2nd SMASHING Workshop 2026 · Vipava

Anomaly Detection for Faint Galactic Substructure with EagleEye

Speaker: Sven Põder

SISSA, Trieste · NICPB, Tallinn





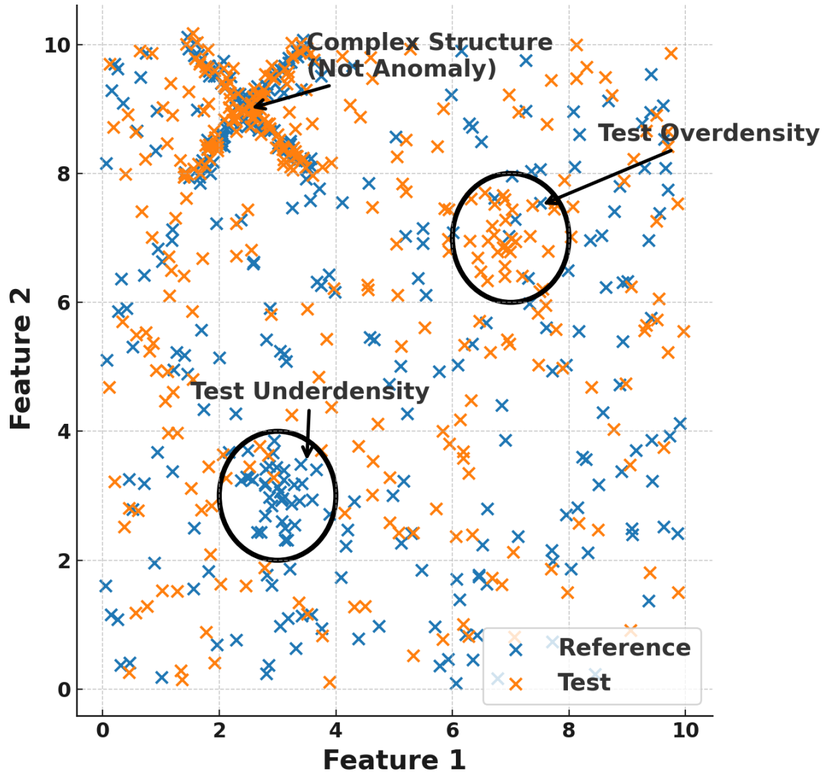


EagleEye is a new kNN-based density comparison method



X = ‘Reference’ set $\mathcal{X}$

X = ‘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

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

Full procedure explained here

arXiv:2503.23927

PREPRINT

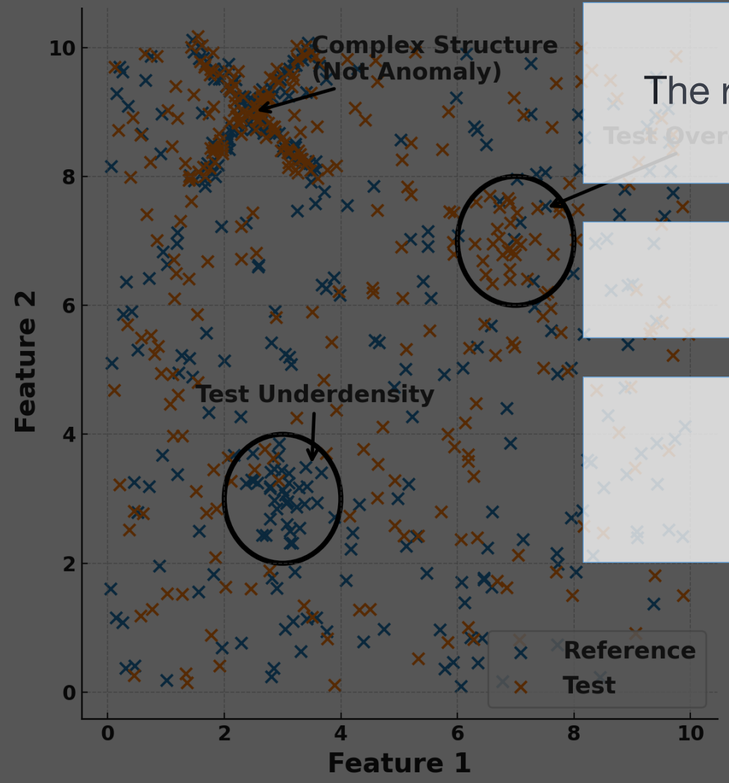
CODE

EagleEye is a new kNN-based density comparison method



X = ‘Reference’ set $\mathcal{X}$

X = ‘Test’ set $\mathcal{Y}$



BASED ON A SIMPLE PREMISE

KEY STRENGTHS

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

Count — how many are test vs. reference?

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

Flag — a skewed local count marks a potential anomaly

Full procedure explained here

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

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

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

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

arXiv:2503.23927

PREPRINT

arXiv

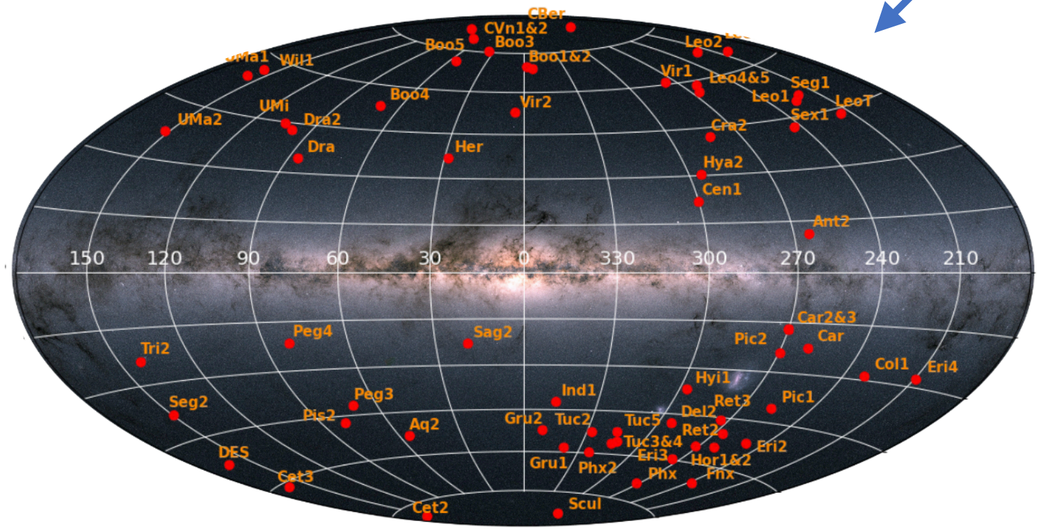
CODE

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)

