

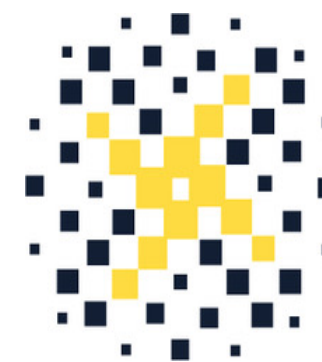
EuCAIFCon 2026 · Heidelberg

Anomaly Detection for Faint Galactic Substructure with EagleEye

Speaker: Sven Pöder

SISSA, Trieste · NICPB, Tallinn

Co-authors: Andre Scaffidi,
Gabriella Contardo,
Roberto Trotta



SISSA
DATASCIENCE
Machine Learning for the Natural Sciences



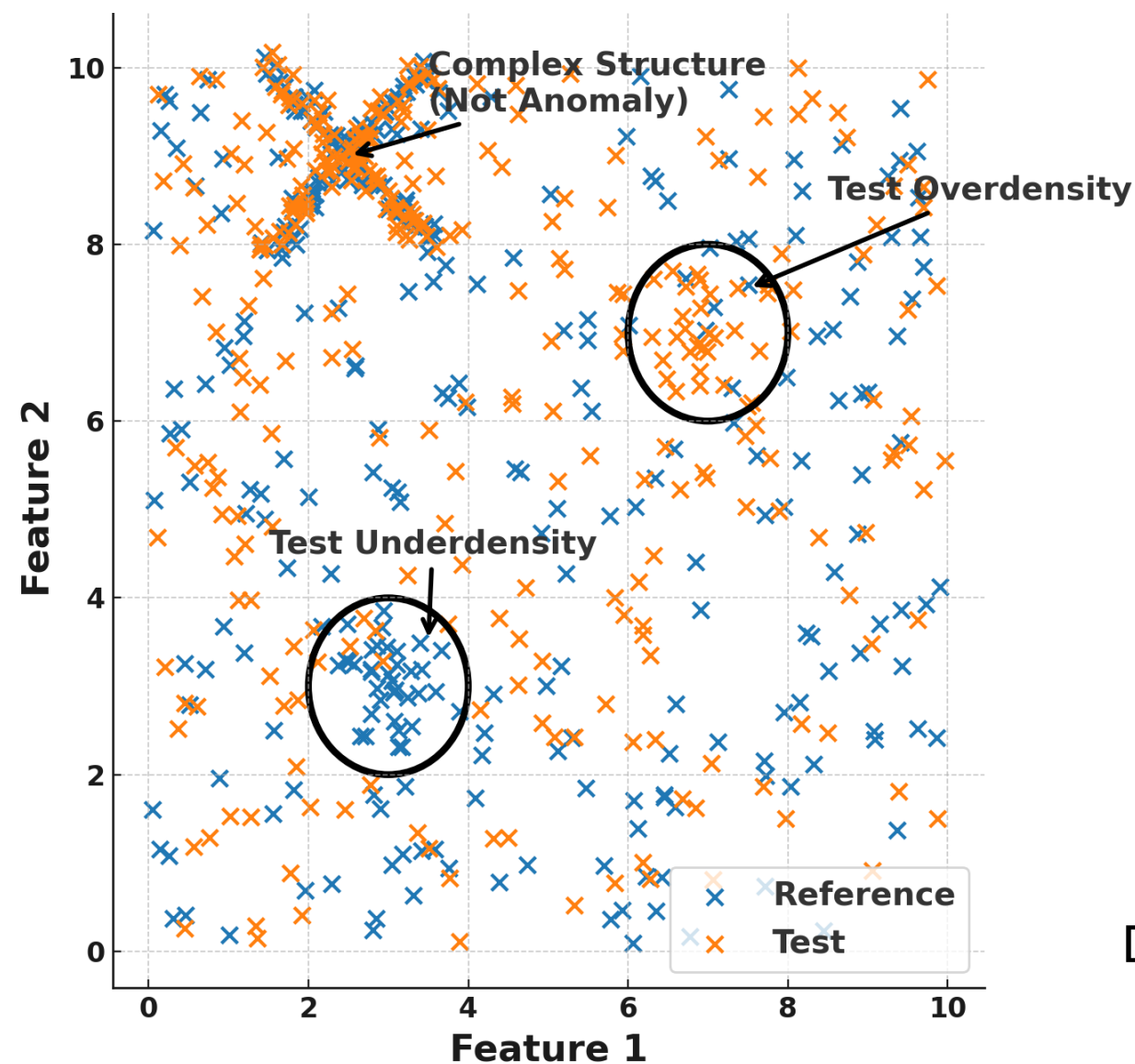
NICPB
KBF



EagleEye is a new kNN-based density comparison method

$\mathcal{X}$ = 'Reference' set $\mathcal{X}$

$\mathcal{Y}$ = 'Test' set $\mathcal{Y}$



BASED ON A SIMPLE PREMISE

Look — check each point's nearest neighbours in combined set $\mathcal{X} \cup \mathcal{Y}$

Count — how many are test vs. reference?

$$b_i^k = \begin{cases} 0, & \text{if } k\text{-th neighbor of } Y_i \in \mathcal{X}, \\ 1, & \text{if } k\text{-th neighbor of } Y_i \in \mathcal{Y} \end{cases}$$

Compare — under $\mathcal{H}_0$, this split should look like a fair coin toss

$$\mathcal{H}_0 : \sum_{k=1}^K b_i^k \sim \text{Binomial}$$

Flag — a skewed local count marks a potential anomaly

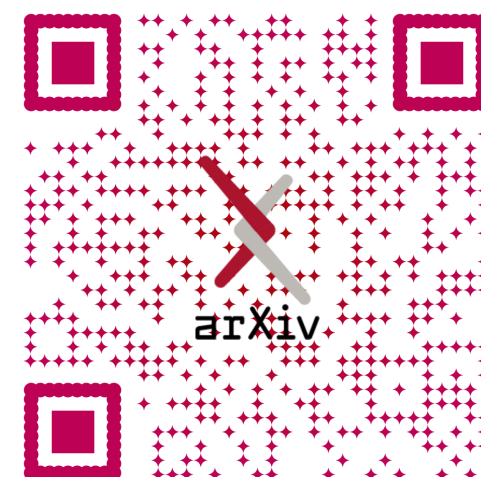
$$\Upsilon_i^+ = \max_{K \leq K_M} \left\{ -\log \left(\text{pval} \left(\sum_{k=1}^K b_i^k \right) \right) \right\}$$

Full procedure explained here

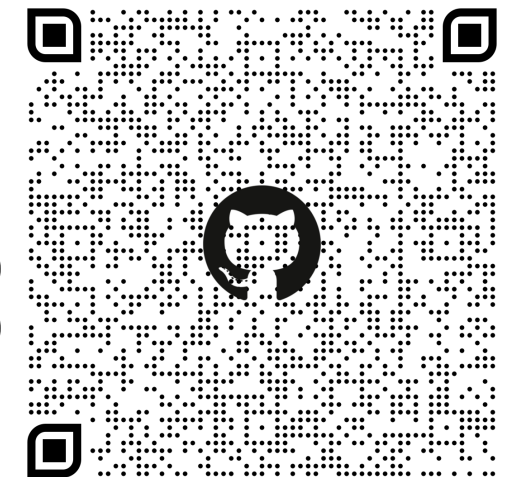
Data-driven signal-to-background estimate

