Cosmological Inference with Machine Learning

Leander Thiele

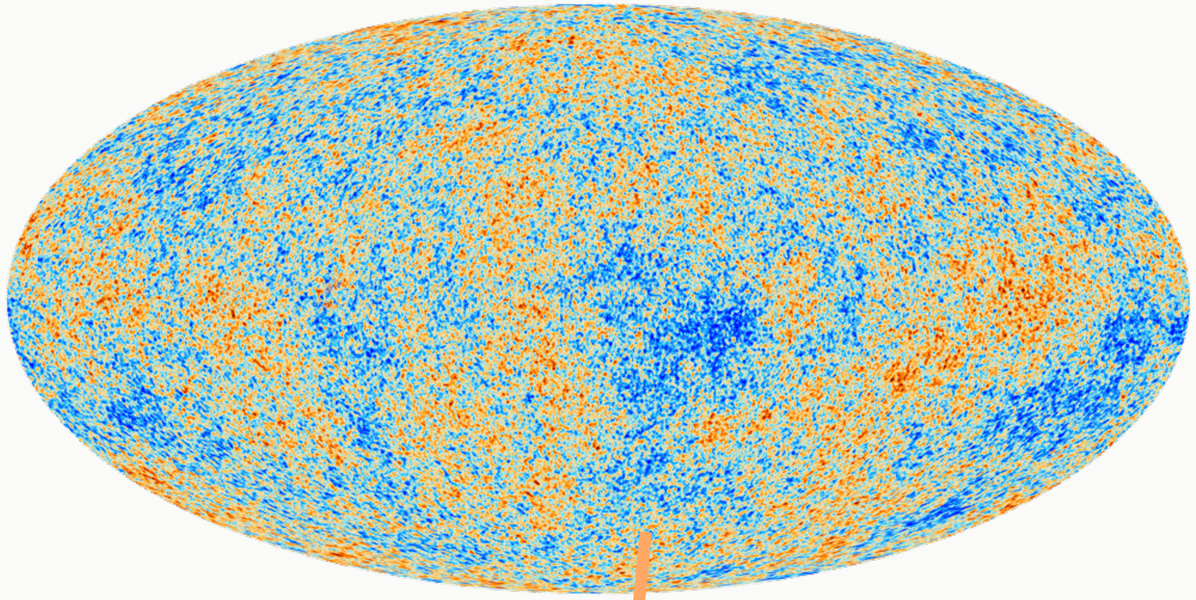
3rd Workshop on AI + Astronomy, Zhejiang Lab, Hangzhou
October 19 2026

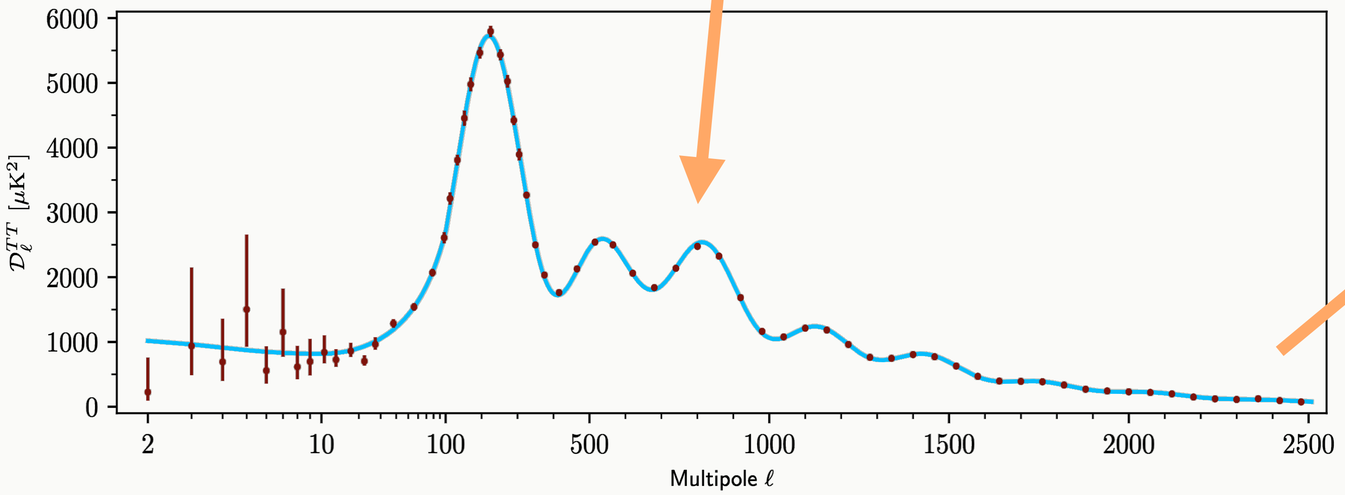
KAVLI IPMU INSTITUTE FOR THE PHYSICS AND MATHEMATICS OF THE UNIVERSE
東京大学 THE UNIVERSITY OF TOKYO

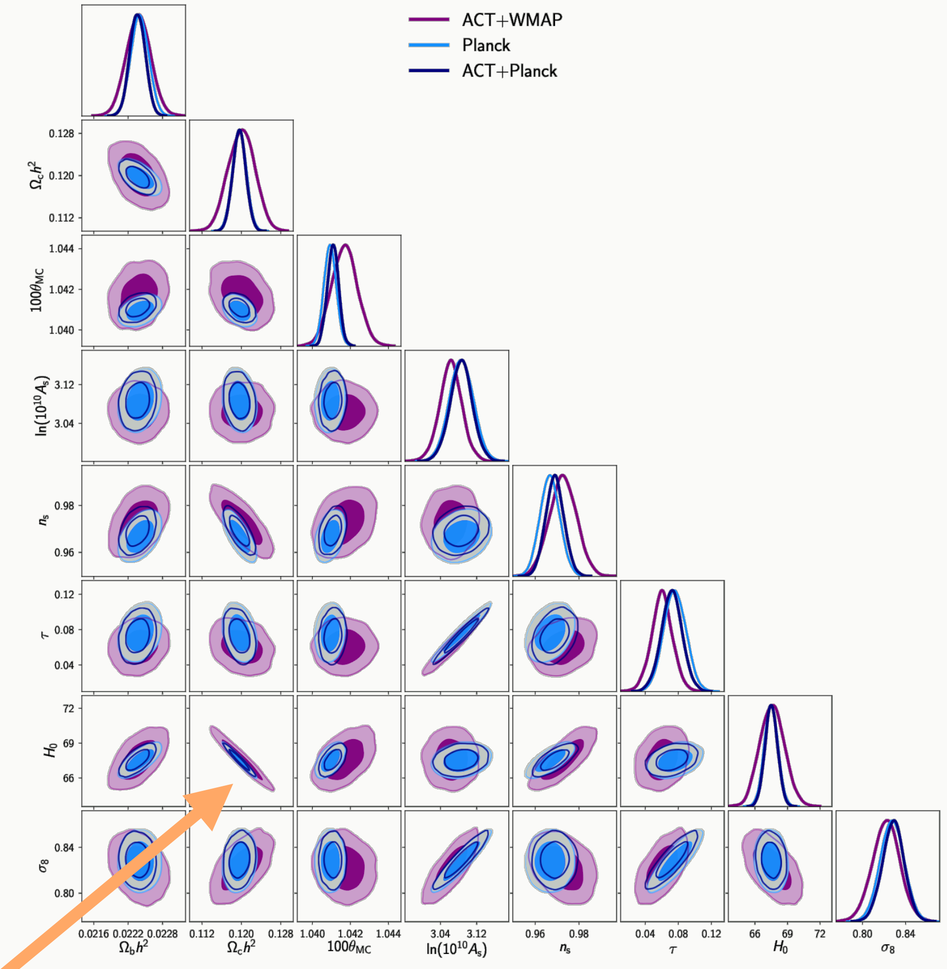
Overview

- implicit likelihood inference (= simulation-based inference)
- multi-fidelity training
- constrained learnable compressions (void finding, filament finding)

Likelihood-based cosmology







Explicit Likelihood

$$-2\log p(x\,|\,\theta) \propto (x-\mu(\theta))^T\Sigma^{-1}(x-\mu(\theta))$$

Implicit Likelihood Inference (ILI)

- Why should nature have conspired to put all information into (quasi)linear modes we happen to be able to describe with perturbation theory?