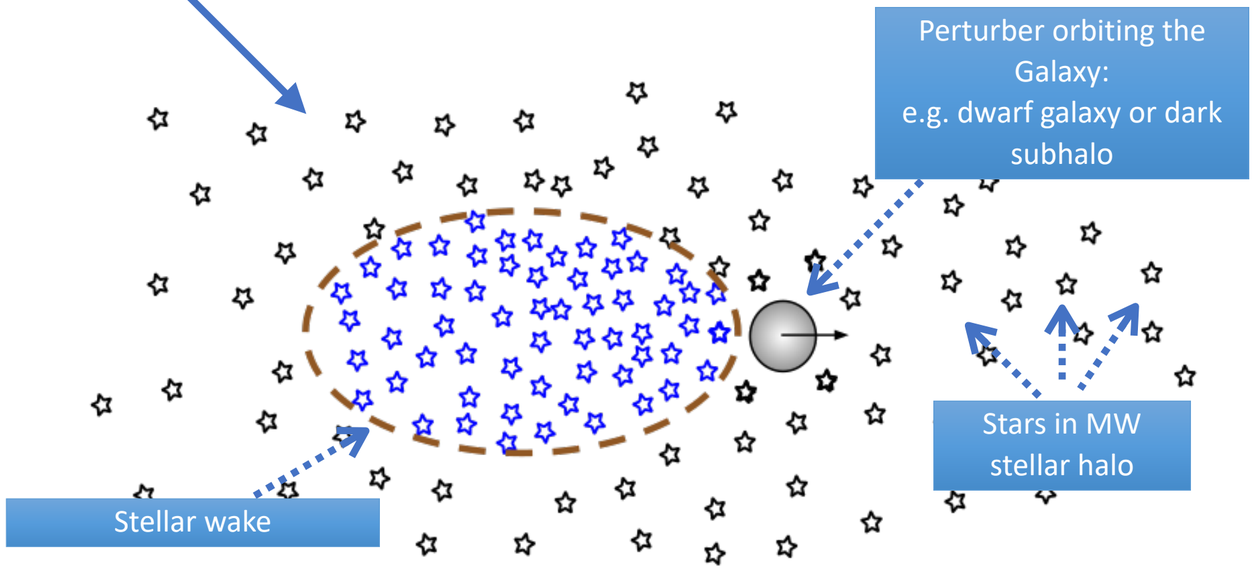


STELLAR WAKES

Indirect probe of DM (sub)structure —> wake response depends on orbit and perturber properties

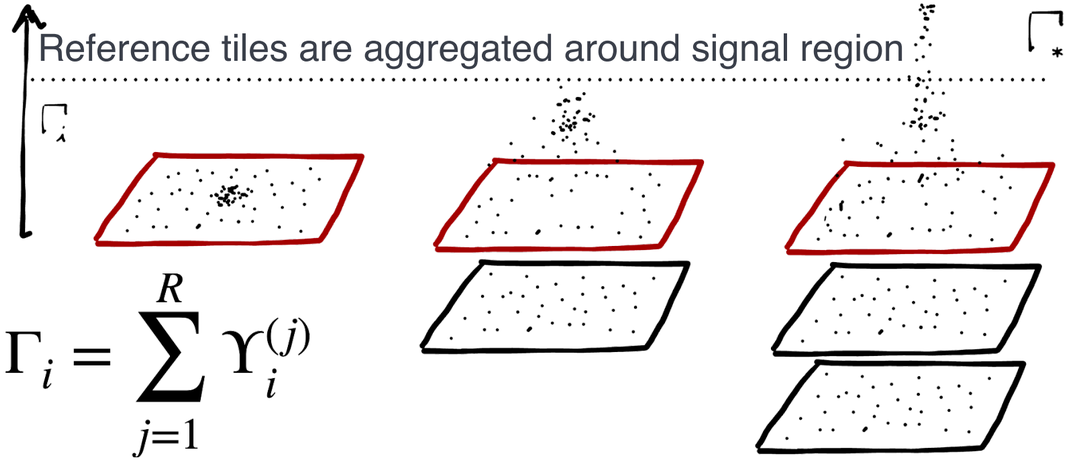
Difficult to model —> steady-state analytics fail to capture what simulations show

We are also working on this.
See talk of **María Benito**!

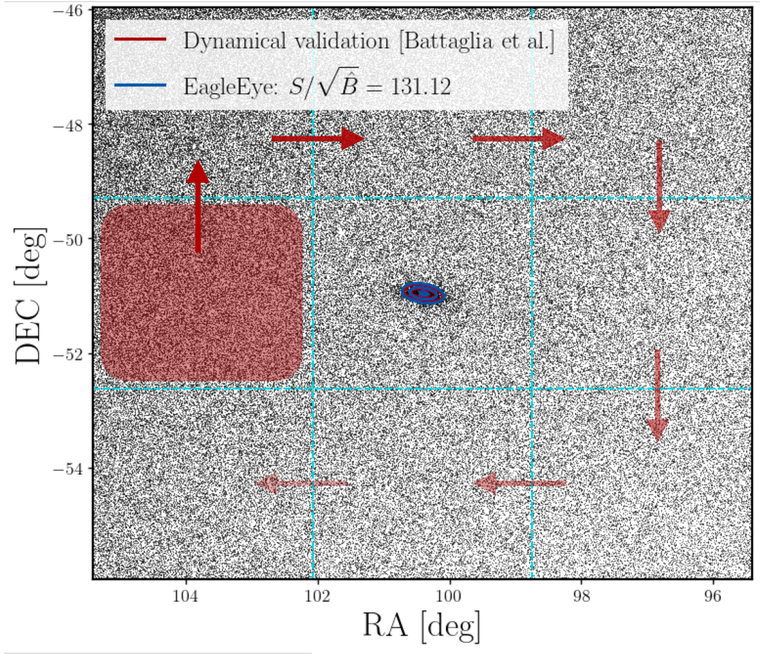


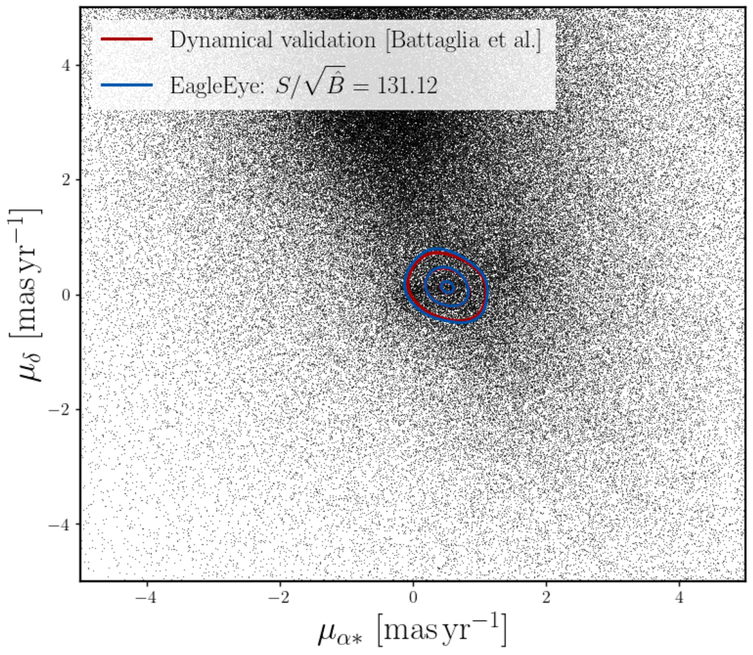
Combining multiple reference sets

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





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}	RA_{cent} [deg]	Dec_{cent} [deg]	Δ_{cent} [deg]
0	131.12	0.96	3184	2272	912	0.29	100.374	-50.972	0.015

