# THE COSMIC GRAPH: WHERE IS THE INFORMATION IN LARGE-SCALE STRUCTURE ?

KICC Focus Meeting on AstroStatistics & Astro-ML, Cambridge

June 1-2, 2023

T. Lucas Mäkinen, *Imperial College London*

# The Cosmic Graph: Optimal Information Extraction from Large-Scale Structure using Catalogues

T. Lucas Mäkinen

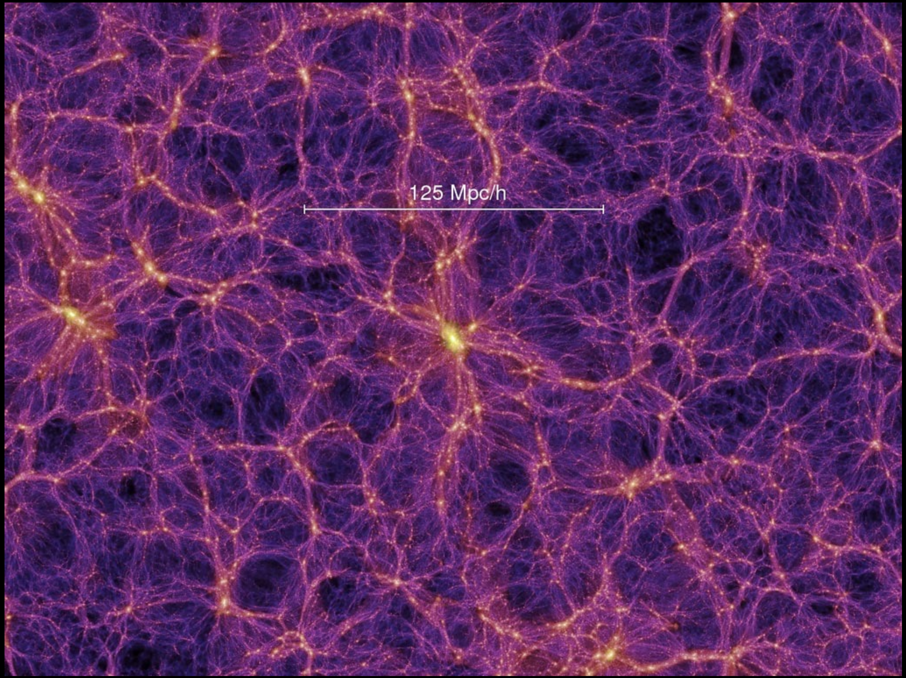
arXiv:2207.05202, arXiv:2107.07405

KICC Focus Meeting on Astrostatistics and Astro-ML

With Tom Charnock (IAP), Ben Wandelt (IAP), and Pablo Lemos (U. Sussex), Alan Heavens & Natalia Porqueres (Imperial), & Ben Wandelt (IAP)

# 01

# Why Compression ?



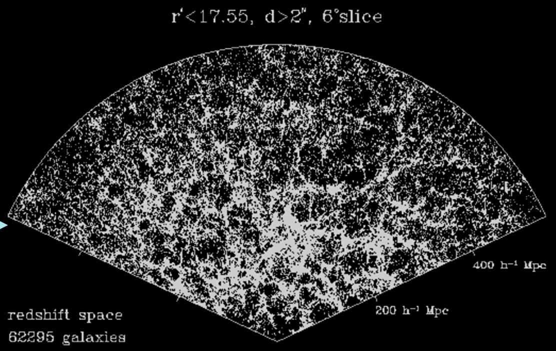

$1\ \mathrm{Mpc} = 3.2 \times 10^6$ light years

# Cosmological Inference from large-scale structure

**Inverse problem: What is the probability of a given parameter, $\theta$, being a good descriptor of observed large scale structure ?**

$\theta$

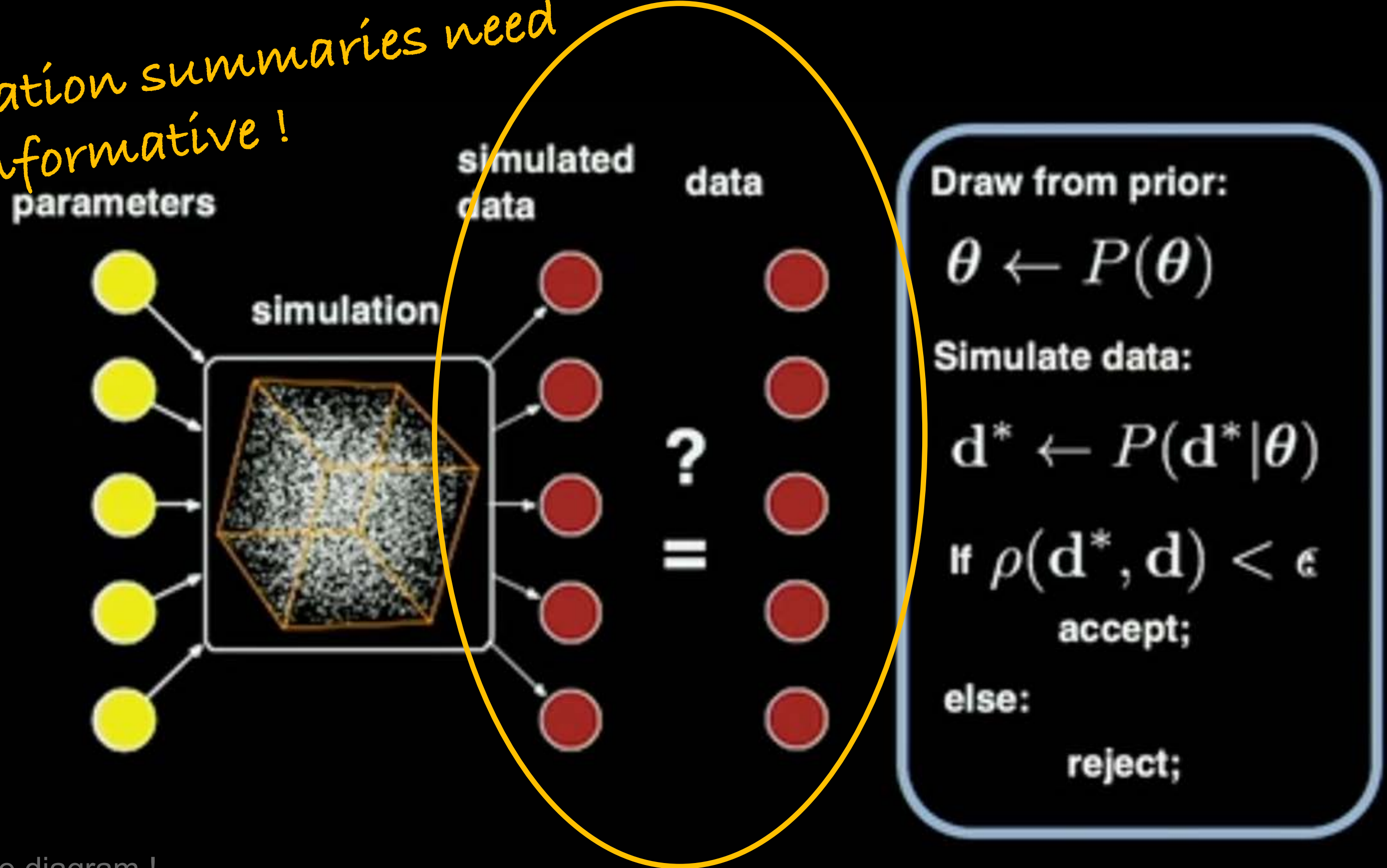
model (likelihood / simulator)




