Cosmology with galaxy clusters, simulation-based inference and AI agents

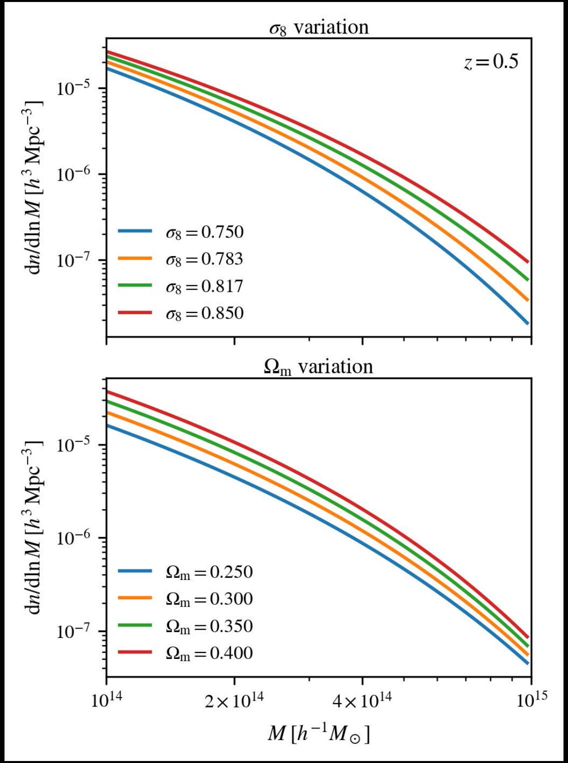
Iñigo Zubeldia
UC Berkeley
3/2/2026

Kavli Institute for Cosmology Cambridge
KICC
UNIVERSITY OF CAMBRIDGE
Department of Applied Mathematics and Theoretical Physics (DAMTP)

Cluster cosmology in a nutshell

➢ **Most massive collapsed objects in the Universe.**

➢ **Number as a function of mass and redshift powerful cosmological probe**

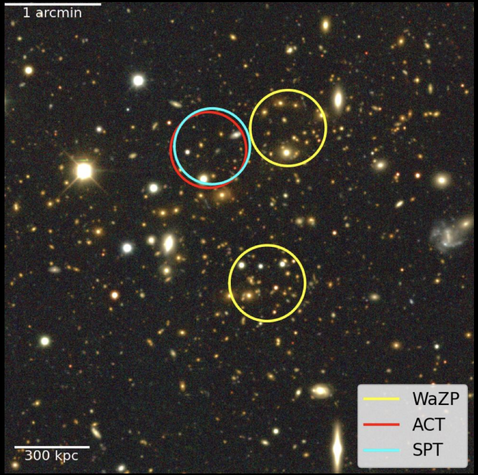


Cluster cosmology in a nutshell

Standard recipe for cluster cosmology

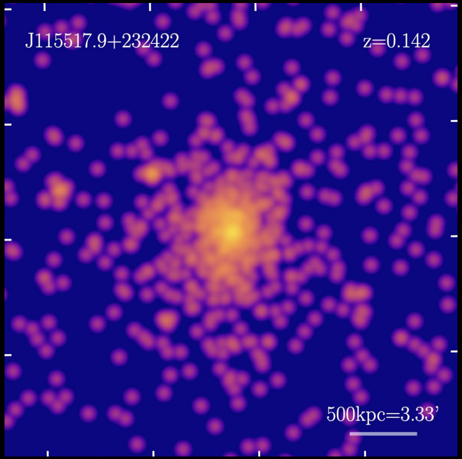
1. Find clusters.
2. Forward model number counts (likelihood, SBI).
   - ➢ Halo mass function.
   - ➢ Mass proxy, not mass.
   - ➢ Need to calibrate relation: lensing.

Optical



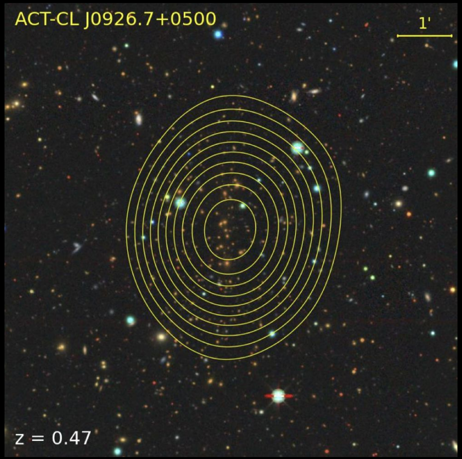
Benoist+ 2025

X-ray



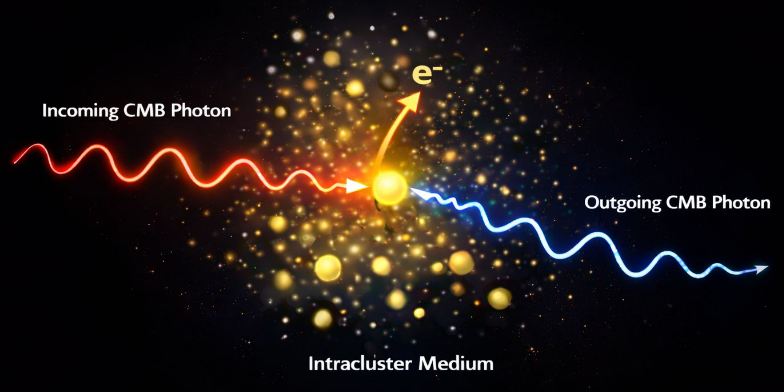
Bulbul+ 2024

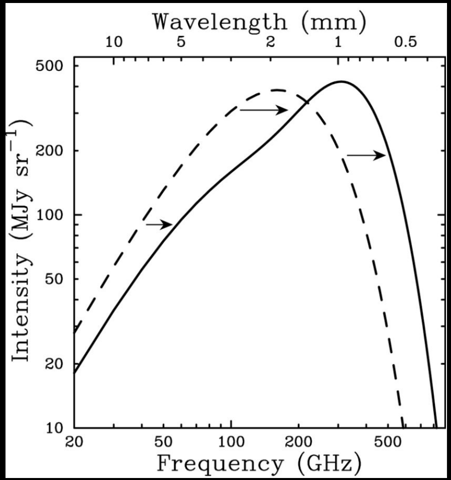
mm



Hilton+ 2025

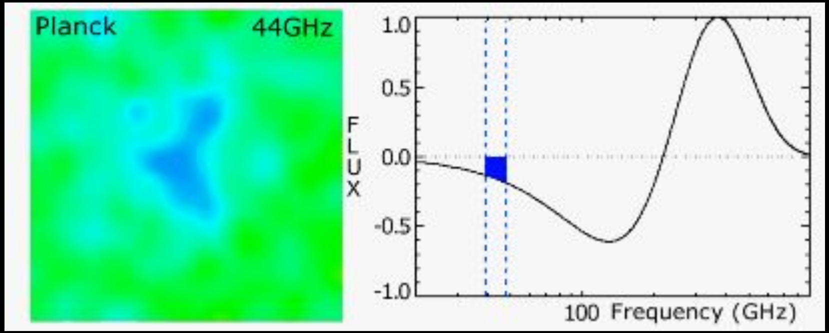
The thermal Sunyaev-Zeldovich (tSZ) effect





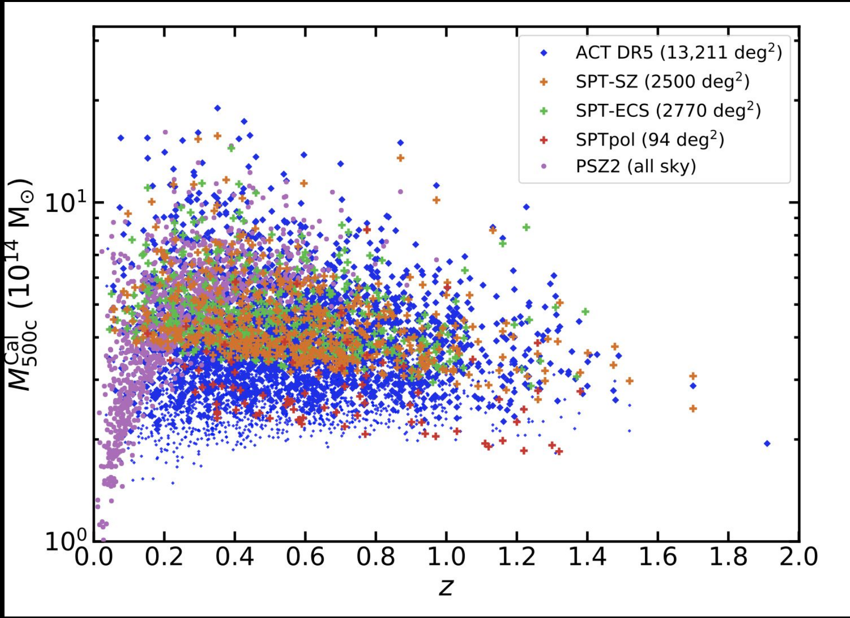
Sunyaev & Zeldovich 1970

- ✓ High purity (>95% SNR 5; >98% SNR 6).
- ✓ Redshift-independent.
- ✓ Selection easy to model.
- ✓ tSZ signal (or SNR) is mass proxy.



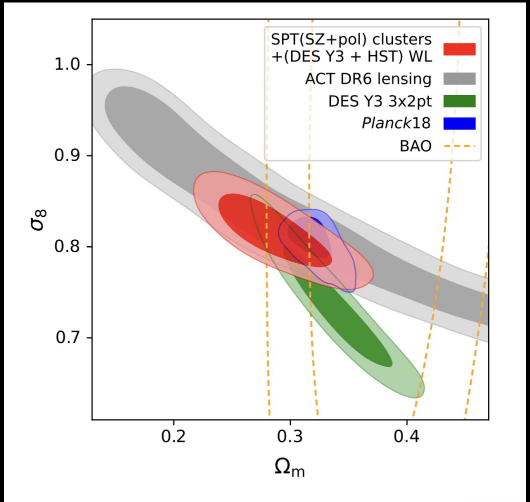
Credit: ESA/Planck Collaboration

Cosmology with tSZ cluster number counts



Hilton+ 2021

- **Planck:** Planck Collab. 2015, IZ & Challinor 2019, Aymerich+ 2025, **439** clusters.
- **SPT:** Bocquet+ 2024, **1005** clusters.