Beyond known dwarfs



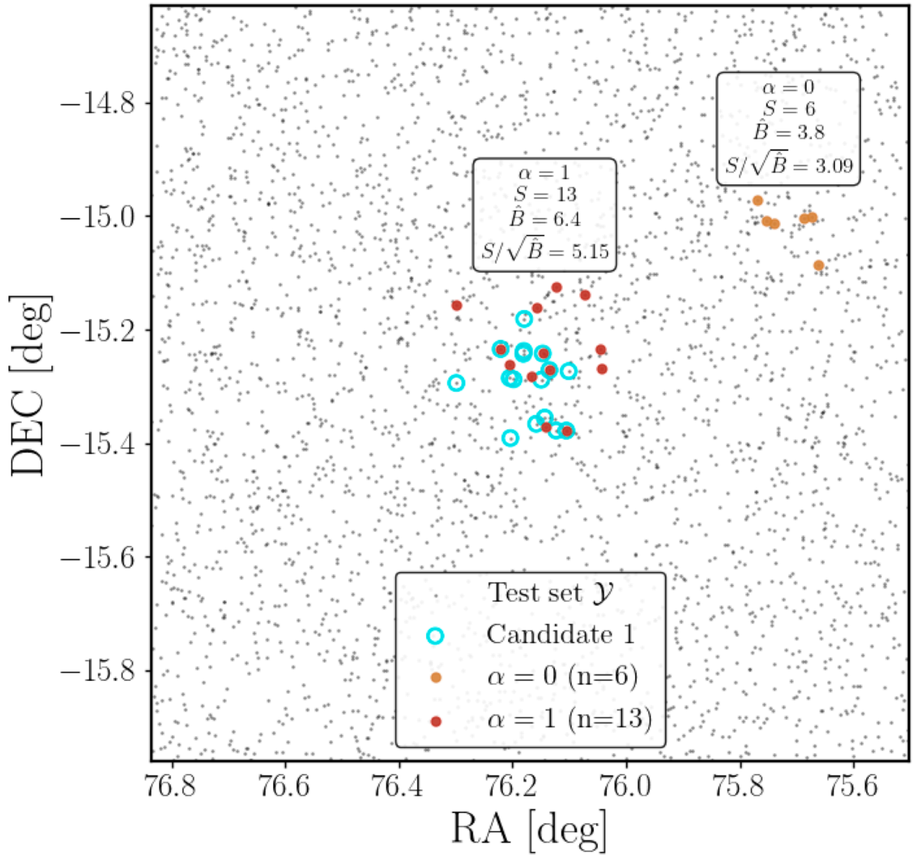
Candidate 1 found by a wavelet search with no confirmed dwarf association in Darragh-Ford et al. (2021)

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.

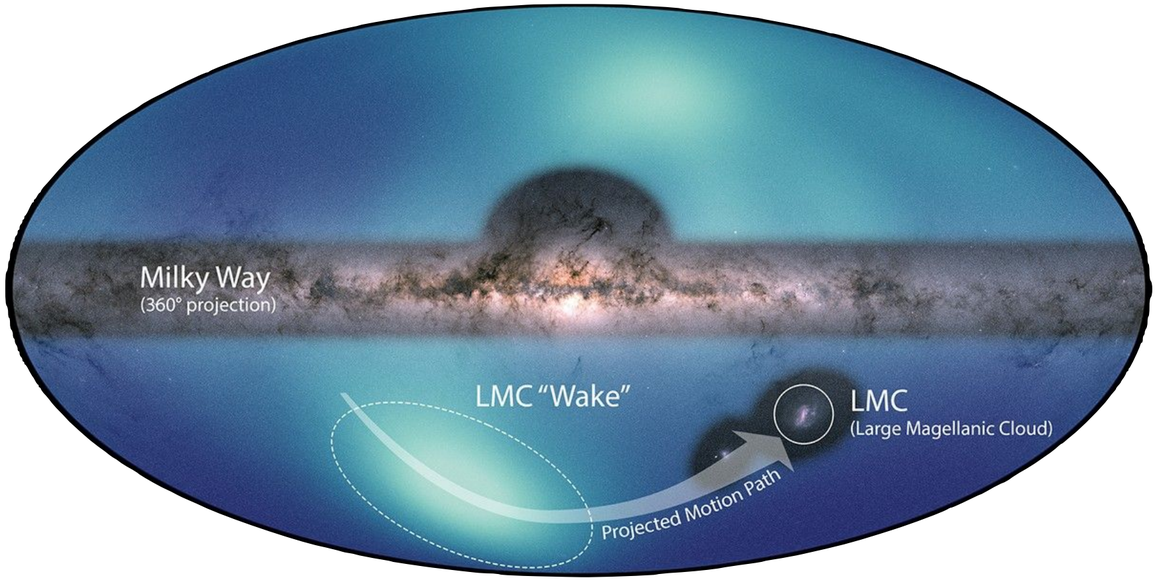
Γ	$S/\sqrt{\hat{B}}$	TPR	N_Γ	$N_{overlap}$	N_{extra}	f_{imp}	RA_{cent} [deg]	Dec_{cent} [deg]	Δ_{cent} [deg]
0	3.09	0.00	6	0	6	1.00	75.714	-15.014	0.535
1	5.15	0.25	13	4	9	0.69	76.143	-15.240	0.061



Simulating the wake of the Large Magellanic Cloud (LMC)

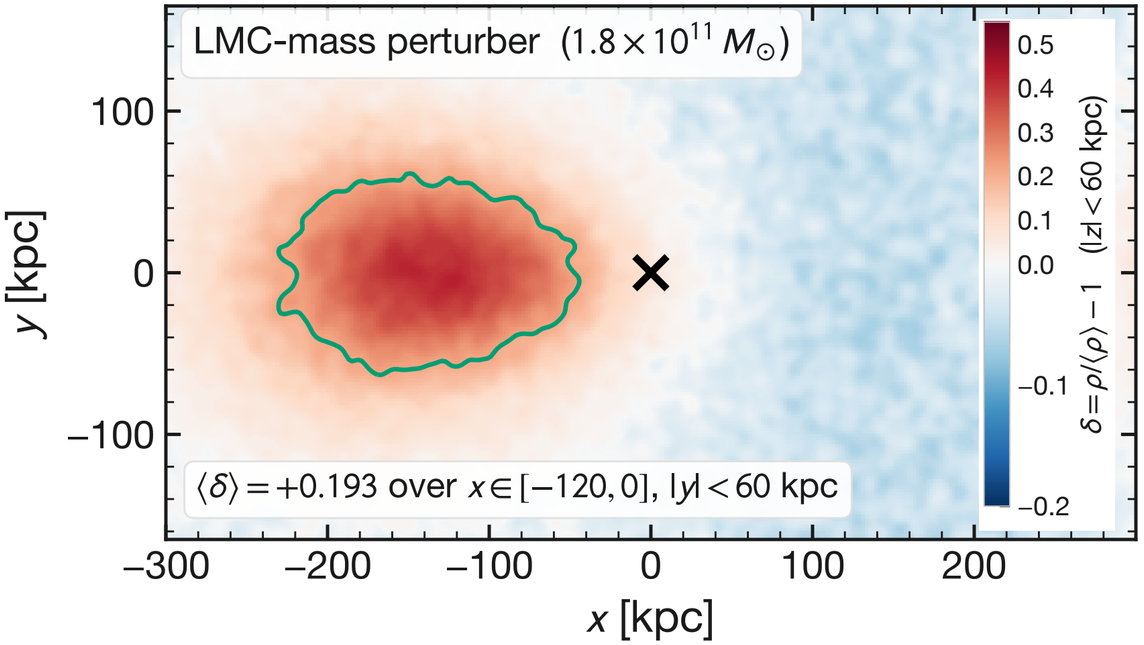
LMC and background parameters chosen to mimic MW stellar halo at 70 kpc [Foote et al. (2023)]

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)]