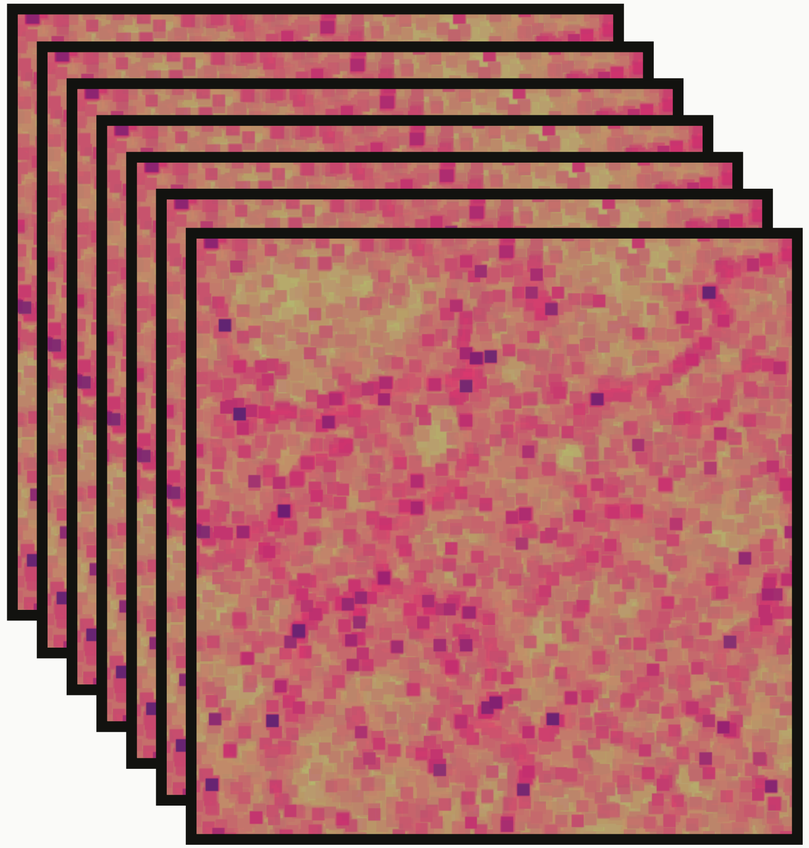
- Accessing information in non-perturbative regime typically requires simulation-based methods

- Implicit Likelihood Inference: techniques to solve inverse problems which are implicitly defined through simulations. Typically using deep neural networks.

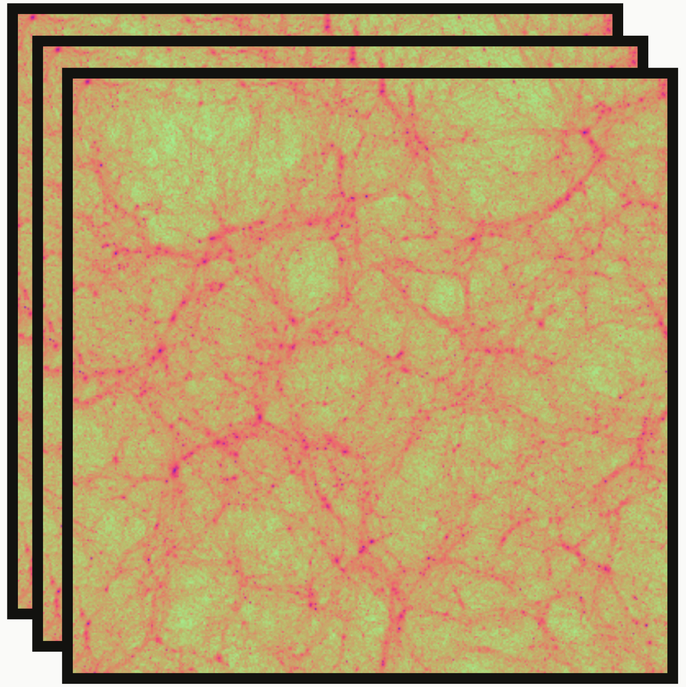
Implicit Likelihood Inference: Status

- increasingly, applications to survey data
- not the same level of trust as we have in traditional inference
- issues:
  - sample efficiency (how can we train with reasonable simulation budget?)
  - robustness (how can we be sure only information is used that is accurately reproduced by simulations?)