PREPRINT

arXiv:2503.23927



CODE

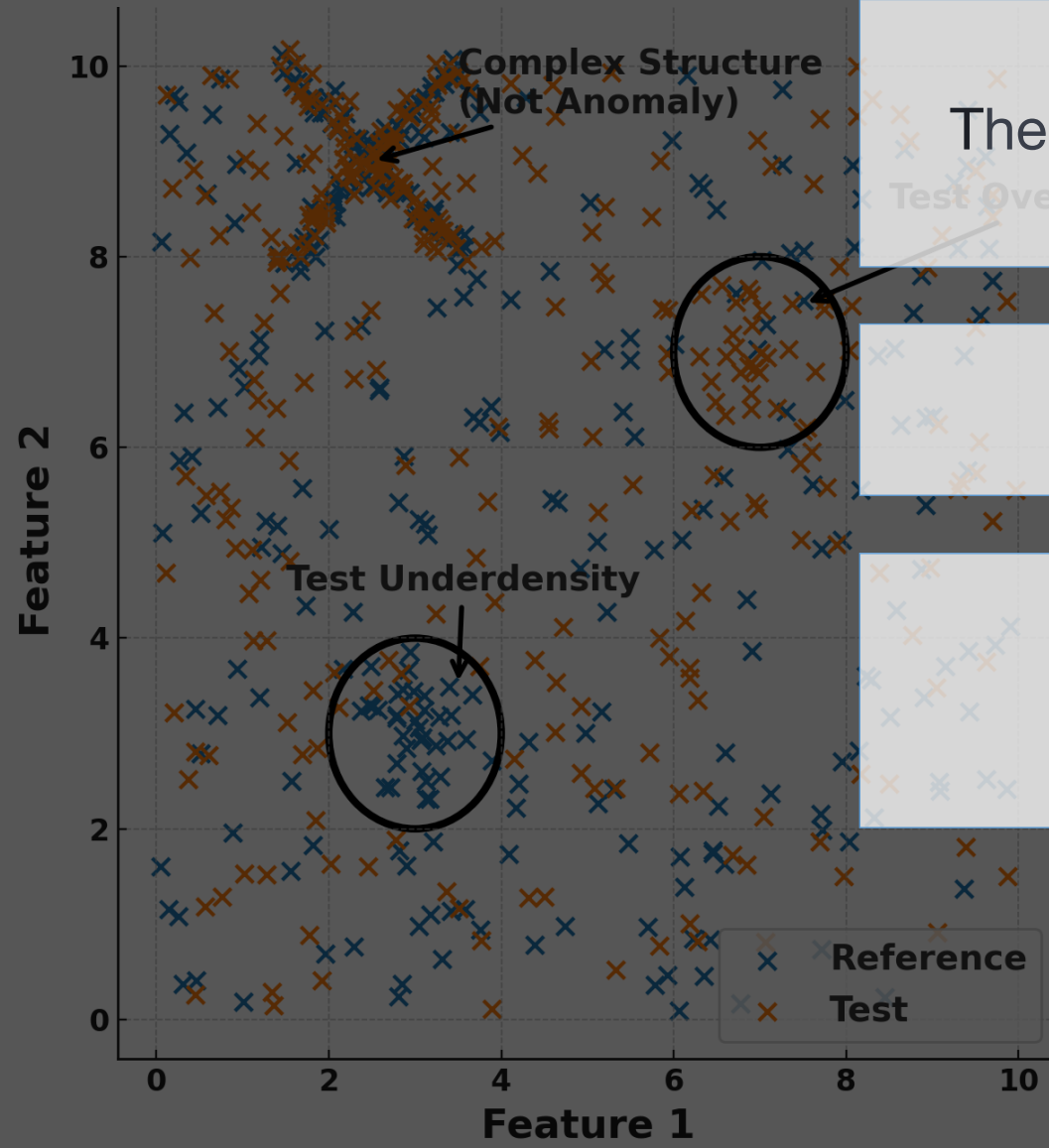




EagleEye is a new kNN-based density comparison method

$\mathcal{X}$ = 'Reference' set

$\mathcal{Y}$ = 'Test' set



BASED ON A SIMPLE PREMISE

KEY STRENGTHS

Look — check each point's nearest

neighbours in combined set $\mathcal{X} \cup \mathcal{Y}$

Data-driven background

The reference is learned from the data itself, not a predefined background model

Signal-agnostic

No assumptions about what the signal looks like

Localized

In addition to asking whether distributions differ, EagleEye also asks where

Data-driven signal-to-background estimate

if k -th neighbor of $Y_i \in \mathcal{X}$,
if k -th neighbor of $Y_i \in \mathcal{Y}$

$$\mathcal{H}_0: \sum_{k=1}^K b_i^k \sim \text{Binomial}$$

$$\gamma_i^+ = \max_{K \leq K_M} \left\{ -\log \left(\text{pval} \left(\sum_{k=1}^K b_i^k \right) \right) \right\}$$

arXiv:2503.23927

PREPRINT



CODE

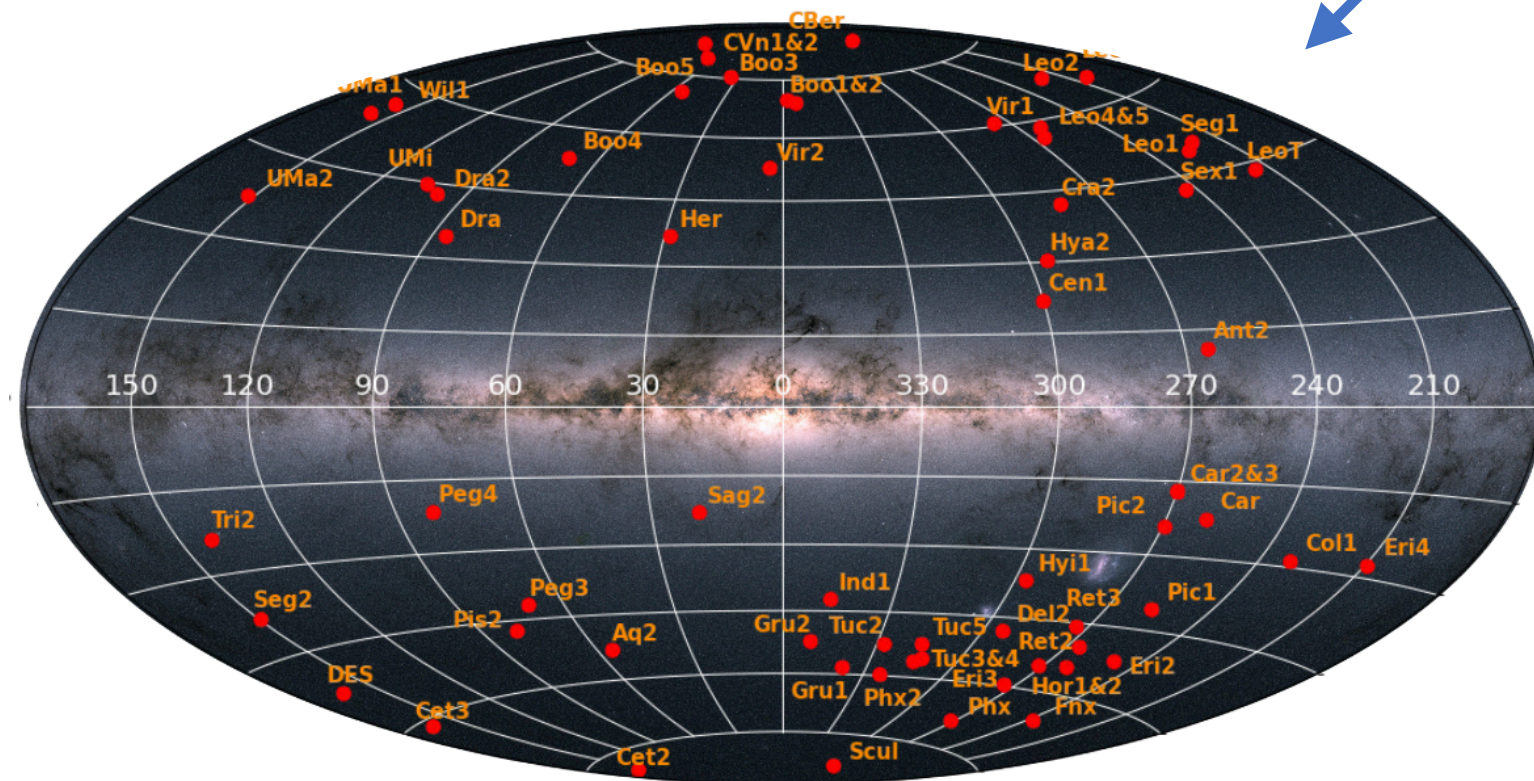
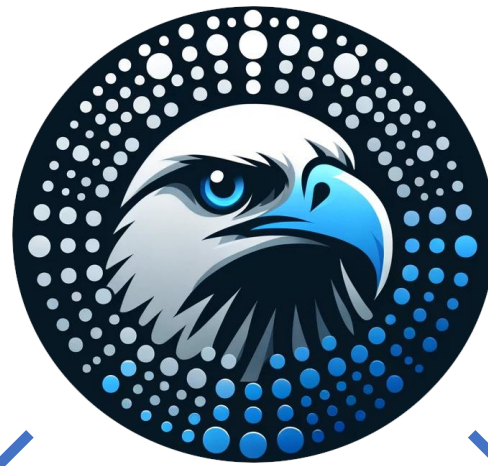


In search of two faint signals

DWARF GALAXIES

Dark matter-dominated systems —> prime targets for testing the nature of dark matter

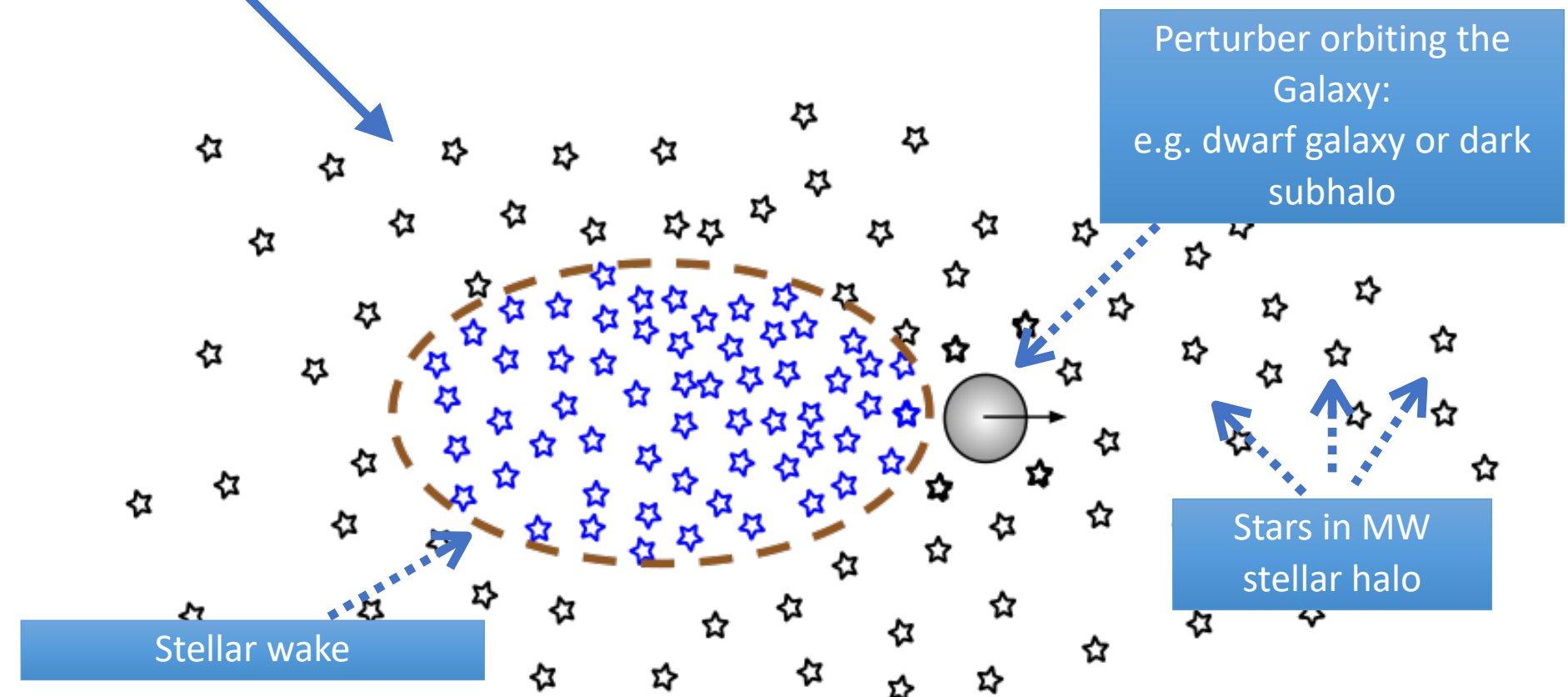
Difficult to detect —> show up as faint overdensities against a large fore- and background (~ 60 confirmed so far)