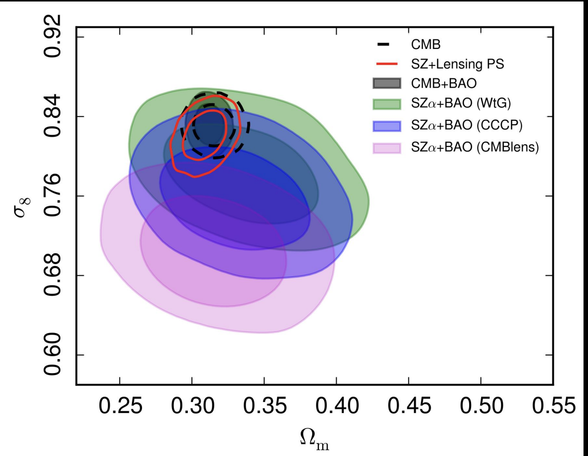


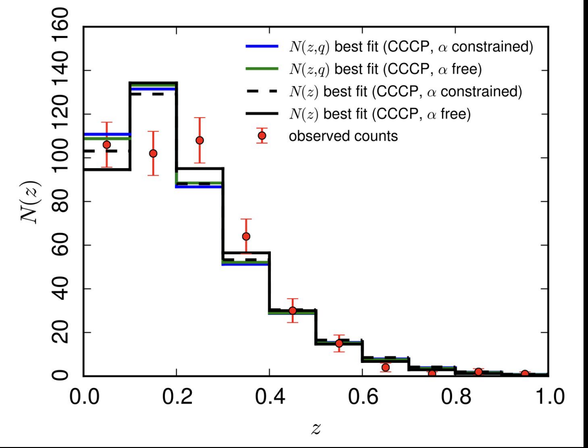
- **ACT DR5**: ~ **3000** clusters, in progress!

Planck cluster cosmology: and independent end-to-end reanalysis

with J-B Melin, J Chluba, R Battye, J Mohr
S Bocquet, A Singh, A Challinor, B Bolliet

	Planck 2015	Our reanalysis
SNR cut	6	5
Cross-channel and fgd. tests	No	Yes
Noise covariance	Biased	Unbiased
Cosmological analysis	Piece-wise	Fully forward modelled





Planck SZiFi cluster catalogues

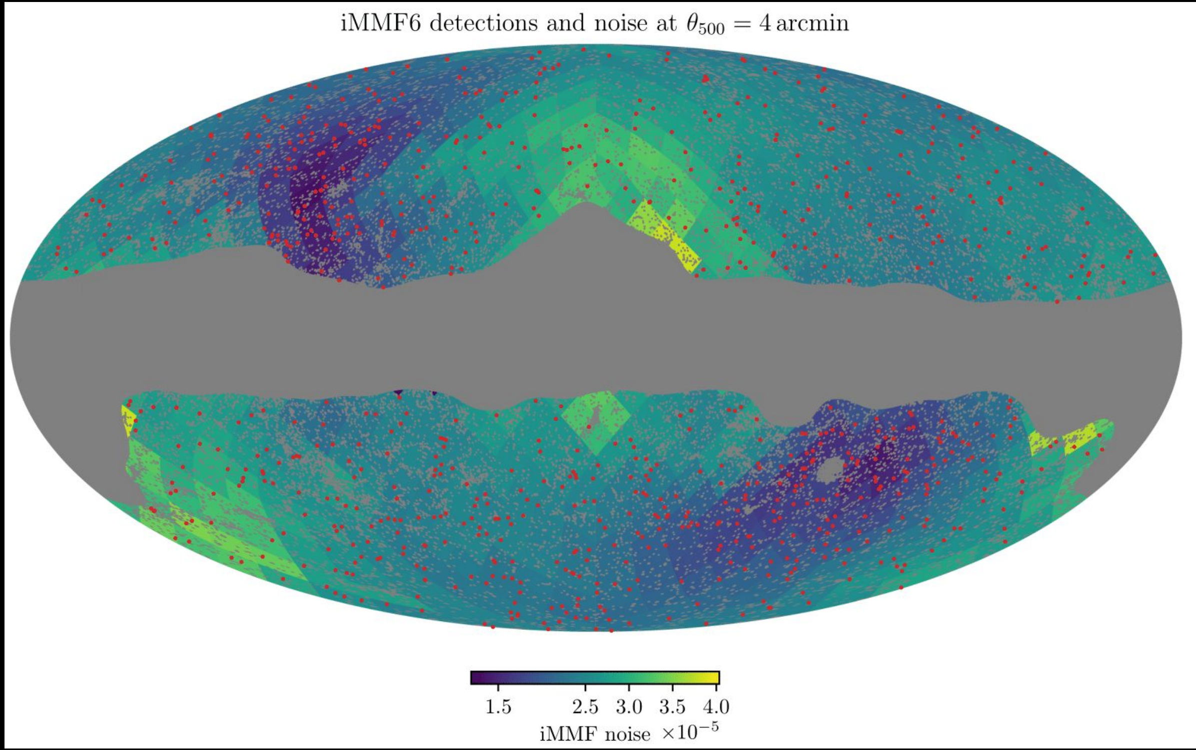
with J-B Melin, J Chluba, Richard Battye

10 catalogues with SZiFi:

2408.06189

- ➢ 3 channel combinations.
- ➢ Deprojecting CIB (7 catalogues, 4 with moments).
- ➢ Baseline catalogue: 833 detections.

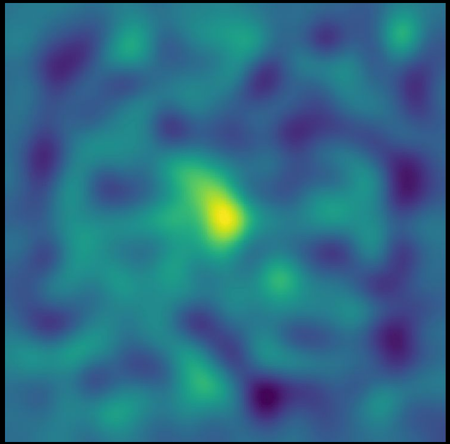
- ✓ **Good consistency** across channel and dep. combinations (first for a *Planck* catalogue)
- ✓ **Cluster-correlated CIB probably negligible** (first).
- ✓ **rSZ negligible** (first).



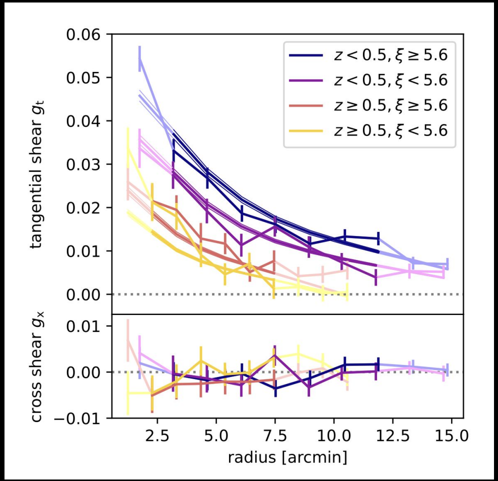
github.com/inigozubeldia/szifi

Planck SZiFi cosmological analysis

Mass calibration of tSZ SNR



Planck CMB lensing
SNR = 4.7



DES Y3 cluster weak lensing
165 clusters, SNR = 33

Likelihood: forward model all observables (`cosmocnc`)

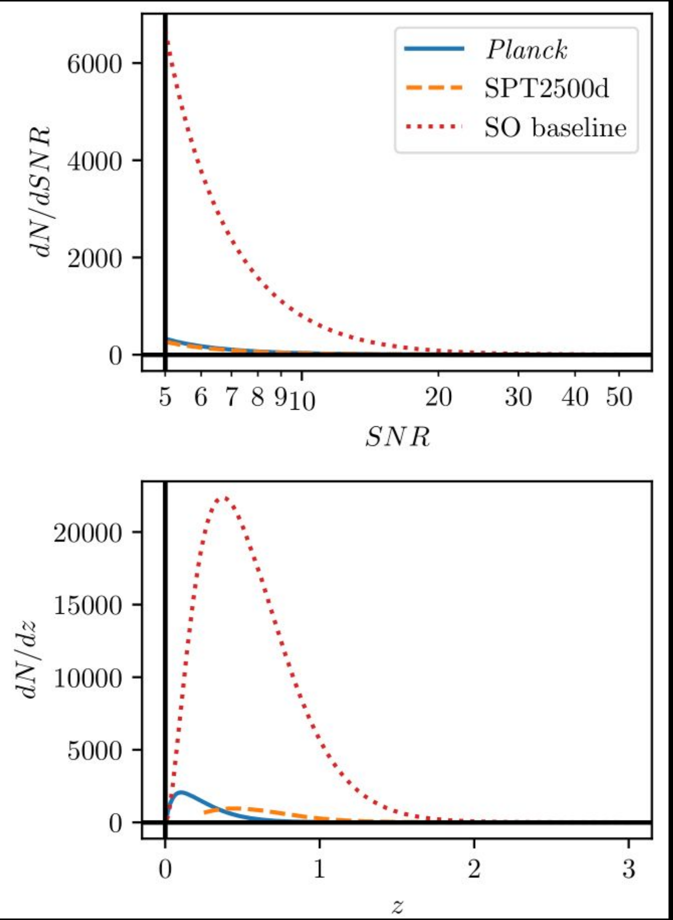
Stay tuned!

Likelihood code: cosmocnc

with B Bolliet 2403.09589

Cluster number count likelihood code, general and versatile:

- ✓ **Binned**, **unbinned**, and extreme value Poisson like.
- ✓ Arbitrary number of mass observables.
- ✓ Vector observables (e.g., lensing profiles).
- ✓ **Stacked** observables.
- ✓ **Correlated scatter**.
- ✓ Redshift uncertainties.
- ✓ **Unconfirmed** detections.
- ✓ Python, interfaced with **Cobaya**.
- ✓ Fast and accurate: **good enough for SO**.