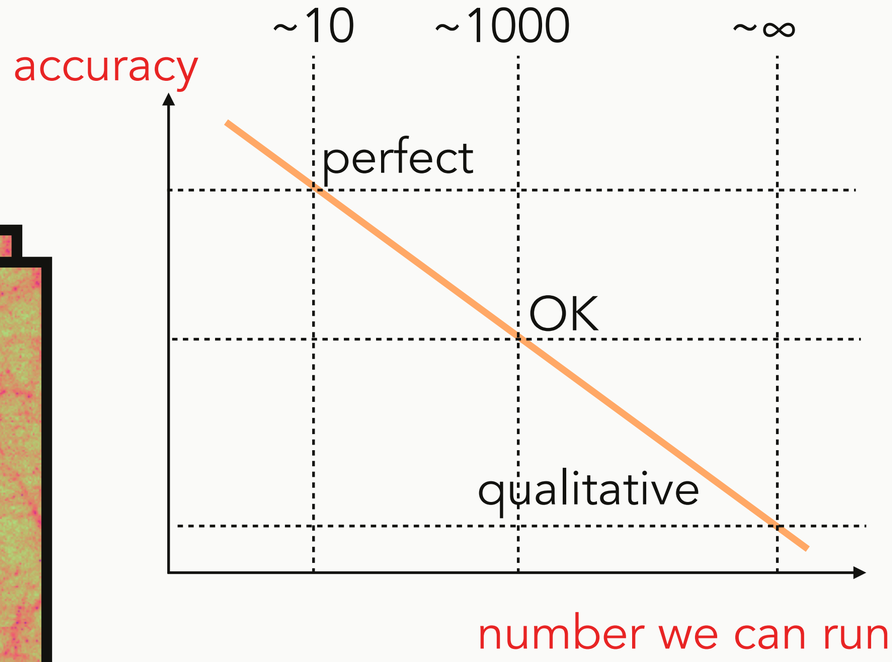
Multi-Fidelity ILI: Motivation



large number of low-fidelity simulations

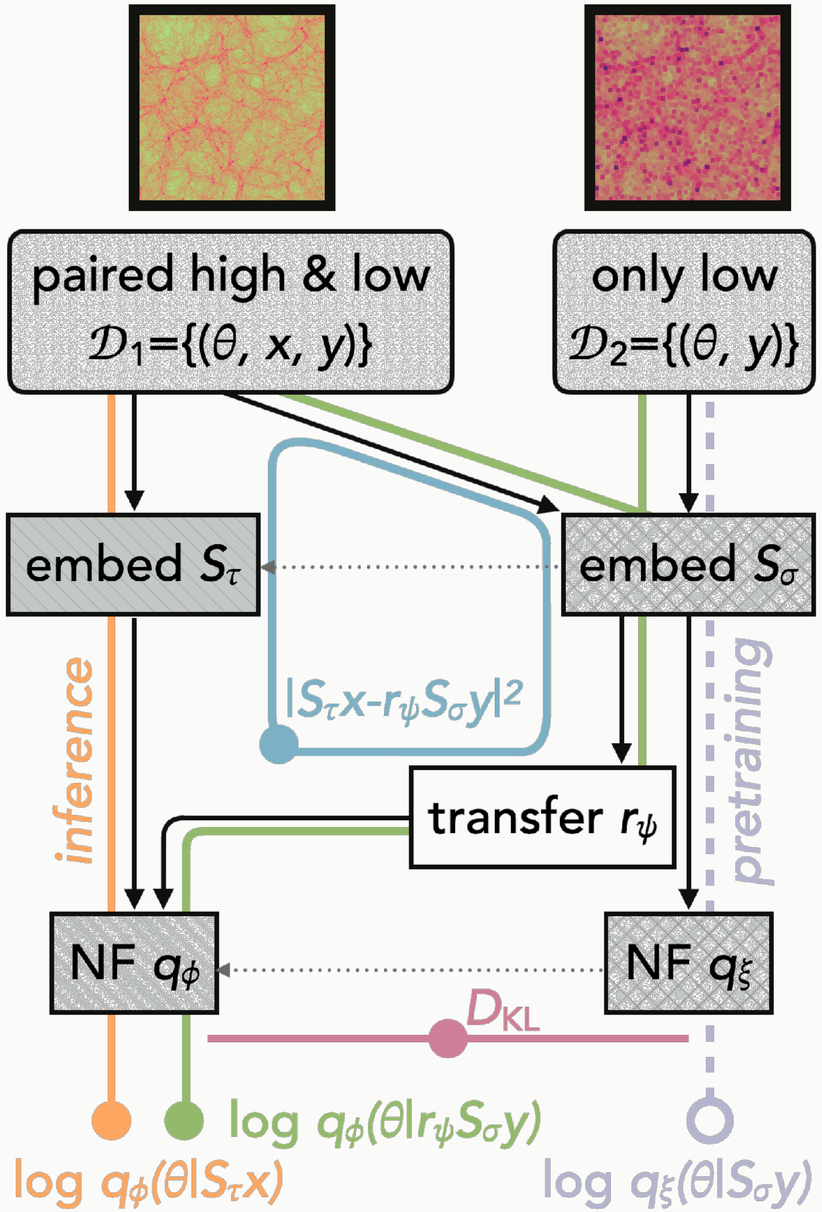


small number of high-fidelity simulations



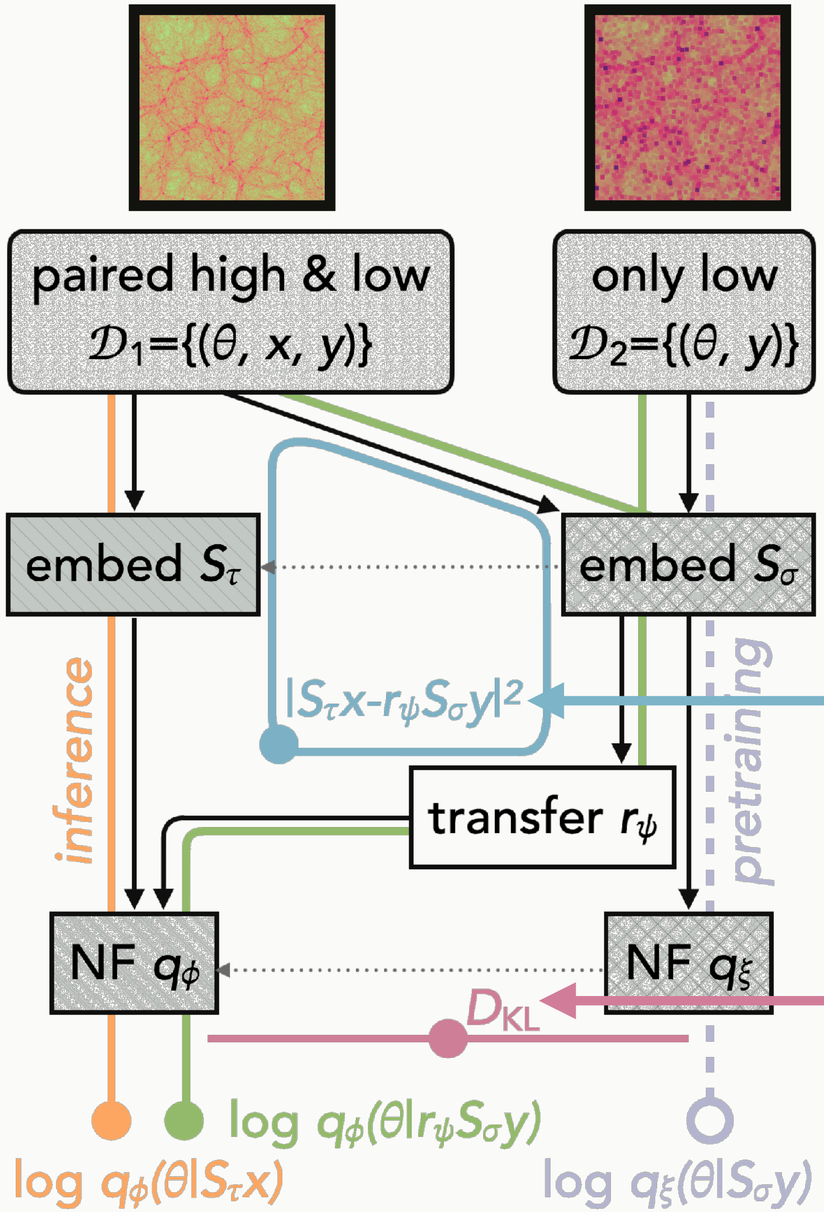
[Thiele, Bayer, Takeishi 2025]

Multi-Fidelity ILI



[Thiele, Bayer, Takeishi 2025]

Multi-Fidelity ILI

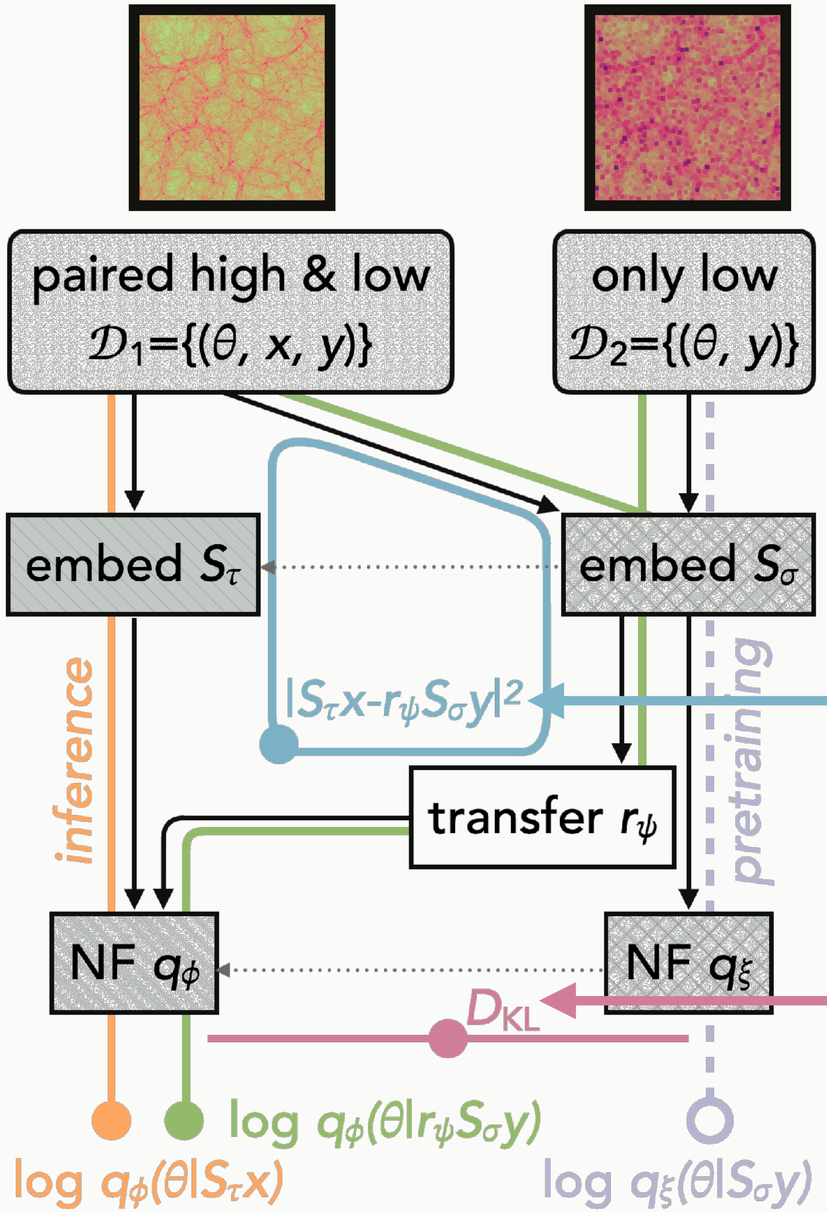


feature matching in latent space

knowledge distillation - exploit similar information content between different simulation qualities

[Thiele, Bayer, Takeishi 2025]