STELLAR WAKES

Indirect probe of DM substructure —> wake response depends on orbit and perturber properties

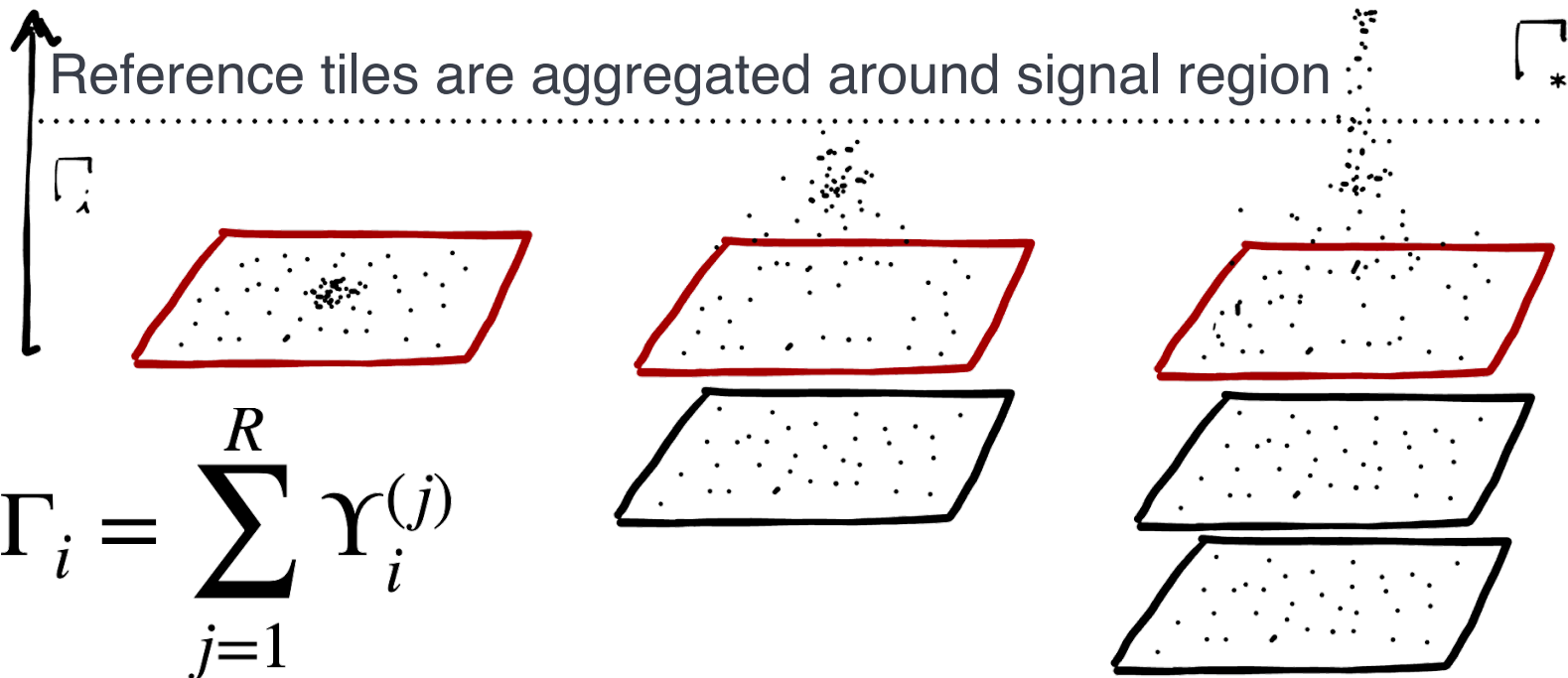
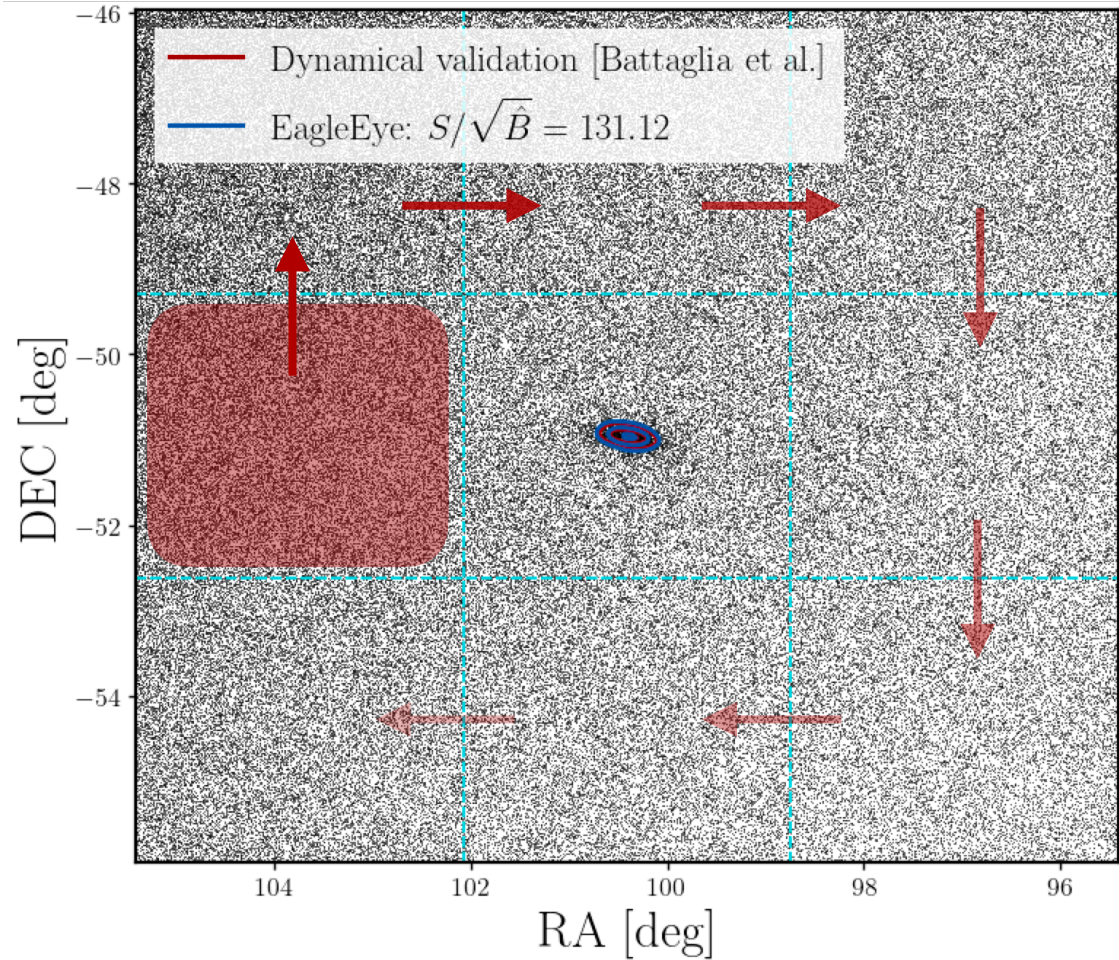
Difficult to model —> steady-state analytics fail to capture what simulations show