observed universe

# ILI: Implicit Likelihood Inference

Simulation summaries need to be informative !



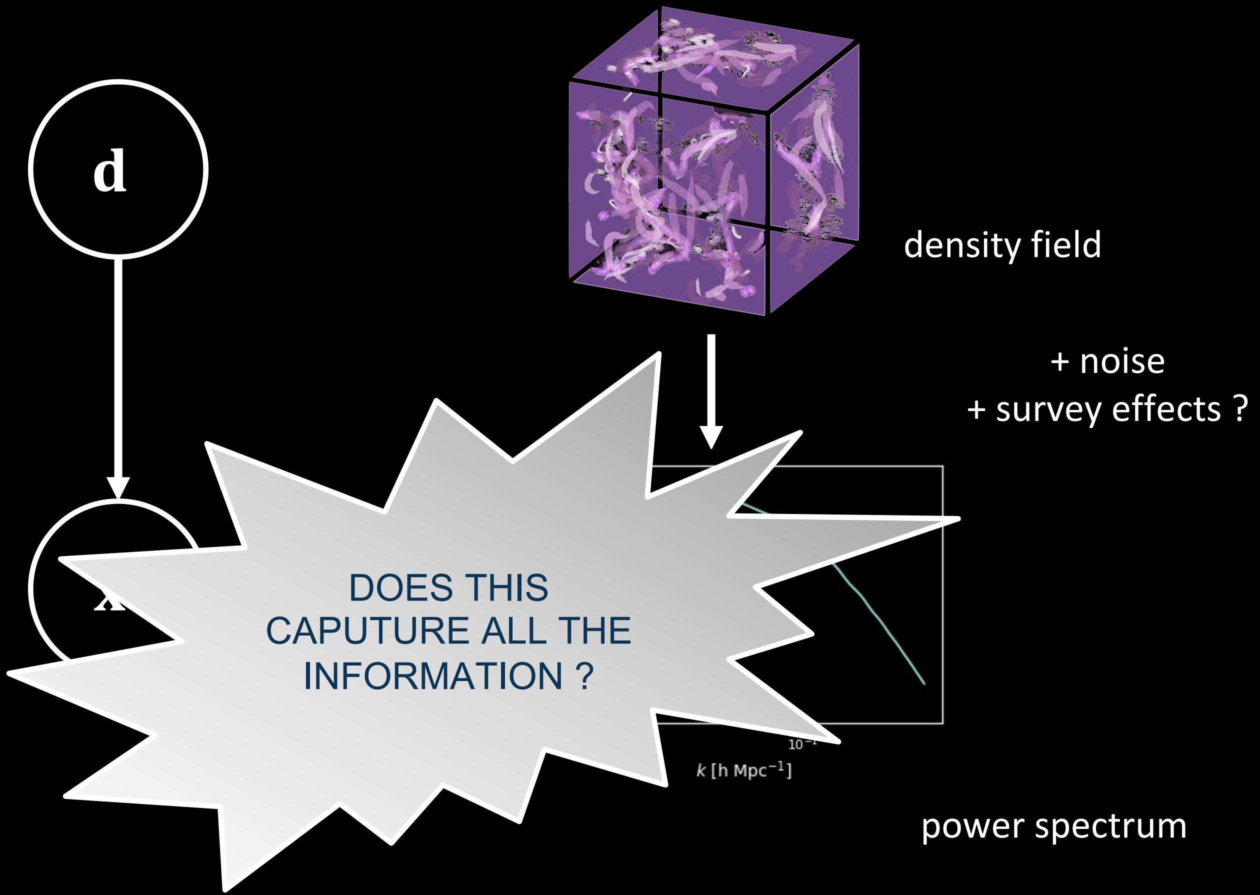

Thanks to Ben for the diagram !

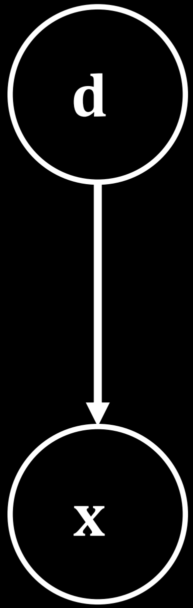
# Cosmology: an Optimization Problem

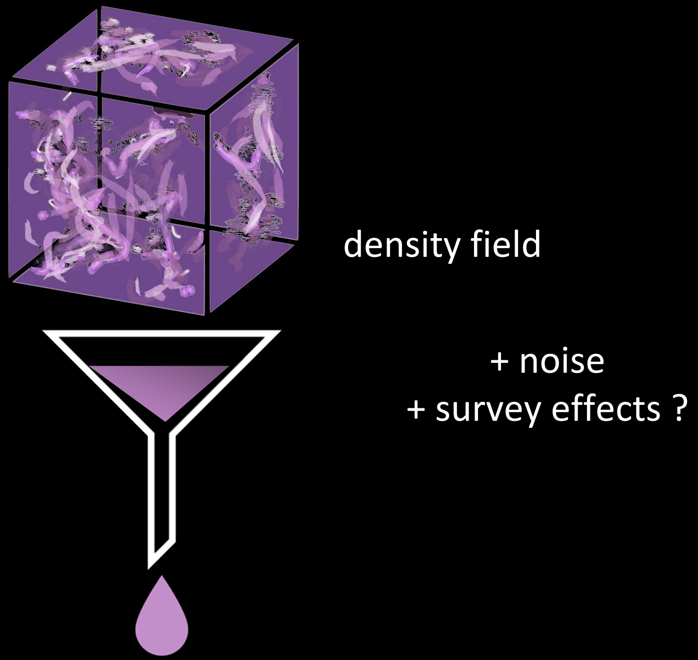
**Objective**: constraints on cosmological parameters