Multi-Fidelity ILI



If you think of different simulation qualities as different data modalities, there might be some interesting overlap with foundation models…

feature matching in latent space

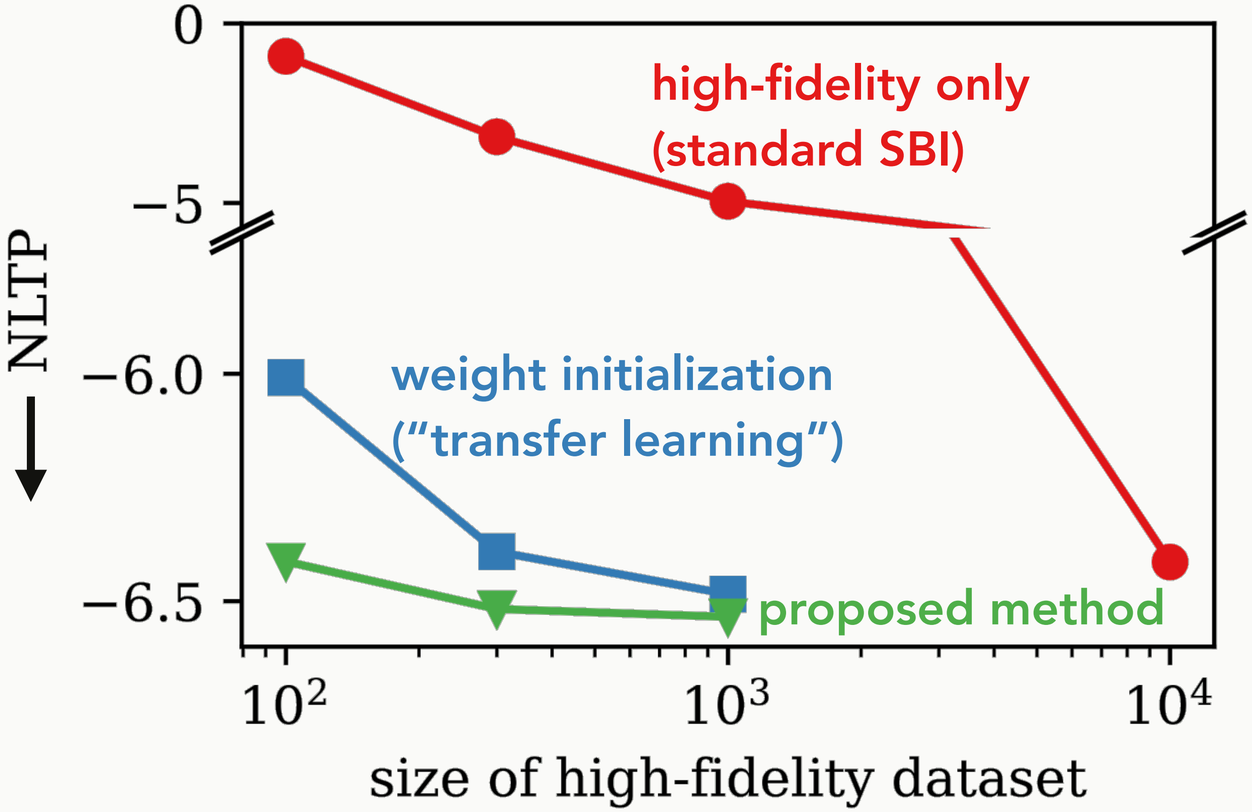
knowledge distillation - exploit similar information content between different simulation qualities

[Thiele, Bayer, Takeishi 2025]

Multi-Fidelity ILI

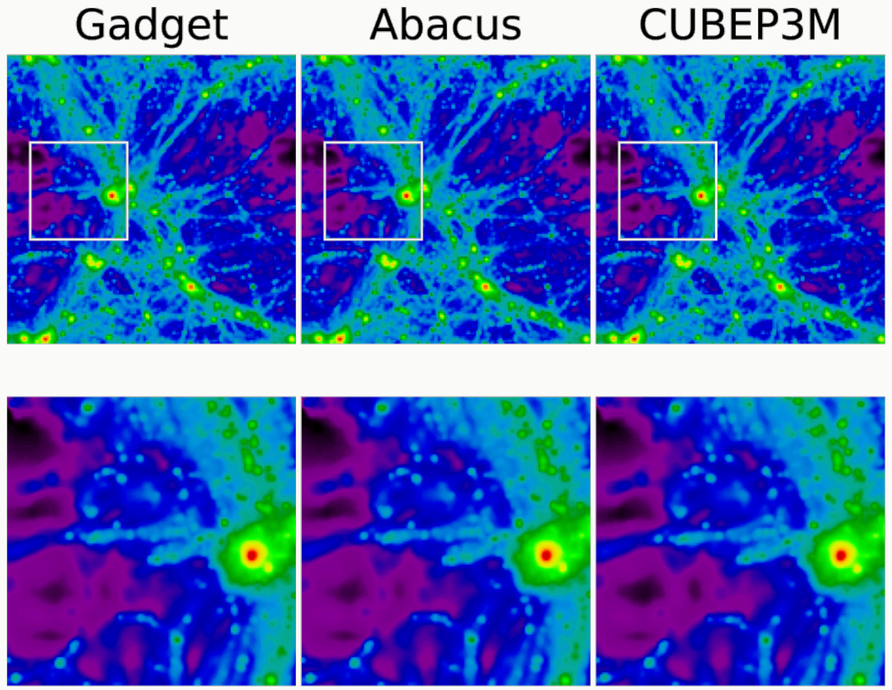
Evaluation: ΛCDM from $P_m(k)$

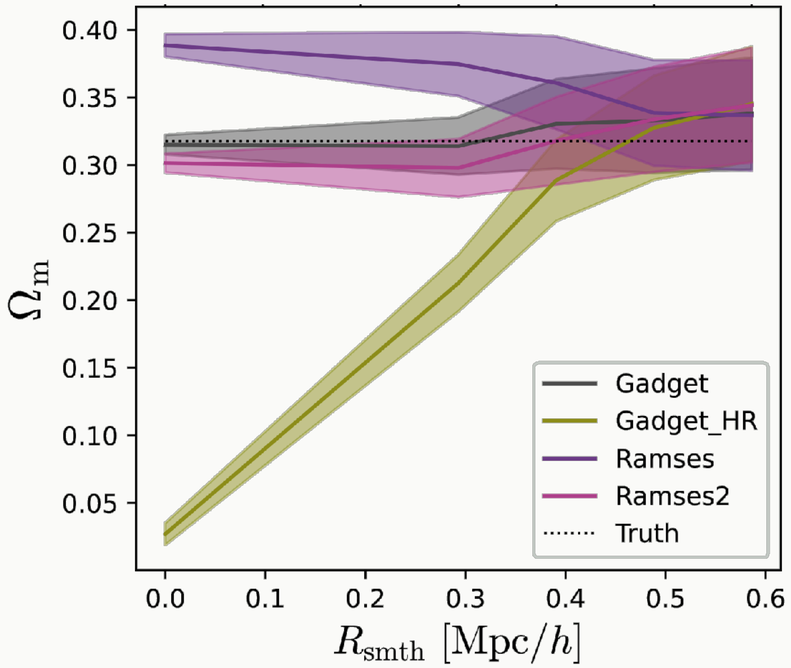
High-fidelity = Quijote
Low-fidelity = FastPM



Maximizing signal-to-noise?