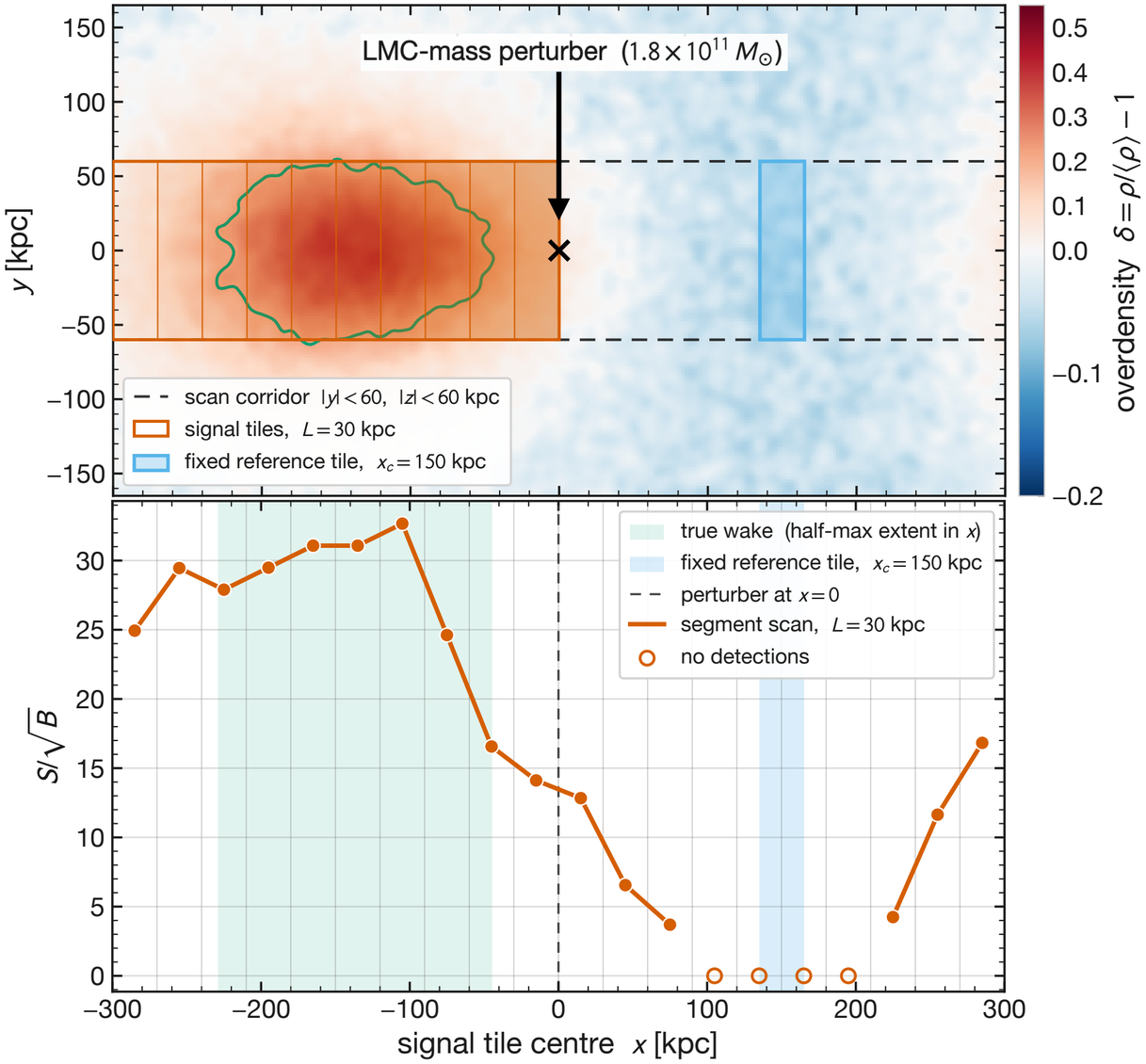


Testing in ideal laboratory conditions

A simple signal-reference strategy to find the wake region

Path of the LMC divided into signal tiles, compared with a reference that does not contain the wake

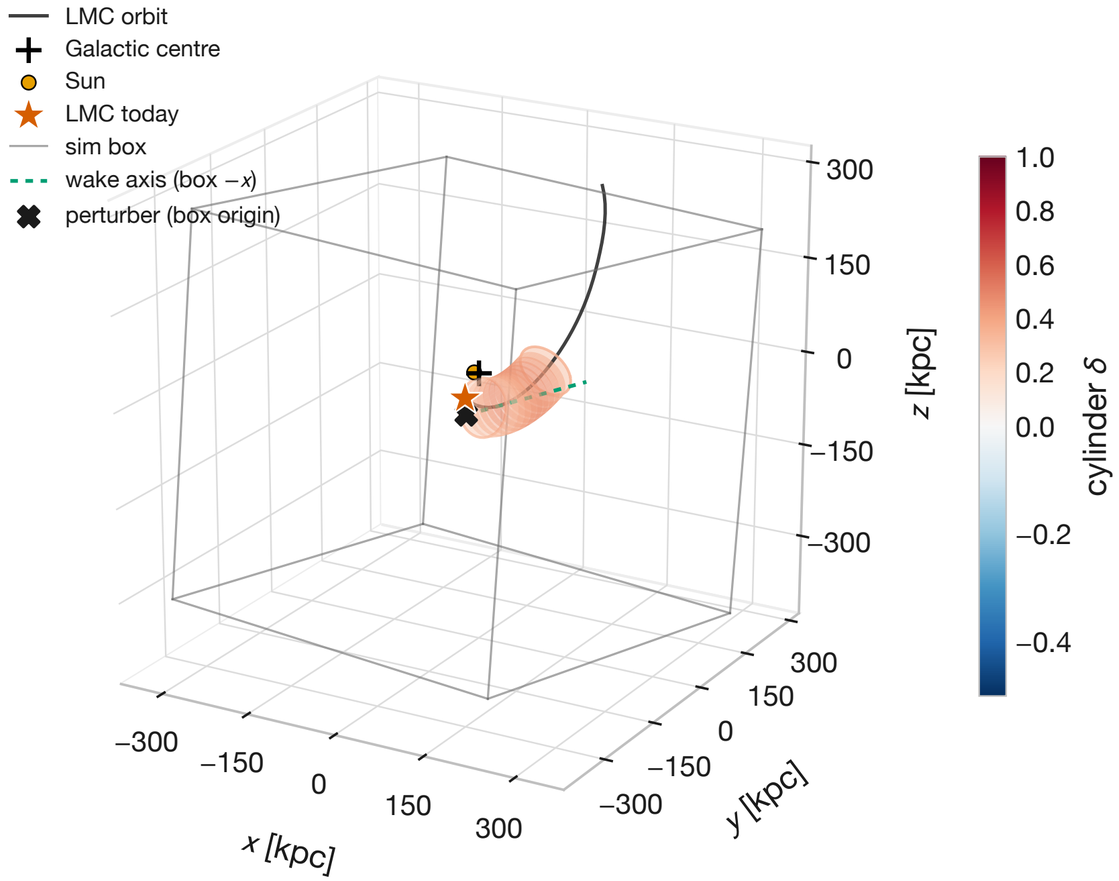
In each signal segment, EagleEye asks: "If I look locally around each star, do I see more stars than expected, compared to a reference?"



