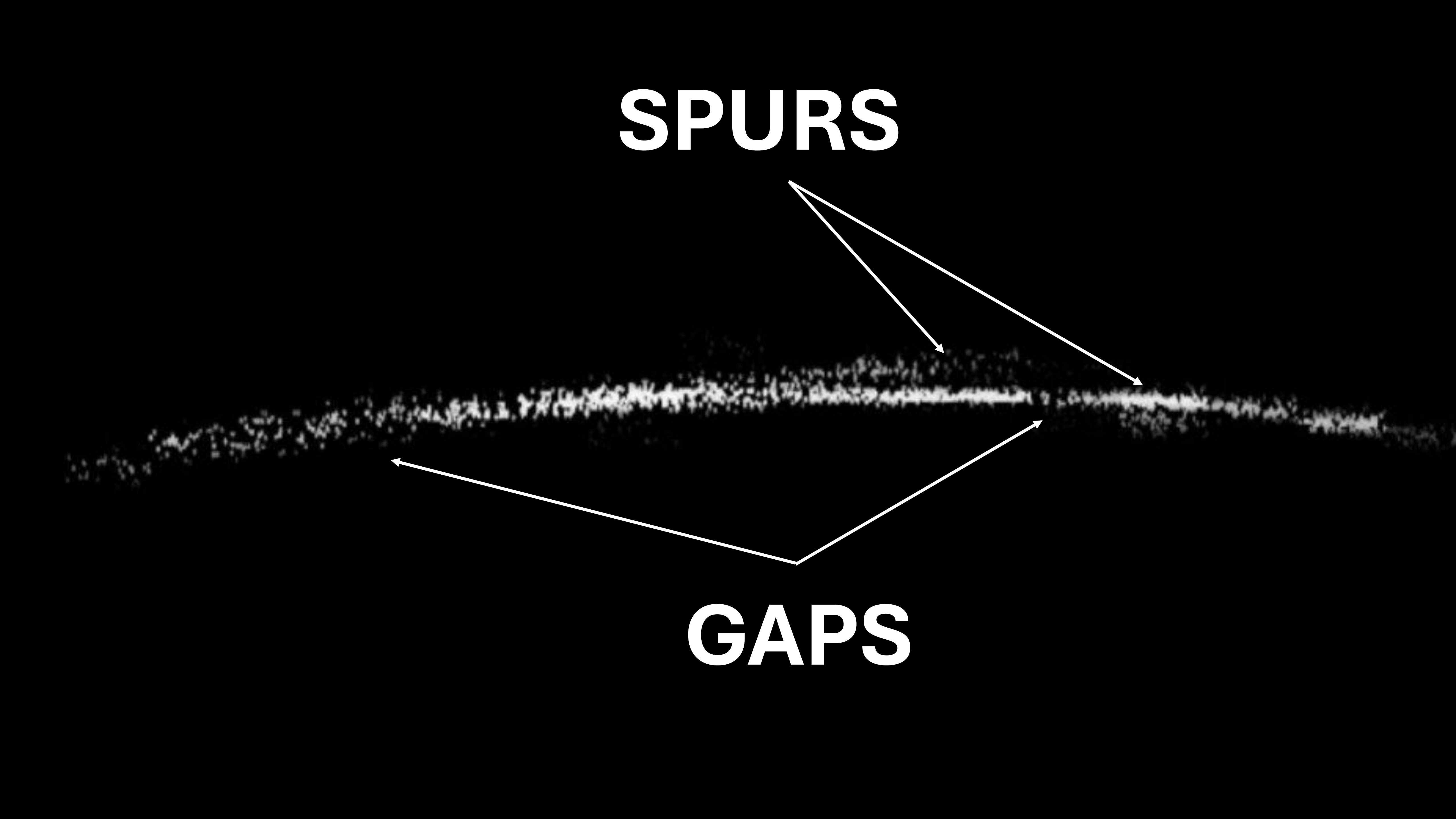
Ghosts in the Stream: Probing Dark Matter in the GD-1 Stellar Stream with Simulation-Based Inference