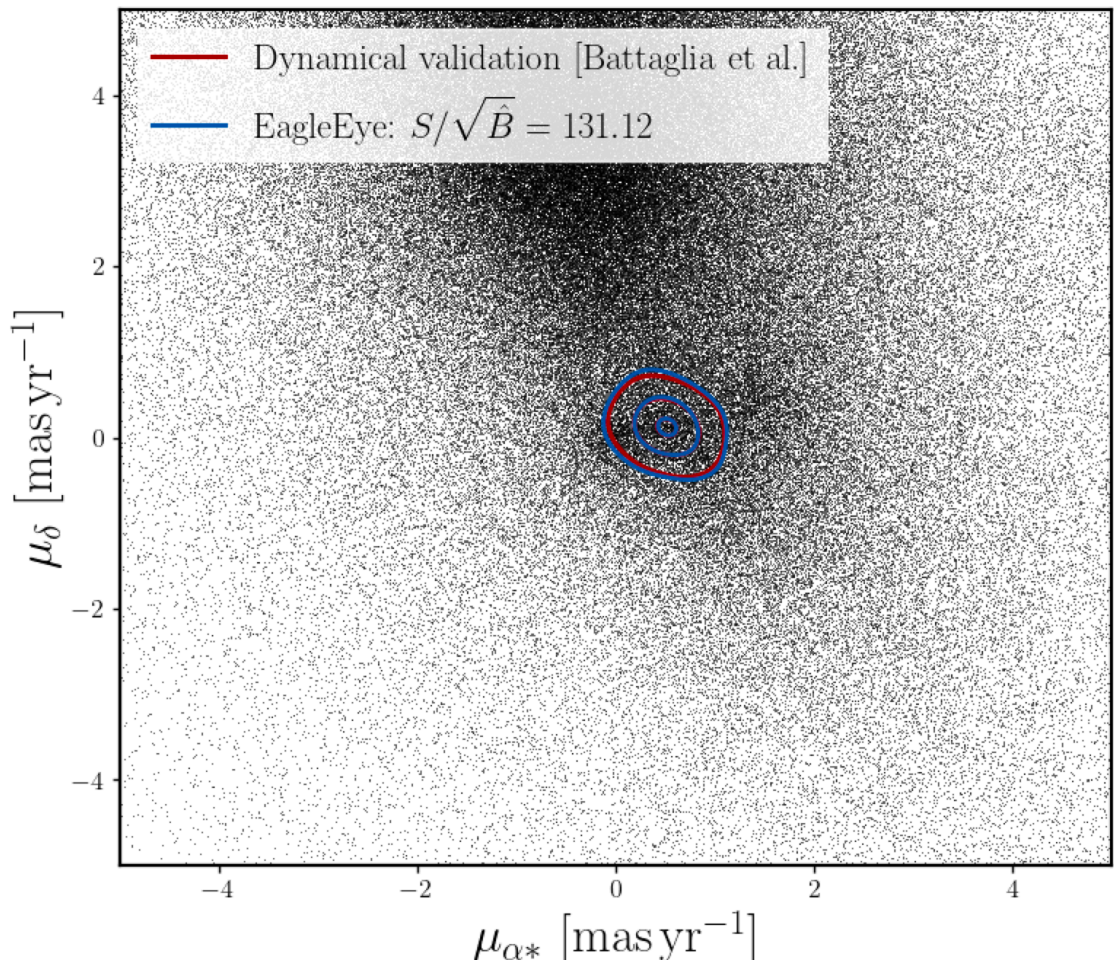
Combining multiple reference sets

Using Gaia DR3 ($\alpha, \delta, \mu_{\alpha*}, \mu_{\delta}$) and the dwarf catalogue of Battaglia et al. (2022), we validate EagleEye on known dwarfs

EXAMPLE A:
CARINA



$$\Gamma_i = \sum_{j=1}^R \Upsilon_i^{(j)}$$



Background
learned
empirically
rather than
assumed from
a model!

Summary statistics for the EagleEye Γ -cluster associated with Carina.

Γ	$S/\sqrt{\hat{B}}$	TPR	N_{Γ}	N_{overlap}	N_{extra}	f_{imp}	$\mathbf{RA}_{\text{cent}}$ [deg]	$\mathbf{Dec}_{\text{cent}}$ [deg]	Δ_{cent} [deg]
0	131.12	0.96	3184	2272	912	0.29	100.374	-50.972	0.015

Beyond known dwarfs



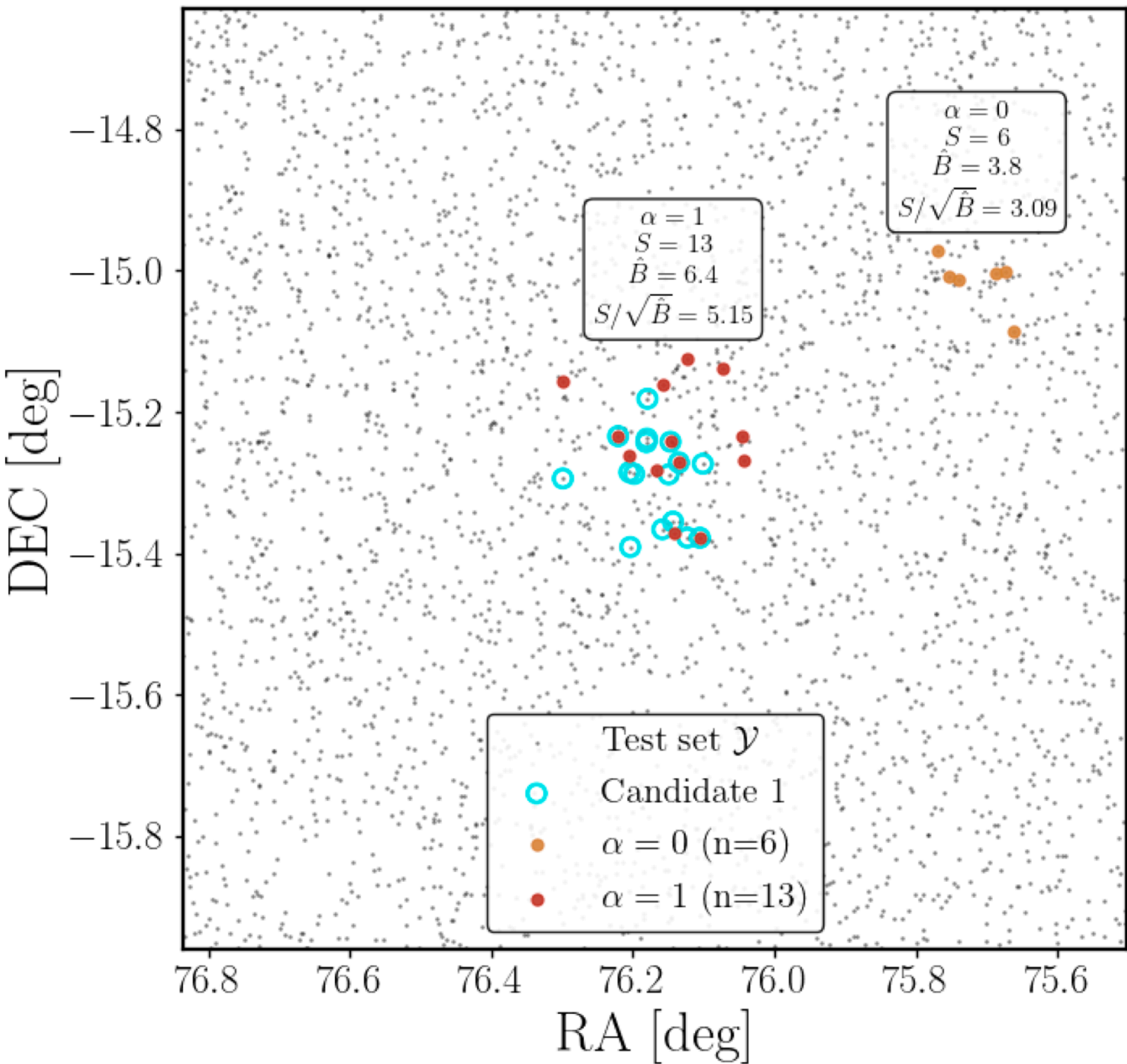
Candidate 1 found by a wavelet search with no confirmed dwarf association in Darragh-Ford et al. (2021)

Centroid location: $(\alpha_{cent}, \delta_{cent}) = (76.17, -15.30)$

Recovered by EagleEye as a significant local overdensity
(an independent, model-free check)

Summary statistics for the EagleEye Γ -cluster associated with Candidate 1.

Γ	$S/\sqrt{\hat{B}}$	TPR	N_Γ	N_{overlap}	N_{extra}	f_{imp}	RA_{cent} [deg]	Dec_{cent} [deg]	Δ_{cent} [deg]
0	3.09	0.00	6	0	6	1.00	75.714	-15.014	0.535
1	5.15	0.25	13	4	9	0.69	76.143	-15.240	0.061



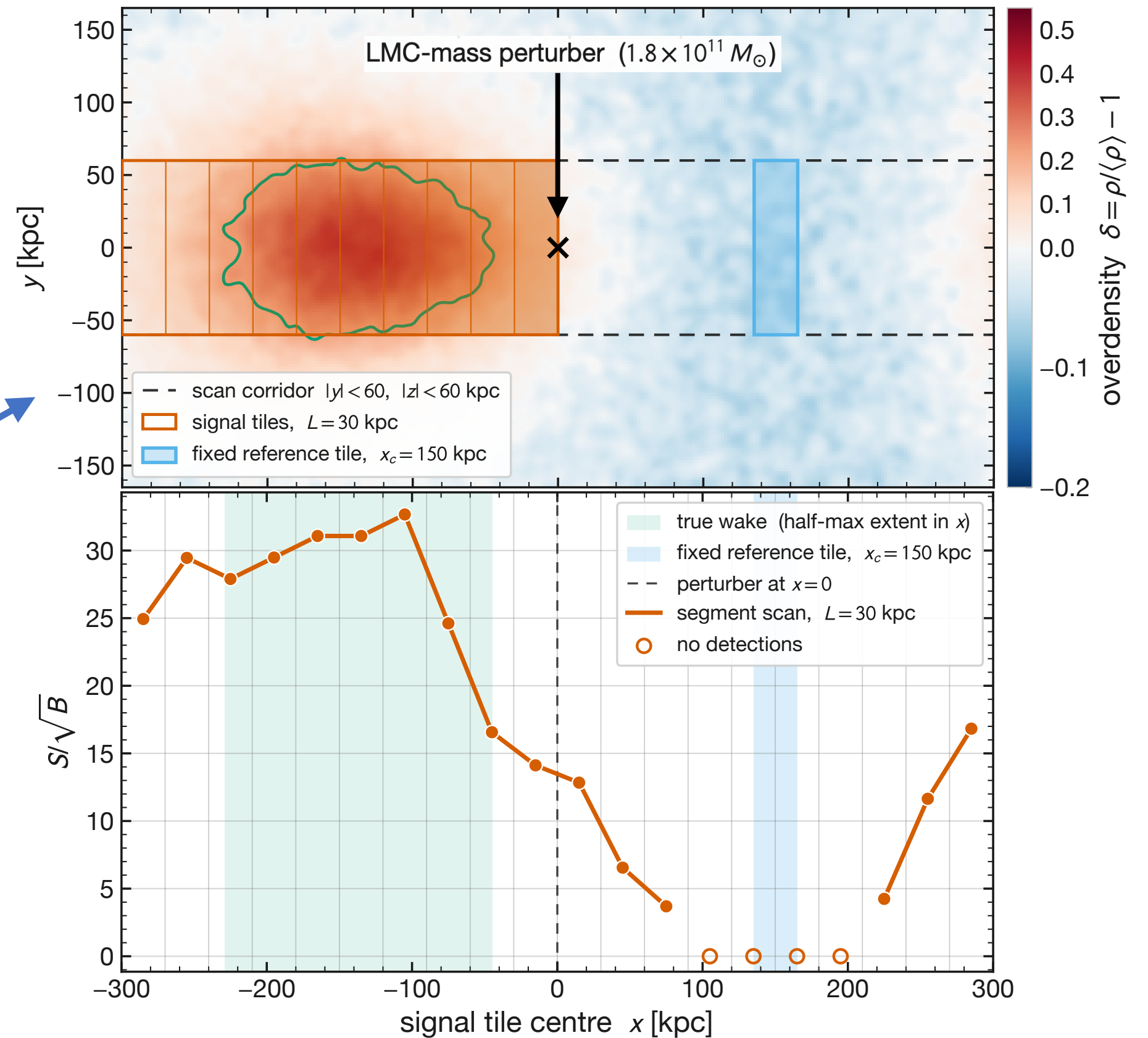
III — STELLAR WAKES

Simulating the wake of the Large Magellanic Cloud (LMC)