- Even “trivial” physics is not trivial on computers, not to speak of astrophysics
- unconstrained machine learning highly sensitive to small-scale details
- simply maximizing S/N is typically not meaningful

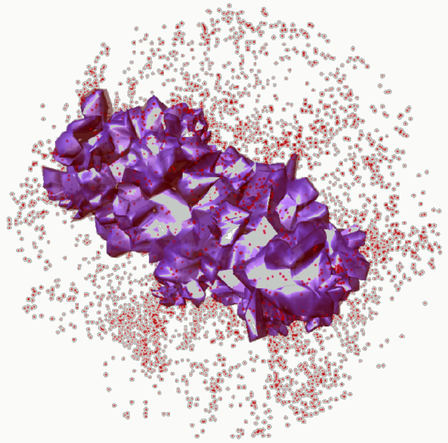




[Bayer+2025]

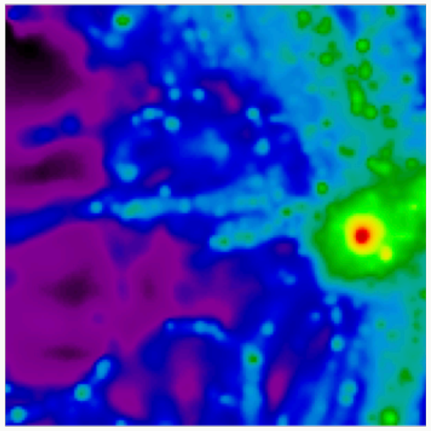
Maximizing signal-to-noise: constrained version

traditional void finding algorithm



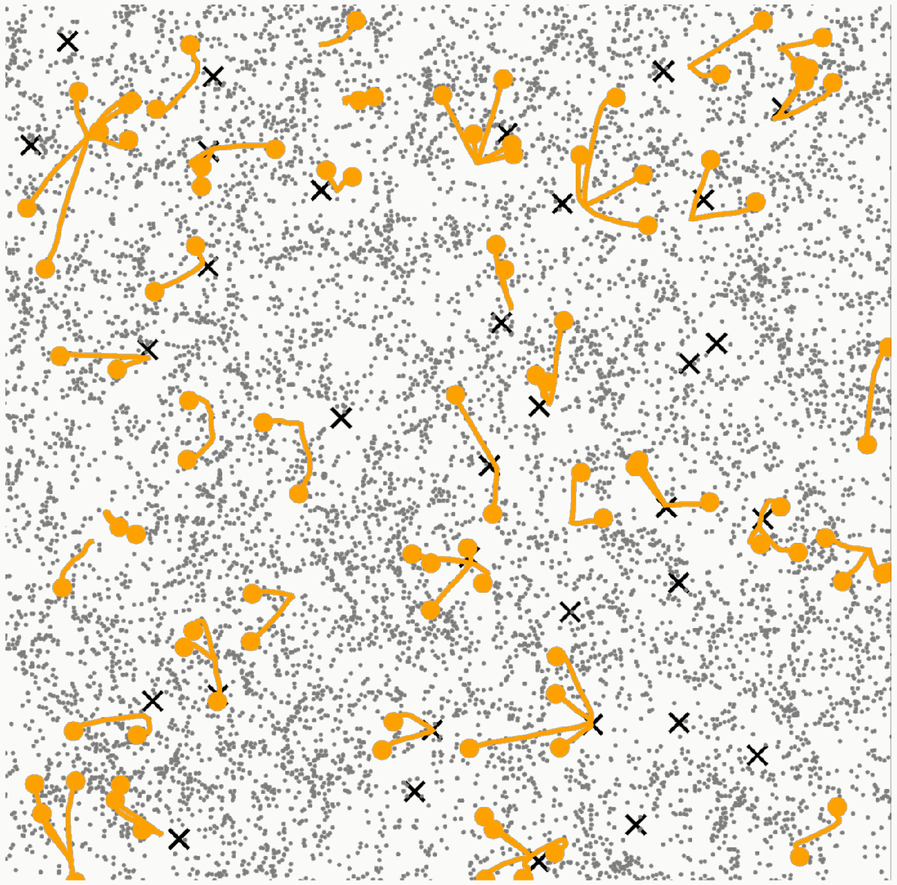
[Sutter+2014]

unconstrained S/N maximization with ML



[Bayer+2025]

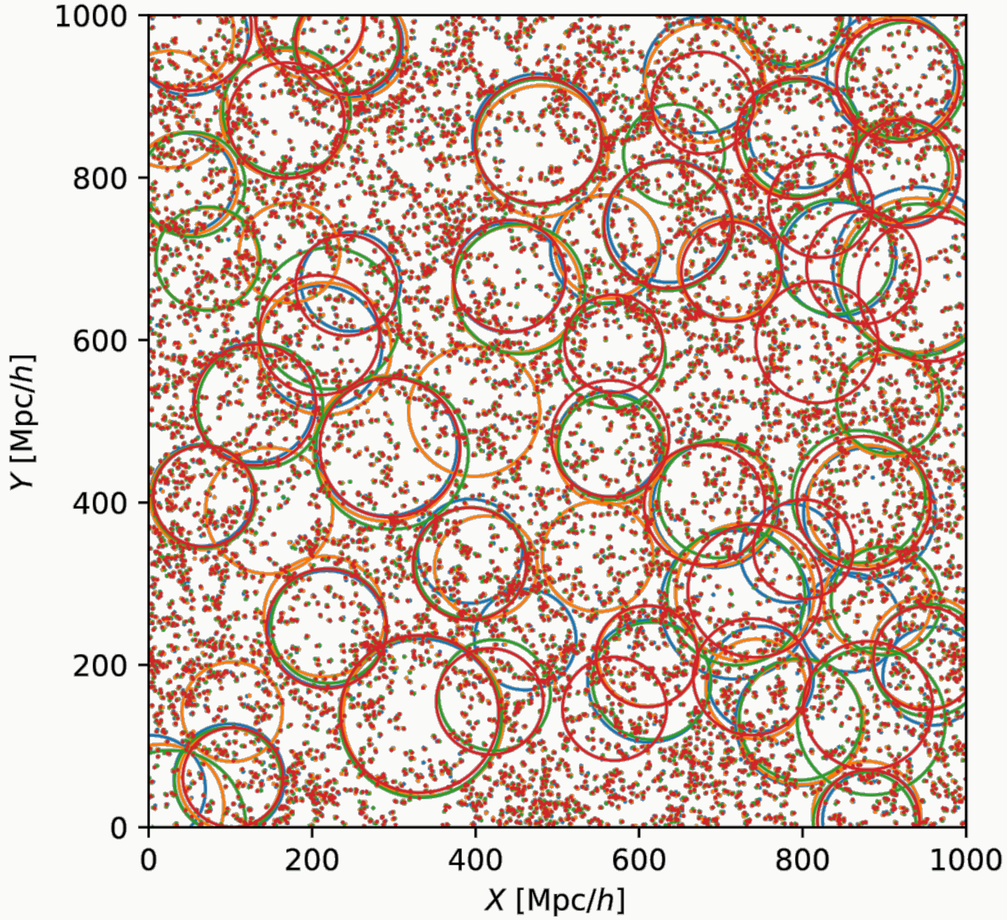
probabilistic, machine-learned void finding:
use ML within specified guardrails



Deterministic void finding is wrong

- Most existing void finders are deterministic
- This design leads to instability: resulting void catalogs very sensitive to small perturbations to the galaxy positions
- One can already trivially improve constraining power by averaging void statistic over random subsamples [Yu Liu+2026]