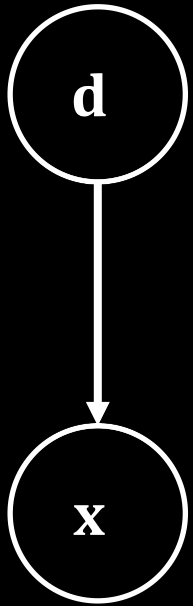
**Path**: find statistic that captures the most relevant cosmological information

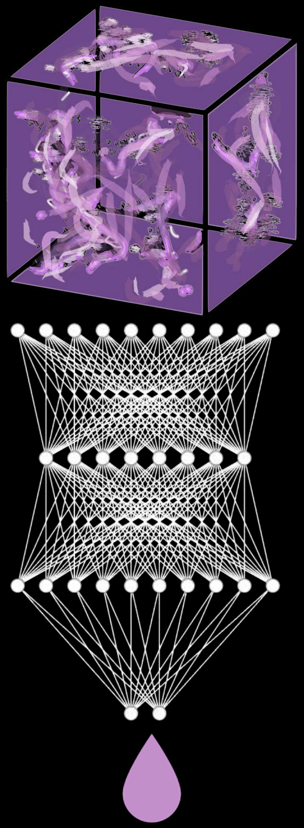
**Question:** Can we *learn* this path by minimizing (or maximizing) the objective ?


d
density field
+ noise
+ survey effects ?
DOES THIS
CAPUTURE ALL THE
INFORMATION ?
k [h Mpc⁻¹]
power spectrum


d
x


density field
+ noise
+ survey effects ?



density field

+ noise
+ survey effects ?

neural compression to
maximally-informative
summaries

02

# The Fisher Information

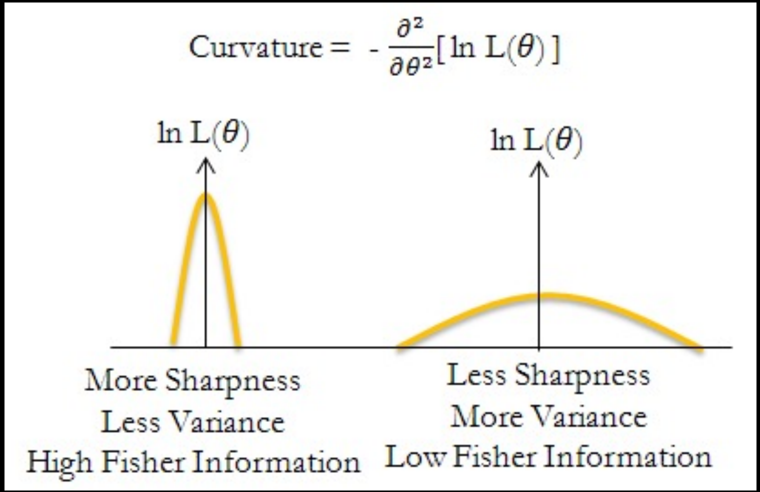
Defining the Optimization Objective

**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$

$$\mathbf{F}_{\alpha\beta} = -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle_{\theta=\theta_{\mathrm{fid}}}$$

Think of this as the *curvature* of the log-likelihood, $\ln \mathcal{L}$ at $\theta_{\mathrm{fid}}$

**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$



**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$

Cramér-Rao bound:

$$\langle(\theta_\alpha - \langle\theta_\alpha\rangle)\left(\theta_\beta - \left\langle\theta_\beta\right\rangle\right)\rangle \geq \mathbf{F}_{\boldsymbol{\alpha\beta}}^{\mathbf{-1}}$$

Gives us a lower bound for the (average) variance of a parameter estimate

**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$

**Example:** draw $\mathrm{n}_{\mathbf{d}}$ independent datapoints from a normal distribution, $\mathcal{N}(\mu,\sigma)$ . Then the likelihood is:

$$\mathcal{L}(\mathbf{d}|\mu,\sigma) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(d_i-\mu)^2}{\sigma^2}\right)$$

And the Fisher matrix is:

$$F = -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle_{\theta_{\mathrm{fid}}} = \begin{pmatrix} \frac{-n_d}{\sigma^2} & 0 \\ 0 & \frac{-n_d}{2\sigma^4} \end{pmatrix}_{\sigma_{\mathrm{fid}}}$$

**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$

What if we can't differentiate through our likelihood / statistic ?

For an arbitrary statistic Q:

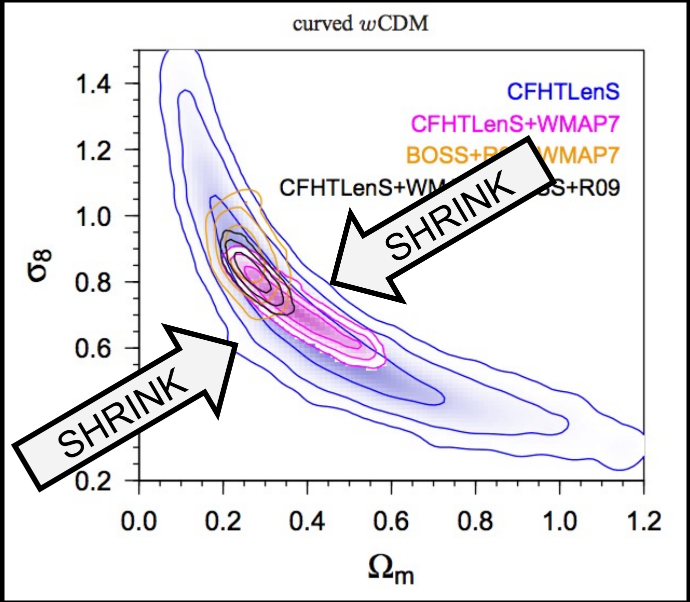
$$F_{ij} = \frac{\partial Q_\alpha}{\partial \theta_i} C^{-1}_{\alpha\beta} \frac{\partial Q_\beta}{\partial \theta_j}$$

where

$$\frac{\partial Q_\alpha}{\partial \theta_i} \approx \frac{Q(\theta_i^+) - Q(\theta_i^-)}{\theta^+ - \theta^-}$$

**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$

this is our objective !



**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$

optimize !