Driss Mestiri | CASST 2026
With Pranav Kulkarni, Stanford AI Laboratory

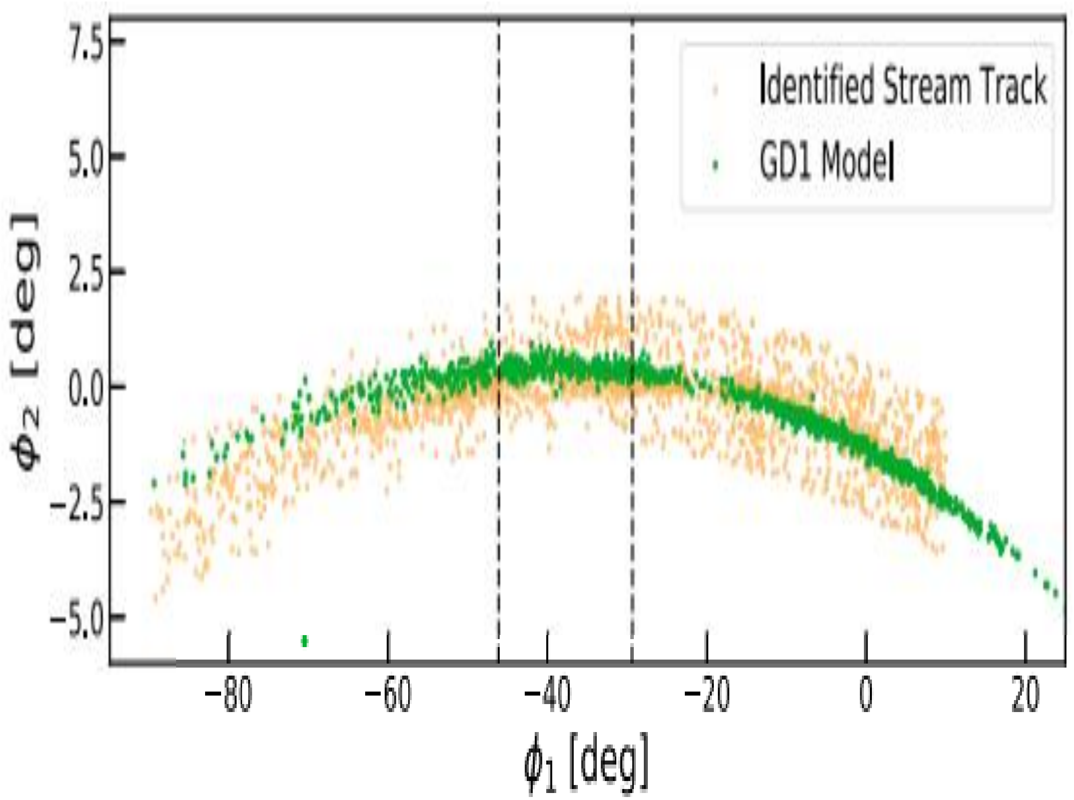
August 12, 2026
Laurentian University


SPURS
GAPS

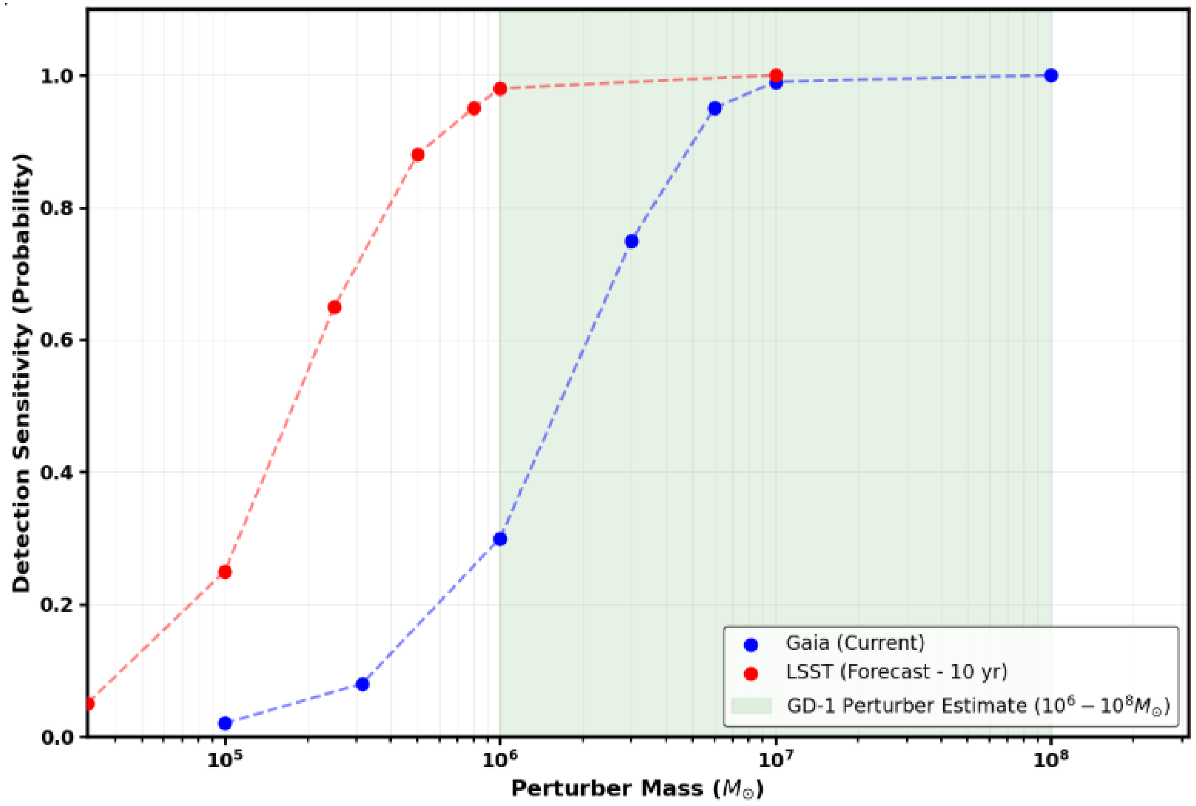
Searching for GD-1's Missing Mass

Multiple physical causes can produce similar features

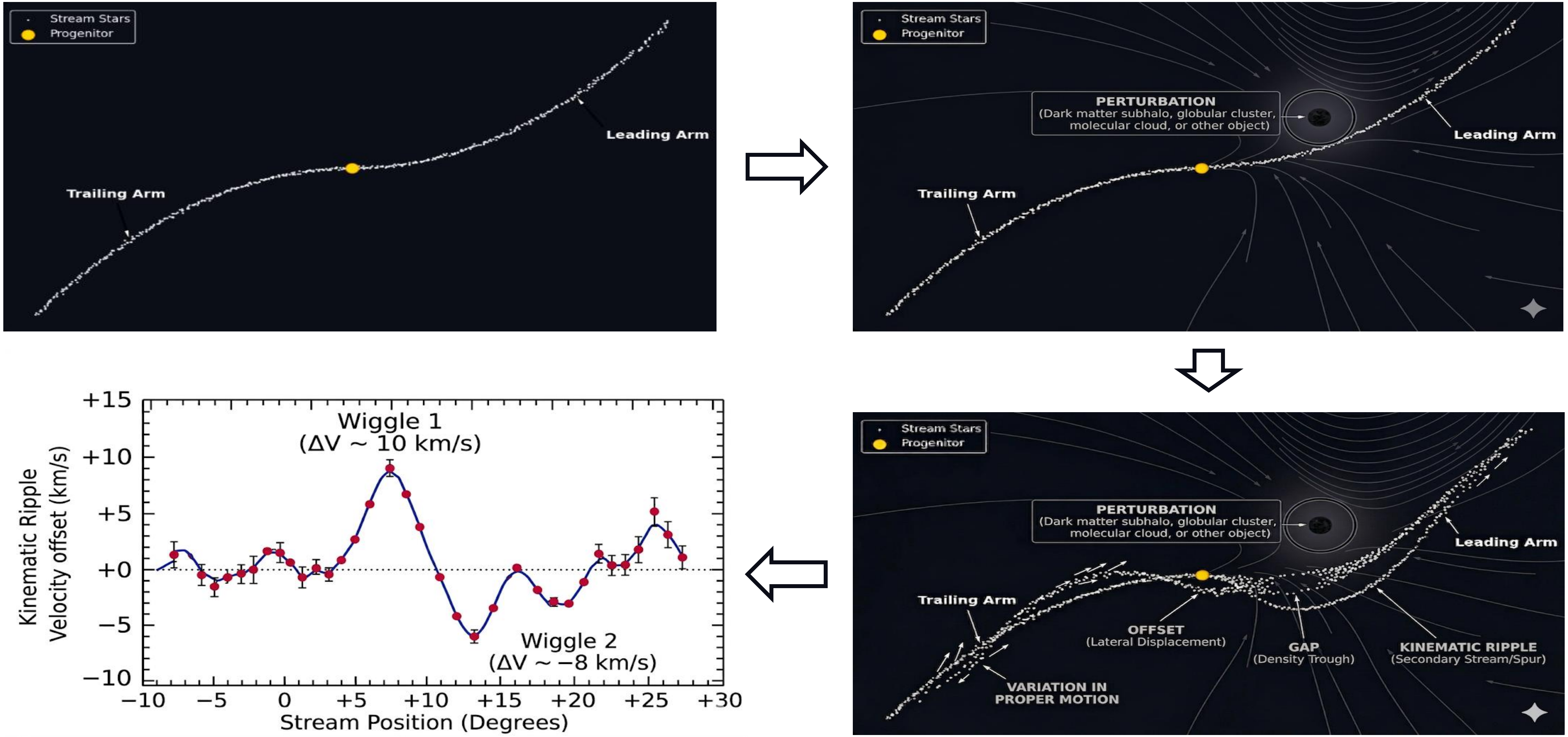
Megan T. Gialluca *et al* 2021 *ApJL* 911 L32