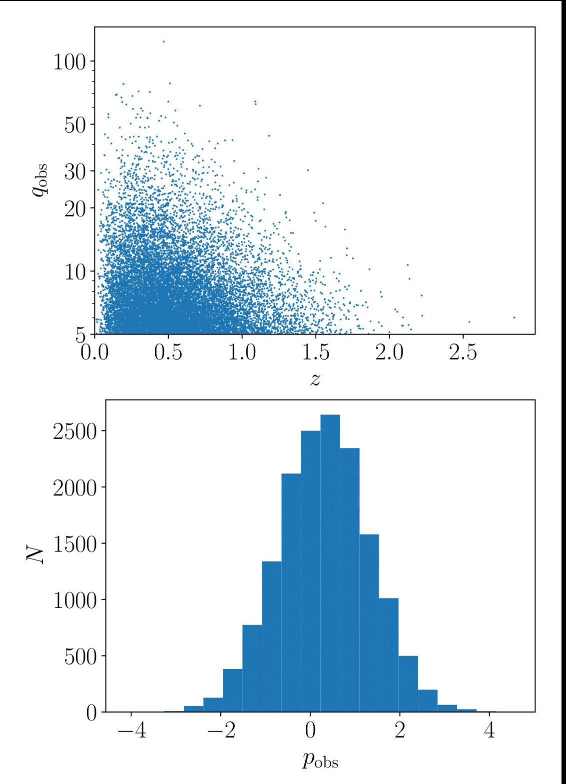
github.com/inigozubeldia/cosmocnc

Simulation-based inference (SBI) for cluster cosmology

with B Bolliet, A Challinor and W Handley

2504.10230

SO-like catalogue



Can we use synthetic catalogues for inference?

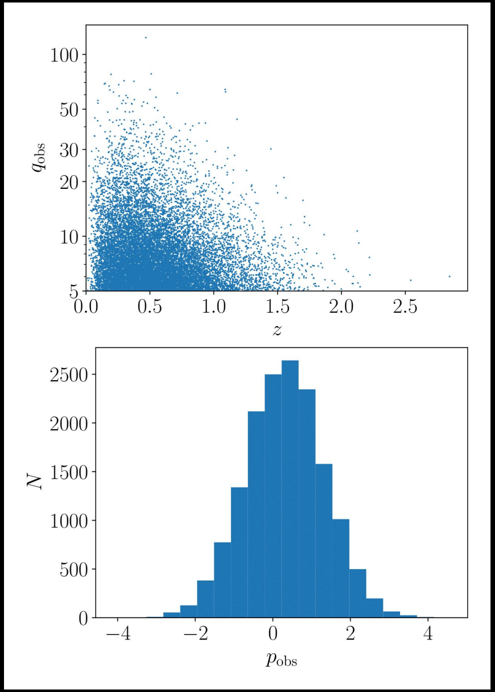
Yes! With SBI.

Recipe to generate Poisson cluster catalogue:

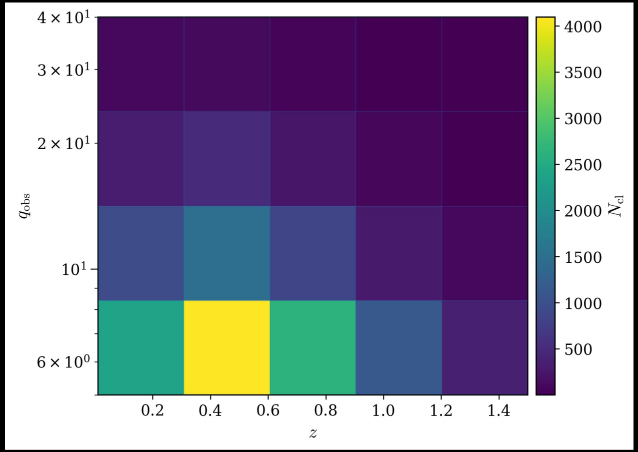
1. Sample *M-z* pairs from HMF.
2. Iterate over layers in mass observable model:
   a. Apply scaling relation(s) (Mean SNR = *f(M,z)*)
   b. Add scatter (log-normal, Gaussian).
3. Apply selection criterion (SNR > 5).

Simulation-based inference (SBI) for cluster cosmology

Data compression



~ 3 x 16 000 dimensions



$$p_{\rm stacked} = \frac{1}{n_{\rm tot}} \sum_{i=1}^{n_{\rm tot}} p_{{\rm obs},i}$$

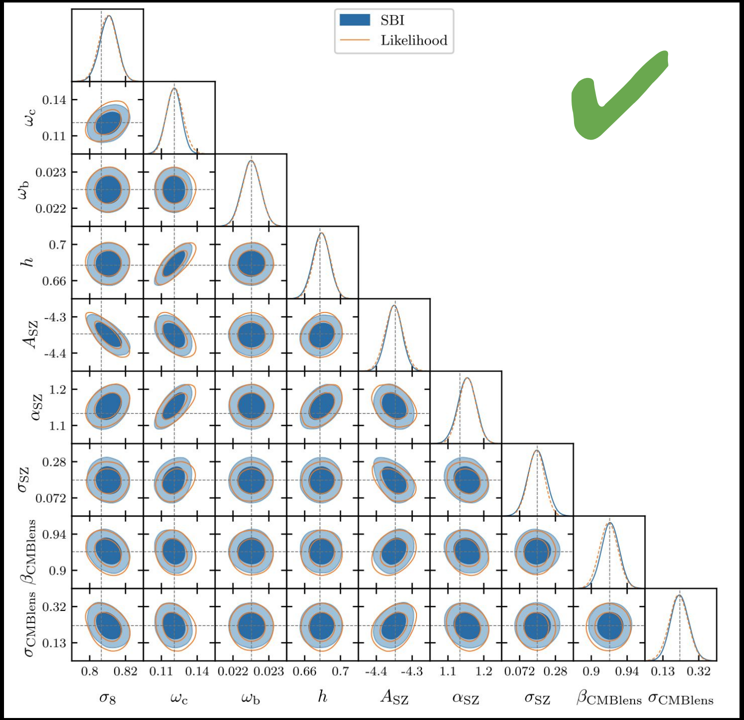
20 + 1 dimensions

Simulation-based inference (SBI) for cluster cosmology

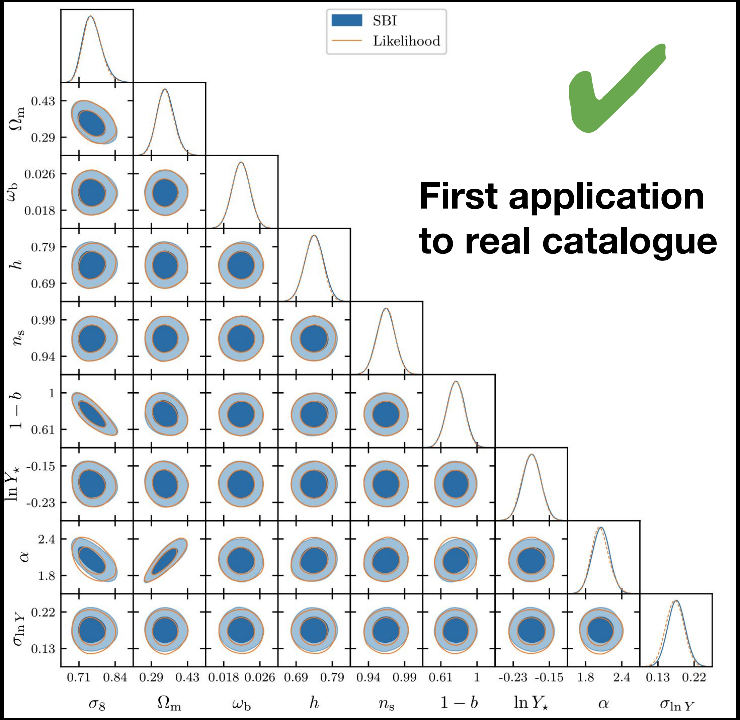
Neural Posterior Estimation (NPE)

1. Sample from prior.
2. Generate catalogues at samples and compress them.
3. Neural density estimation (normalising flow, `sbi`) to learn the posterior from param-data pairs.
4. Sample posterior for real dataset.

SO-like catalogue



Real *Planck* catalogue



First application to real catalogue

Goodness of fit assessed with only synthetic catalogues

Motivation, limitations and extensions

✓ Much easier to develop (including complexities like correlated scatter).

✓ Faster and embarrassingly parallelisable.

✓ Easier to combine with correlated data (e.g., cosmic shear).

☝️ Demonstrated with Poisson catalogues and same modelling assumptions as likelihood. To go beyond these:

- Synthetic catalogues from cosmo. sims.
- Catalogues from finders on cosmo. sims (get rid of HMF?)

tSZ cosmology beyond counts: power spectrum

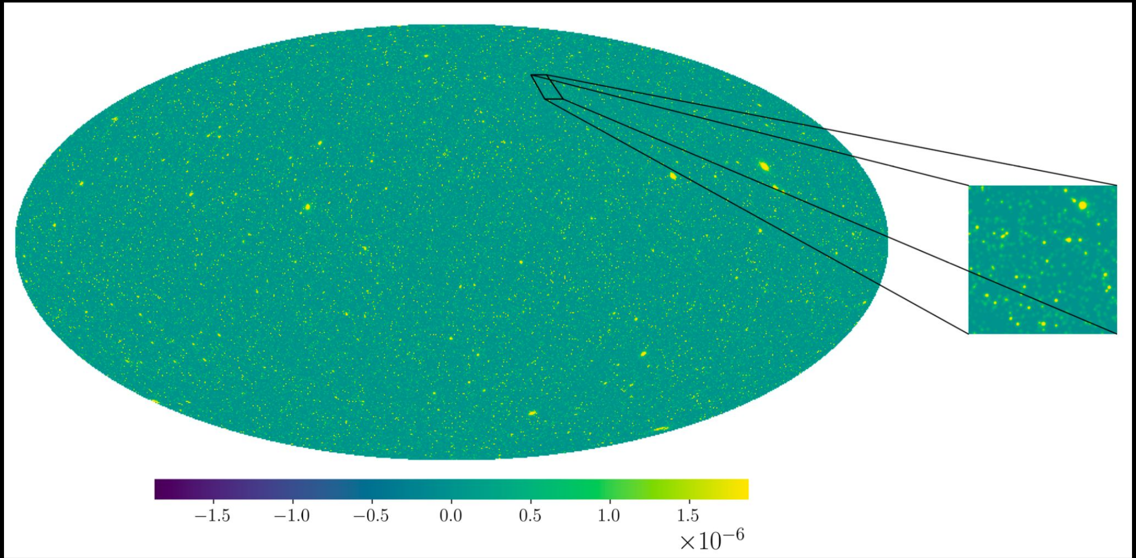
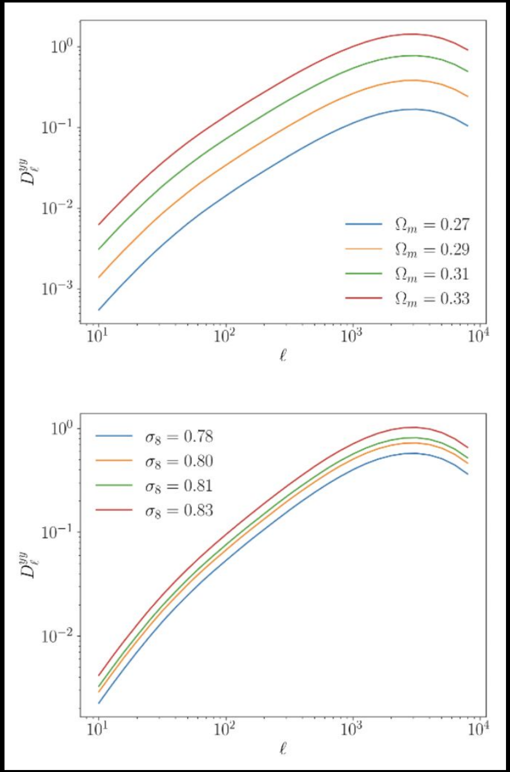


