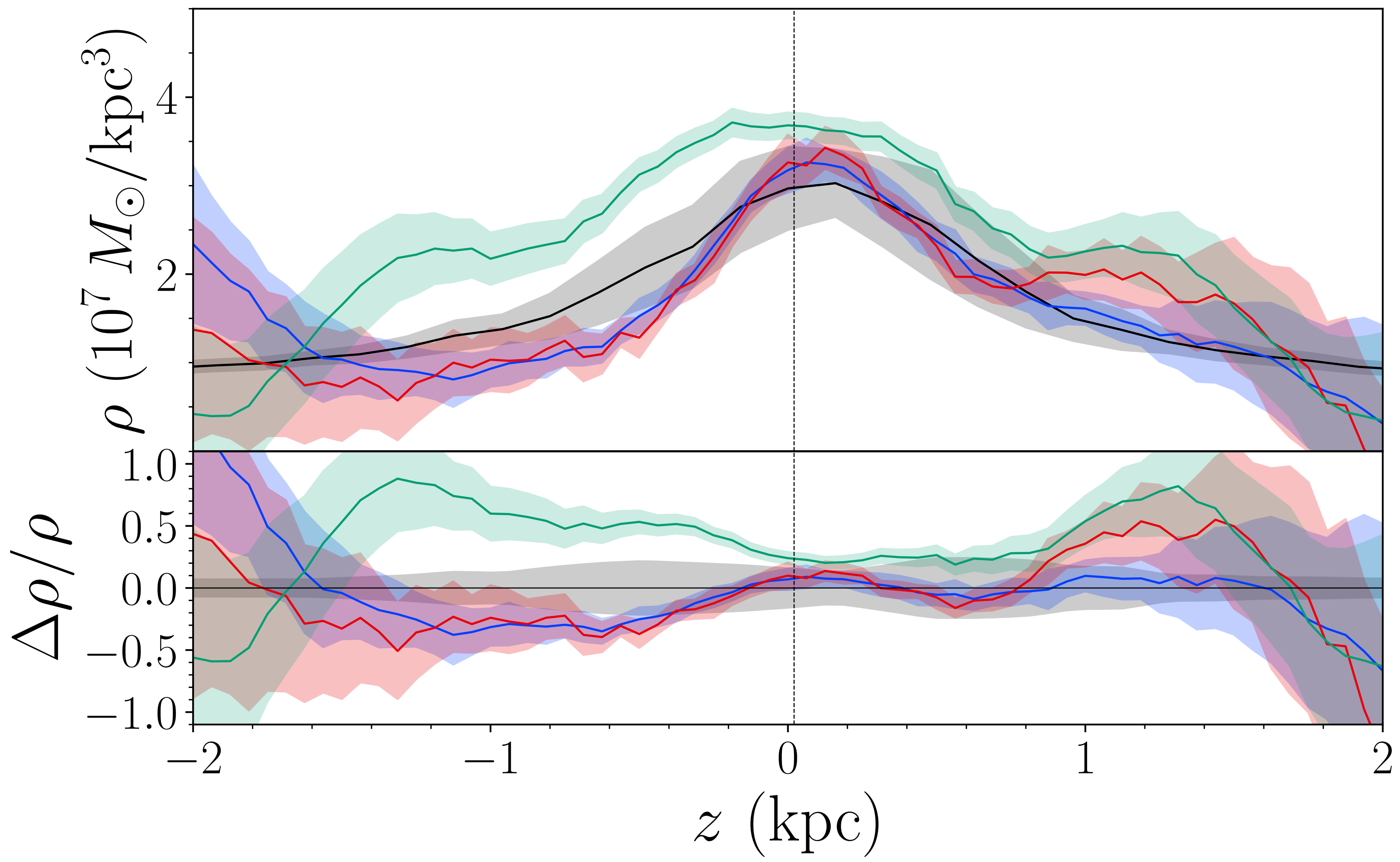
- random training initialization
- finite training data statistics (bootstrap)
- measurement error