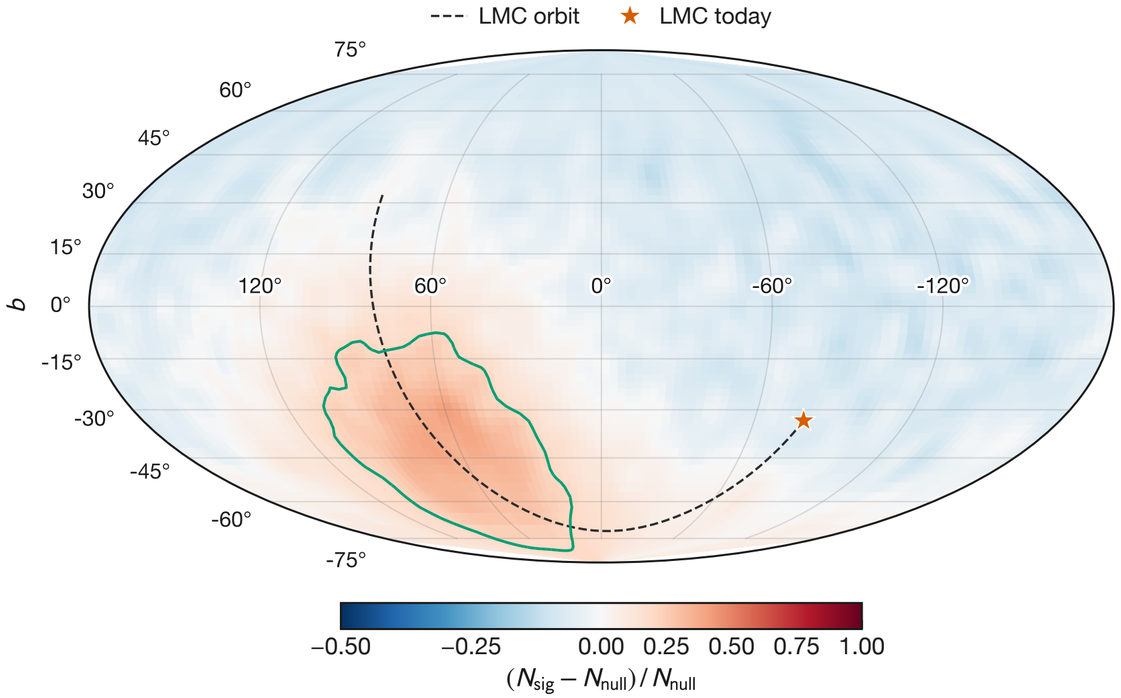
Modifying the signal-reference configuration to target LMC orbit

Transformation into mock observations

Simulation box is rotated along the LMC orbit as prescribed in Foote et al. (2023)

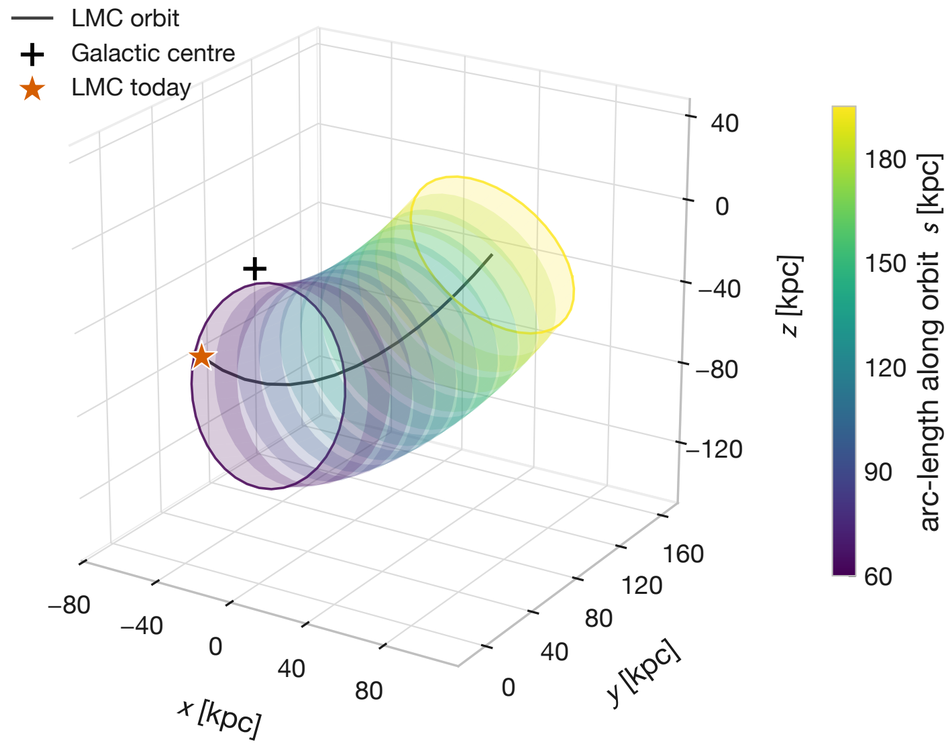


Mollweide projection of the rotated wake: **as observed from the Sun**



Extending reference strategy to LMC orbit

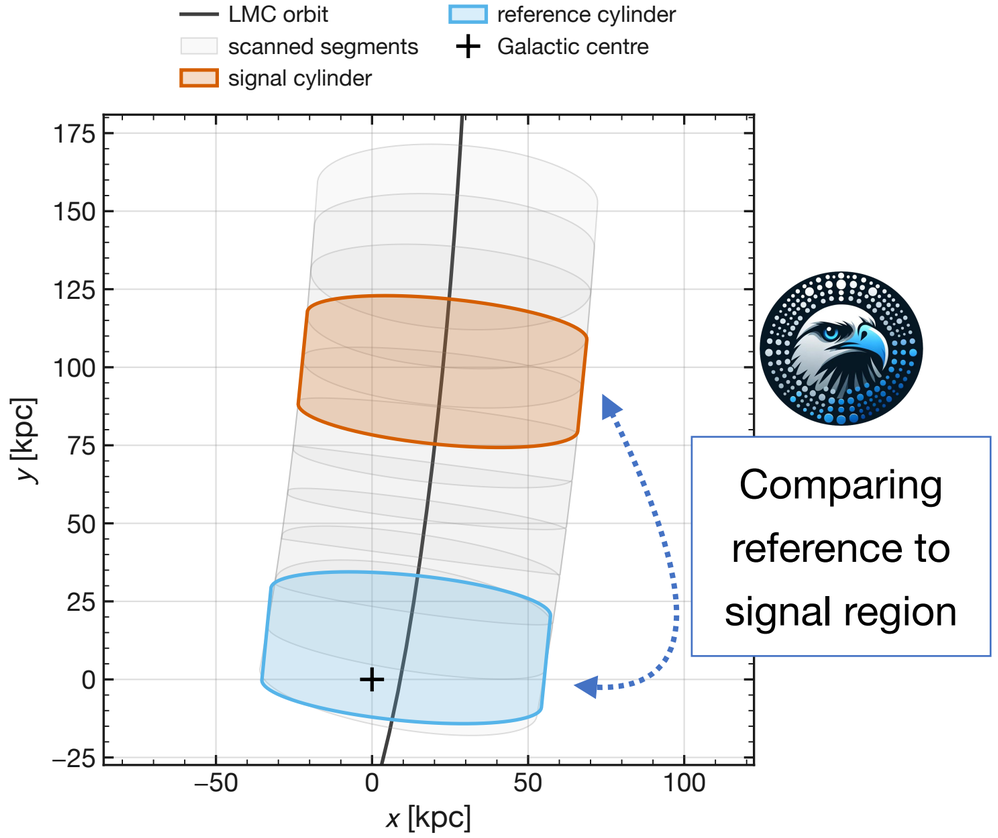
We construct a cylindrical corridor along the LMC's orbit