Figure credit: L Xu

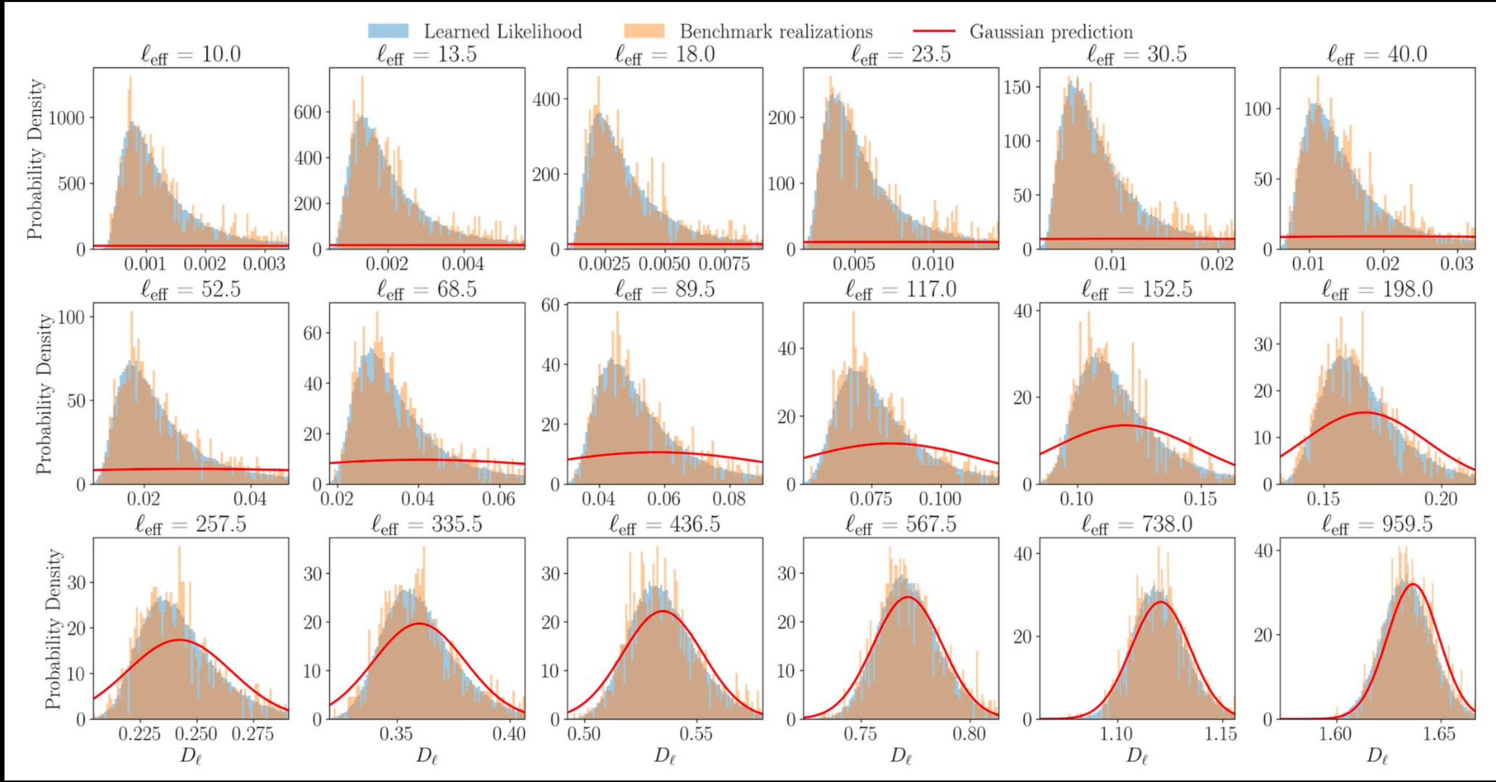


Highly complementary to number counts: undetected **lower mass halos.**

tSZ cosmology beyond counts: power spectrum

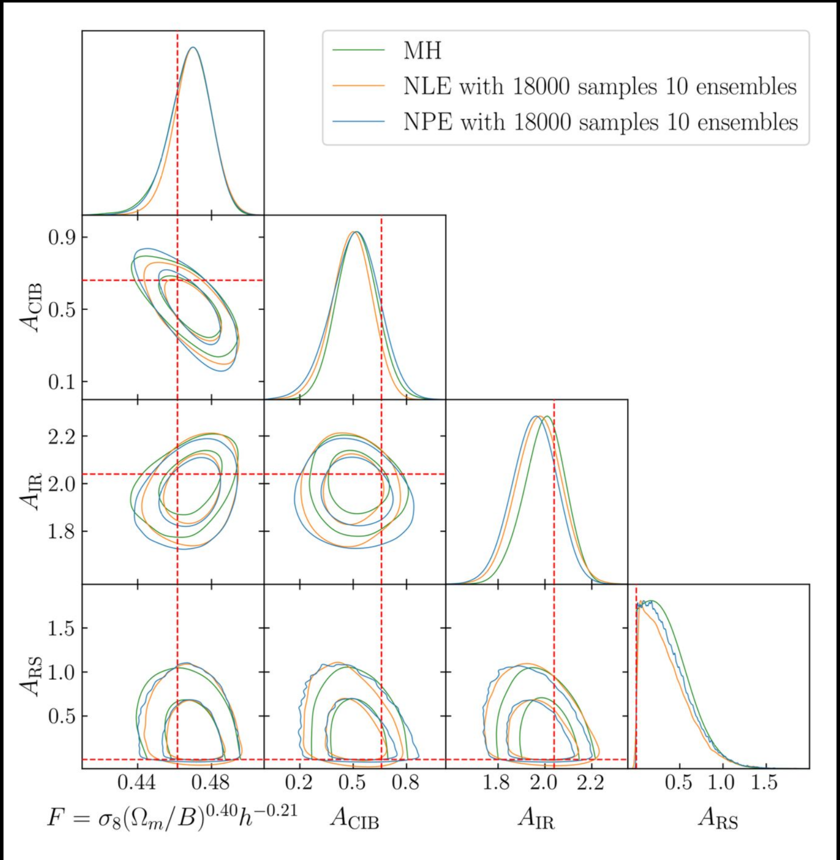
Work led by **Licong Xu** (PhD at Cambridge), with IZ, B Bolliet, A Challinor and J Alvey

Likelihood of *y* power spectrum typically assumed to be Gaussian: but it is **very** non-Gaussian.



Planck-like sim. data

tSZ cosmology beyond counts: power spectrum



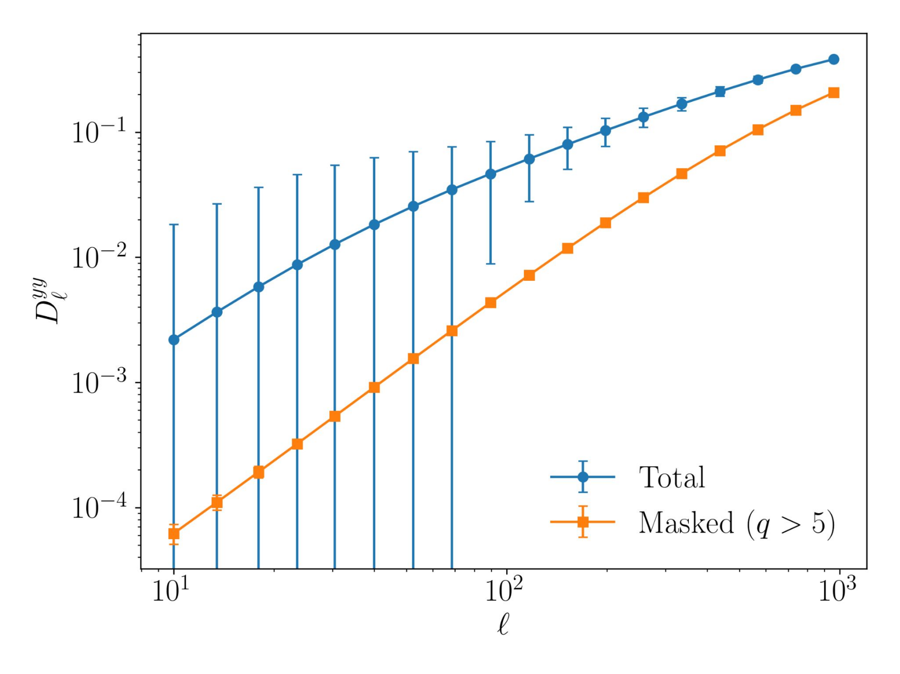
Planck-like

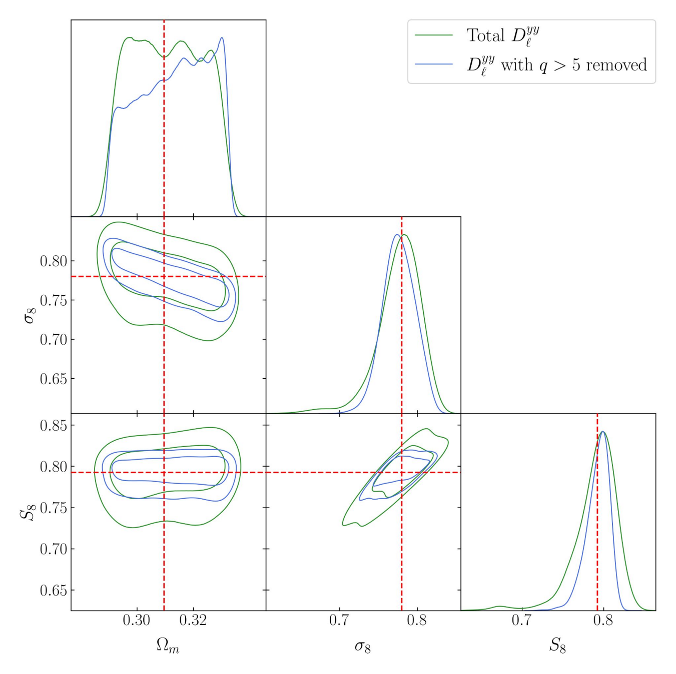
data

Paper out very soon!