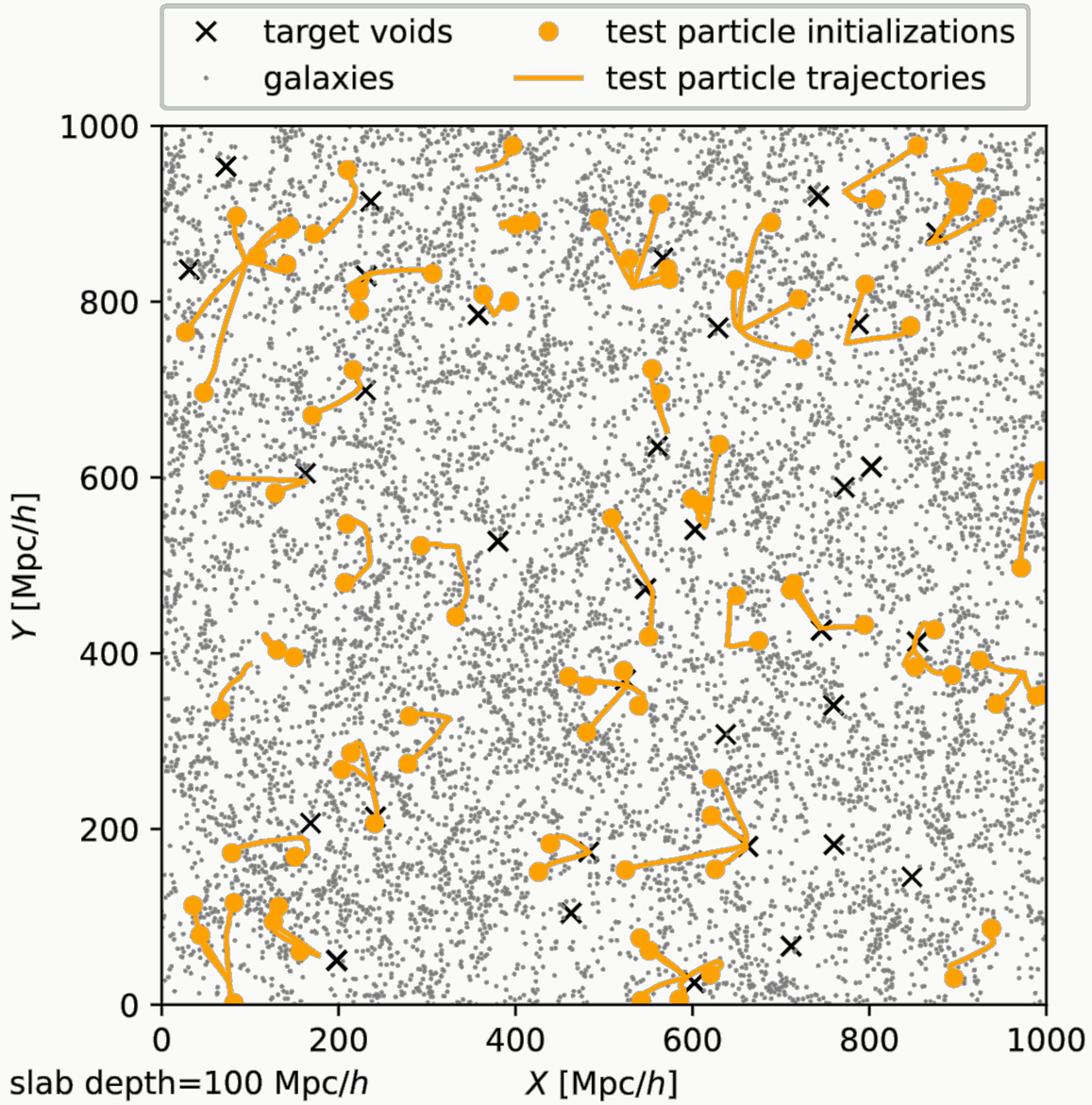
jittering galaxy positions changes void catalog



[Thiele 2026]

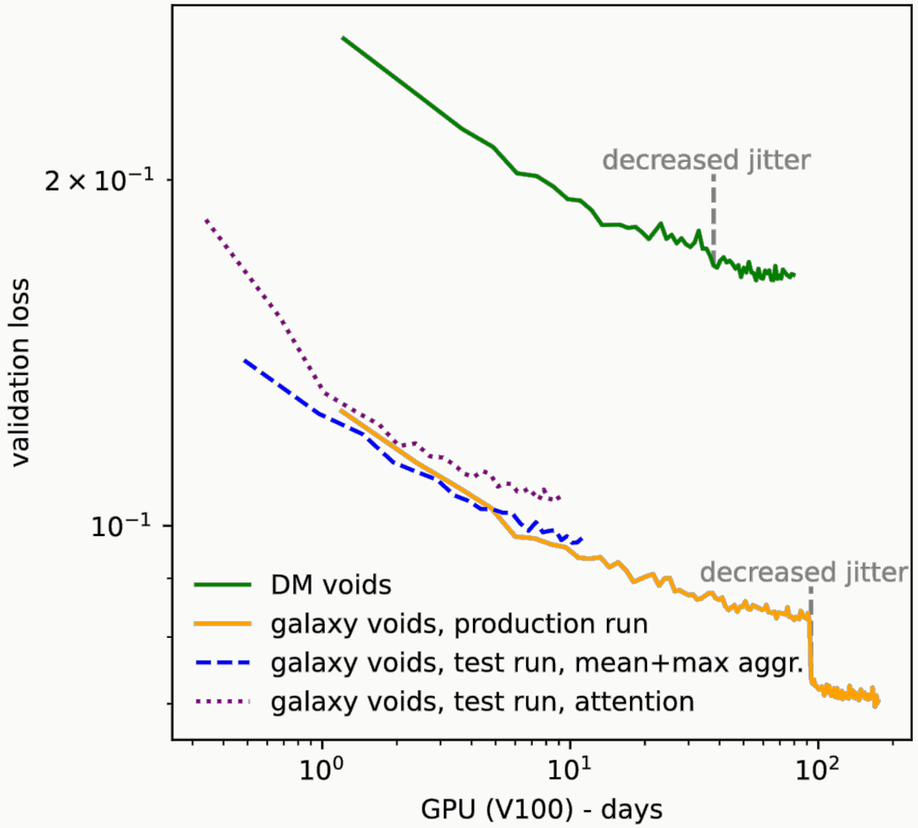
Solving void finding with machine learning

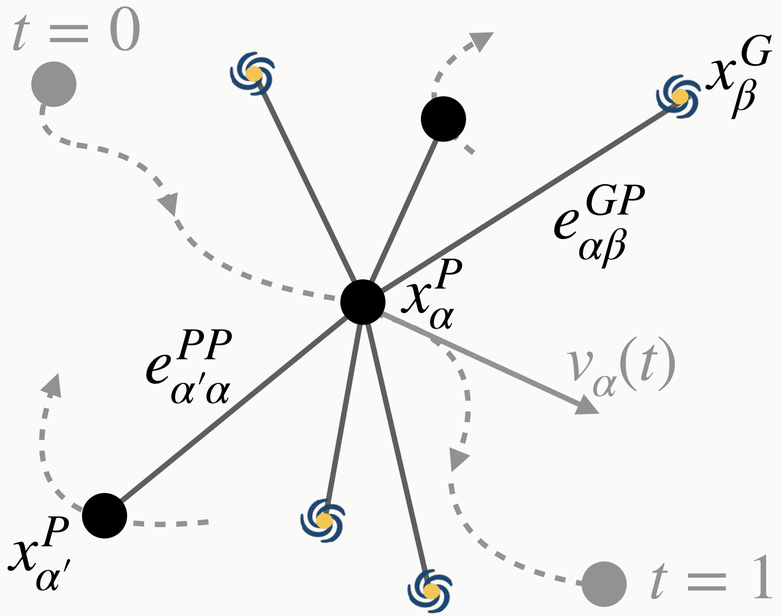
- Task: compute samples from P(void catalogs | galaxy catalog)
- outputs are sets of large unknown size
- more symmetry than protein generation
- solution: test particles evolved with flow ODE, followed by clustering