Cramér-Rao bound:

$$\langle(\theta_\alpha - \langle\theta_\alpha\rangle)\left(\theta_\beta - \langle\theta_\beta\rangle\right)\rangle \geq \mathbf{F}^{-1}_{\boldsymbol{\alpha\beta}}$$

Gives us a lower bound for the (average) variance of a parameter estimate

# 03

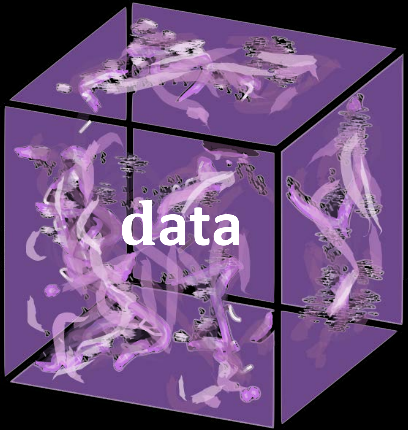
# Extracting Info with IMNNS

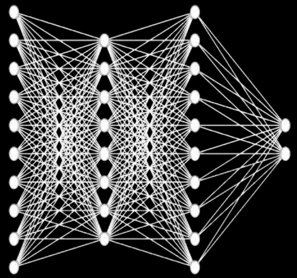
# Information Maximising Neural Networks

*Can we train a neural network to compress a universe simulation down to a couple of numbers ?*

$$f : \mathbf{d} \mapsto \mathbf{x}$$

**params** →



**summaries**

See Alsing & Wandelt (2018) arXiv:1712.00012 for why summary space is taken to be the same dimension as parameter space

$$\boldsymbol{\theta} \in \mathbb{R}^2 \qquad \mathbf{d} \in \mathbb{R}^{N_{\mathbf{pix}}} \qquad \mapsto \qquad \mathbf{x} \in \mathbb{R}^2$$

# Information Maximising Neural Networks

1) adopt a Gaussian likelihood form to compute our Fisher information:

$$-2\ln\mathcal{L}(\mathbf{x}|\mathbf{d}) = \left(\mathbf{x} - \boldsymbol{\mu}_f(\boldsymbol{\theta})\right)^T \boldsymbol{C}_f^{-1}(\mathbf{x} - \boldsymbol{\mu}_f(\boldsymbol{\theta}))$$

mean and covariance of network outputs

Charnock et al (2018) arXiv:1802.03537

# Information Maximising Neural Networks

1) adopt a Gaussian likelihood form to compute our Fisher information:

$$-2\ \ln\mathcal{L}(\mathbf{x}|\mathbf{d}) = \left(\mathbf{x} - \boldsymbol{\mu}_f(\boldsymbol{\theta})\right)^T \ \boldsymbol{C}_f^{-1}(\mathbf{x} - \ \boldsymbol{\mu}_f(\boldsymbol{\theta}))$$

2) Compute IMNN Fisher:

$$\mathrm{F}_{\alpha\beta} = \mathrm{tr}[\boldsymbol{\mu}_{f,\alpha}^T\, C_f^{-1} \boldsymbol{\mu}_{f,\beta}]$$

3) train until Fisher information is maximised *at a fiducial model*

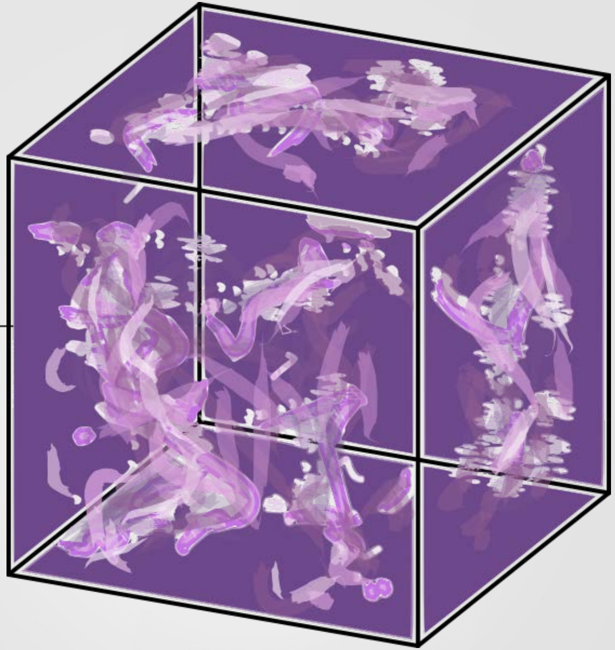
Charnock et al (2018) arXiv:1802.03537

# Main IMNN Scheme


completely differentiable in Jax !