The LMC provides an ideal testbed:
large mass = strong response

We simulate the LMC wake
using PKDGRAV3 Potter et al. (2016)

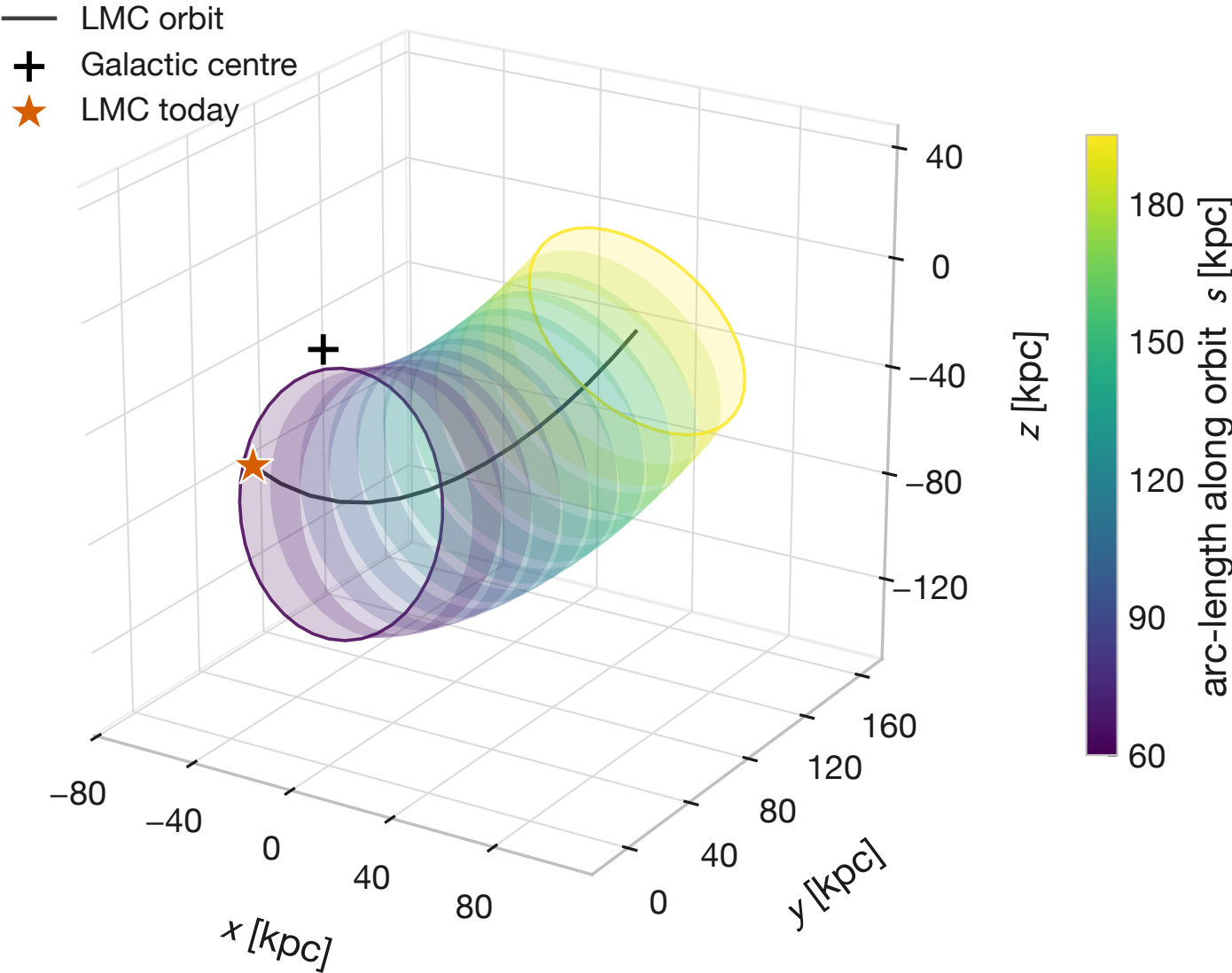
*EagleEye asks: “If I look locally
around each star, do I see more
stars than expected, compared
to a reference?”*



Modifying the signal-reference
configuration to target LMC orbit

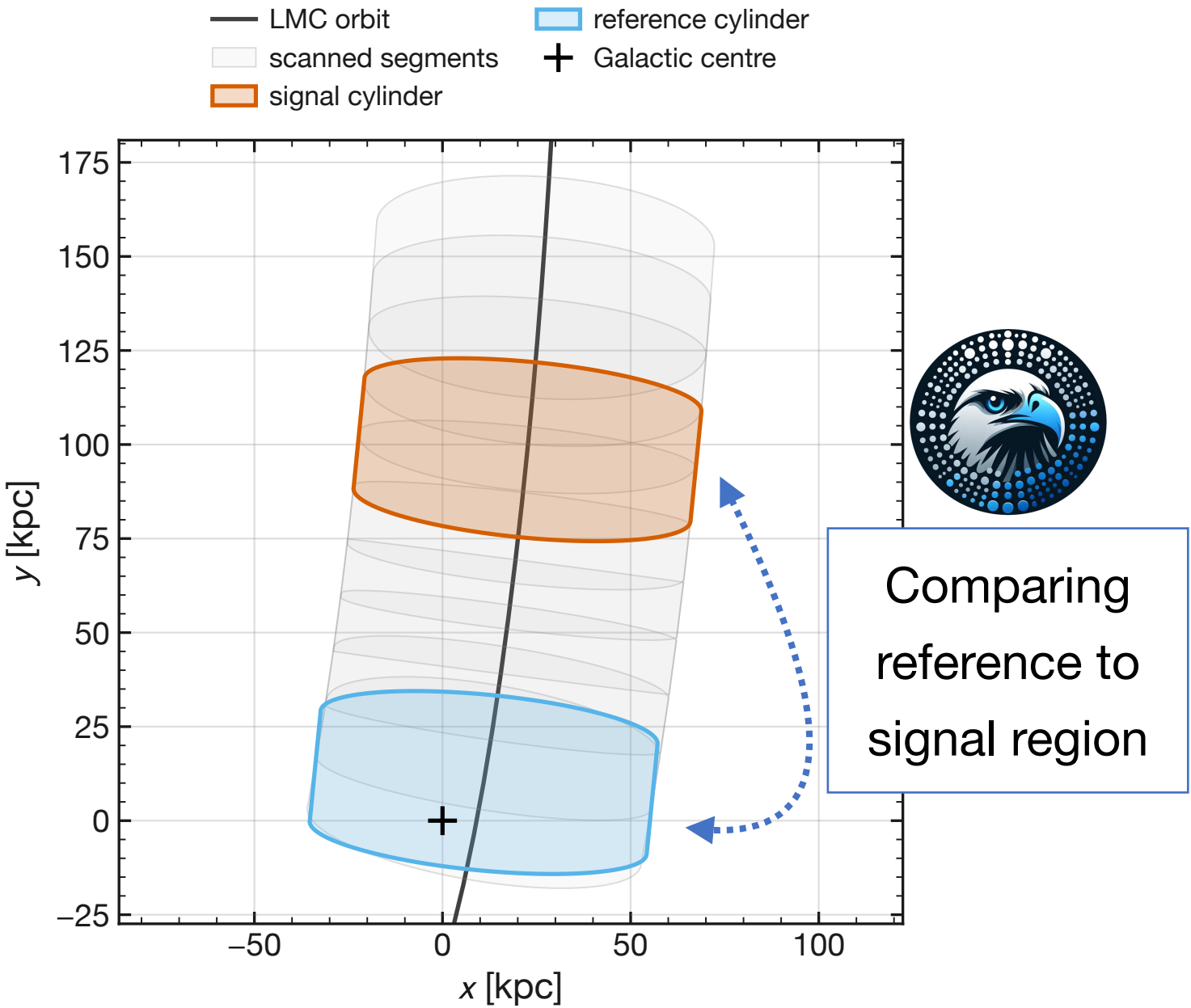
Extending reference strategy to LMC orbit