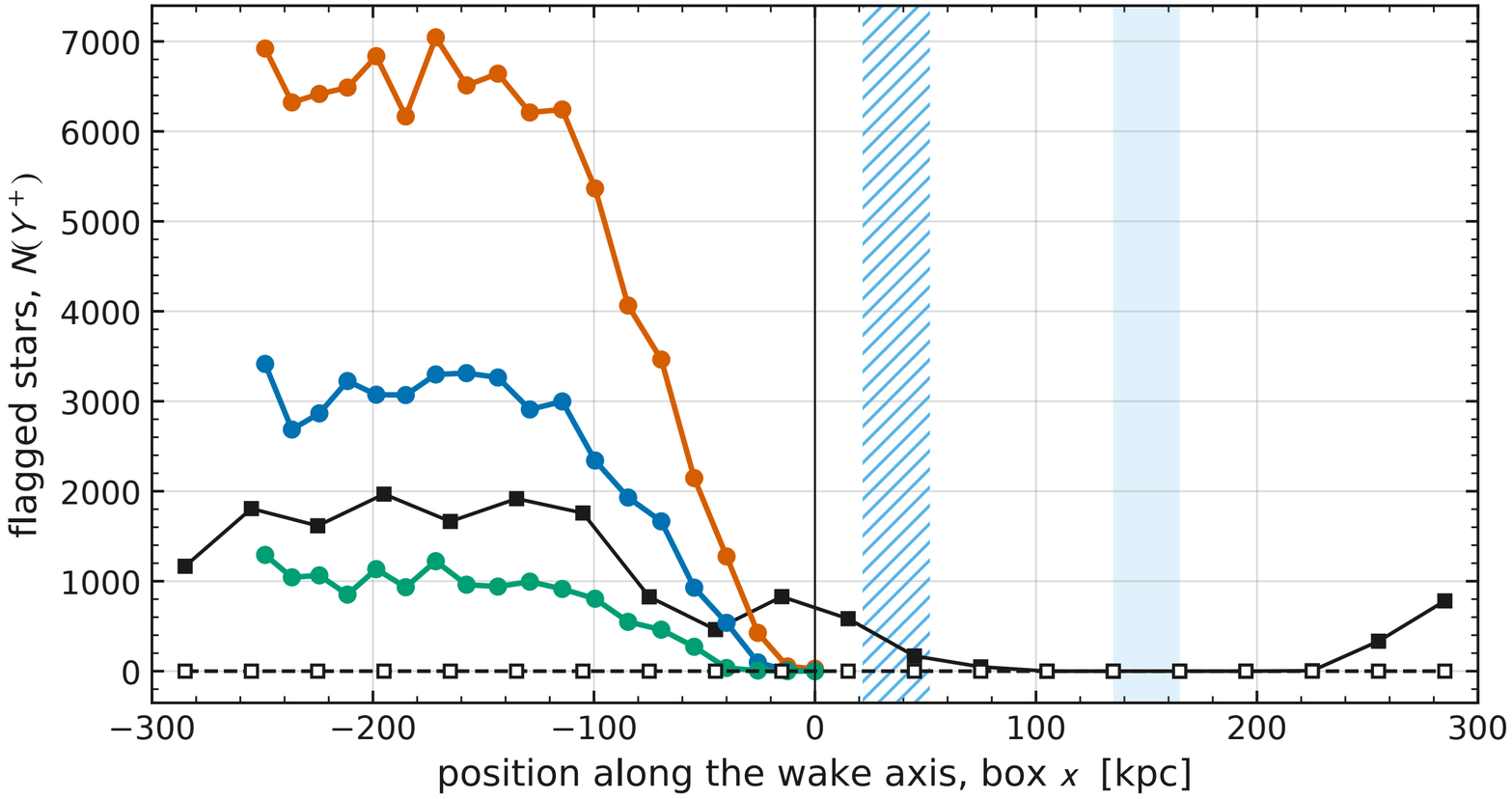
Orbit from Vasiliev et al. (2021)

TEST: Orbit corridor divided into signal regions (cylinders along orbit)

REFERENCE: Orbit segment where no wake is expected



Following the orbit recovers the wake, the wake-free control flags nothing



solid line, filled markers: signal run; dashed line, open markers: the corridor scan on the wake-free control

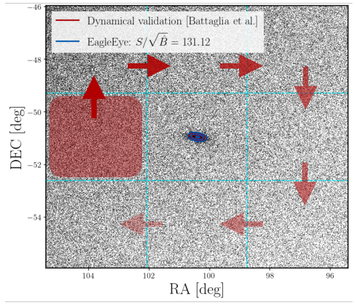
corridor's reference tile
worm's reference cylinder (72 kpc off the axis)
corridor scan, 50k, 6D
worm, 50k
worm, 25k
worm, 10k

TAKEAWAY

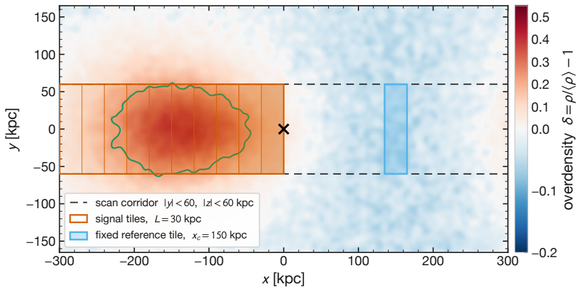
One model-independent framework, two faint signals

EagleEye is a new kNN-based density comparison method

DWARF GALAXIES



STELLAR WAKES



Applied as …

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

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**.

Sven Põder spoder@sissa.it

Bonus slides

Toward observations

Galactocentric transformation of RGB stars from Gaia DR3:

$$(\alpha, \delta, \varpi)_{GAIA} -> (x, y, z)_{GC}$$

Given by Gaia

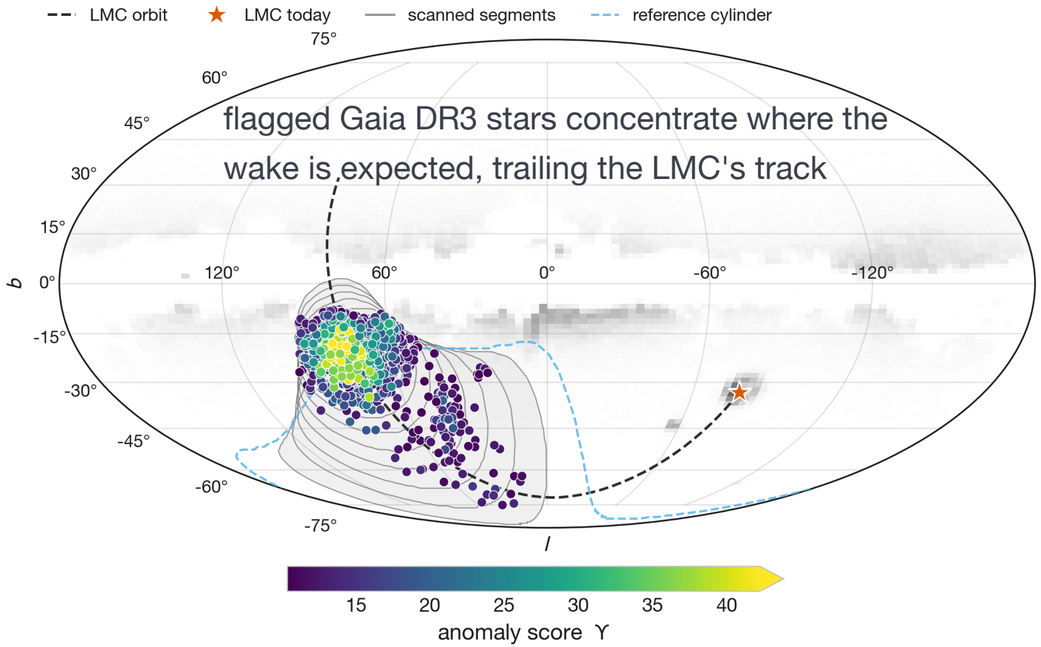
Segmentation + EagleEye

Similar to Põder et al. (2023)