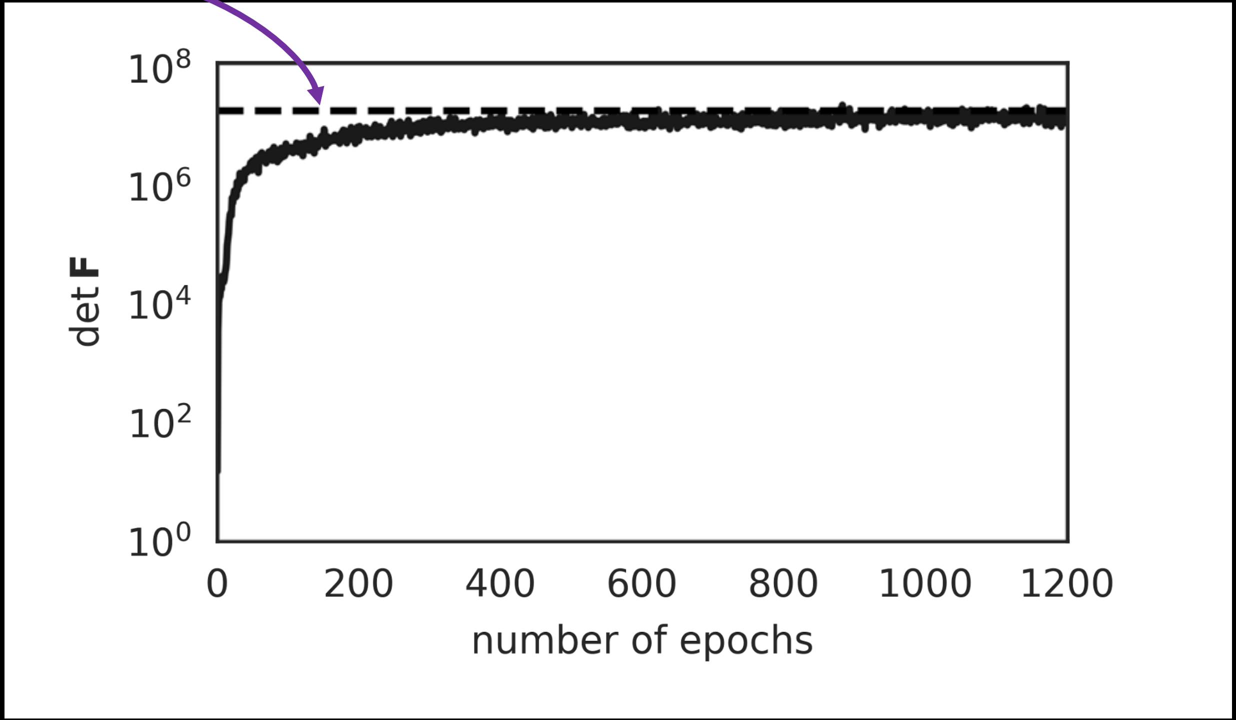
Makinen et al (2021) arXiv:2107.07405

# Fisher information at the Field.  level



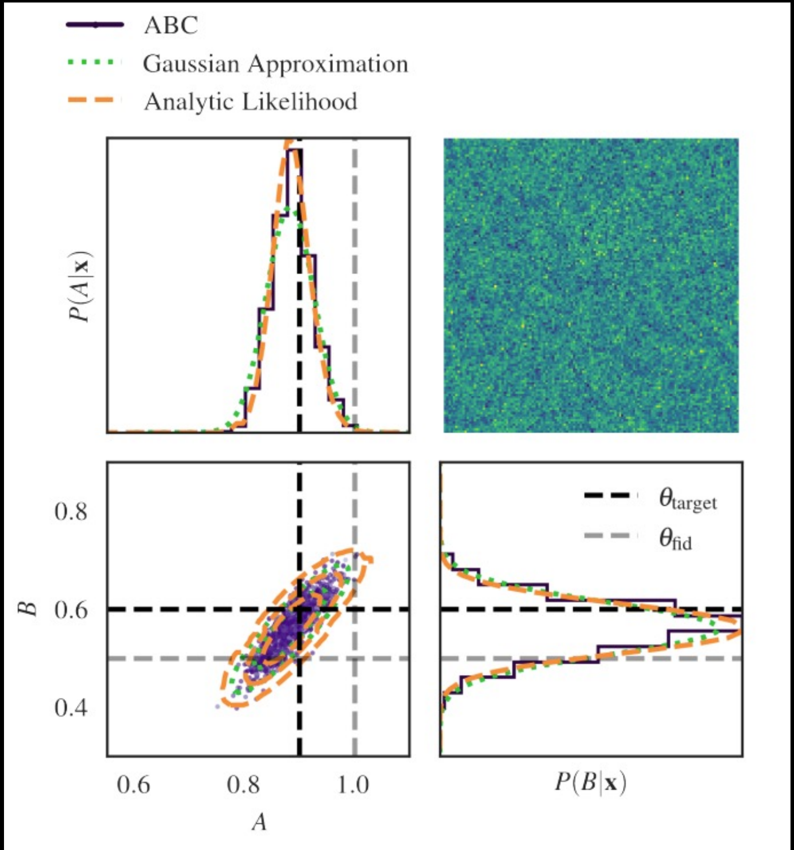
**Overdensity Field**

Field-level information is found in *fluctuations at the pixel level* (known) theoretical field information content (all pixels) !




Makinen et al (2021) arXiv:2107.07405

obtain an exact posterior with compressed simulations !