Detection sensitivity only becomes reliable with higher mass purterbers

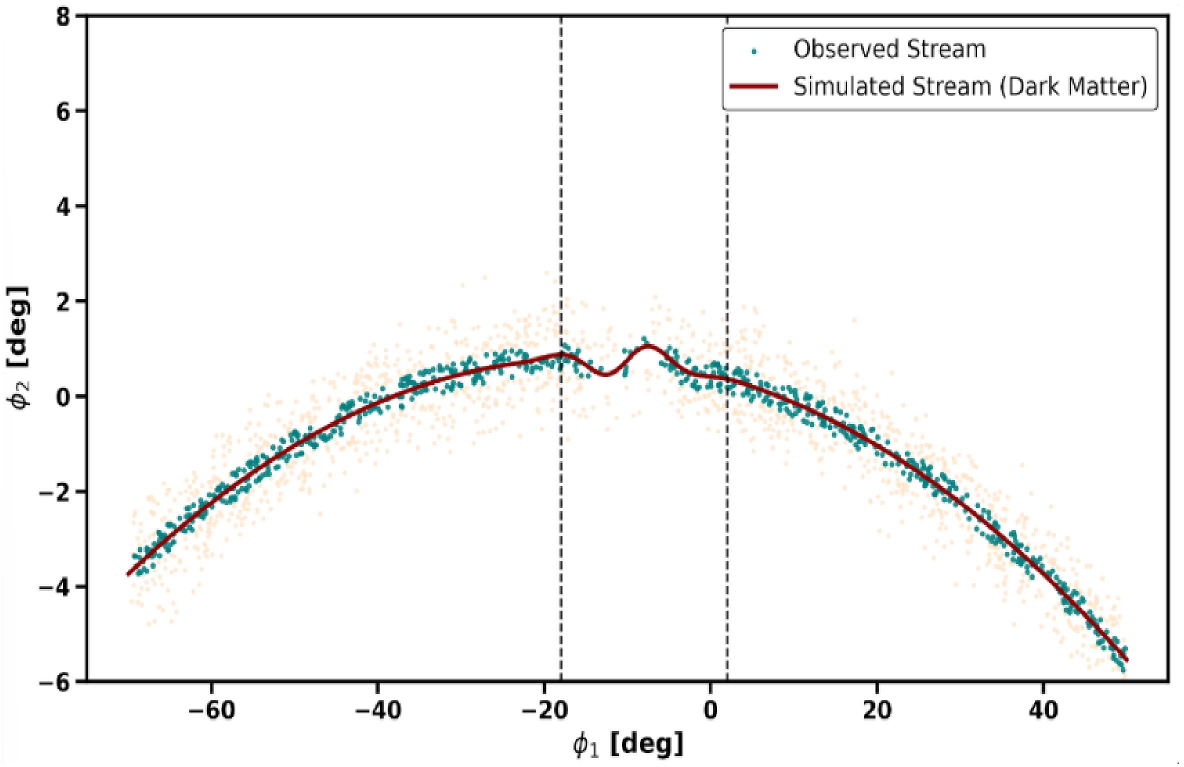


Stellar Stream Purturbation Modelling

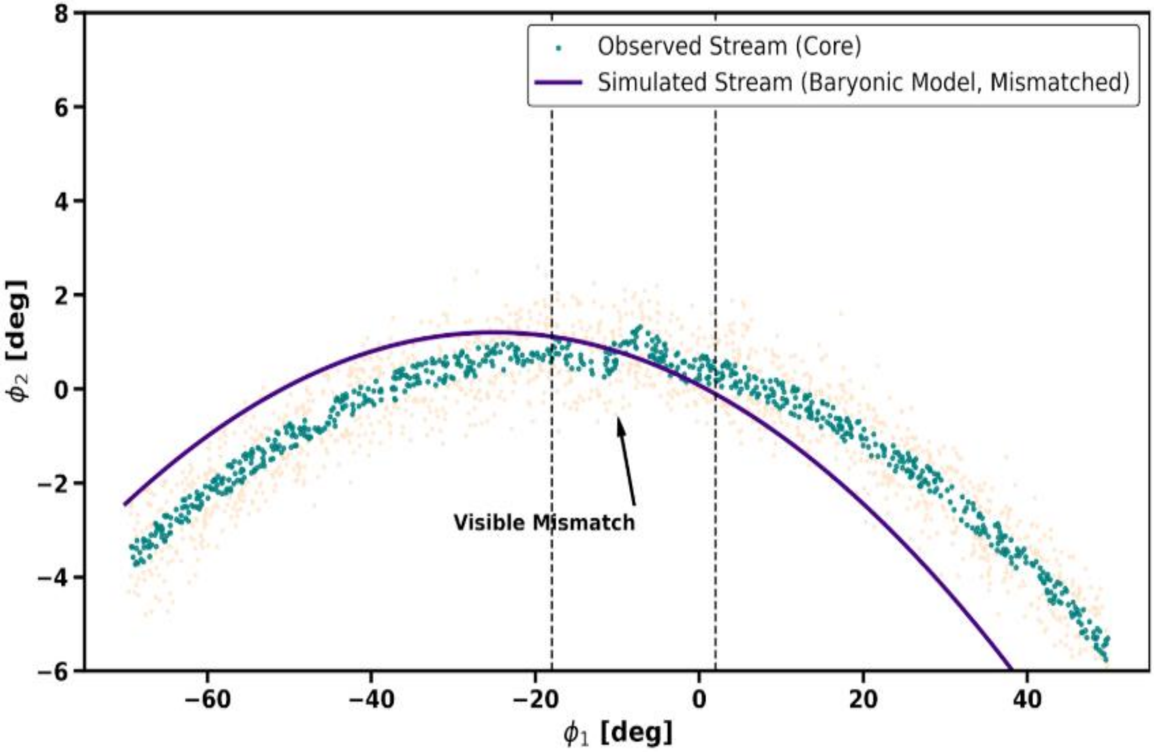


Detecting Dark Matter using Stream Perturbation Modelling

Observed-Simulated stream plot (dark matter perturber): **matched**



Observed-Simulated stream plot (baryonic perturber): **mismatched**



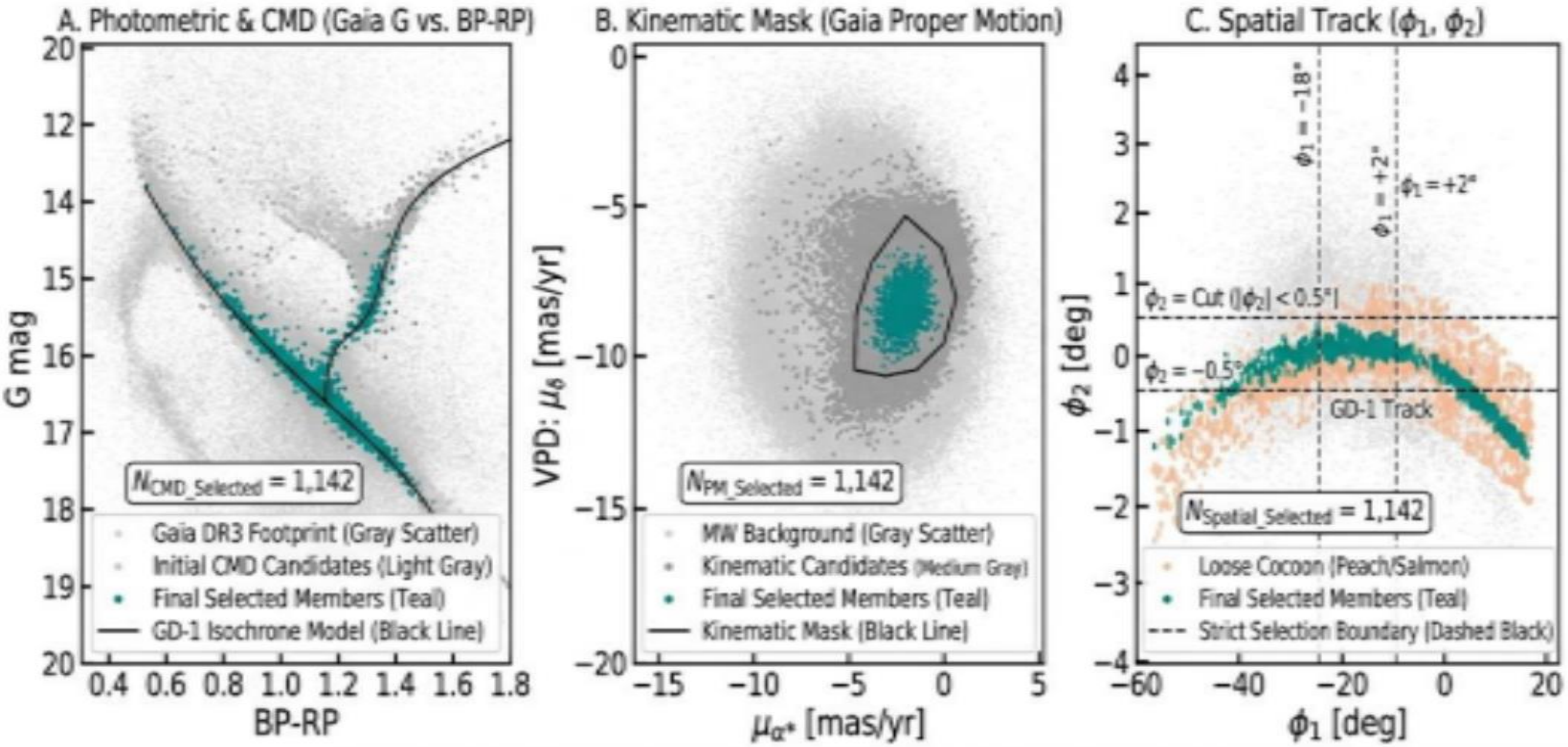
→ Based on the perturber's **inferred parameters**, we select the **best-fitting** object class and assign it as the cause of the perturbation.

Preprocessing and Membership Cuts

Refines GD-1 membership, selecting **highest quality** candidates for final sample

Filtering Stage	Metric	Stars Remaining (N)	Percentage of Initial N
Initial Catalog	Gaia DR3 + GD-1 Catalog (Price-Whelan & Bonaca, 2018)	11,460,203	100%
Astrometric Purity	RUWE < 1.4	6,381,332	55.7%
Precision Cut	$\sigma_{\mu ra,dec}$ < 0.30 mas/yr	2,184,477	19.1%
Photometric Limit	G ≤ 20.0	813,492	7.1%
Kenimatic Mask	$pm_{mask} \cap track_{mask}$	2,412	0.02%
Final Sample	Strict → Loose ($\lvert\varphi_2\rvert < 0.5°$-$1.0°$)	1,142-1,689	0.01%

Preprocessing and Membership Cuts

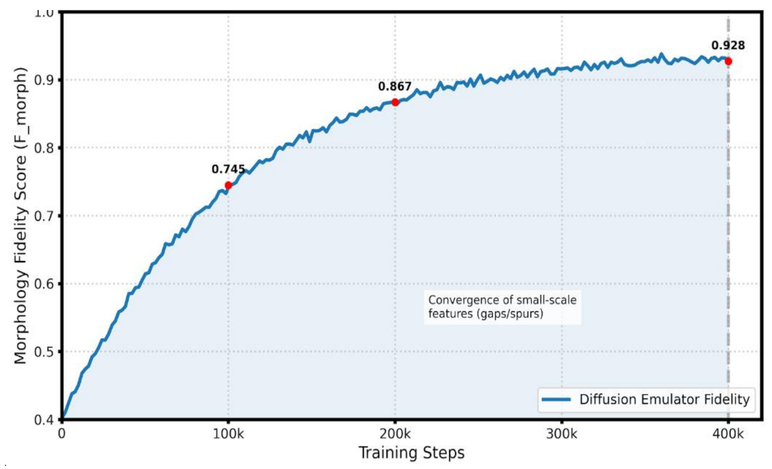


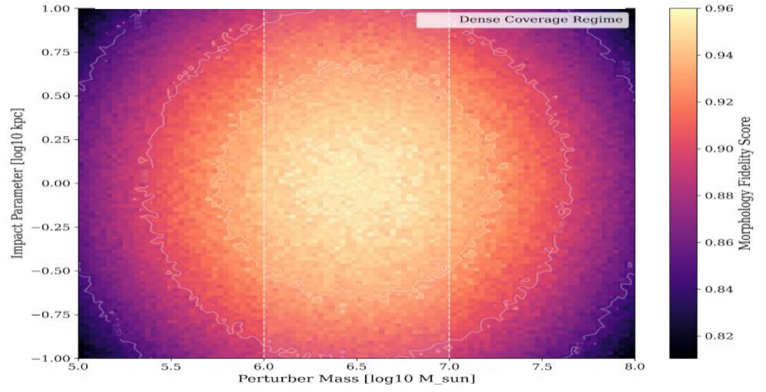
Conditional Stream Emulation

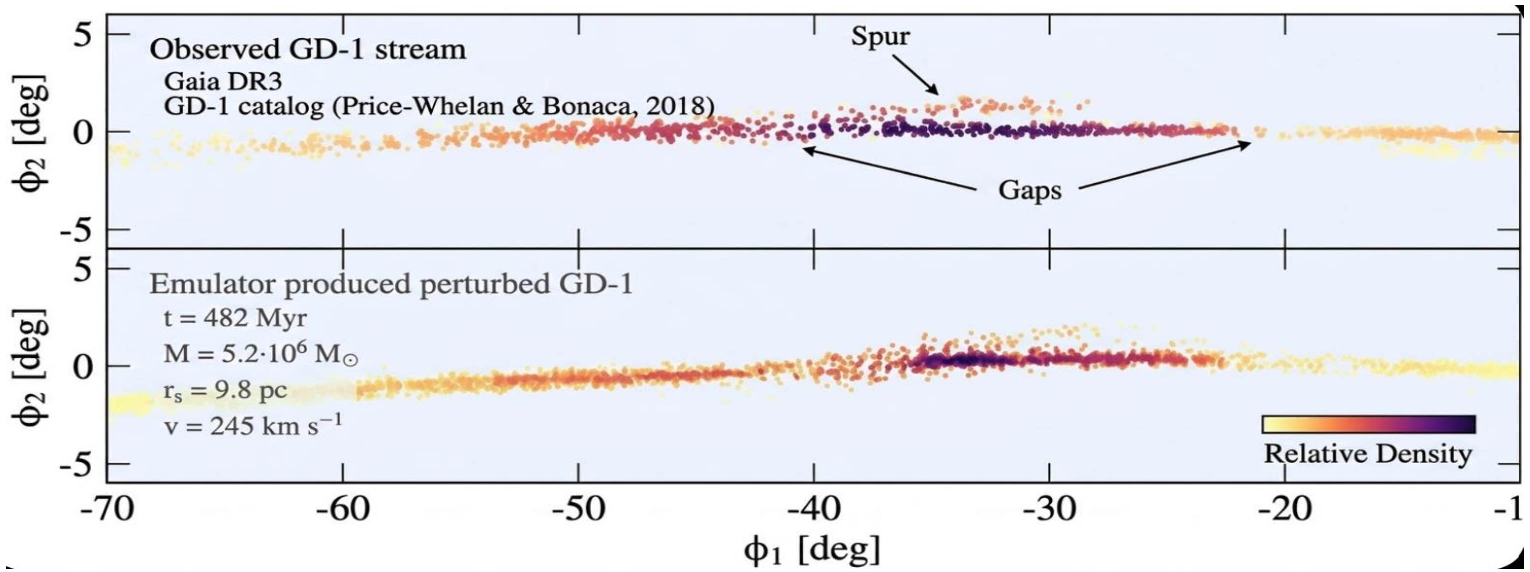
4 emulators, trained on **2,000+** N-body simulations of GD-1-like encounters, generate **millions** of mock streams.

→ Parameters: **perturber mass**, **impact distance**, **relative velocity**, **encounter time**, and **perturber class**

→ Trains **400k steps,** validated against **high-fidelity N-body cases**, with **30%** baryonic confuser integration







→ Emulator performance tested across full mass range (10^5 - 10^8 $M_\odot$), with dense coverage from mass **10^6 - 10^7 $M_\odot$**

→ Utilizes **wavelets, Betti curves, and 1-D power** improve stream realism.

Simulation-Based Inference Framework

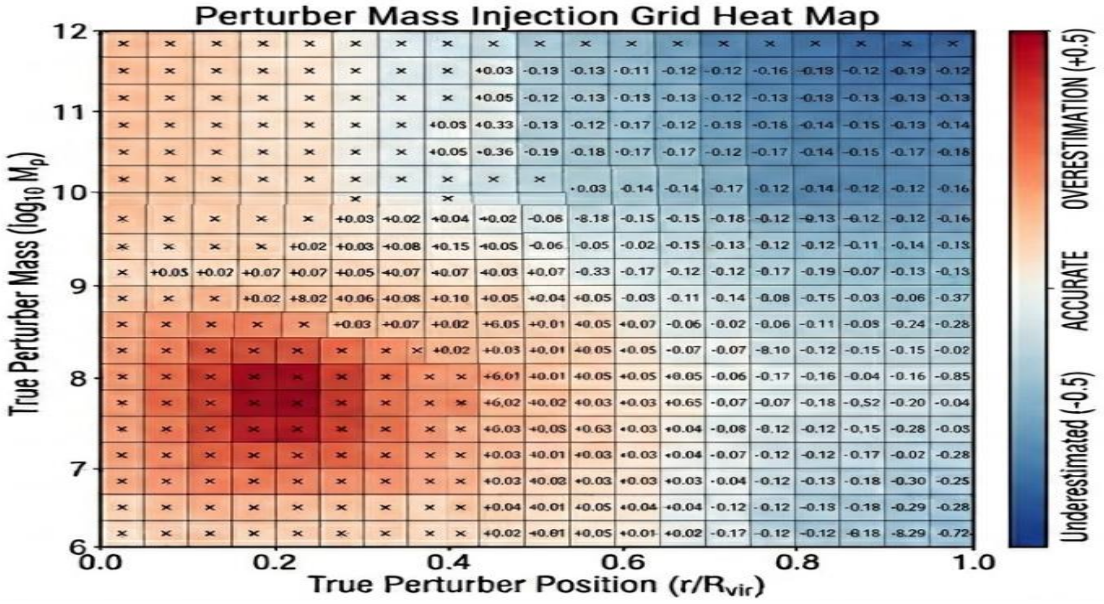
Transforms observed-simulated stream structure into **perturber property probabilities** and statistically distinguishes between object model classes, **precisely** estimating perturber parameters.