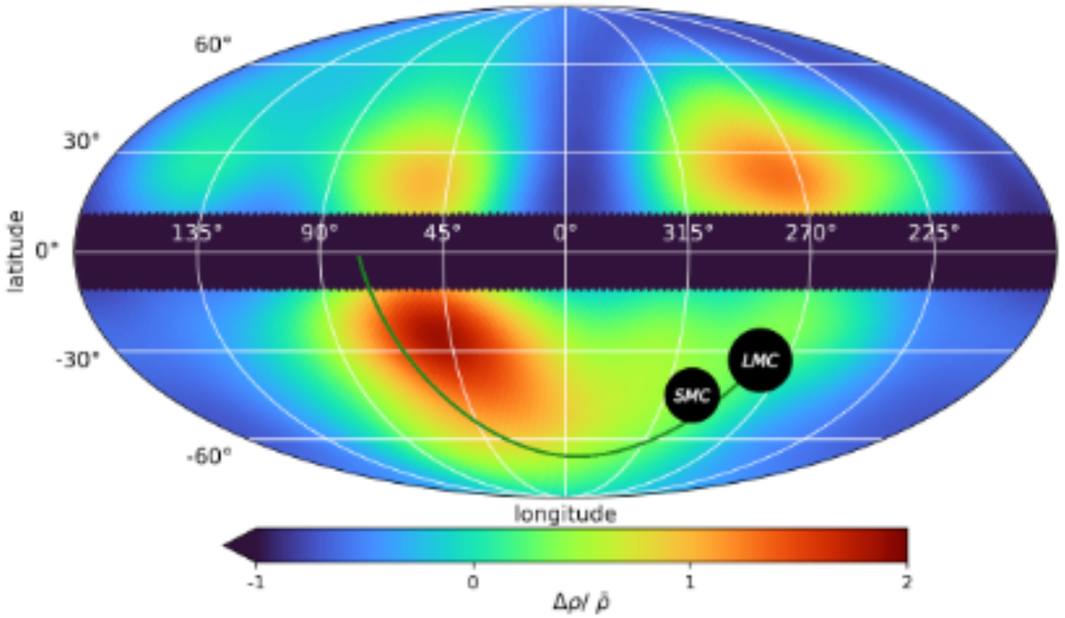
Ongoing work: kinematics beyond positions and improved distances, possibly with Gaia DR4

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)



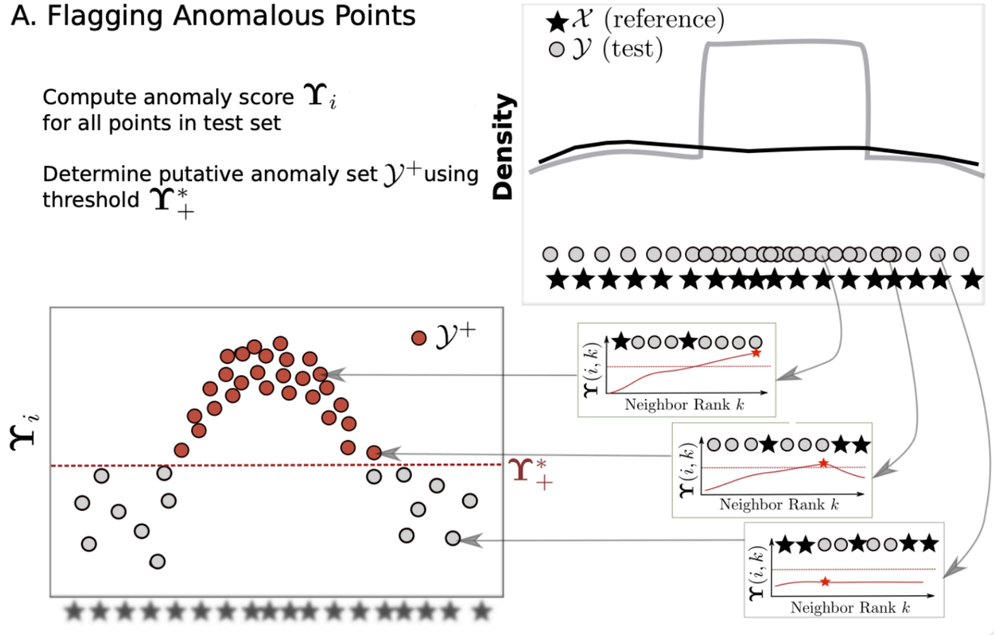
BACKUP

EagleEye full pipeline

A. Flagging Anomalous Points

Compute anomaly score Υ_i for all points in test set

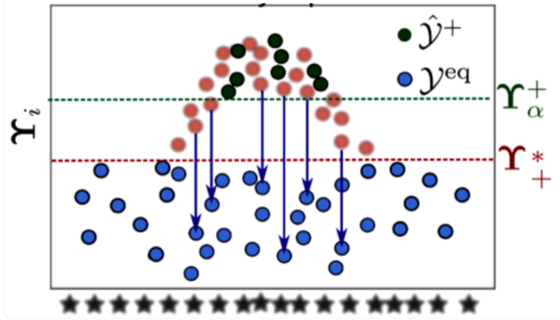
Determine putative anomaly set $\mathcal{Y}^+$ using threshold Υ^*_+



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

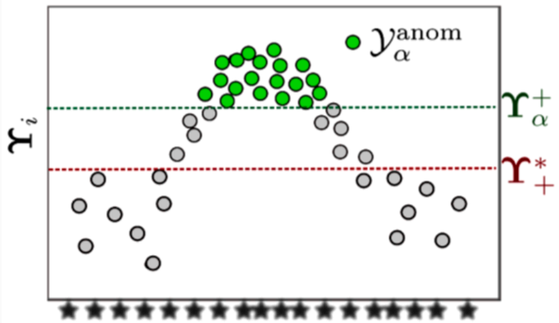
Figures adapted from arXiv:2503.23927

B. Iterative Density Equalization (IDE)



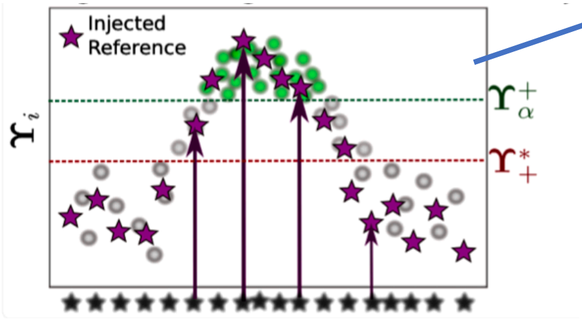
Iteratively remove points (dark green) starting with the most anomalous to equalize test and reference