We construct a cylindrical corridor along the LMC's orbit



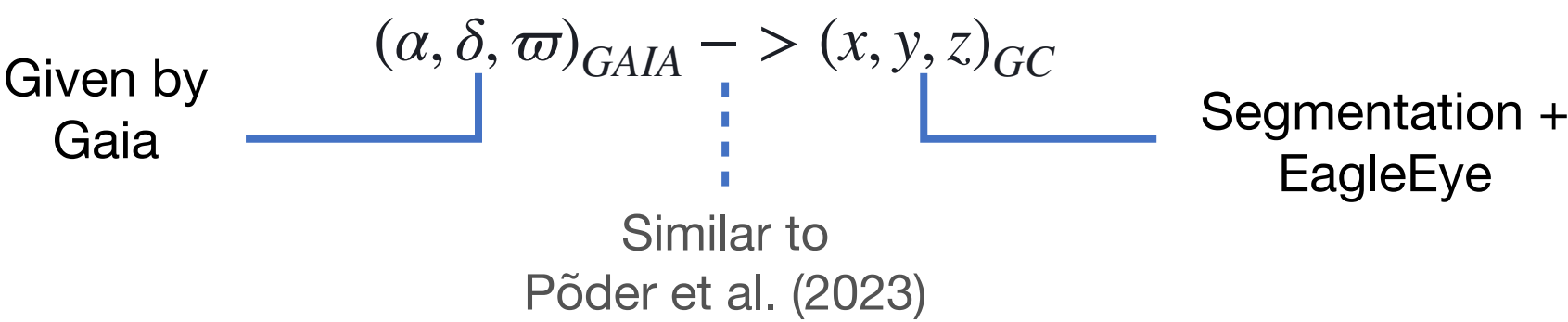
Orbit from Vasiliev et al. (2021)

TEST: Orbit corridor divided into signal regions (cylinders along orbit)
REFERENCE: Orbit segment where no wake is expected



Toward observations

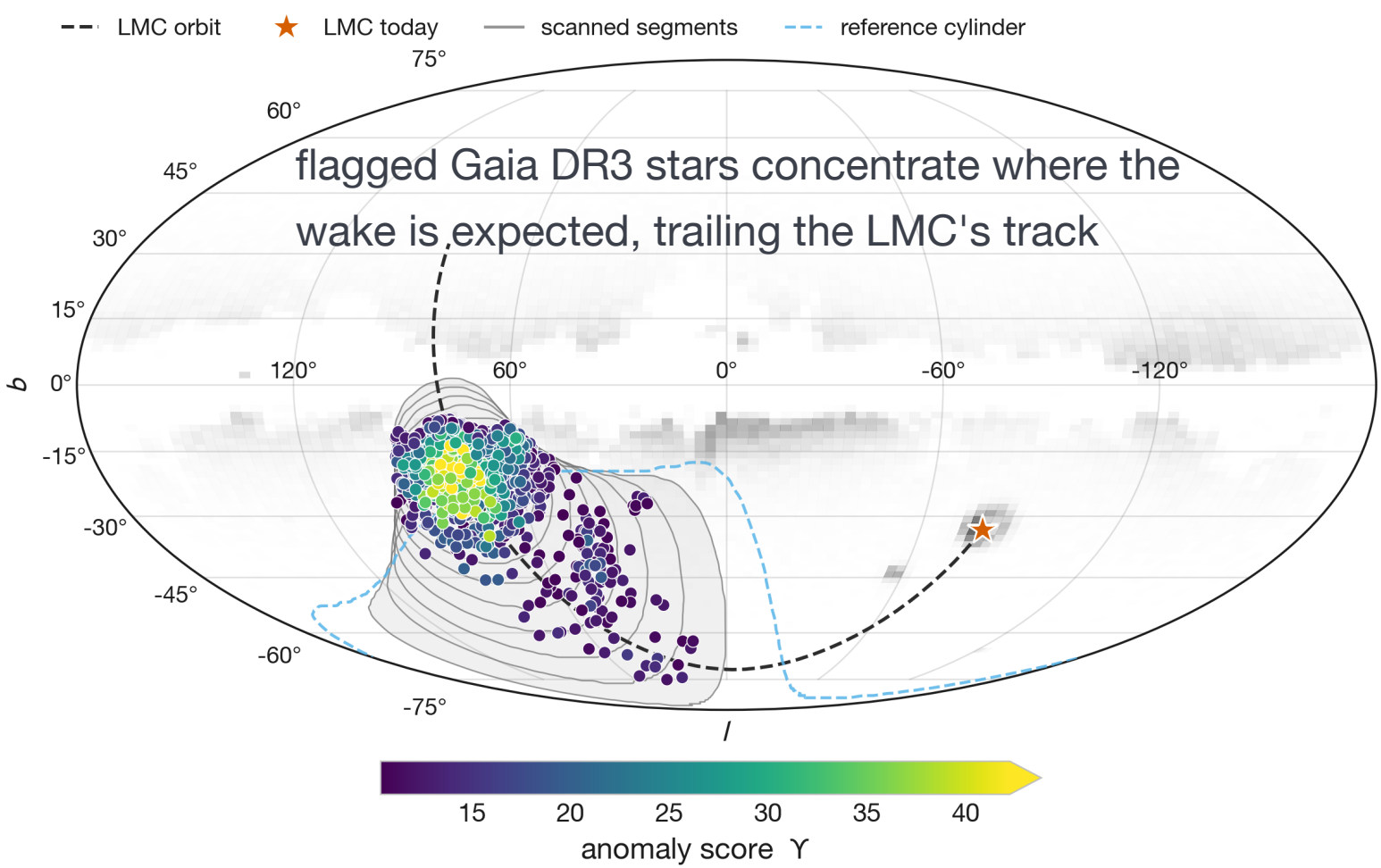
Galactocentric transformation of RGB stars from Gaia DR3:



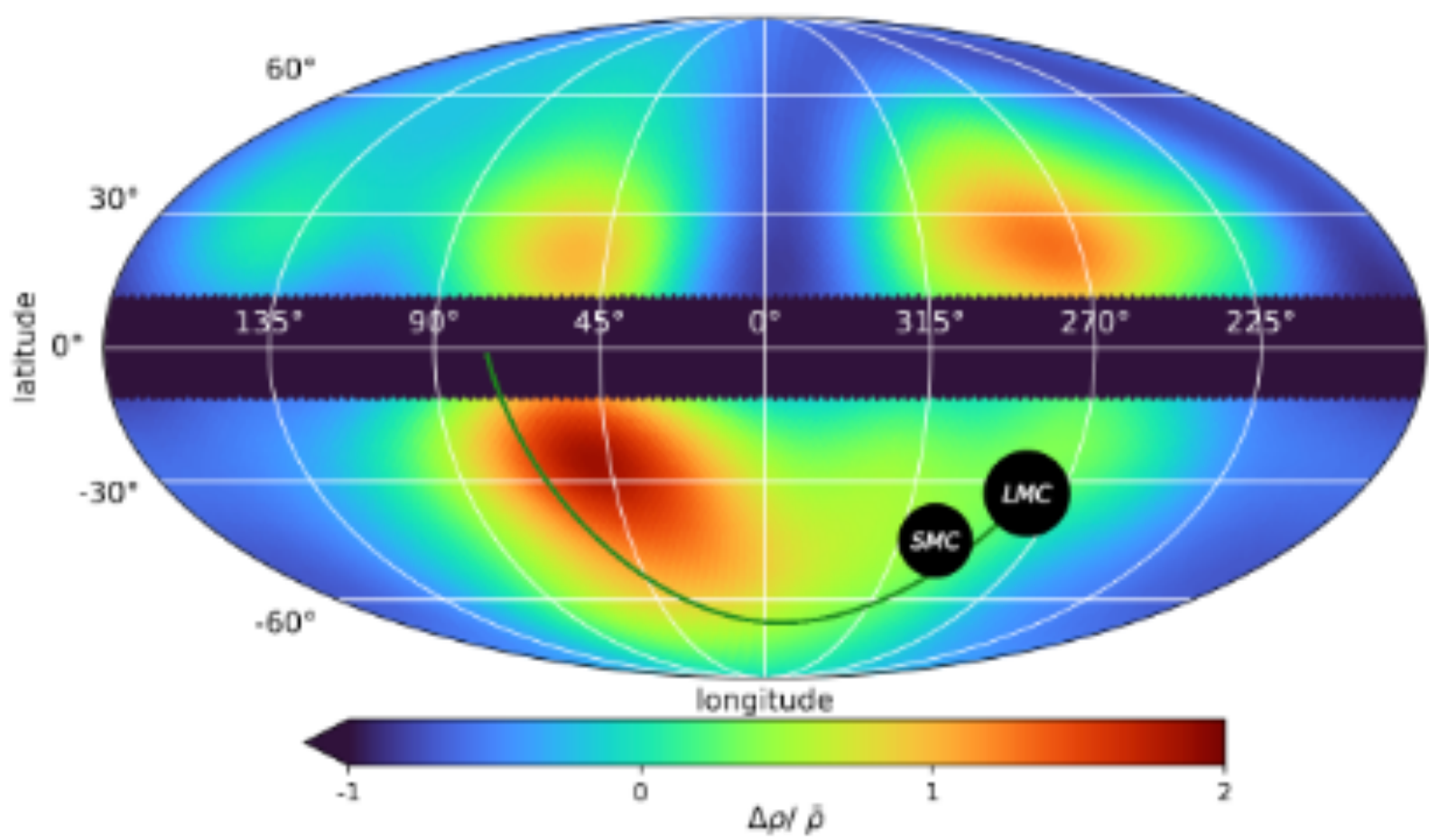
Ongoing work: kinematics beyond positions and improved distances, possibly with Gaia DR4

Ultimately: LMC only the first step → Targeting smaller perturbers → fainter wakes

After running scan, flagged stars plotted in (l, b)



Detected wake in Fushimi et al. (2023)