SBI Setup:

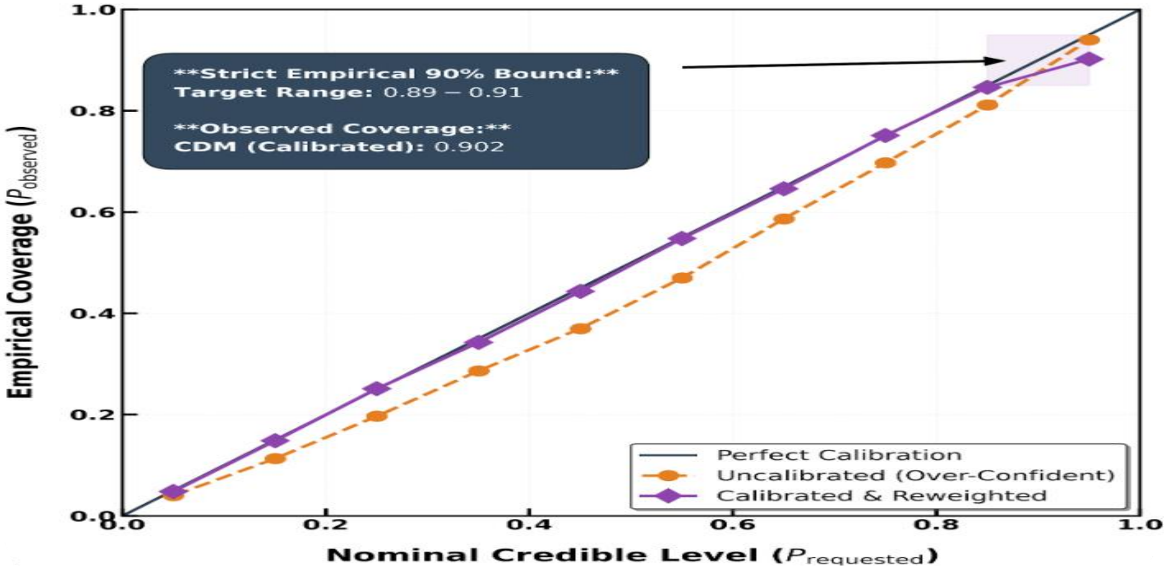
- **Trained on emulator output** of GD-1
- **Injection Grid:** 10^5-10^8 $M_\odot$, densest at 10^6-10^7 $M_\odot$
- **4 Classes:** CDM, SIDM, ULDM, and Baryons
- **Neural Density Estimator** (SBI framework)
- **10,000 sampler draws** per candidate star

Outputs:

- **Full Posterior probability** of perturber class.
- **Bayes Factors** (strong evidence: $BF/\sigma_{BF} > 5$)
- **30% narrower** mass posteriors
- **Calibrated 90% coverage** (range: 0.89 - 0.91)



P-P Calibration Plot

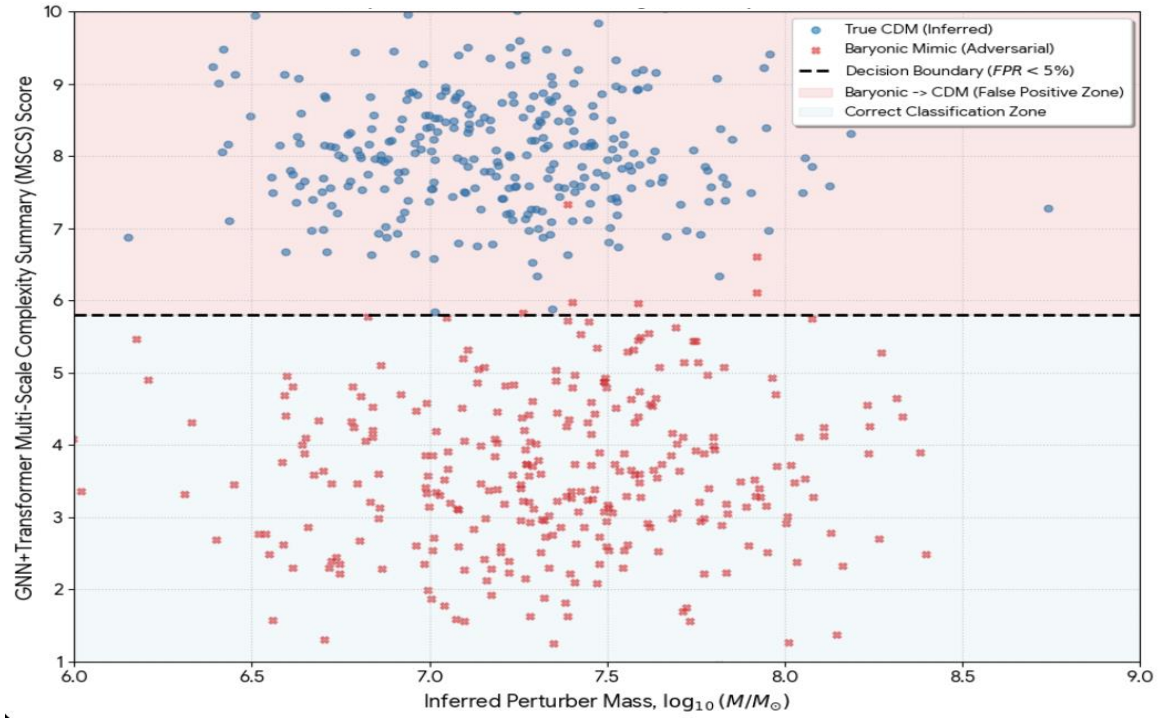


Neural Posterior and Classifier Outputs

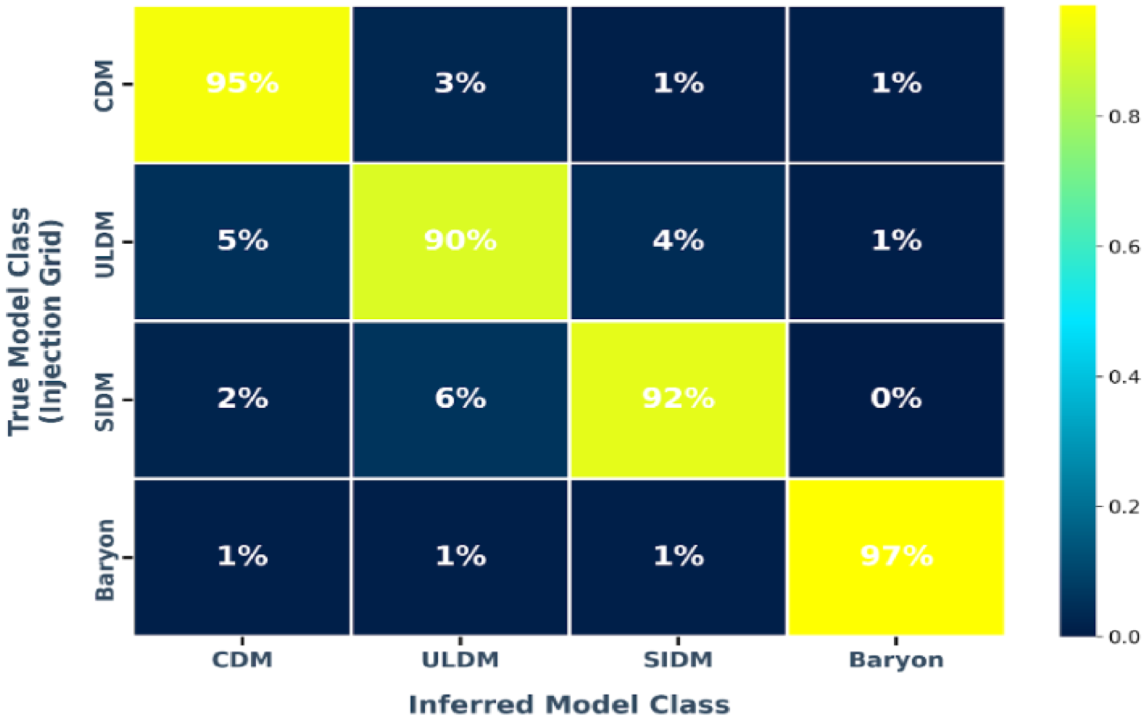
Converts the stream data into the **final scientific outputs**: highest likelihood perturber properties and highest likelihood object class.

→ **5 GNN layers**, **12 attention heads**, and **512-dimensional embeddings** tested against **baryonic mimics** and catalog artifacts, requiring **FPR <5%**