Next steps: power spectrum + number counts

- *y* power spectrum: very bad summary statistic, especially at low *l*.
- Can mask detected clusters and take power spectrum of that map (e.g., Rotti+ 2020).





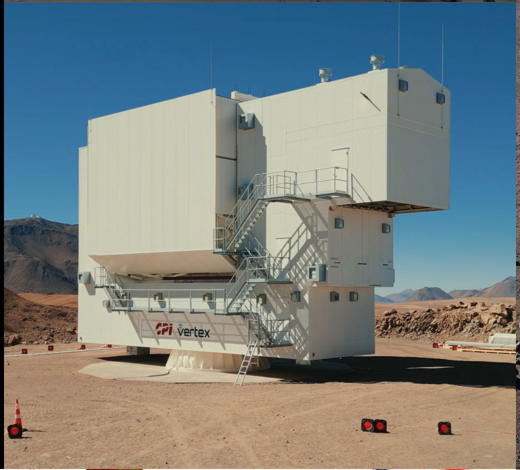
Simons Observatory (SO)

Fig credits: N Galitzki + SO Collaboration

Ground-based CMB experiment.

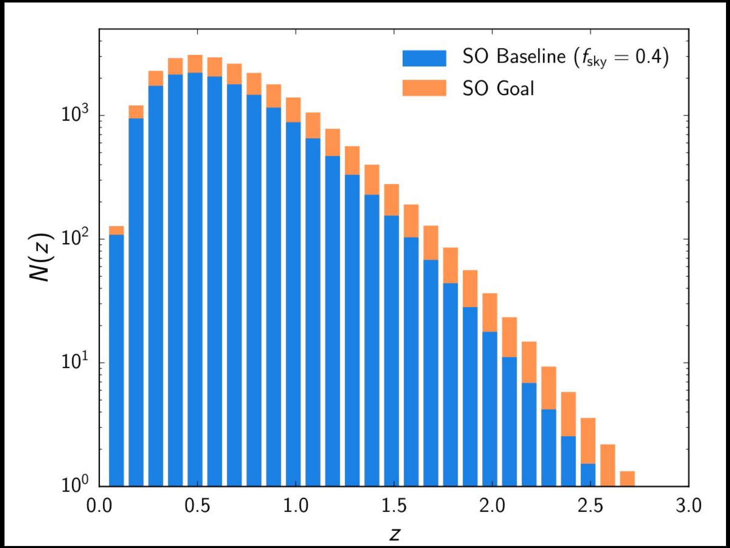
Cerro Toco, Chile (5300 m).

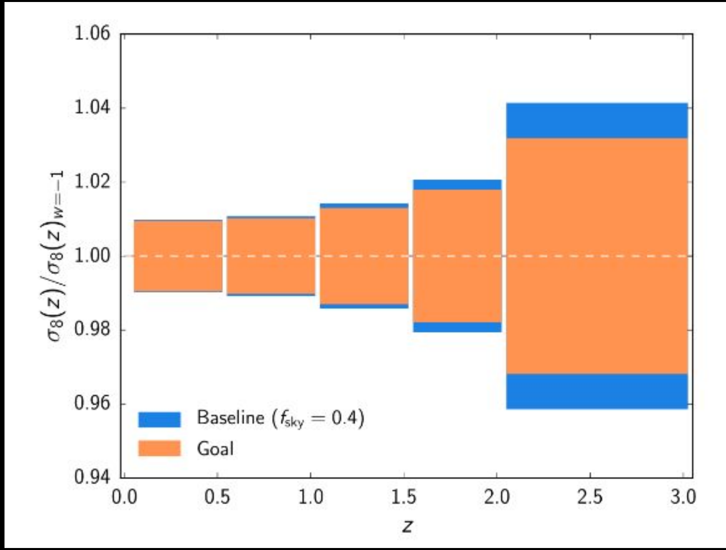
- 6 **SATs** (3 + 2 SO:UK + 1 SO:Japan).
  - 0.4 m.
  - *B*-modes.
- 1 **LAT**:
  - 6 m, ~1 arcmin res.
  - 50% of the sky.
  - SZ, CMB lensing.
  - Y1 survey starting soon.



Cluster cosmology with SO

~ 20 000 clusters in 3 years of LAT observations





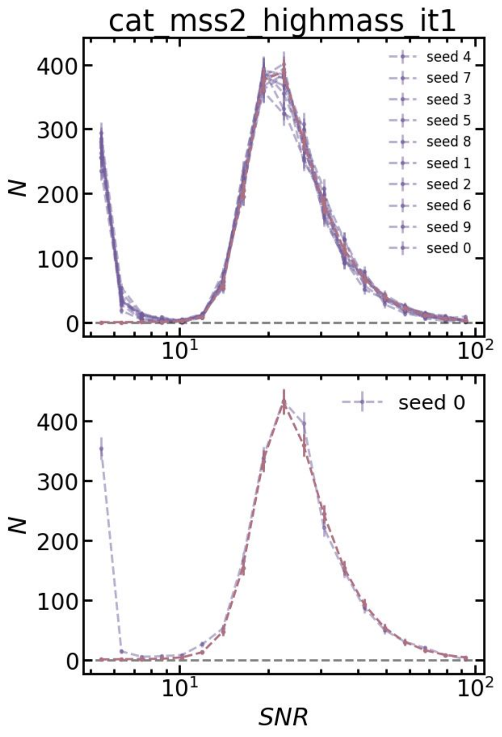
(Simons Observatory Collaboration forecast paper 2018)

Cluster detection and forward modelling must be understood at unprecedented level!

SO cluster cosmology efforts

Z Chen, A Nicola, E Lee, JB Melin, E Rosenberg, B Bolliet,
M Hilton, R Battye, J Chluba, C Sifon, N Battaglia, IZ ++

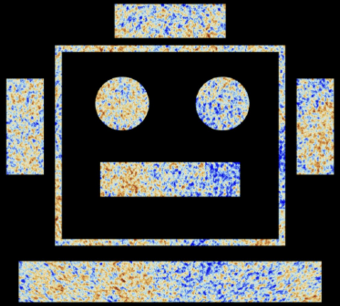
- Cluster finder challenge: MMF3, Nemo, SZiFi
- Likelihood + catalogue-based SBI.
- Cluster lensing for mass calibration
- Things we worry about:
  - Cluster clustering (Matilde Abreu).
  - Point sources.
  - Blending.



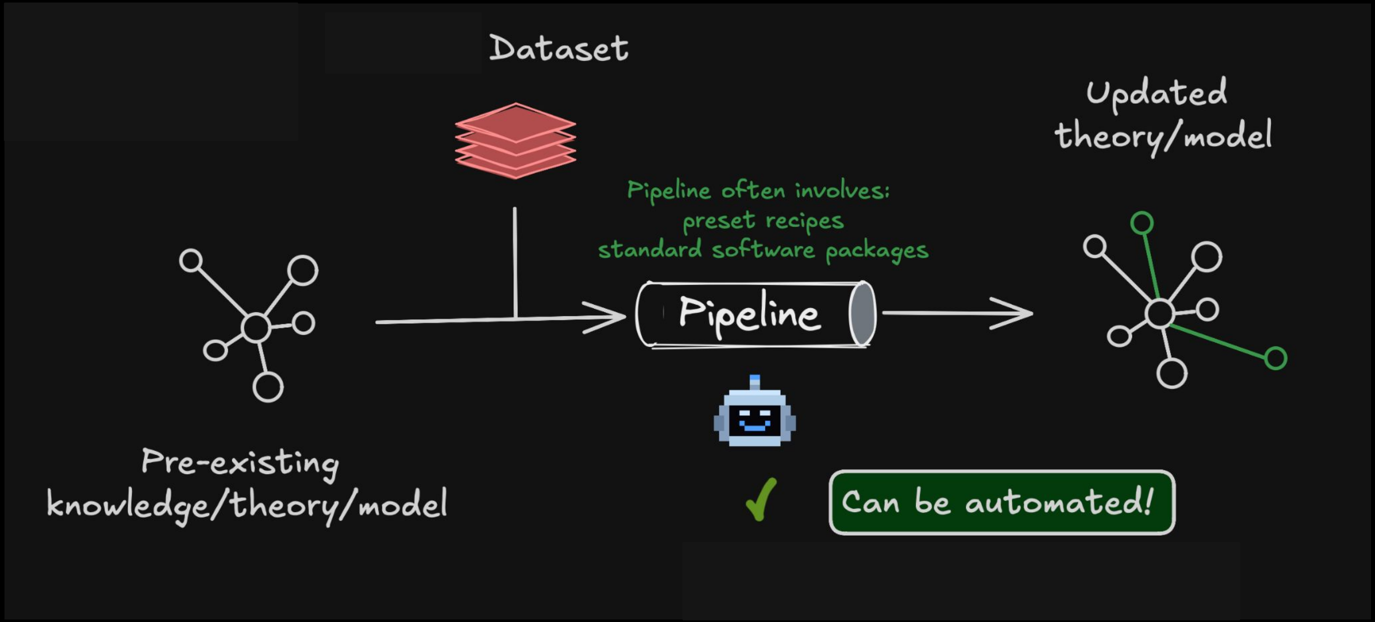
Should we trust all of this?