C. Cluster Refinement



Define final anomalous set (light green points) by retrieving points with anomaly score higher than the minimum score in $\hat{\mathcal{Y}}^+$

D. Signal-to-background estimation via injection



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

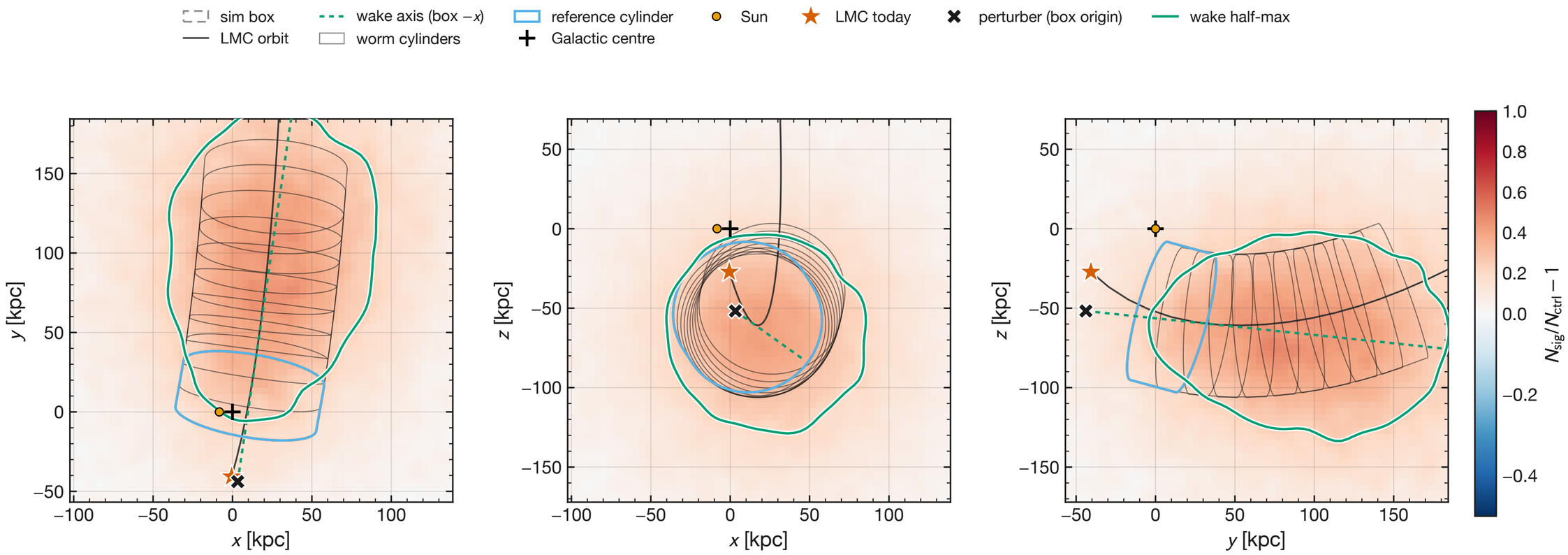
S_α — Expected number of signal points

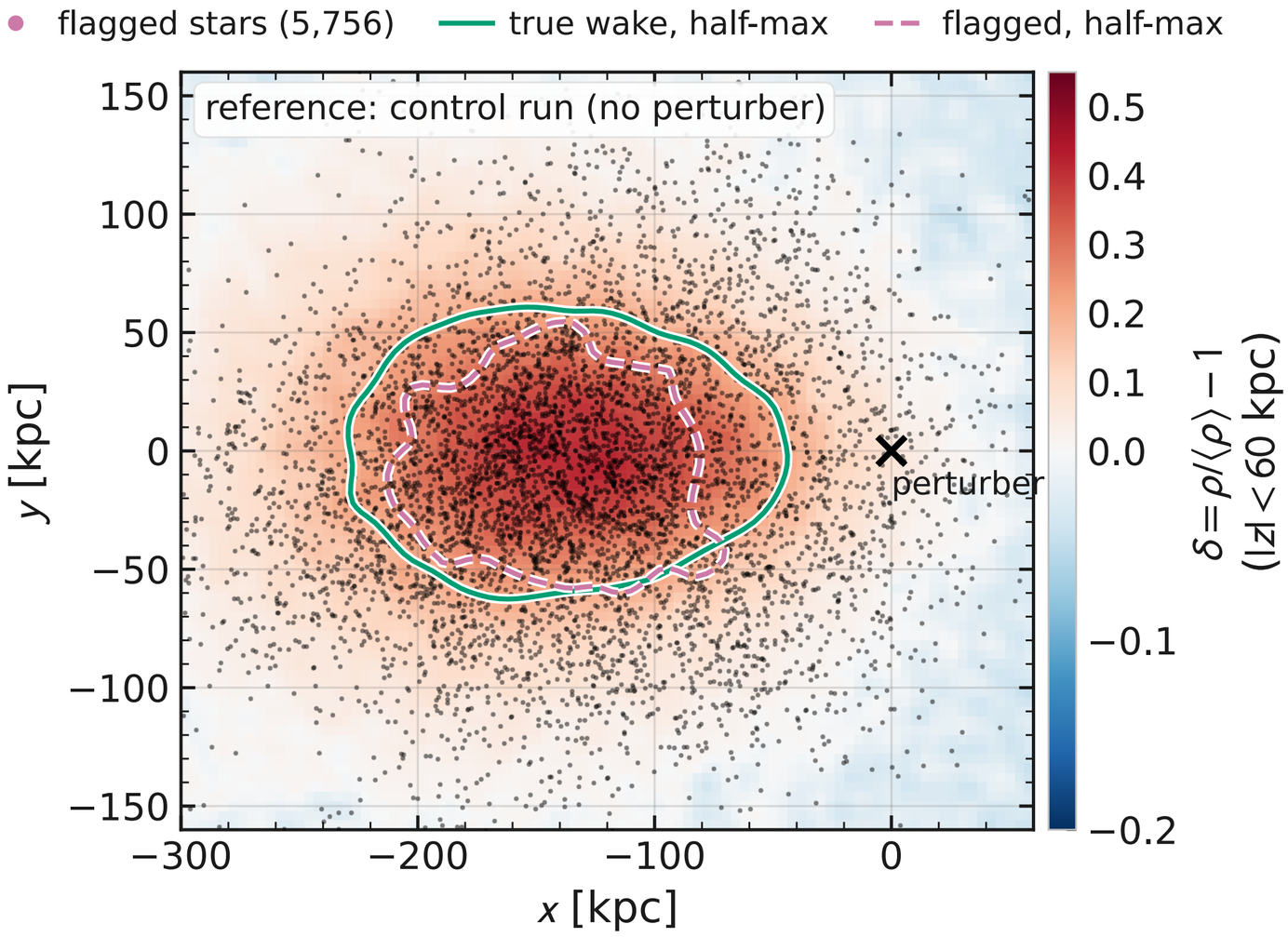
B_α — Expected number of background points contributing to $\mathcal{Y}^{anom}_\alpha$

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

$$\mathcal{Y}^{\text{inj}}_\alpha := \left\{ X_i \in \underline{\mathcal{Y}}^+ \,\middle|\, \Upsilon_i \geq \Upsilon^+_\alpha \right\}$$

How well does the rotation match the geometry of the orbit?





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