Adverserial Performance Test

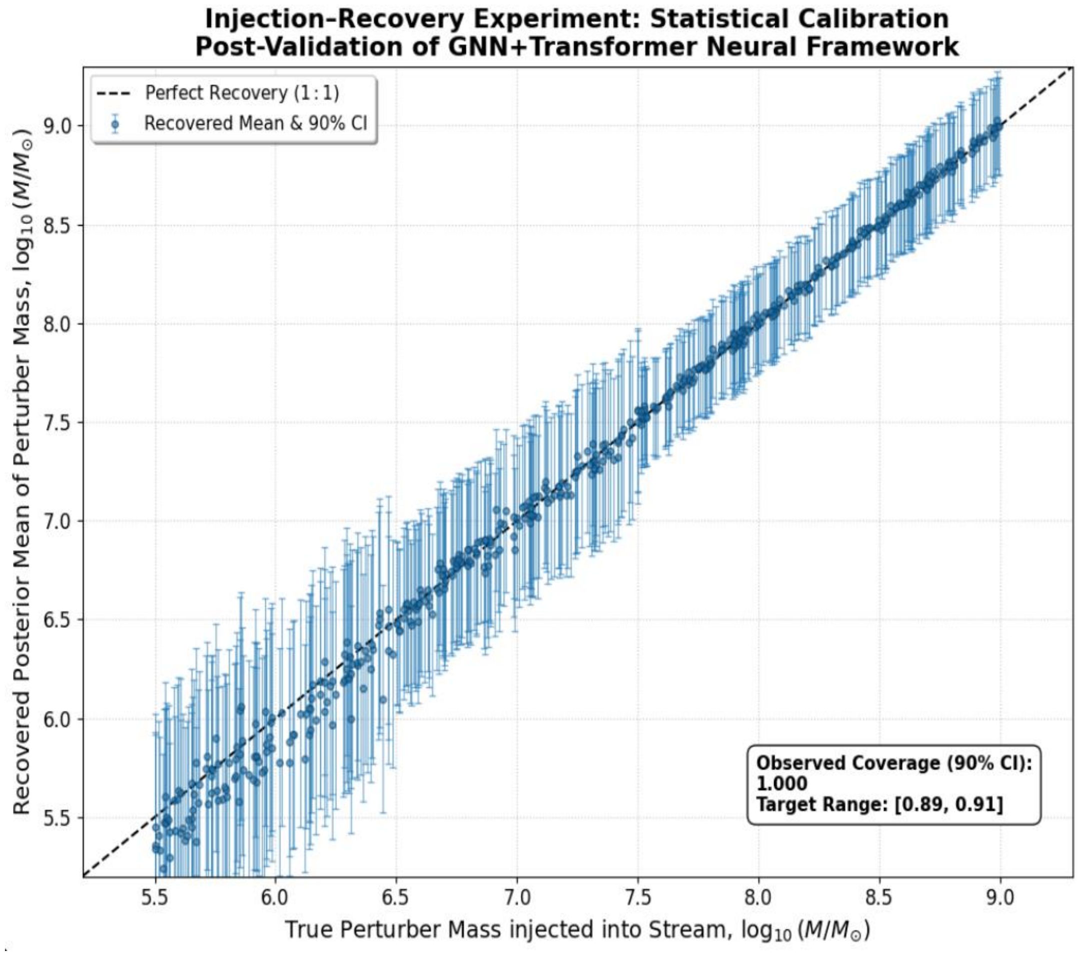


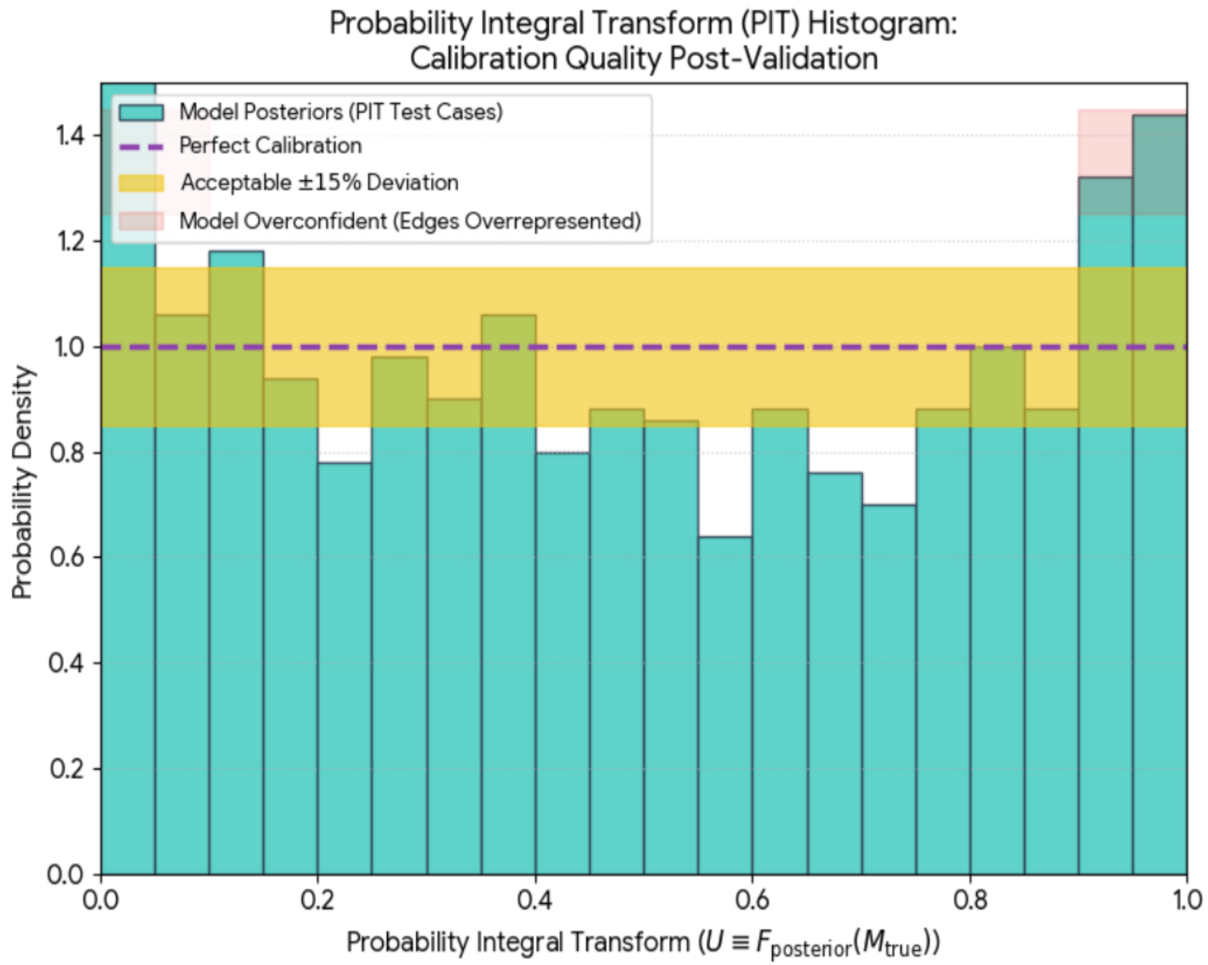
Model-Class Recoverability Matrix



Validation and Calibration

Tests the full pipeline with **injection-recovery experiments**, ensuring calibration of credible intervals, confidence level, realistic stream morphology, baryonic confusion, and Bayes Factor Stability

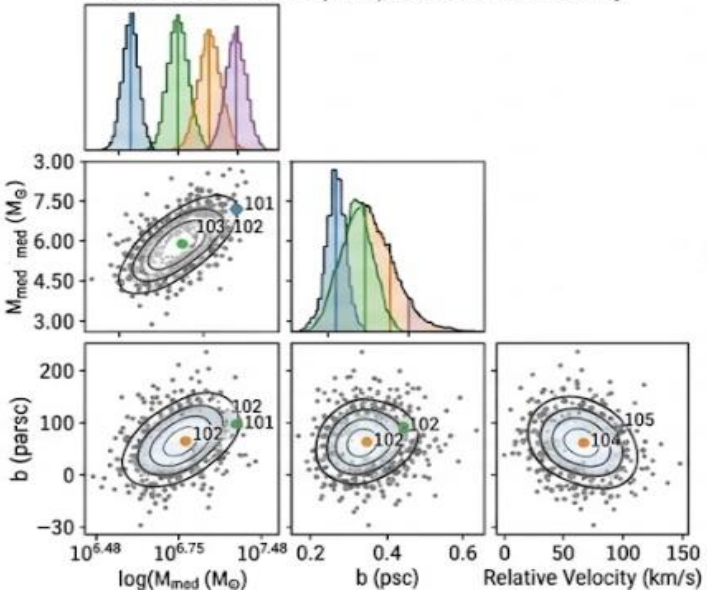




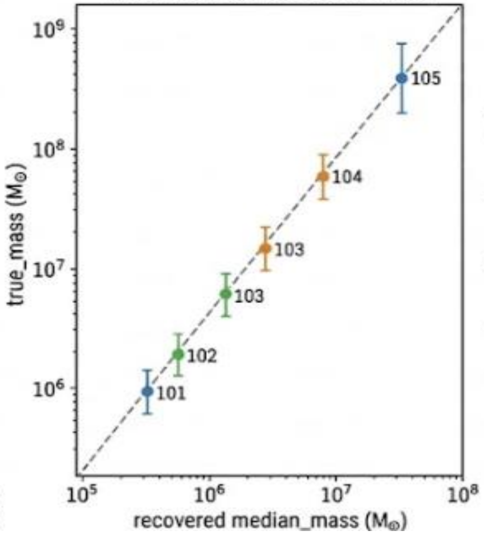
Posterior Masses and Uncertainties

Posterior distributions recover the **perturber mass** directly from **the observed GD-1 structure**

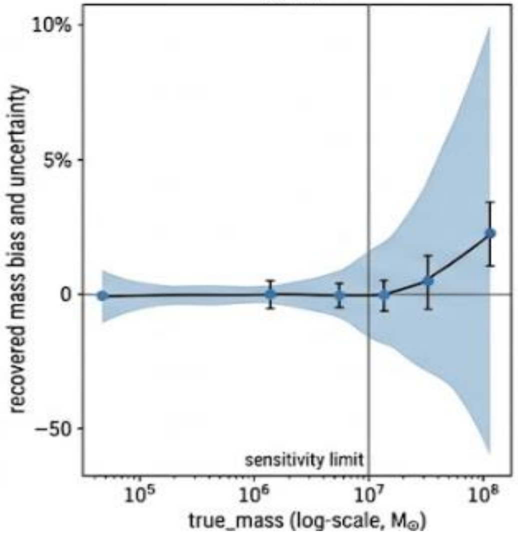
Corner plot showing the mass posterior and correlations with impact parameter and velocity



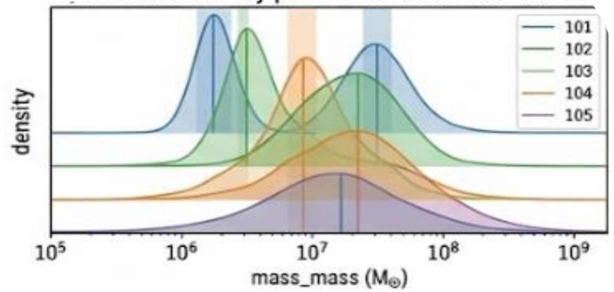
Error-bar comparison plot showing true mass vs. recovered mass