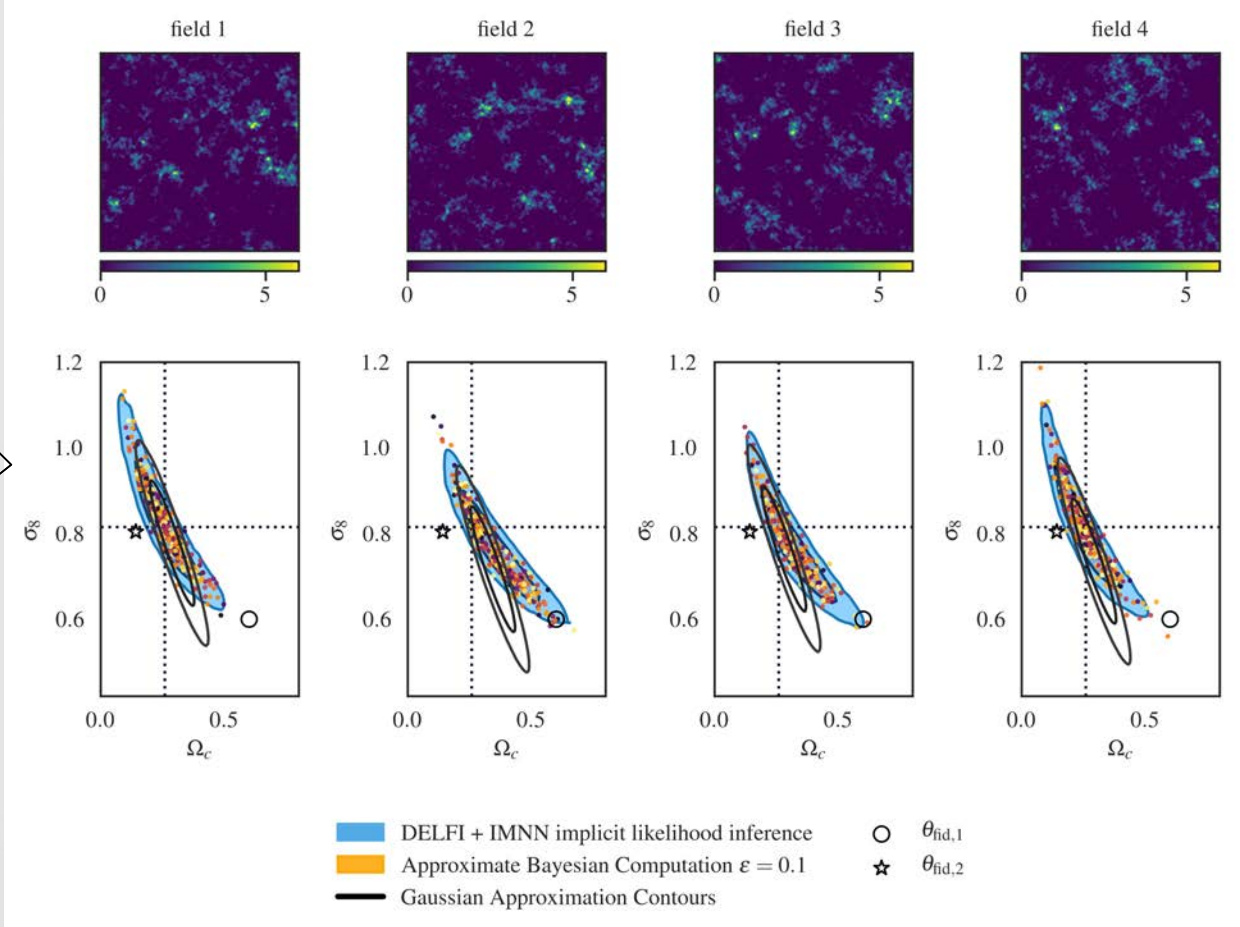



Makinen et al (2021) arXiv:2107.07405

# Final Inference

-ABC requires 12,000 simulations over prior to obtain 350 accepted points

-DELFI requires 4000 simulations sampled in batches of 1000 from posterior

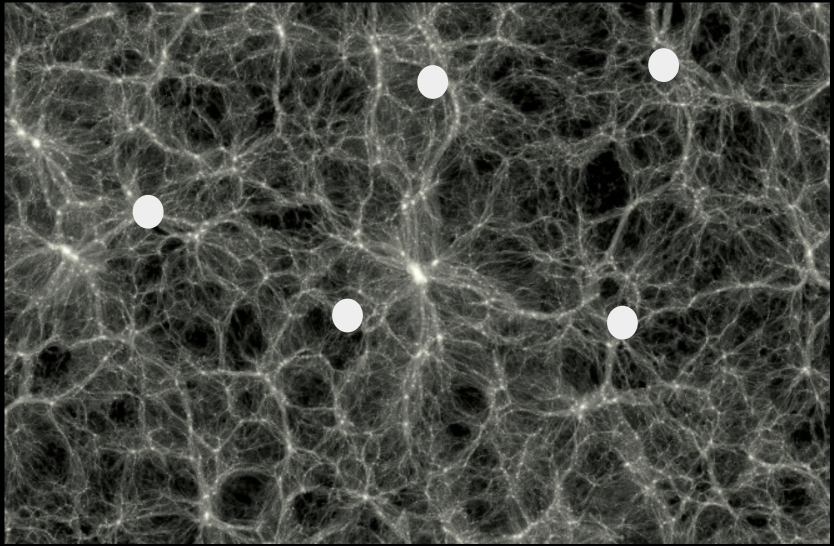


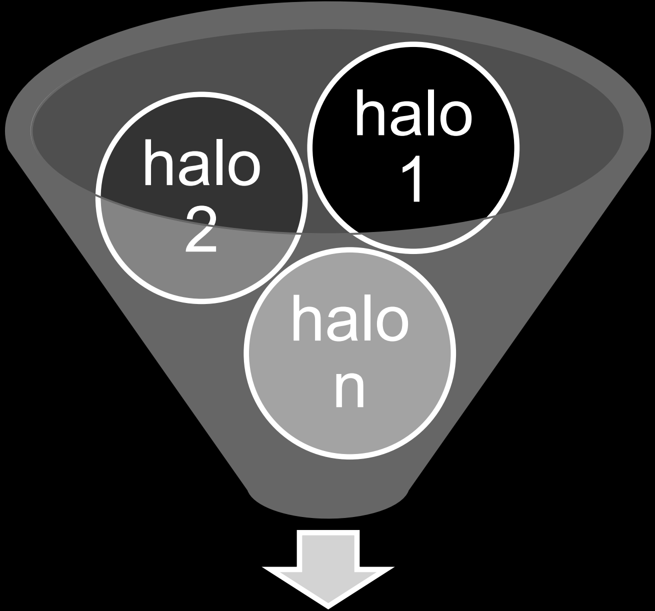

Makinen et al (2021) arXiv:2107.07405

# FAQ: WHERE IS THE INFORMATION HIDING ?

# 04

## Cosmic Graphs



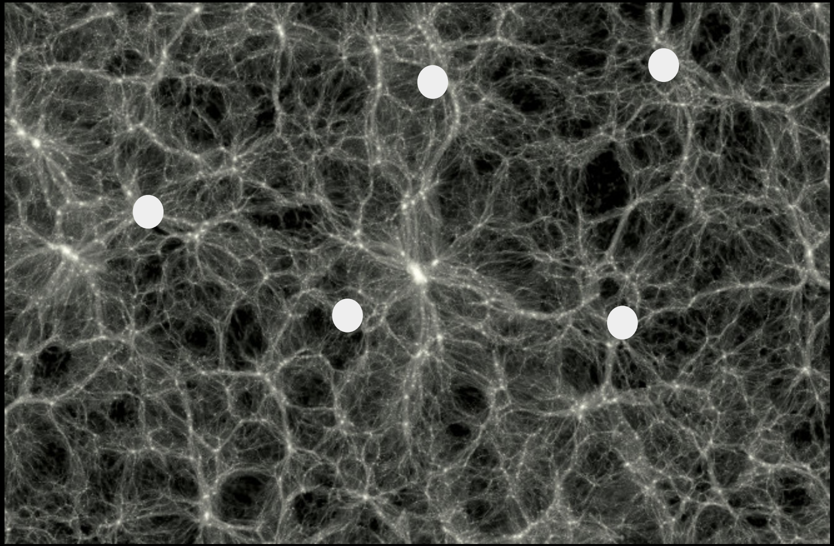


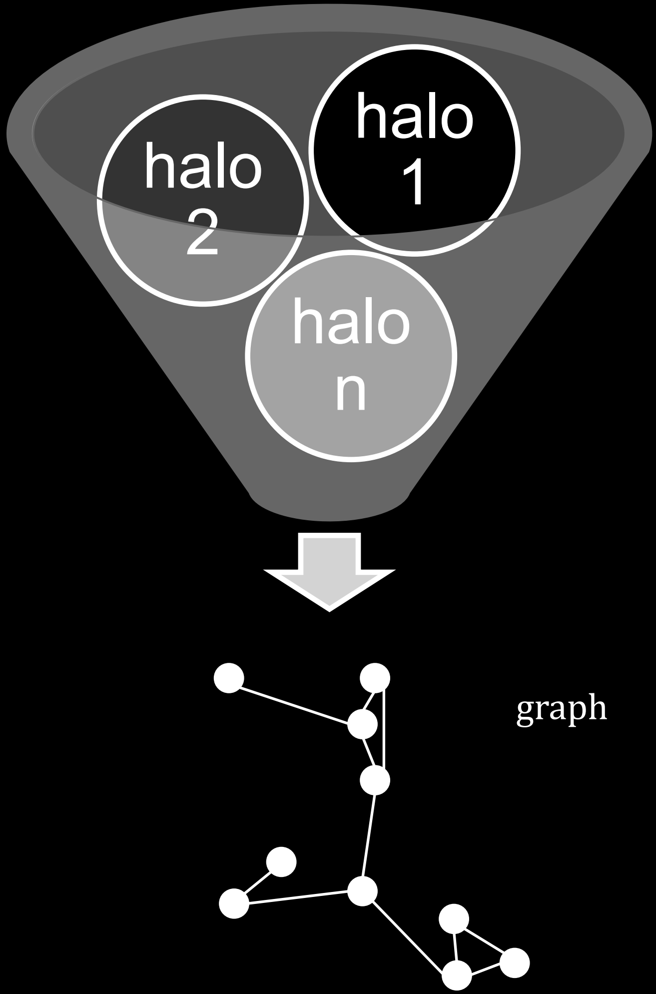

Catalogs: usually a bad idea

| LOREOM IPSUMAT CONSTRUM ETRIM KIST | Lorem ipsum dolor | Amister umarkl finish | Gatolep odio un accums | Tortores remus justica |
|---|---|---|---|---|
| LOREM DOLOR SIAMET | 8 288 | 123 % | YES | $89 |
| CONSECTER ODIO | 123 | 87 % | NO | $129 |
| GATOQUE ACCUMS | 1 005 | 12 % | NO | $99 |
| SED HAC ENIM REM | 56 | 69 % | N/A | $199 |
| REMPUS TORTOR JUST | 5 554 | 18 % | NO | $999 |
| FCELISQUE SED MORBI | 12 569 | 112 % | NO | $123 |
| SENECTUS URNA MOSTUM | 779 | 33 % | N/A | $56 |