Results: mass density

Buckley, Lim, Putney & **DS** 2205.01129







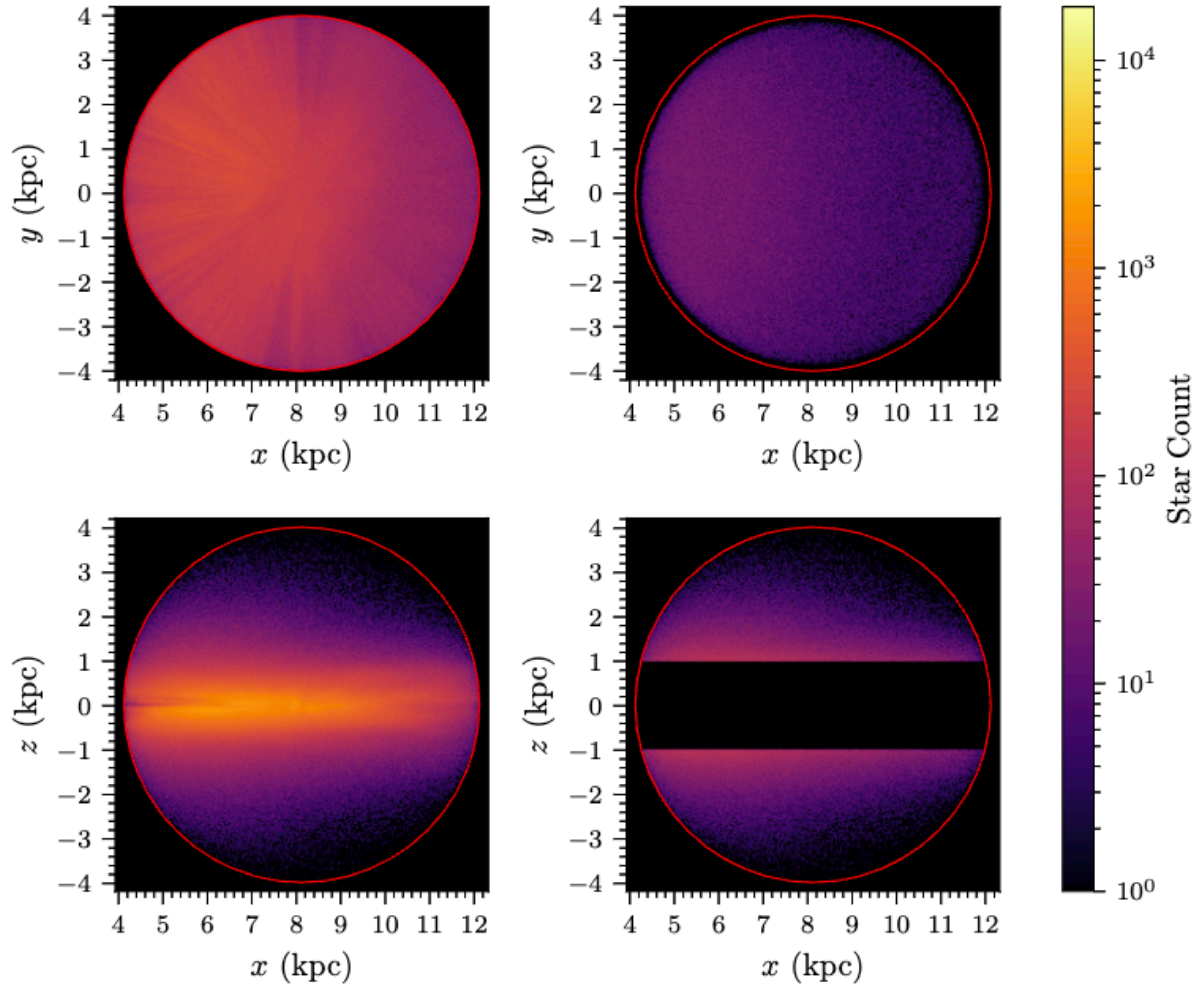
Mass density to within 10-20% accuracy!