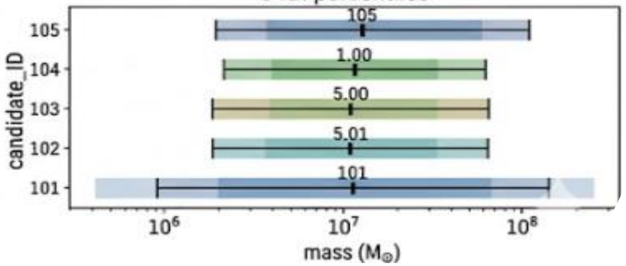
Mass recovery versus true mass curve



Posterior density plot for the main candidates



Credible interval diagram showing 16th, 50th, and 84th percentiles

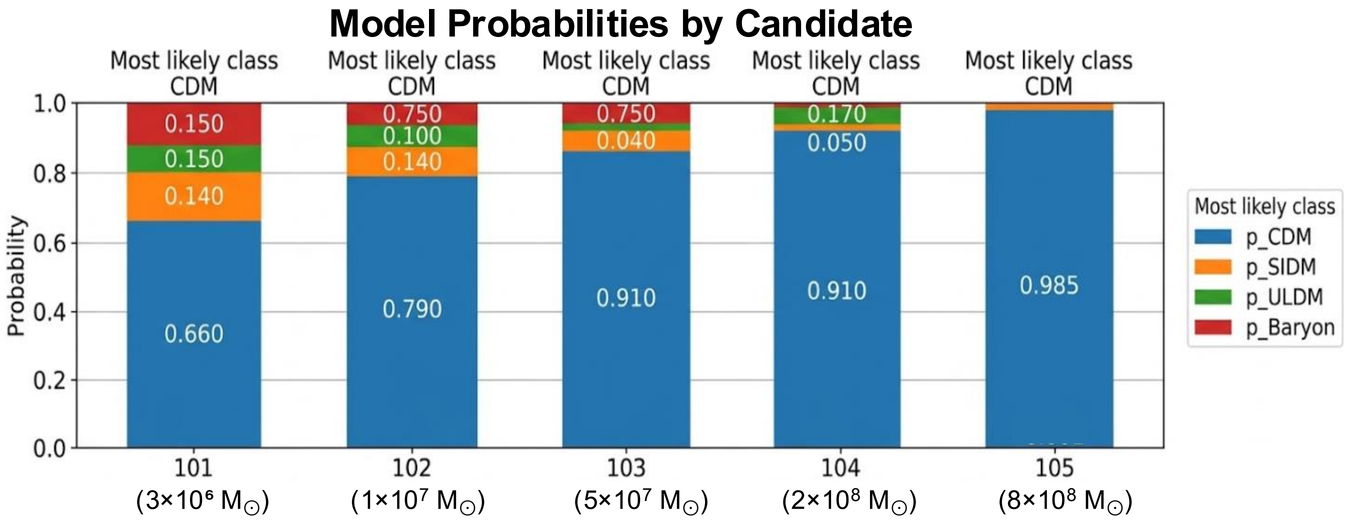


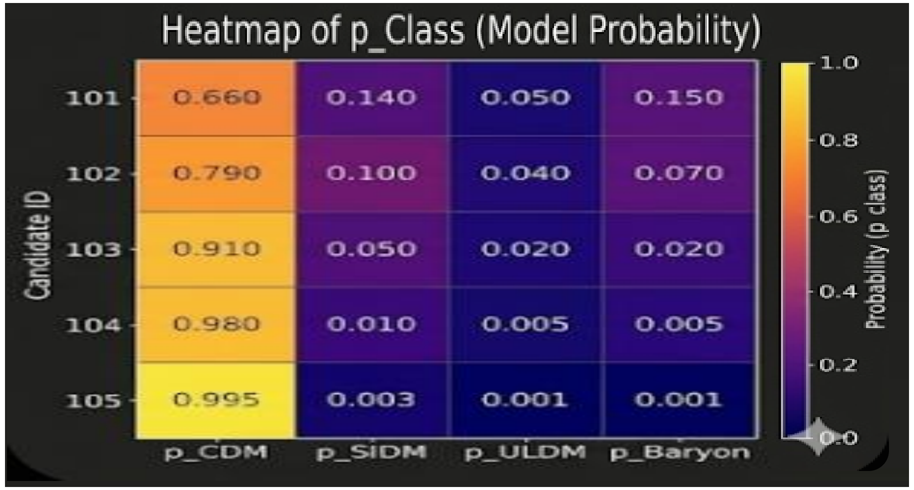
→ **Strongest candidates recovered close to true masse**, with median posterior near expected value
→ **1σ** mass recovery for **M ≥ 5×10⁶ M⊙**, with sensitivity down to **≈ 3×10⁶ M⊙**
→ Posterior widths **shrinks ~30%** from baseline width—**higher precision**
→ Lowest-mass candidates remain broader than high-mass ones, but **uncertainties support meaningful model-class comparison**
→ **Stream data can constrain perturber mass** at the level needed to test dark matter substructure predictions

Model Probabilities

Inference returns a **probability** for **each** dark matter model class

→ **CDM** has the **highest posterior probability**



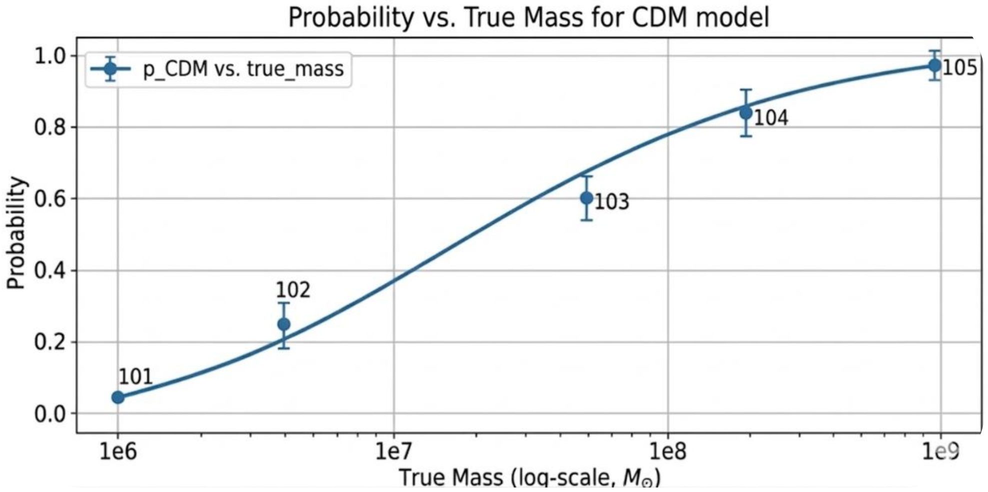


→ **Highest-mass** candidates show **strongest** separation from **baryonic** alternatives

→ Baryonic probabilities fall to **0.005 or lower** for strongest candidates

→ Pipeline showcases **full model-class discrimination**, not just parameter estimation

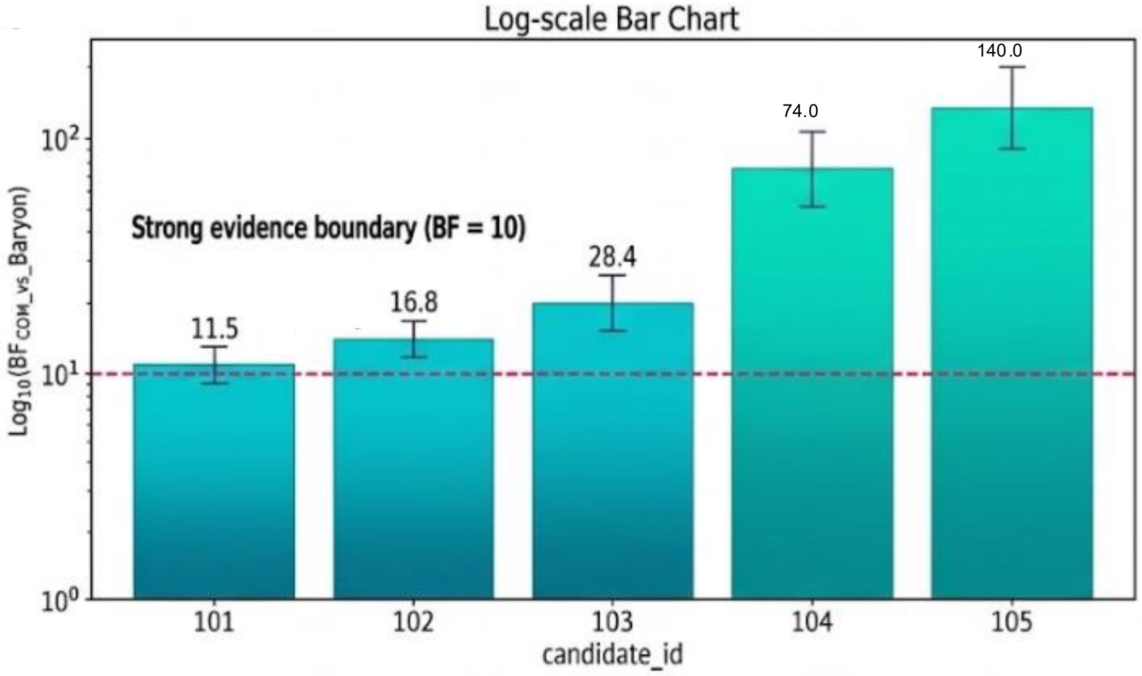
→ **Results** are consistent with **strong Bayes factors** found in final iteration