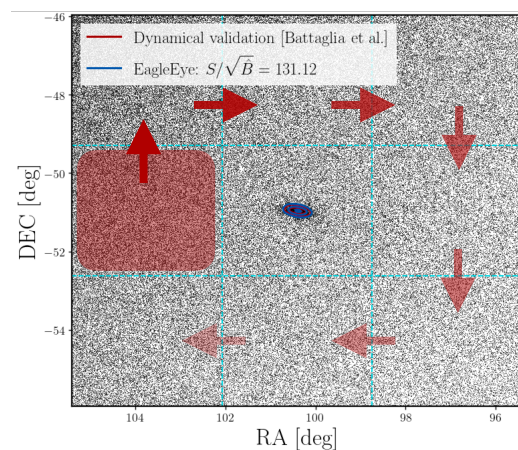


TAKEAWAY

One model-independent framework, two faint signals

EagleEye is a new kNN-based density comparison method

DWARF GALAXIES

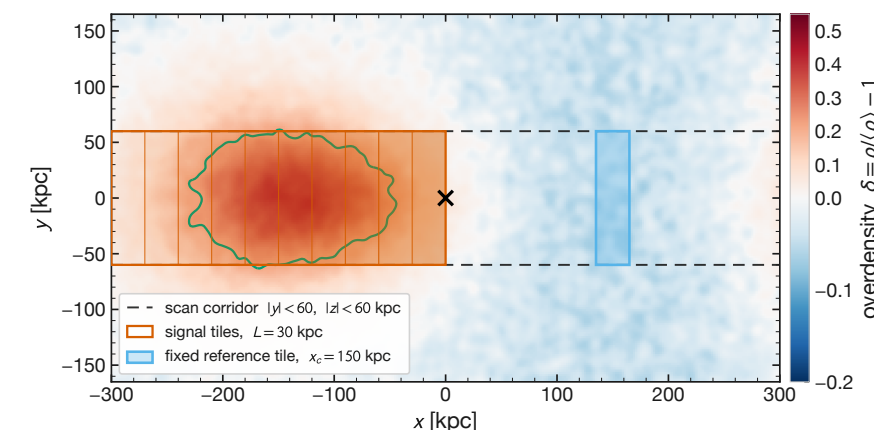


A dwarf detector that flags where stellar density in sky position and proper motion deviates from the surrounding field.



Applied as ...

STELLAR WAKES



A wake detector that flags where the stellar phase-space density behind a moving perturber deviates from a smooth halo.

Reference set strategy adopted according to problem:
backgrounds derived **empirically**.

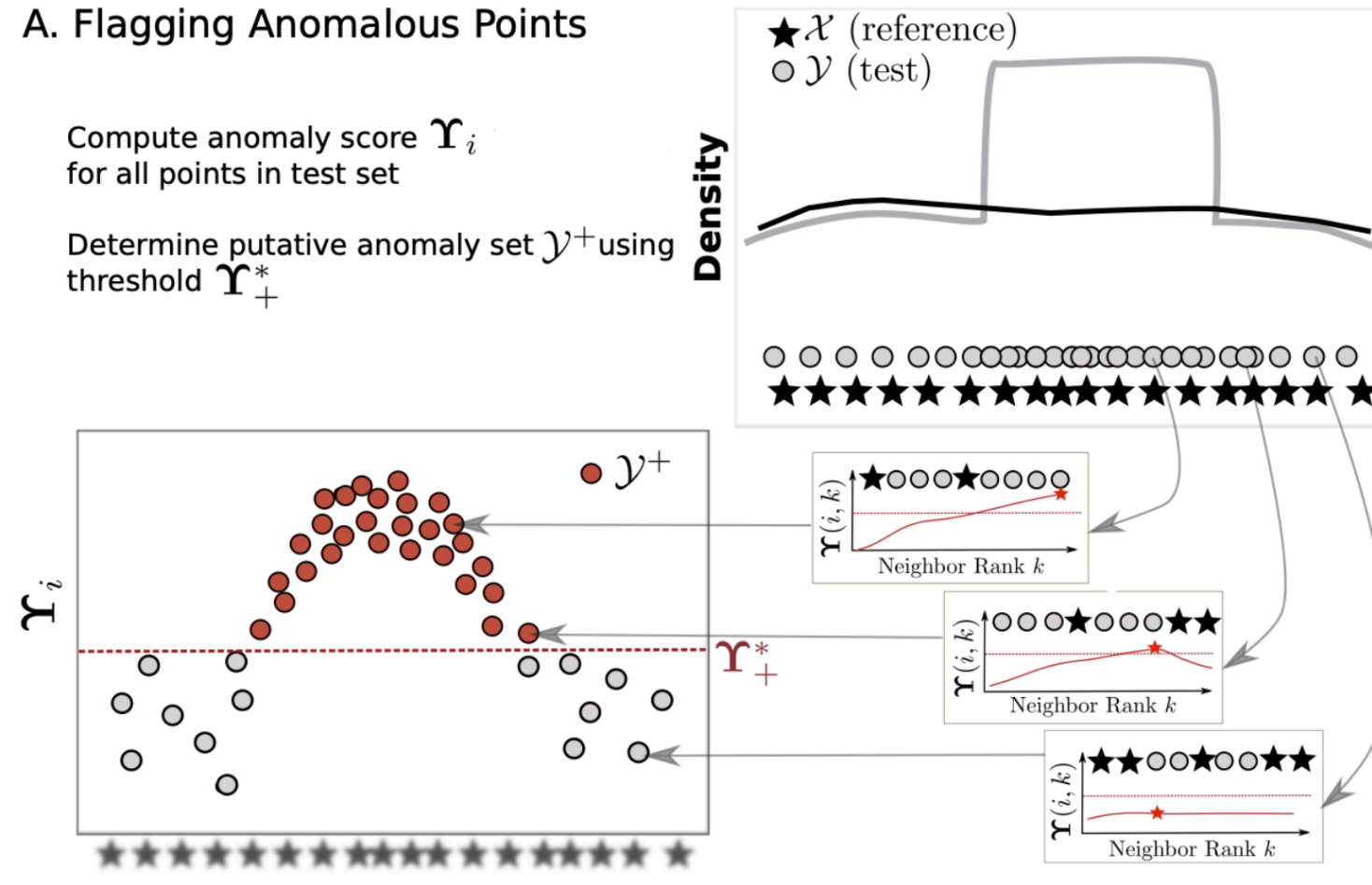
BACKUP

EagleEye full pipeline

A. Flagging Anomalous Points

Compute anomaly score Υ_i for all points in test set

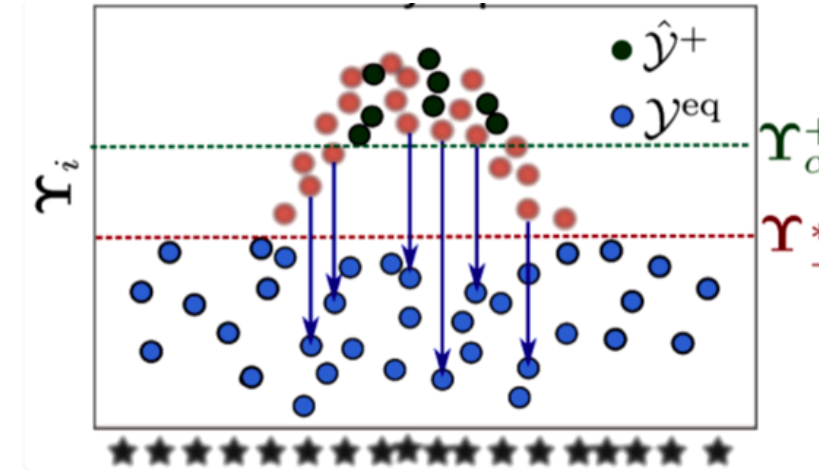
Determine putative anomaly set $\mathcal{Y}^+$ using threshold Υ_+^*



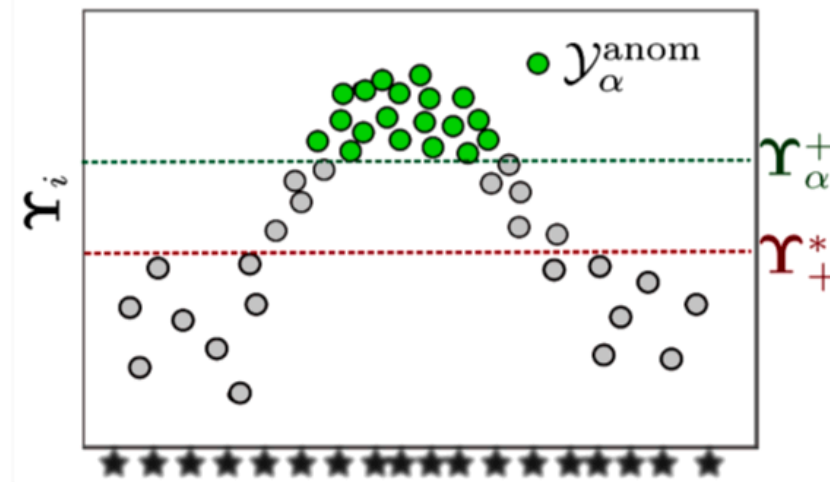
Figures adapted from
arXiv:2503.23927

Use injected reference samples to estimate signal to background ratio in anomalous modes

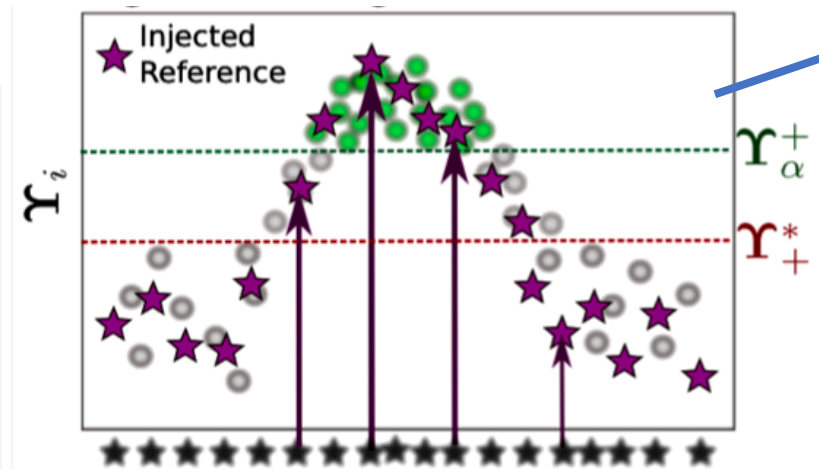
B. Iterative Density Equalization (IDE)