| VESIBU LORIS SET MURTIL | 6 112 | 27 % | YES | $684 |

VectorStock® VectorStock.com/26829535

catalog





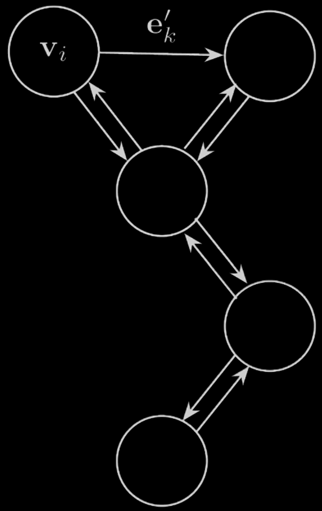

**Catalogs: usually a bad idea**

# Cosmic Graphs

1 Gpc

Makinen et al (2022) arXiv:2207.05202

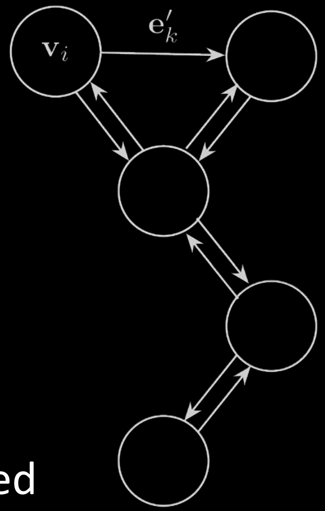
# Graphs 101




$$G = (V, E, u)$$

- A graph G is a *tuple* of nodes $V = \{\mathrm{v_i}\}$ , edges, $E = \{\mathrm{e}_k, s_k, r_k\}$, and global features $u$
- Each node and edge is a *vector*
- Edges propagate information to nodes, via senders $\boldsymbol{s_k}$ and receivers $\boldsymbol{r_k}$

# Graphs 101




Neural Networks also work on graphs !

Functions of edges and nodes can be learned with simple connected networks:

$$\mathbf{e}'_k \leftarrow \phi^e(\mathrm{e}_k, \mathbf{u})$$

$$\mathbf{v}'_i \leftarrow \phi^v(\mathrm{v}_i, \mathrm{e}'_k, \mathbf{u})$$

$$\mathbf{u}' \leftarrow \phi^u(\mathbf{v}'_\mathbf{i}, \mathbf{e}'_\mathbf{k}, \mathbf{u})$$

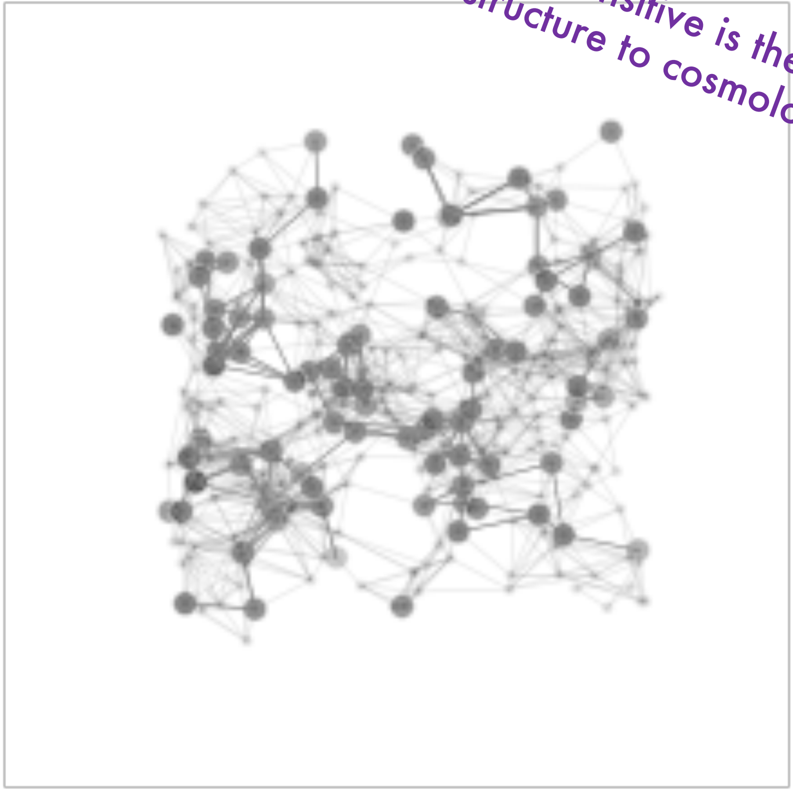
# Halo graph representation

how sensitive is the graph structure to cosmology ?