- tSZ catalogues highly **pure** (sims + cross-matching).
- **tSZ modelling uncertain** (tSZ itself + correlated foregrounds), especially at low masses, but can get around through **empirical mass calibration** with lensing.
- **Blending** / **cluster definition**: simulations, SBI.

AI agents for scientific discovery

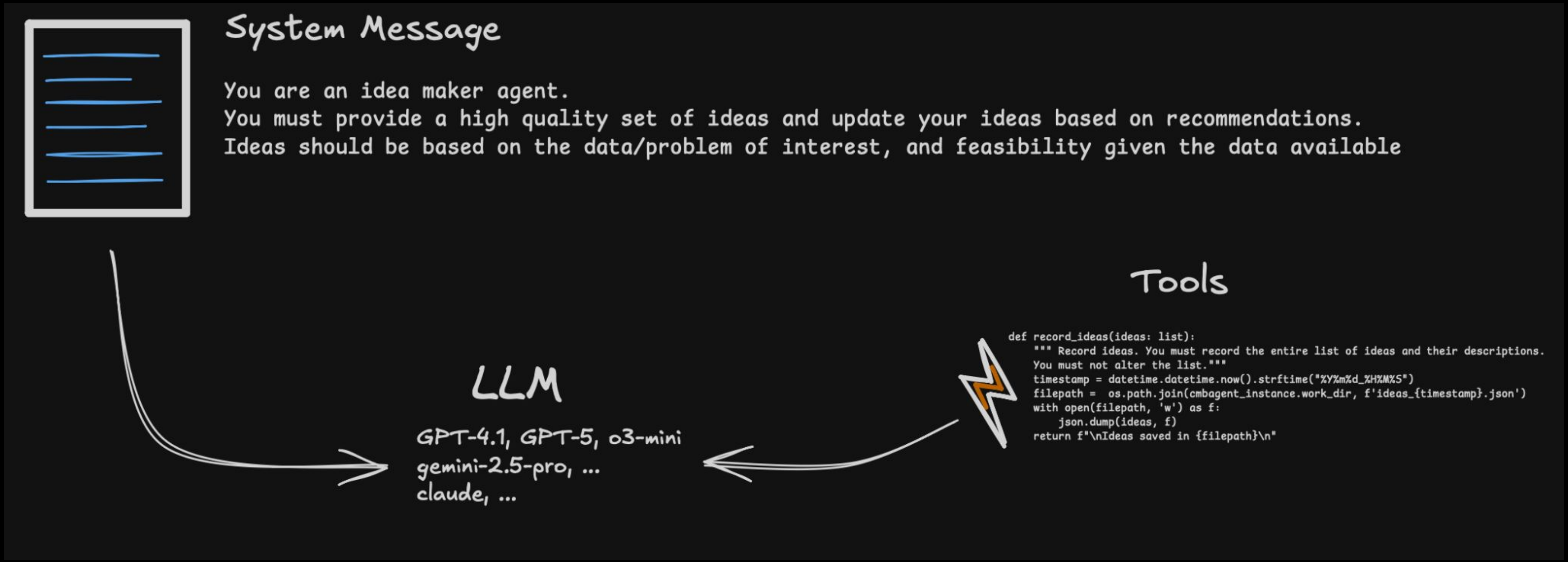


Discovery flow in data-driven fields

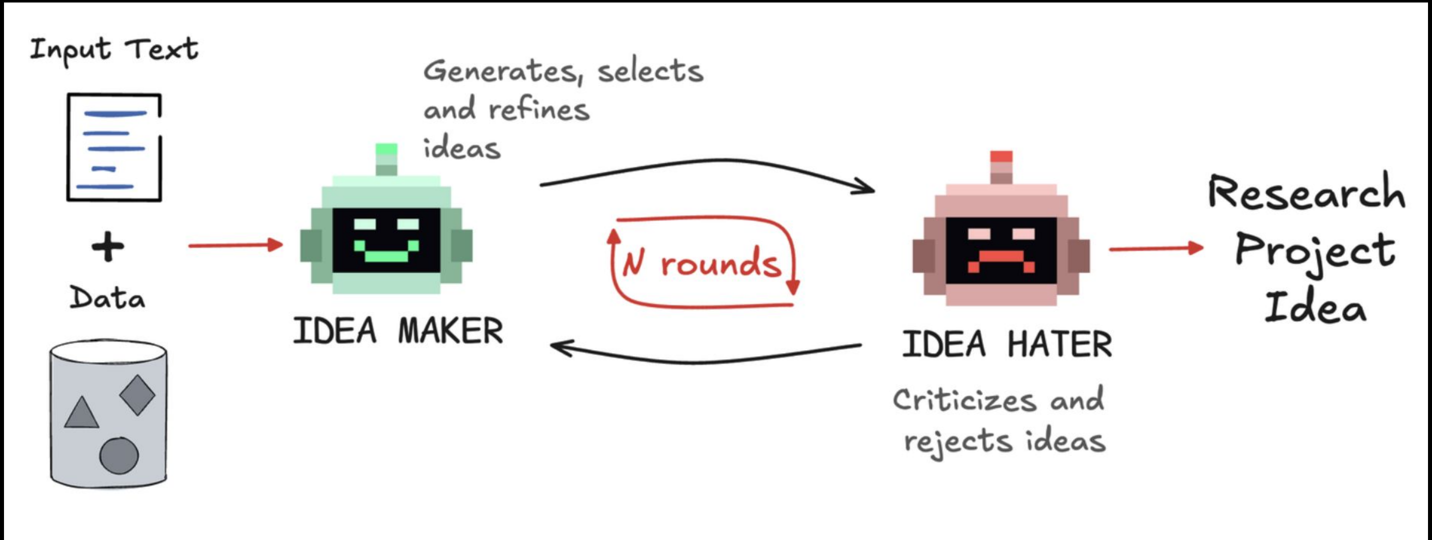


What is an agent?

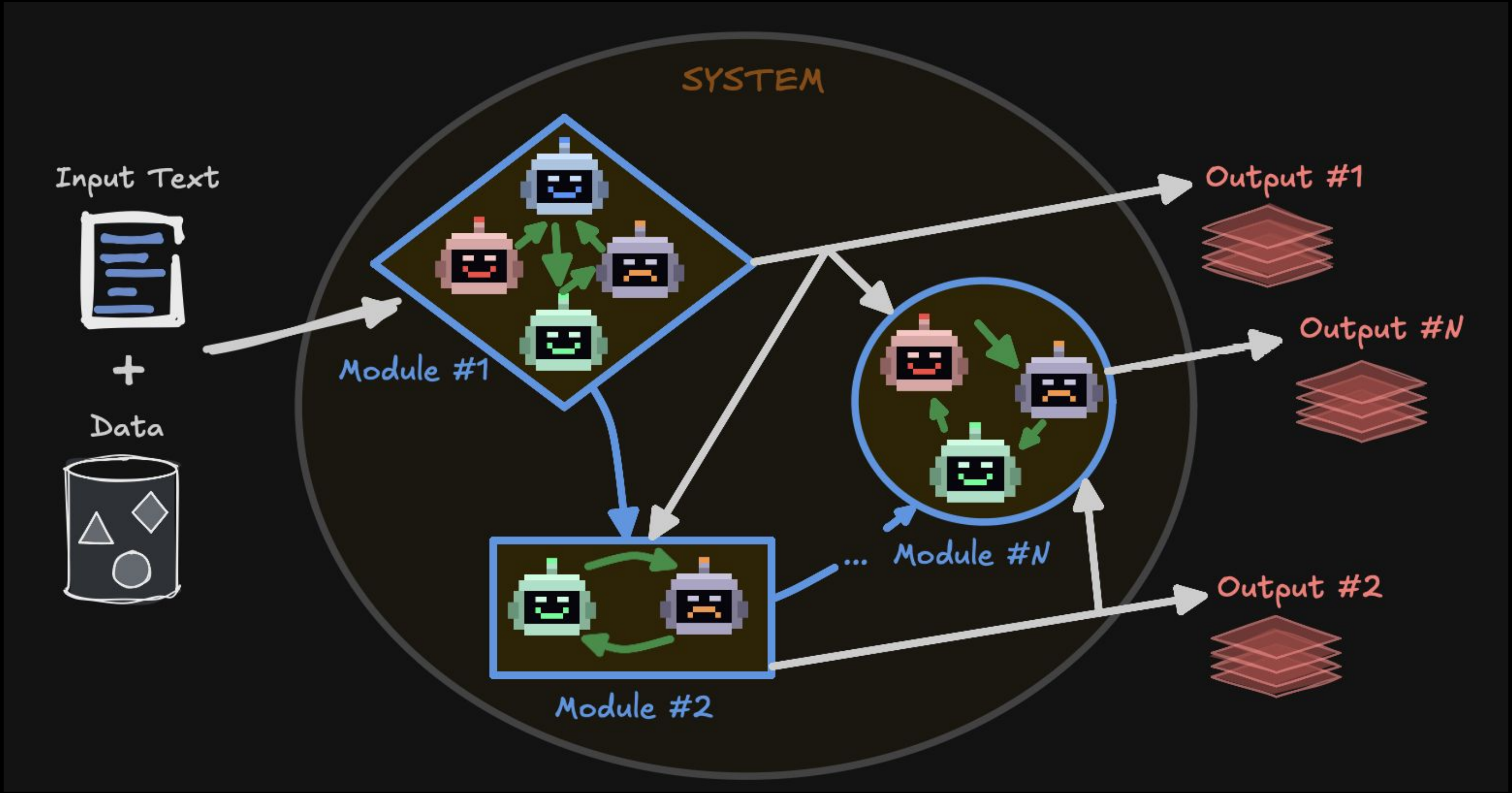
An **agent** is an LLM instructed to play a **role** and that can use **tools**



A simple multi-agent system: scientific project idea generation




Literature Agent
Referee Agent
Idea Agent
Paper writing Agent
Simulation agent
•••