C. Cluster Refinement



D. Signal-to-background estimation via injection



$$\Lambda_\alpha := \frac{\widehat{S_\alpha}}{\sqrt{B_\alpha}} = \frac{|\mathcal{Y}_\alpha^{\text{anom}}| - \left(|\mathcal{Y}_\alpha^{\text{inj}}| \frac{(n_{\mathcal{Y}} - |\hat{\mathcal{Y}}^+|)}{(n_{\mathcal{X}} - |\hat{\mathcal{X}}^+|)} \right)}{\sqrt{|\mathcal{Y}_\alpha^{\text{inj}}| \frac{(n_{\mathcal{Y}} - |\hat{\mathcal{Y}}^+|)}{(n_{\mathcal{X}} - |\hat{\mathcal{X}}^+|)}}}$$

S_α — Expected number of signal points

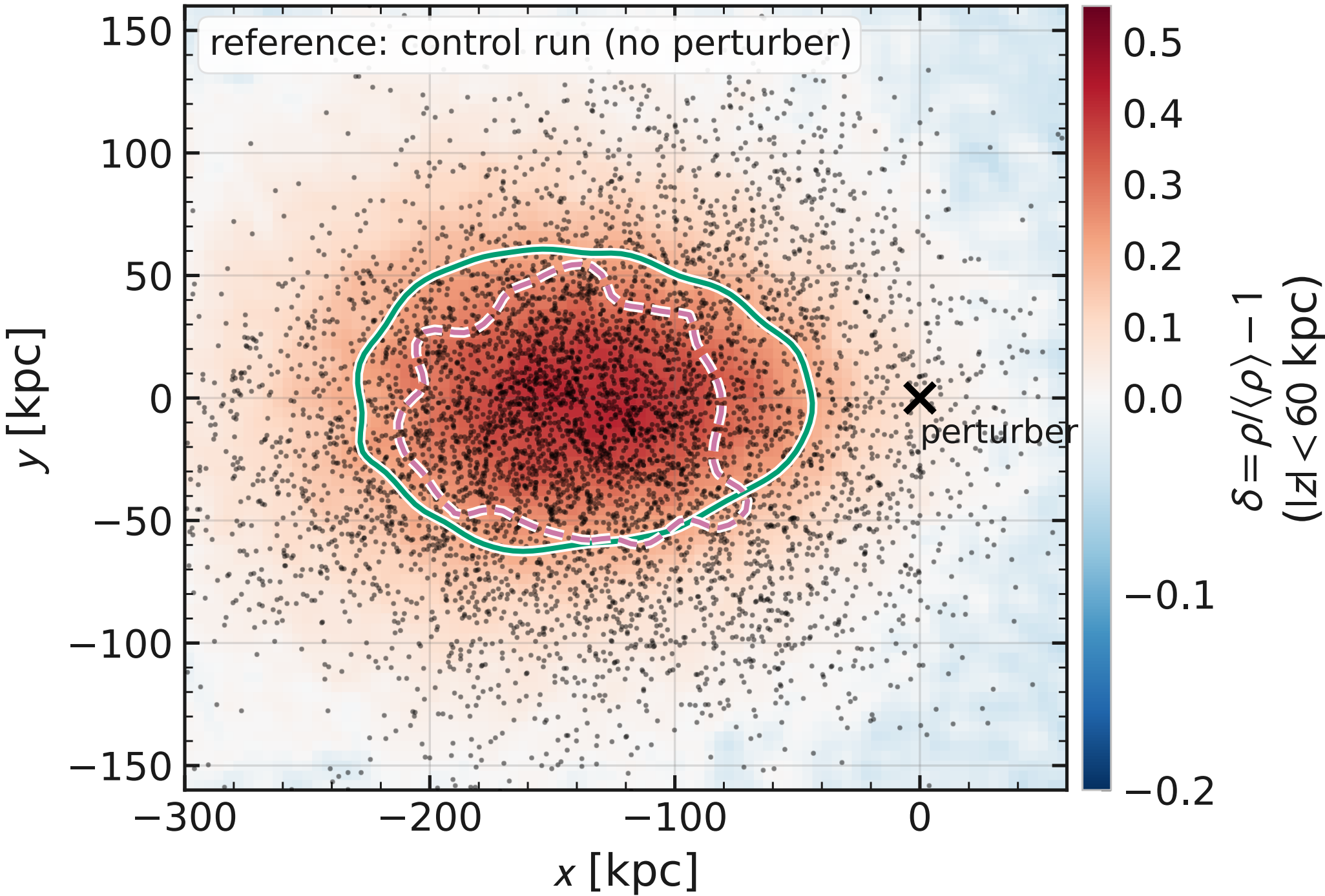
B_α — Expected number of background points contributing to $\mathcal{Y}_\alpha^{\text{anom}}$

$\mathcal{Y}_{\text{inj}}^\alpha$ — Injected set

$$\mathcal{Y}_\alpha^{\text{inj}} := \left\{ X_i \in \underline{\mathcal{Y}}^+ \mid \Upsilon_i \geq \Upsilon_\alpha^+ \right\}.$$

III — STELLAR WAKES

● flagged stars (5,756) — true wake, half-max - - - flagged, half-max



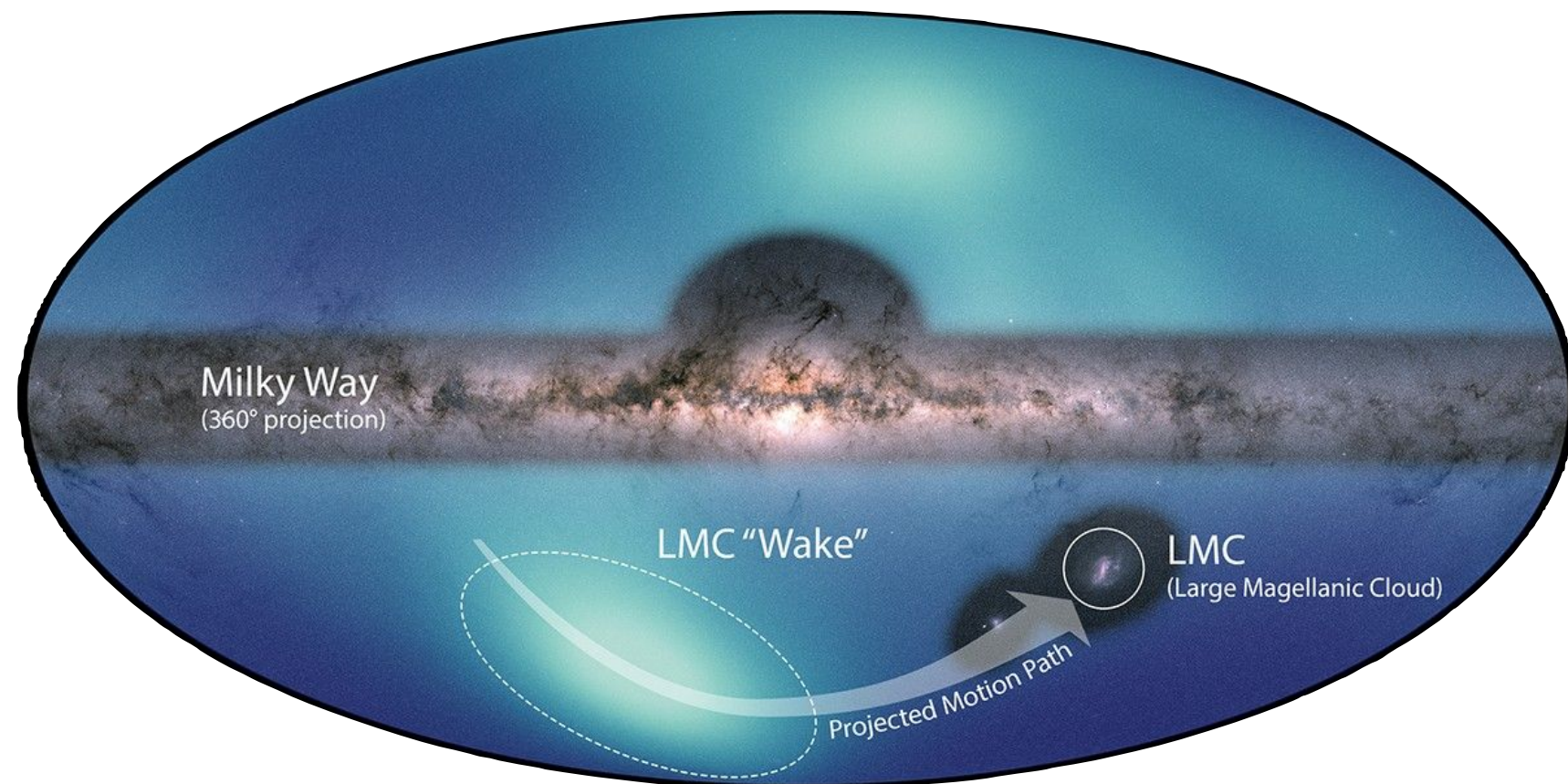
N-body validation against a
genuinely wake-free control run

III — STELLAR WAKES

Simulating the wake of the Large Magellanic Cloud (LMC)

LMC and background
parameters chosen to mimic
MW stellar halo at 70 kpc
[Foote et al. (2023)]

The LMC provides an ideal testbed:
large mass = strong response



Credits: NASA/ESA/JPL-Caltech/Conroy et al. 2021

We simulate the LMC wake
using PKDGRAV3 [Potter et al. (2016)]