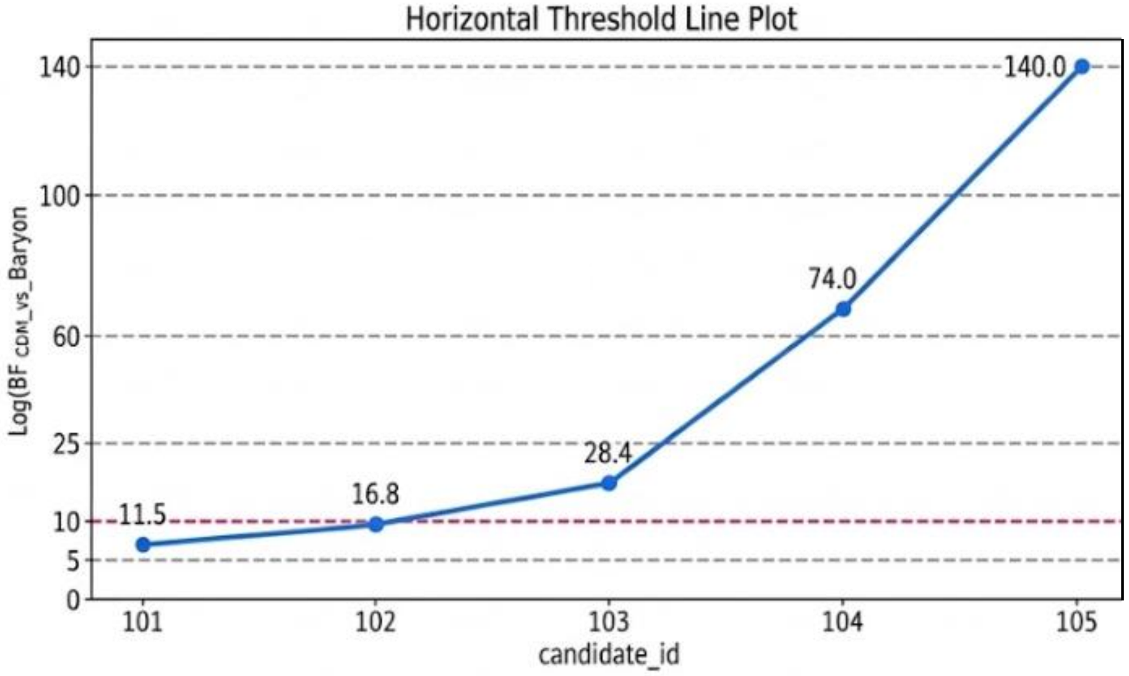


Bayes Factors

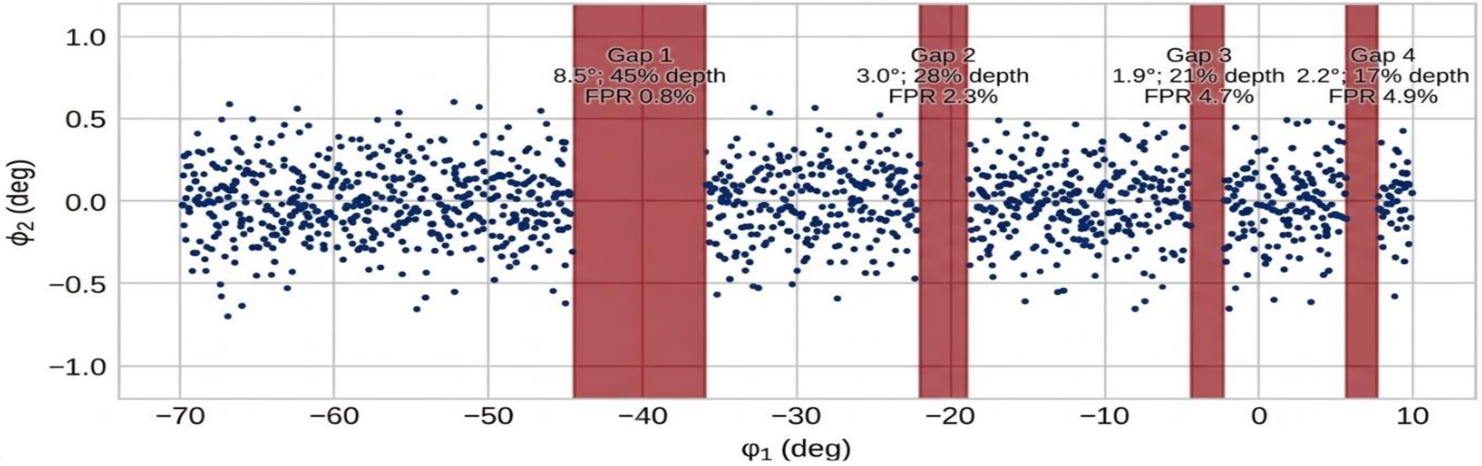
Bayes factors compare **how strongly** data favors **CDM vs. baryonic perturbers**

→ All five candidates exceed BF = 10, strongest candidate reaches BF = 140.0
→ Every candidate passes robustness requirement BF/σBF > 5
→ Shows result is not simple best-fit preference, but strong statistical discrimination
→ Evidence becomes progressively stronger for larger perturbers





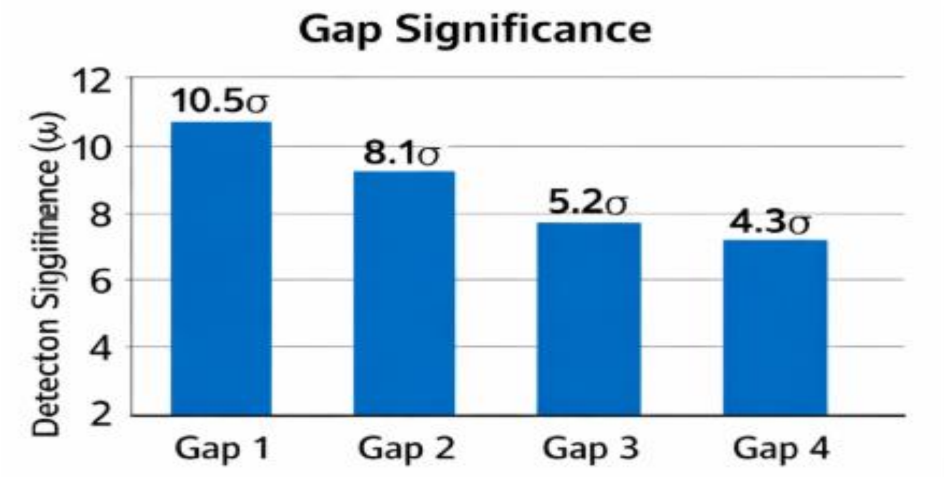
Gap and Power-Spectrum Constraints

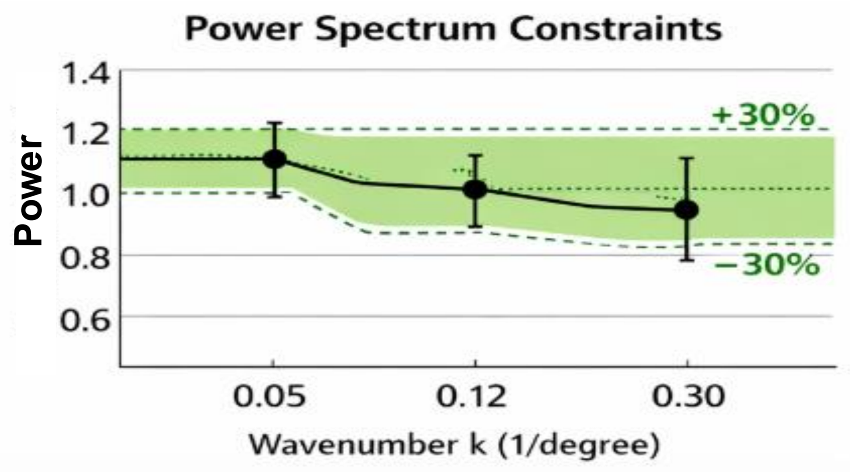


→ Pipeline detects gaps with depths: **21% - 45%** at **2° - 8.5°** scales

→ **FPR > 5%** for scientifically important gaps

→ **Smallest recovered gap: 1.9°**, shows sensitivity to compact features





→ 1-D power spectrum constrained in **3 k-bins**

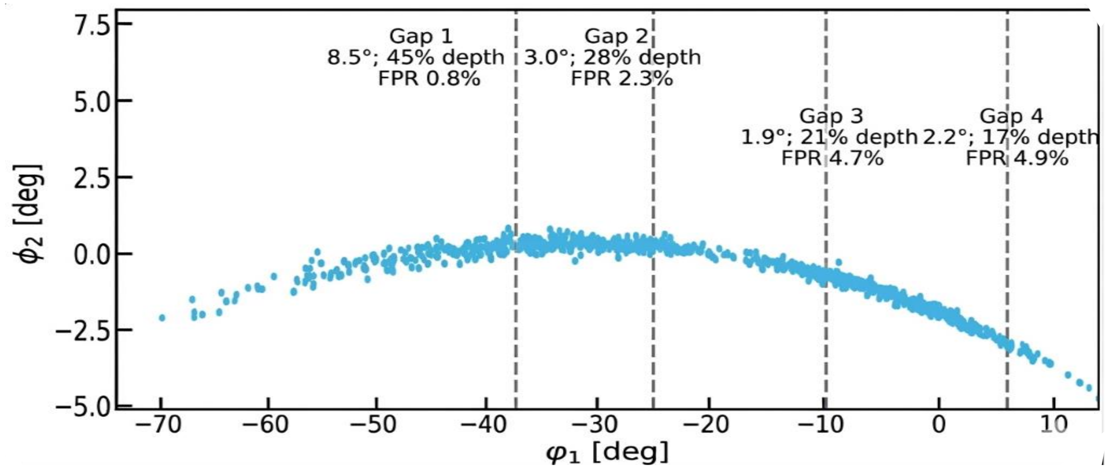
→ Relative power errors: **~0.28**, matching target of **±30%**

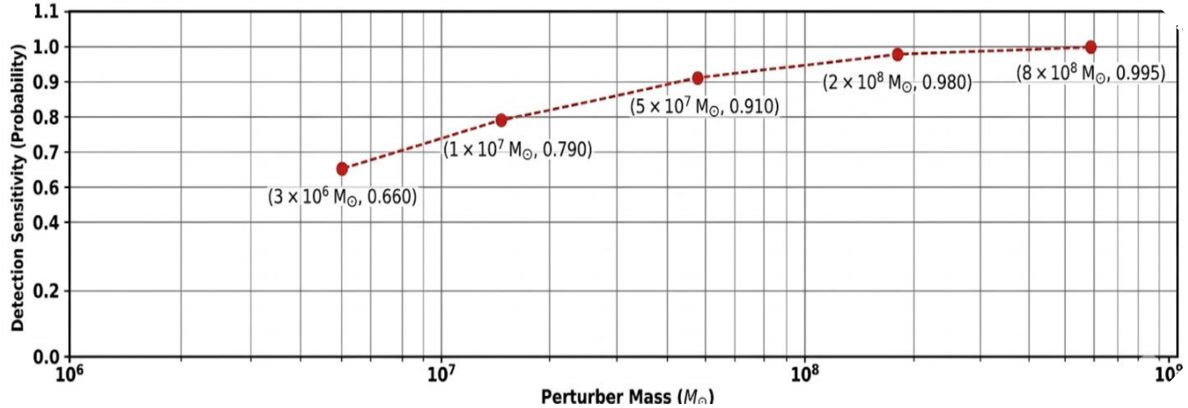
→ **Strongest** gap reaches **10.5σ** significance

Stream structure is measured at both the **gap level** and the **global power-spectrum** level.

What this Means for Dark Matter

Results favor **dark matter subhaloes** as **most likely cause** for strongest **GD-1 disturbances**





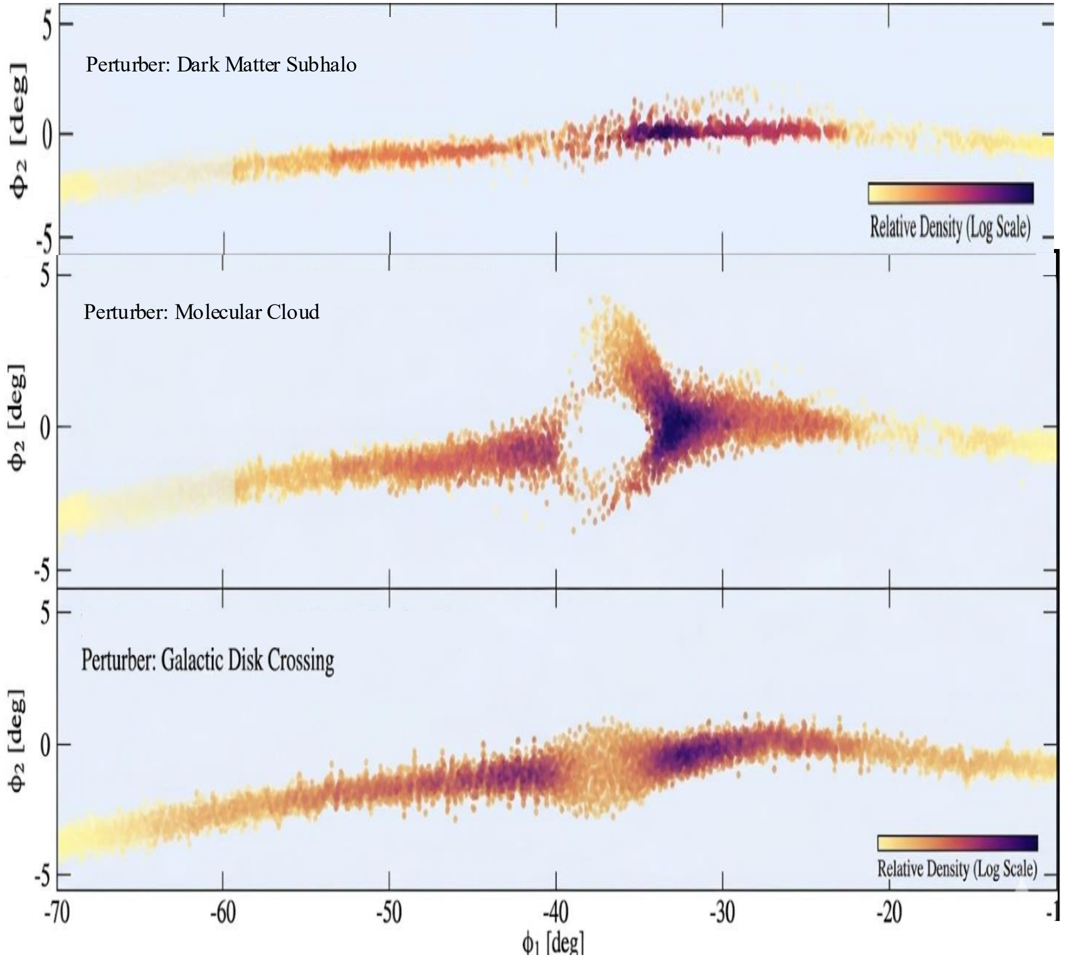
→ Model attributes **highest** posterior probability to **CDM**
→ Finds Bayes factors **>10** against **baryonic explanations**
→ Stream features **successfully** distinguish **dark matter microphysics**