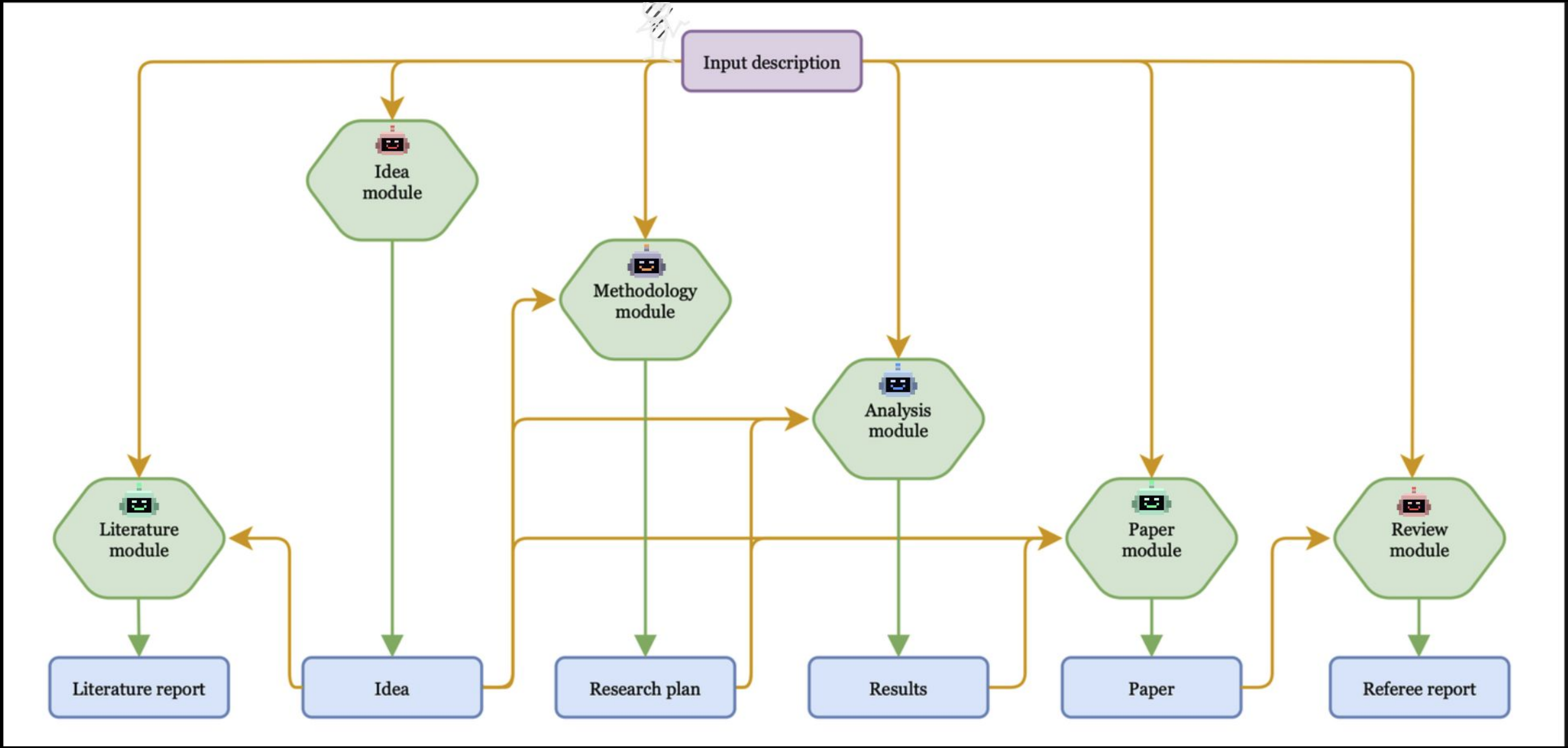

SYSTEM
Input Text
+
Data
Module #1
Module #2
...
Module #N
Output #1
Output #N
Output #2

The Denario project: Deep knowledge AI agents for scientific discovery

Boris Bolliet[*], Pablo Villanueva-Domingo[*], Francisco Villaescusa-Navarro[*], Adrian E. Bayer, Aidan Acquah, Chetana Amancharla, Almog Barzilay Siegal , Pablo Bermejo, Camille Bilodeau, Pablo Cárdenas Ramírez, Miles Cranmer, Urbano L. França, ChangHoon Hahn, Yan-Fei Jiang, Raul Jimenez, Jun-Young Lee, Antonio Lerario, Osman Mamun, Thomas Meier, Anupam Anand Ojha, Pavlos Protopapas, Shimanto Roy, David N. Spergel, Pedro Tarancón-Álvarez, Ujjwal Tiwari, Matteo Viel, Digvijay Wadekar, Chi Wang, Bonny Y. Wang, Licong Xu, Yossi Yovel, Shuwen Yue, Wenhan Zhou, Qiyao Zhu, Jiajun Zou, Íñigo Zubeldia

[*]Equal Contribution. Listing order of BB, PVD, FVN is random.

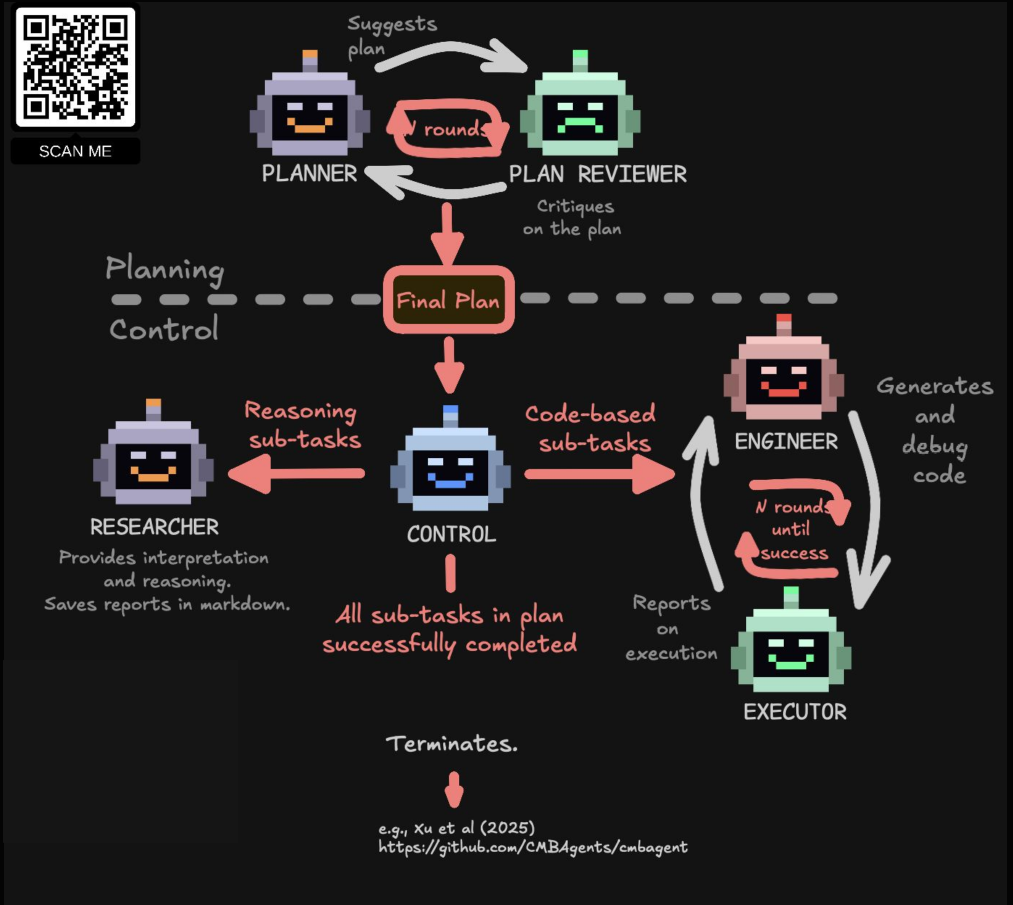
Villaescusa-Navarro + 2025


Input description
Idea module
Methodology module
Analysis module
Literature module
Paper module
Review module
Literature report
Idea
Research plan
Results
Paper
Referee report

Deep research architecture: cmbagent

- **Planning**
  - Decompose main task into subtasks.
  - Propose-critique loop.

- **Control**
  - Carry out each subtask.
  - Generate-evaluate loop.



Open Conference of AI Agents for Science 2025

The 1st open conference where AI serves as both primary authors and reviewers of research papers

Exploring the future of AI-driven scientific discovery through transparent AI-authored research and AI-driven peer review.

Mapping the Diversity of the Black Hole-Stellar Mass Relation: The Role of Feedback and Cosmology in Simulated Galaxies

AstroPilot[1]

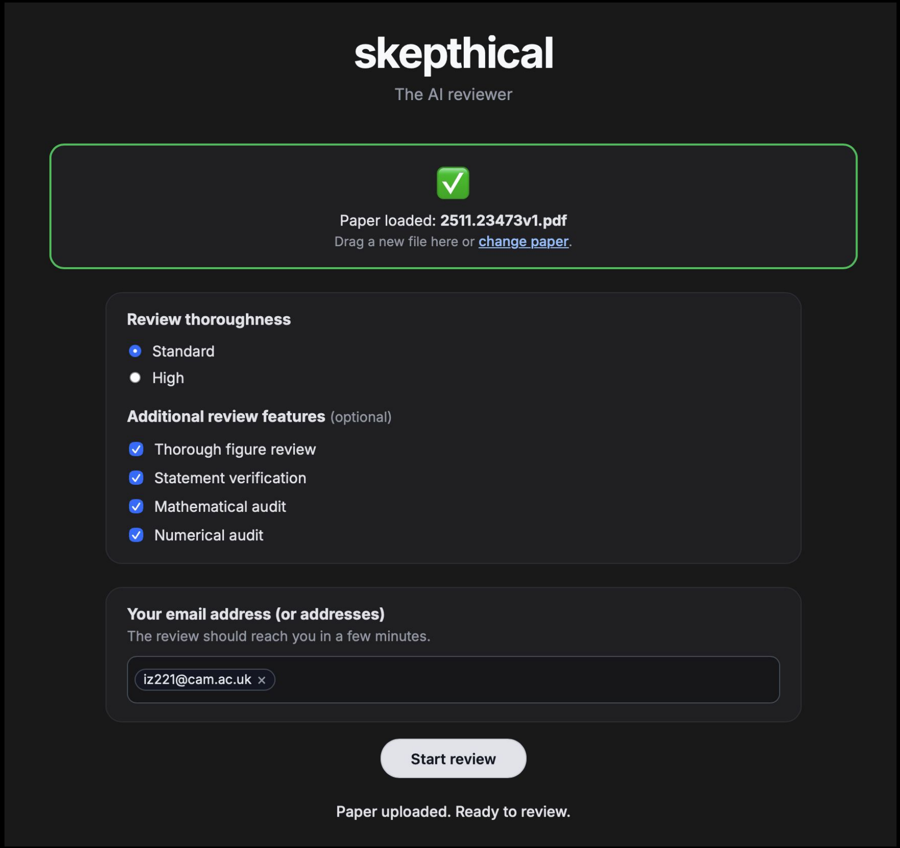
[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