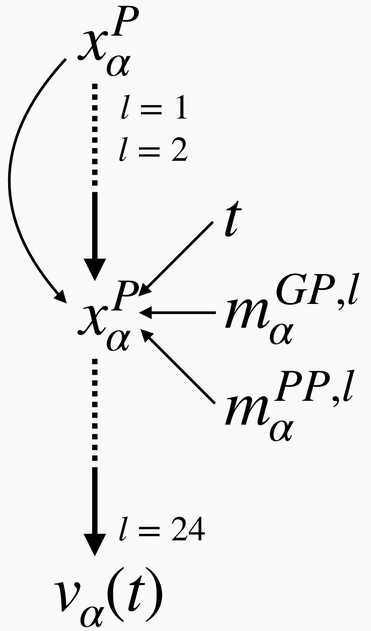
Neural Network

- graph net on k-NN graph
- old-school aggregation beats attention





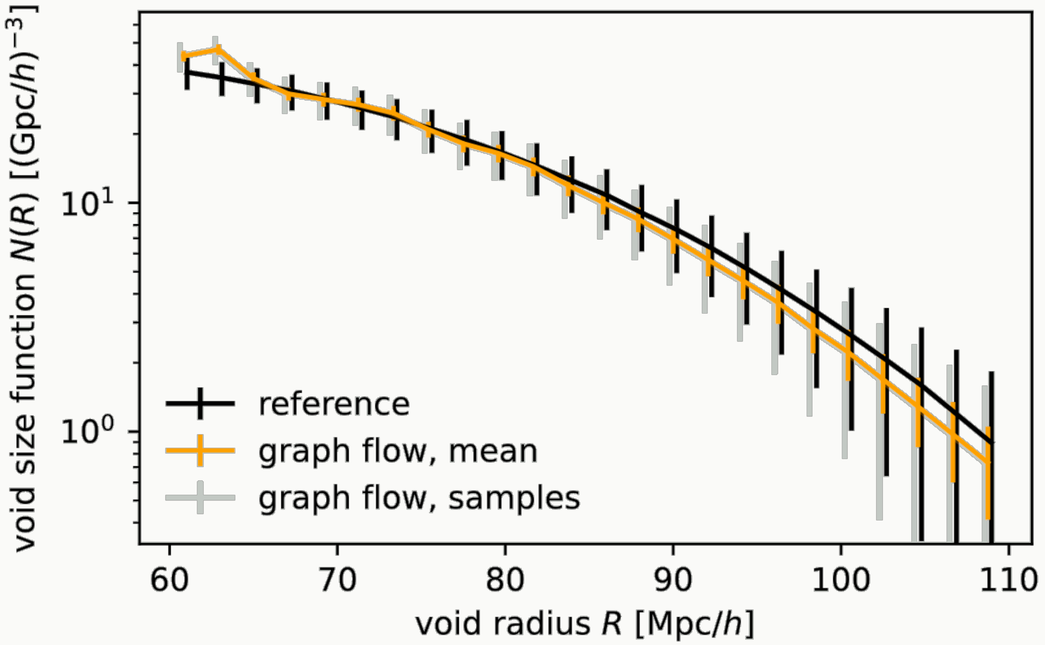
message-passing step

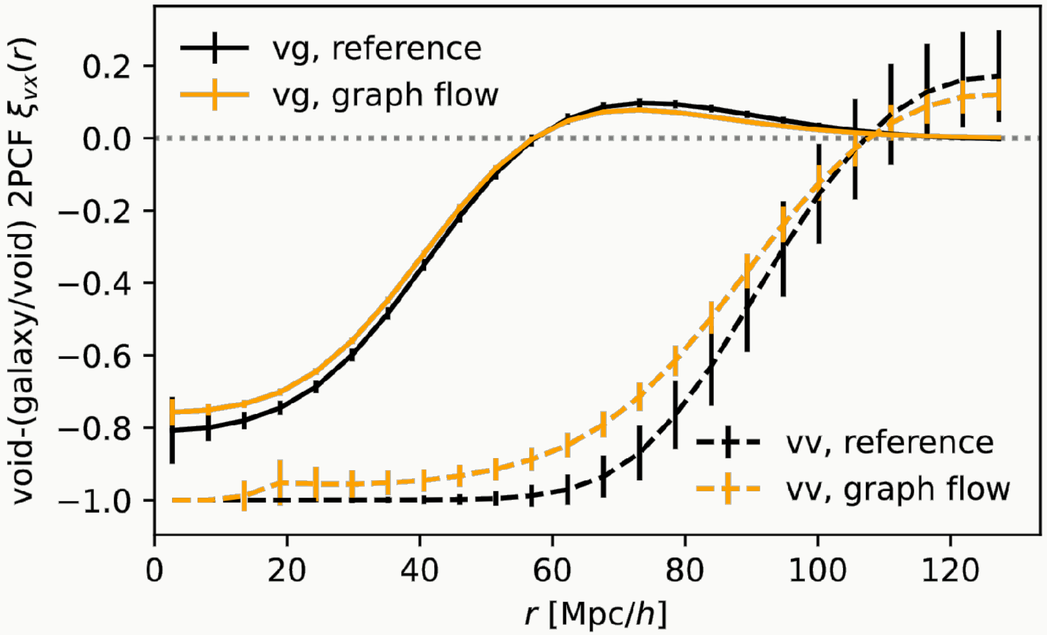


multi-layer
message aggregation

Results: Summary Statistics

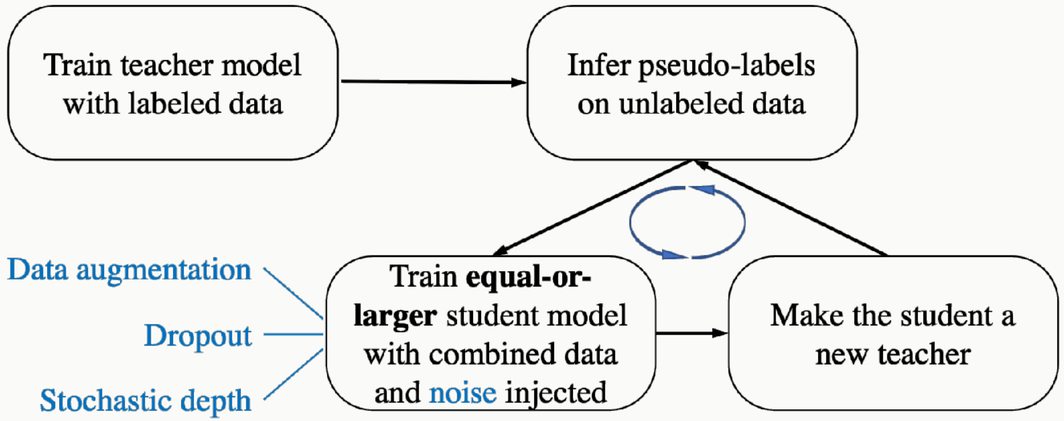
- begin by trying to match existing code (VIDE)
- in principle, the target is deterministic
- with finite expressivity, the network learns some "smoothed" version of this delta function
- Note: target voids can be defined in any way, ongoing work: using full density & velocity fields



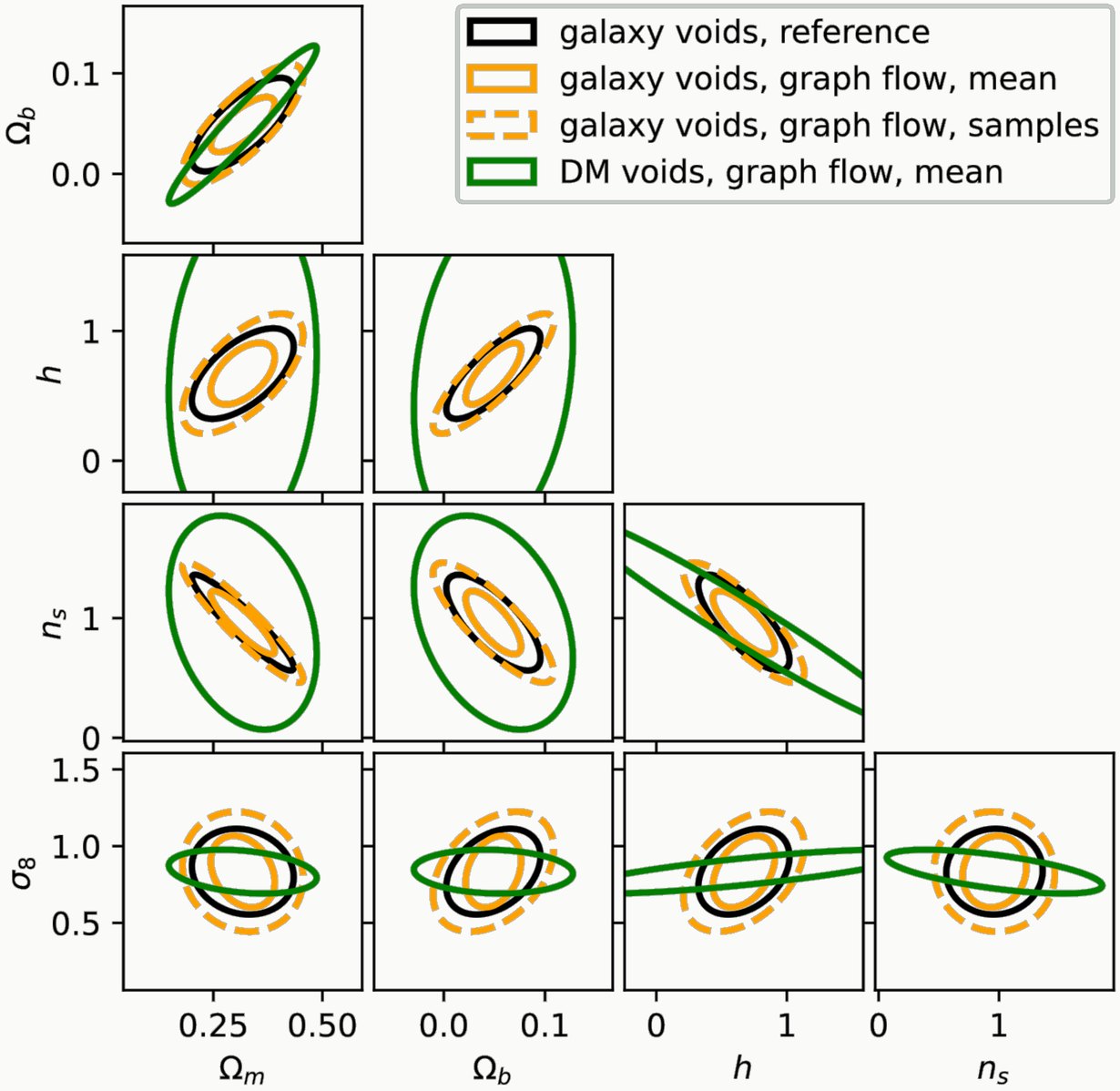


Results: Cosmological Information

- cosmological information of summary statistics averages outperforms VIDE (the target)
- interpret as effect of useful regularization