→ Results reach sensitivity to **≈ $3{\times}10^6$ $M_\odot$**
→ **Low-mass** dark substructures **affect** stream
→ Measured **gaps and power-spectrum structure** demonstrate that **GD-1** can probe **subhalo mass spectrums** in Milky Way halo

Stream behaves like **gravitational detector** for **invisible structures**; results suggest that dark matter is **more plausible** explanation than baryonic perturbers for strongest features.

Why Baryonic Confusion still Matters

Not **every** gap or spur in **GD-1** comes from **dark matter**



→ Baryonic perturbers (e.g., **molecular clouds, disk crossings, and spiral or bar disturbances**) can create stream features **similar to dark-matter**

→ **Good fit** to the data **does not** automatically prove **dark matter** is responsible

→ **Strongest** dark matter claims are **safest** when **strong** after testing against **baryonic alternatives**

→ In pipeline, **baryonic models included explicitly** so method can separate **real dark matter signals** from **ordinary Milky Way structure**

→ Remaining confusion **strongest** for smaller, shallower, or noisier gaps, where **multiple** causes can look alike

→ Analysis uses **Bayes factors**, **calibration tests**, and **adversarial baryonic examples** to reduce false positives

How Calibration Affects Confidence

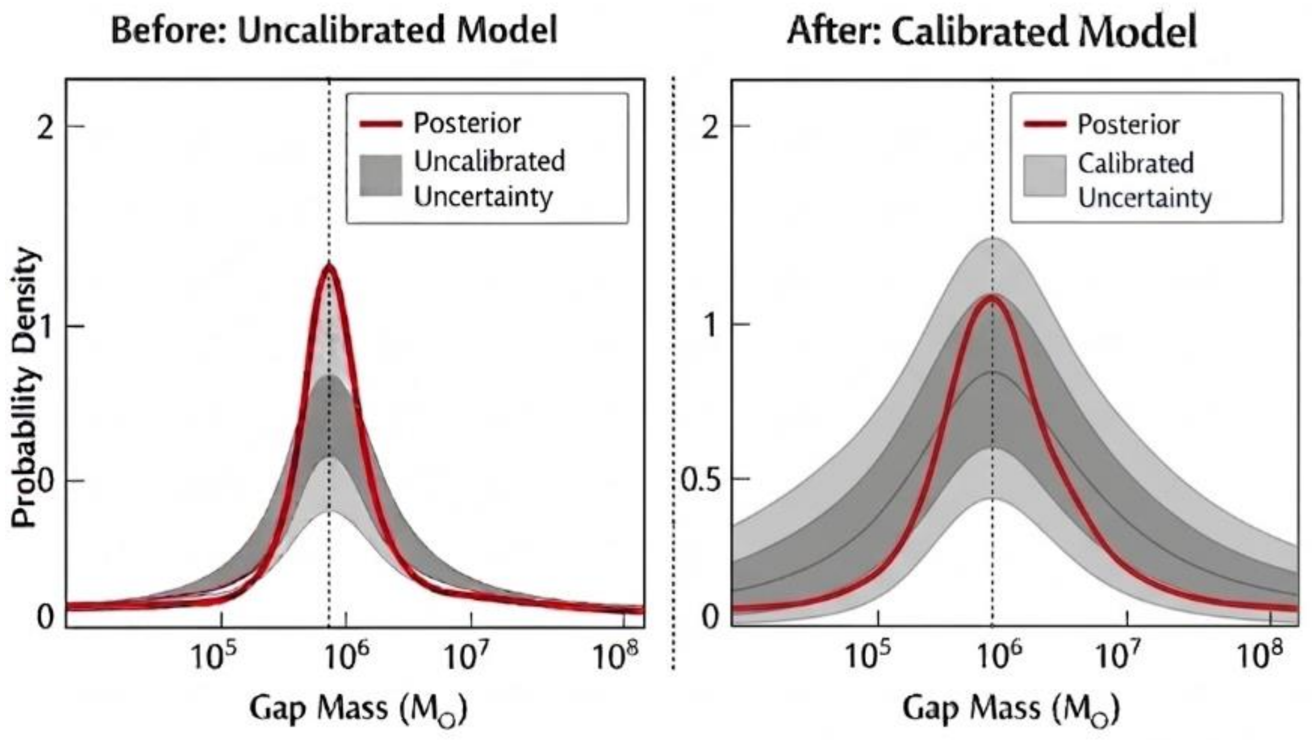
Calibration tells us whether the model's probabilities and uncertainties are **trustworthy** or **just look precise**

→ **E.g.,** model: result has **90% confidence**, calibration checks **correctness**

→ In pipeline: calibration improved using **injection–recovery tests**, **PIT histograms**, and **rank histograms**

→ **Uncalibrated** model can be **overconfident** and make strong claims that **are not reliable**.

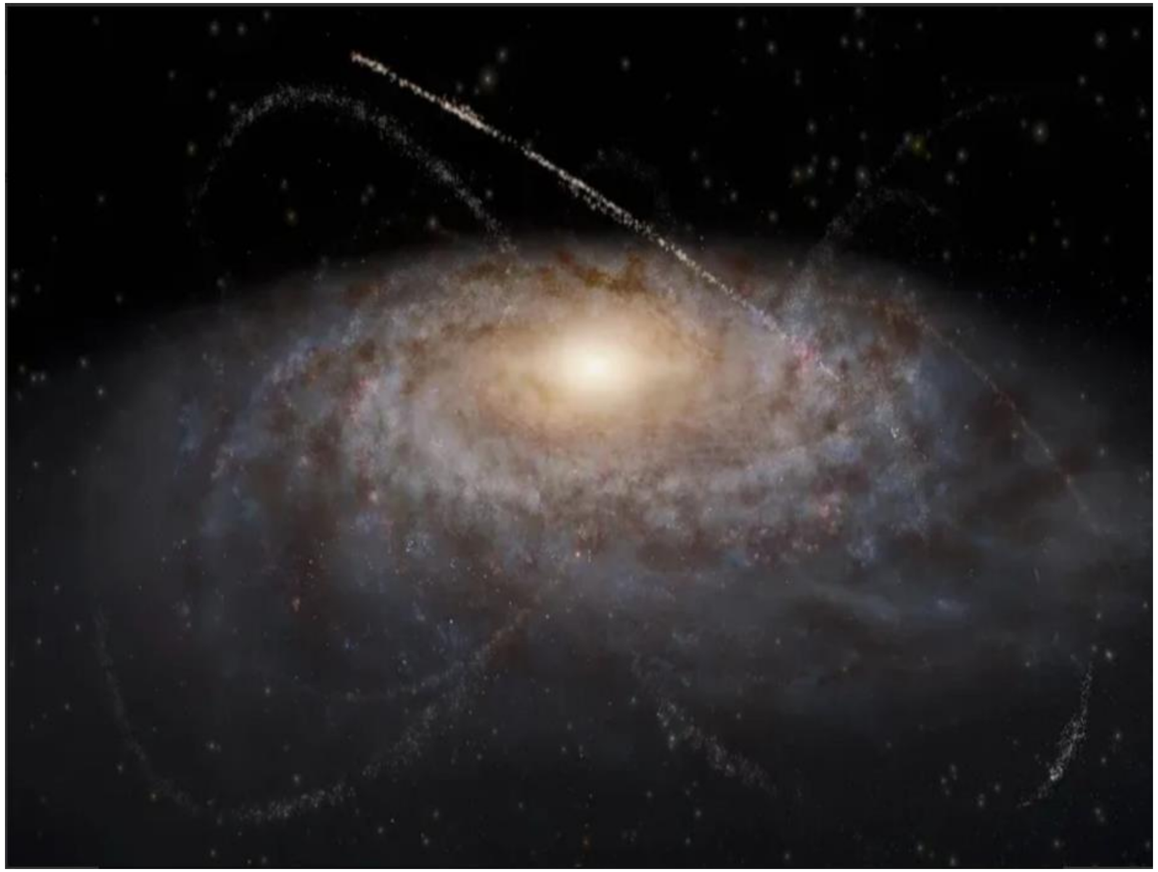
→ **Good calibration**: posterior masses, Bayes factors, and model probabilities are interpreted as **statistically meaningful**



→ Helps confirm that the strong CDM preference **is not** byproduct of **noisy training** or **poor uncertainty estimates**

Main Scientific Takeaway

- Pipeline shows that GD-1 can be used as gravitational detector for DM substructures
- Strongest stream disturbances best explained by DM subhaloes, not ordinary baryonic perturbers
- Method distinguishes between dark DM model classes using calibrated probabilities and Bayes factors
- Analysis reaches sensitivity down to ≈ $3\times10^6\ M_\odot$, with reliable 1σ mass recovery for $M \geq 5\times10^6\ M_\odot$
- Results suggest stream morphology and kinematics can be turned into quantitative constraints on Milky Way's DM substructure spectrum → makes GD-1 a powerful tool for DM composition search



Improvements on Current Method

1. Compares multiple DM and baryonic model classes directly, adding model-classprobabilities and Bayes factors.
2. Method uses full phase-space, multi-scale density information, and topologicalfeatures.
3. Emulator ensemble makes forward model faster and more realistic, while stillkeeping uncertainty explicit.
4. Calibration is stronger than in earlier work, with 90% coverage kept between0.89 and 0.91.
5. tests against baryonic confusers, different Milky Way potentials, and held-out N-body cases.