Understanding the diversity of the black hole–stellar mass ($\mathrm{M_{BH}}$–$\mathrm{M_{star}}$) relation and its physical drivers is crucial for galaxy evolution models but challenging due to the complex interplay of physical processes across large parameter spaces. We systematically quantify how the slope, normalization, intrinsic scatter, and occupation fraction of this relation depend on variations in cosmological and feedback parameters using a large suite of 1,000 simulated galaxy catalogs at redshift zero, comprising approximately 720,000 galaxies. Each catalog is characterized by six global parameters: matter density (Ω_m), density fluctuation amplitude (σ_8), and supernova (A_{SN1}, A_{SN2}) and active galactic nucleus (A_{AGN1}, A_{AGN2}) feedback efficiencies. For each catalog, we fit the $\mathrm{M_{BH}}$–$\mathrm{M_{star}}$ relation using robust linear regression within three stellar mass bins, estimating uncertainties via bootstrapping. We then employ advanced statistical and machine learning techniques, including hierarchical linear modeling, random forest regression, gradient boosting, permutation importance, SHAP values, and partial dependence plots, to map the derived scaling relation parameters and occupation fractions to the global catalog parameters. Our analysis reveals substantial diversity in the $\mathrm{M_{BH}}$–$\mathrm{M_{star}}$ relation across catalogs and mass bins. Supernova feedback (A_{SN1}) is the dominant driver of the relation's properties and scatter at low stellar masses, where the relation is typically shallower and the occupation fraction lower. Active galactic nucleus feedback, particularly A_{AGN1}, becomes the primary driver at intermediate and high masses, with its influence transitioning from promoting black hole growth (steepening the relation) at intermediate masses to potentially suppressing it (changing the sign of dependence) at the highest masses. Cosmological parameters (Ω_m, σ_8) play a secondary, modulating role, primarily affecting normalization and scatter. The black hole occupation fraction increases strongly with stellar mass and is significantly suppressed by feedback at low mass. Simulated relations broadly agree with observational constraints at high mass but predict greater diversity and lower occupation at low mass, consistent with recent dwarf galaxy studies. This work provides a comprehensive, statistically rigorous mapping of the $\mathrm{M_{BH}}$–$\mathrm{M_{star}}$ relation's diversity across a wide parameter space, highlighting the critical, mass-dependent role of feedback physics in shaping black hole-galaxy coevolution and offering a detailed framework for interpreting observed scaling relation diversity.

Keywords: Confidence interval, Cosmological parameters, Dwarf galaxies, Galaxy evolution, Hierarchical models, Linear regression, Matter density, Principal component analysis, Robust regression, Scaling relations

skepthical: scientific review+verification system



Skepthical review: *FastEddington: A Streamlined Python Package for Interactive Eddington Bias Simulation in Power-Law Distributions with Gaussian and Log-Normal Errors*

Summary

The paper presents FastEddington, a Python package to simulate and visualize Eddington bias for lower-truncated, single-slope power-law populations under either additive Gaussian or multiplicative log-normal measurement errors. The described workflow (inverse-transform sampling → noise injection → selection/thresholding → naive truncated power-law MLE → visualization of slope/normalization shifts; Sec. 2–3) is clear and the example experiment (n=100,000, true α=2.5, x_min=1.0, σ=0.3; Sec. 3) demonstrates the expected flattening of the recovered slope when measurement scatter and a hard detection threshold are present. To make the package's results fully interpretable and the manuscript's claims well-supported, the paper needs (i) a more explicit generative/selection model description (what is simulated, what is selected, what is fit), (ii) correction/clarification of the stated power-law likelihood and fitting procedure (including constraints and/or the closed-form estimator), (iii) clearer uncertainty quantification and robustness checks across realizations, and (iv) better grounding of claims about performance, novelty, and generality via benchmarks, parameter sweeps, and improved software availability/reproducibility details.

Strengths

- Clear motivation and explanation of Eddington bias and why naive power-law fits can be misleading when measurement scatter interacts with a hard threshold (Sec. 1, Sec. 3).
- Conceptually standard and easy-to-follow simulation components (inverse-transform sampling; Gaussian and log-normal noise injection) make the tool pedagogically useful (Sec. 2.2–2.3).
- Modular package framing (generation, perturbation, fitting, plotting) supports reuse and extension (Sec. 2.1, Sec. 2.6–2.8).
- The example results show qualitatively and quantitatively large parameter shifts consistent with expectations, and the narrative ties the observed flattening to selection effects (Sec. 3.1–3.5).

skepthical: scientific review+verification system

Key statements and references

- ⚠ **Eddington bias arises when random measurement errors interact with steeply declining power-law distributions so that more sources are scattered above a detection threshold than below it, which artificially flattens the observed distribution and significantly distorts the inferred slope and normalization of the underlying population in astronomical datasets where signal-to-noise ratios are limited and thresholds are set by instrumental sensitivity.**
- *Reference(s):* [1], [2], [3]
- *Justification:* The papers support the core mechanism: when a steep underlying distribution is convolved with random scatter (e.g., scintillation or measurement noise), more low-flux sources are moved above threshold than high-flux sources are moved below, boosting detections (Eddington bias). [1] explicitly demonstrates this and quantifies the normalization increase (a factor Γ(5/2−δ)) while emphasizing that, over the relevant range, the slope of the differential counts is not changed by this effect. [3] likewise notes Eddington-type biases near the sensitivity limit causing overestimation of faint-source counts. However, neither paper directly supports the claim that Eddington bias "artificially flattens" the observed distribution or "significantly distorts" the inferred slope; in fact [1] argues the slope can remain unchanged. Thus the statement is only partially supported.
- ⚠ **FastEddington generates synthetic power-law data using inverse transform sampling, in which for each uniform random variable u~U(0,1)**

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: substantial

The paper's core analytical content defines a truncated continuous power-law distribution, provides an inverse-transform sampler, introduces additive Gaussian and multiplicative log-normal measurement error models, and fits a naive truncated power-law via maximum-likelihood estimation. The main internal mathematical inconsistency is in the stated log-likelihood expression, which conflicts with the earlier PDF definition and would (as written) change optimization behavior. Other issues are primarily clarity/notation and dimensional rigor (log of a quantity with units).

Checked items

1. ✔ **Truncated power-law PDF definition** (Sec. 2.2 "Synthetic Power-Law Data Generation", p.3)
   - **Claim:** Defines p(x)=C x^{-α} for x≥x_min with α>1 and C a normalization constant.
   - **Checks:** definition consistency, support/constraints
   - **Verdict:** PASS; confidence: high; impact: moderate
   - **Assumptions/inputs:** Continuous x with support [x_min, ∞), α>1 for integrability at infinity
   - **Notes:** Definition and stated constraint α>1 are consistent with a normalizable tail on [x_min,∞).
2. ✔ **Normalization constant C** (Sec. 2.5 "Calculation of Power-Law Normalization", p.5)
   - **Claim:** States C=(α−1)x_min^{α−1} ensures the PDF integrates to unity above x_min.
   - **Checks:** algebra/integration, units/dimensions
   - **Verdict:** PASS; confidence: high; impact: critical

skepthical: scientific review+verification system

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

A total of 11 candidate numerical statements were recomputed/validated (formula recomputations, percent-of-total checks, parts-vs-total consistency, difference checks, and percent-change checks). All 11 checks passed within the provided absolute tolerances (typically reflecting rounding to 0.1% or 4 decimal places).

Checked items

1. ✔ **C01_powerlaw_norm_true_case** (Section 3.1 (page 6): 'The inferred normalization at the threshold was C = 1.4979... analytic value for the specified α and xmin.')
    - **Claim:** Check that the reported normalization C matches the stated formula C = (α − 1) * xmin^(α−1) for αfit = 2.4979 and xmin = 1.0.
    - **Checks:** formula_recompute
    - **Verdict:** PASS
    - **Notes:** At xmin=1, C should equal (α-1) exactly; small tolerance for rounding.
2. ✔ **C02_powerlaw_norm_gaussian_case** (Section 3.2 (page 7): 'The inferred normalization at xmin also decreased to C = 1.2936...')

skepthical-ai.org

Being extended to include data

by Dulain Thannippuli

(Master's student at Cambridge)

AIR

Sign in

Research infrastructure for AI-powered science

Run autonomous research agents, access scientific tools, and accelerate discovery — from your terminal or desktop.

Get Started

Python SDK

Install with pip and run research tasks from any Python environment. Simple API, full control.

Desktop App

Native app for macOS and Windows. Execute code locally, manage files, and view results in real time.

REST API

Integrate AI research capabilities into your own tools and workflows via a straightforward API.

air-superintelligence.com

AI agents for science: some thoughts

- ✓ Navigating the **literature**, **cross-disciplinarity**
- ✓ **Automating** tedious parts of the work and **exploring** new ideas **fast**
- ✓ More time for deep thinking.
- ✓ **Not just LLMs:** use tools (Python, Mathematica, LEAN)
- ✓ **Auditable**
- ⚠️ Output must be **verified** and critically **examined**
- ⚠️ Potential for **misuse** (flooding arXiv with low quality or biased papers).
- ☝️ Mostly limited to **theoretical fields** or **existing data** - for now!

AI agents for science: some thoughts

✓ Navigating the **literature**, **cross-disciplinarity**

✓ **Automating** tedious parts of the work and **exploring** new ideas **fast**

✓ More time for deep thinking.

✓ **Not just LLMs:** use tools (Python, Mathematica, LEAN)

✓ **Auditable**

⚠️ Output must be **verified** and critically **examined**

⚠️ Potential for **misuse** (flooding arXiv with low quality or biased papers).

☝️ Mostly limited to **theoretical fields** or **existing data** - for now!

"The goal of integrating these tools should not be full automation or dehumanization, but empowerment: elevating the level at which researchers can engage within the scientific process, enabling them to tackle more ambitious projects while concentrating on fundamental questions, critical interpretation, and creative reasoning."

(Opportunities in AI/ML for the Rubin LSST DESC)