[Xie+2020]

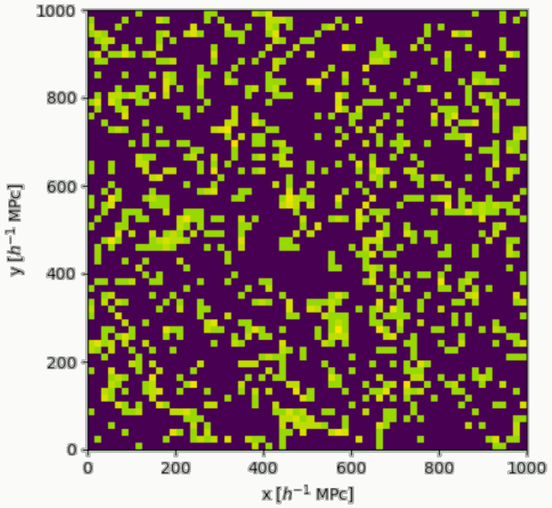


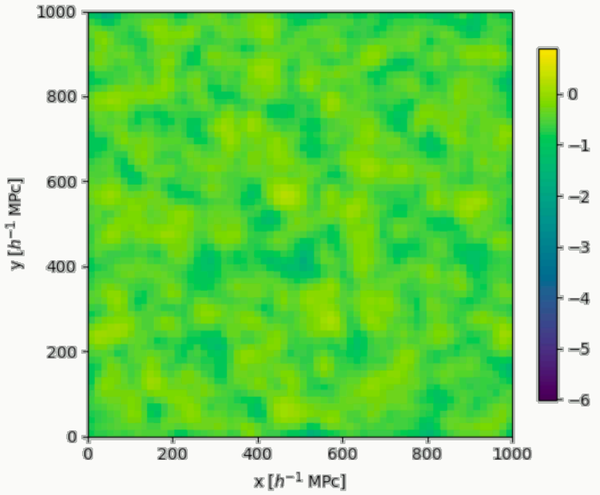
Filament finding

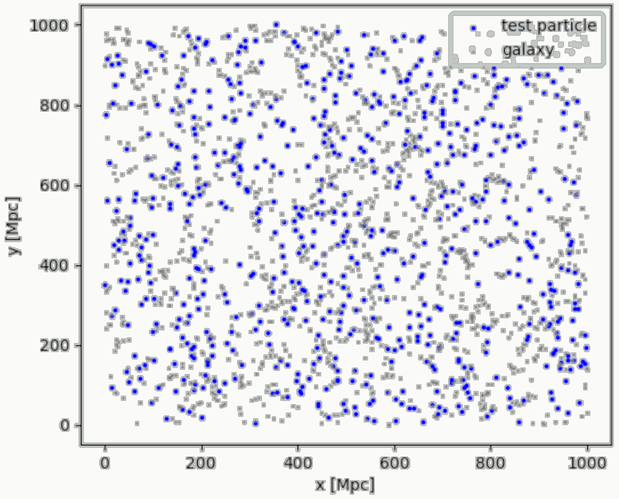
- in principle, similar technique can be used to find filaments
- complications:
  - 0d to 1d (need to link up into filaments)
  - much larger problem (some algorithms don't scale well)



w/ Alexis Tastet (Ecole Polytechnique) & Daniela Galárraga-Espinosa

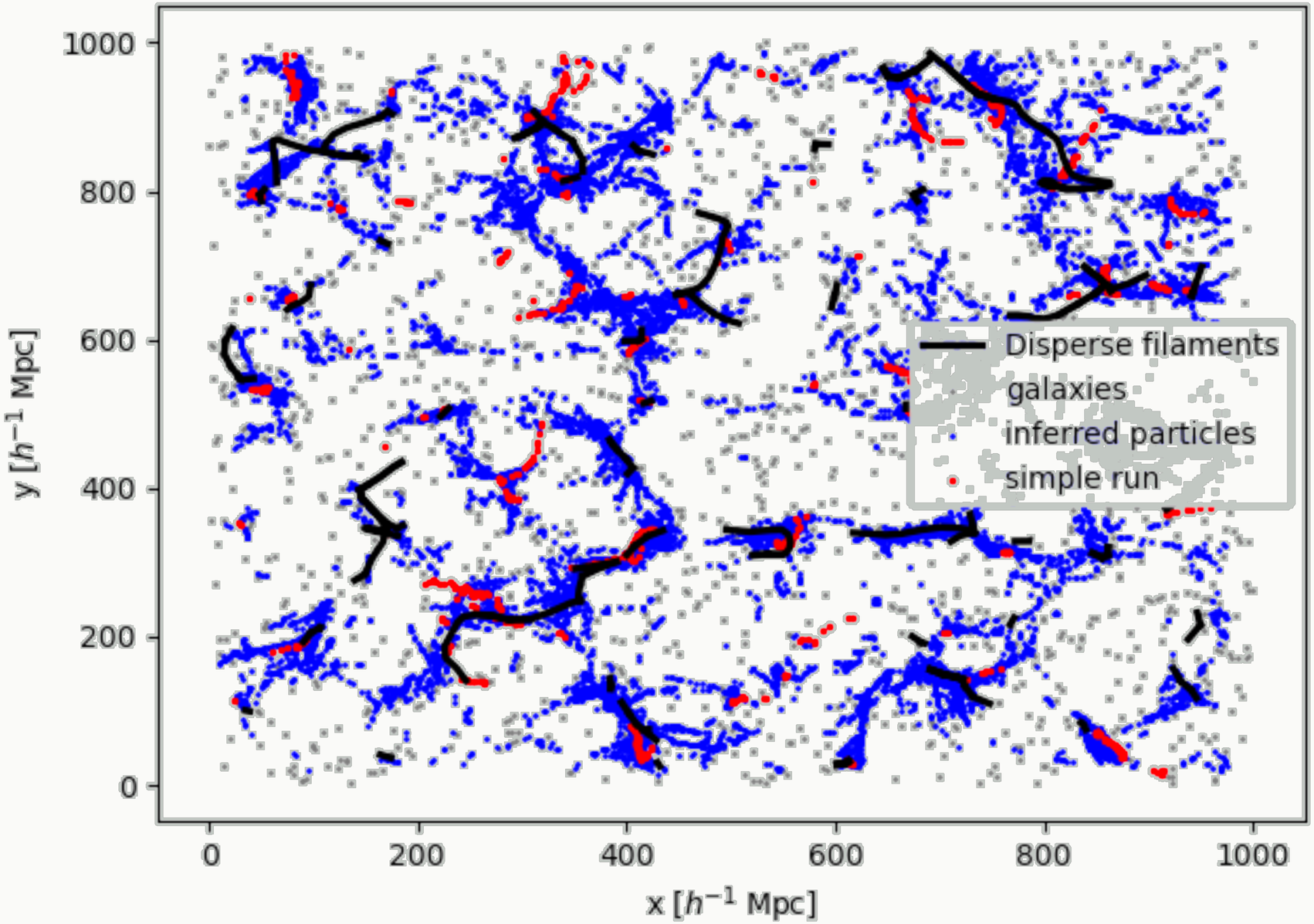






more sophisticated method to construct the base distribution

Filament finding: results



Summary

- Implicit Likelihood Inference (ILI) needed for some inverse problems
- ILI demands methodological advancements:
  - simulation efficiency: multi-fidelity training
  - robustness: constrained learnable compressions