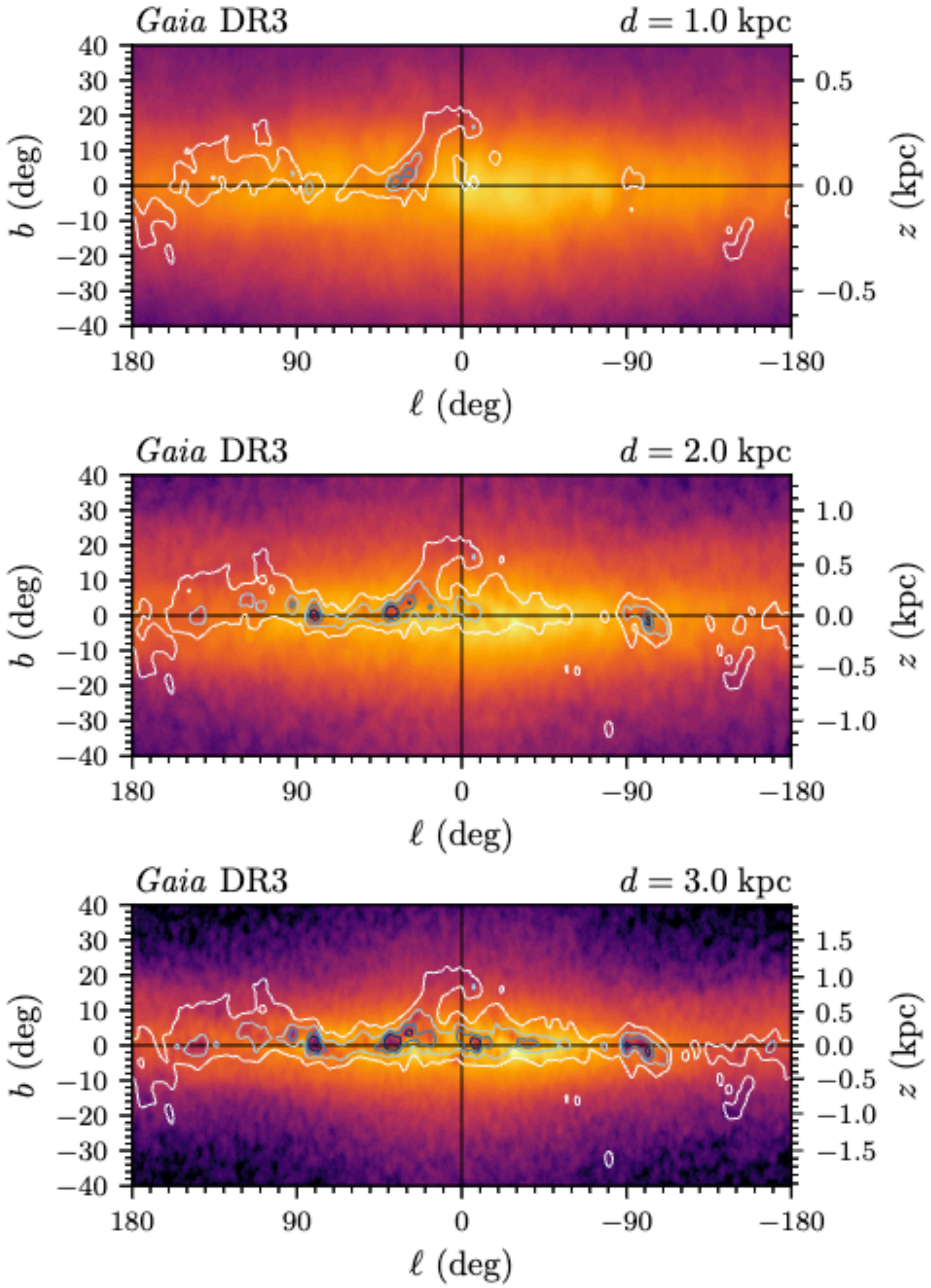
We estimated uncertainties from:

- random training initialization
- finite training data statistics (bootstrap)
- measurement error

The Galaxy is a dusty place

Lim, Putney, Buckley & **DS** 2305.13358





The Galaxy is a dusty place

Lim, Putney, Buckley & **DS** 2305.13358

- In this first work we did not attempt to correct for dust extinction (this is work in progress).
- Rather we explored our results along 1d slices that should avoid the worst of the dust effects.