**Nodes**: masses (positions)

**Edges**: distances and angles between halos

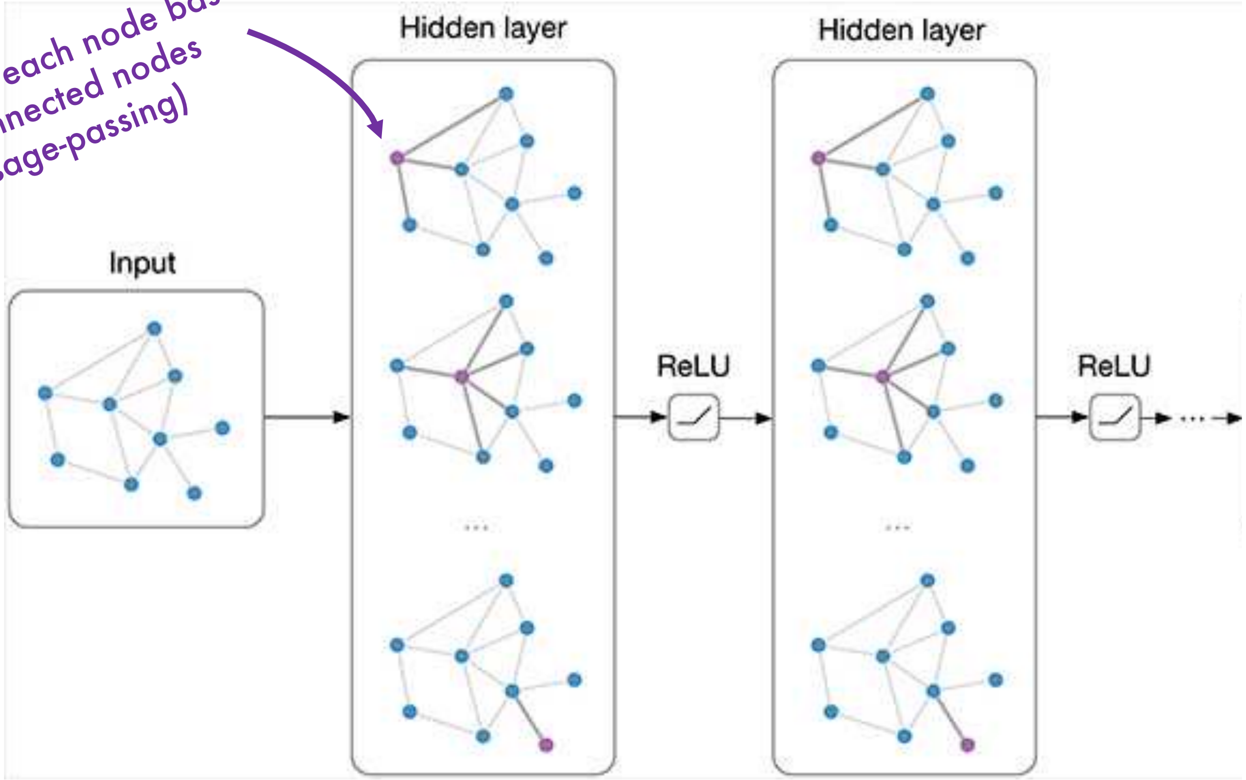
1. Take all halos with $M > 1.5{\times}10^{15} M_{\odot}$ (roughly 100 halos per simulation)
2. Connect all halos within a radius $r_{\mathrm{connect}}$



Halos assembled from *Quijote* simulation suite

# Graph Neural Networks

update each node based on connected nodes (message-passing)




**global summaries x**

**Fisher information**: tells us (on average) how informative some data $\mathbf{d}$ is about a parameter $\theta$ of a distribution, $\mathcal{L}(\mathbf{d}|\theta)$ that models $\mathbf{d}$

What if we can't differentiate through our likelihood / statistic ?

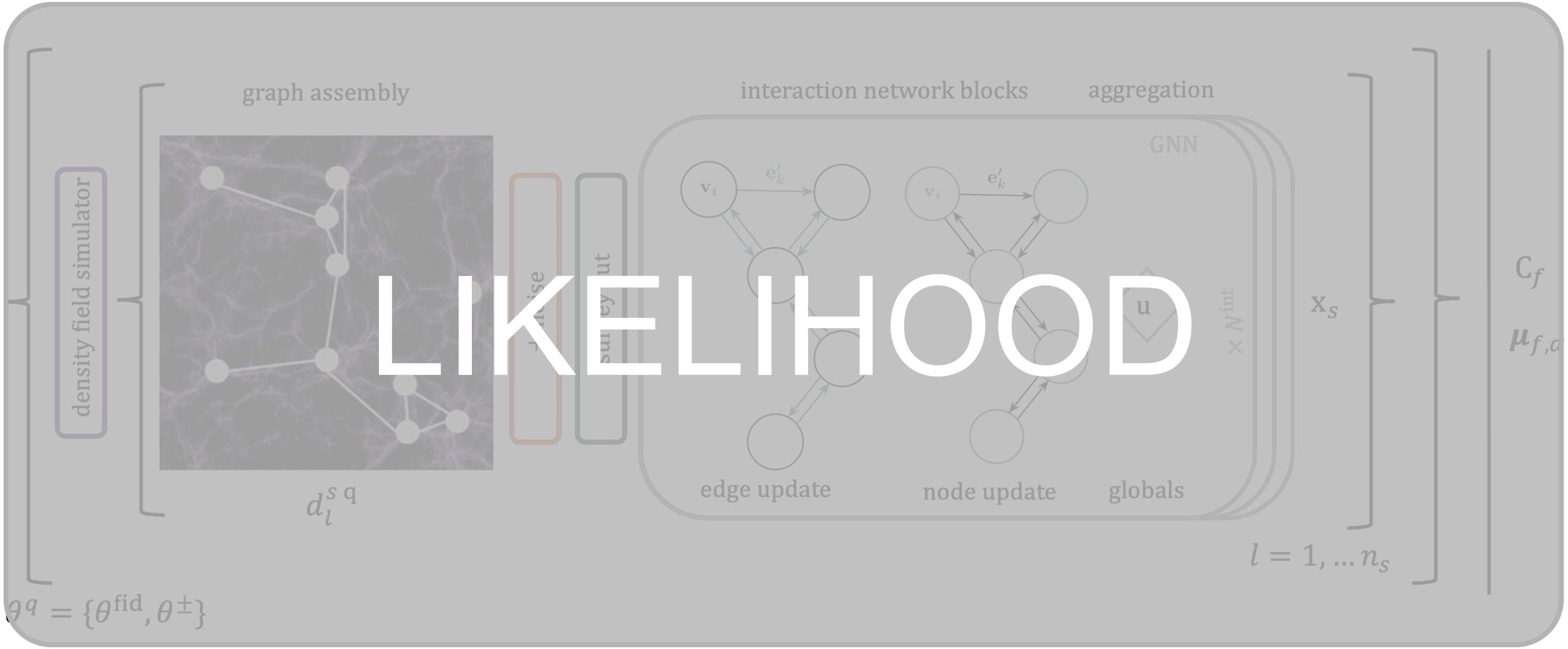
For an arbitrary statistic $Q = \mu_{\text{graph}}$:

$$F_{ij} = \frac{\partial \mu_\alpha}{\partial \theta_i} C^{-1}_{\alpha\beta} \frac{\partial \mu_\beta}{\partial \theta_j}$$

where

$$\frac{\partial \mu_\alpha}{\partial \theta_i} \approx \frac{\mu_{\text{graph}}(\theta_i^+) - \mu_{\text{graph}}(\theta_i^-)}{\theta^+ - \theta^-}$$

# Graphs can be used in the IMNN scheme !

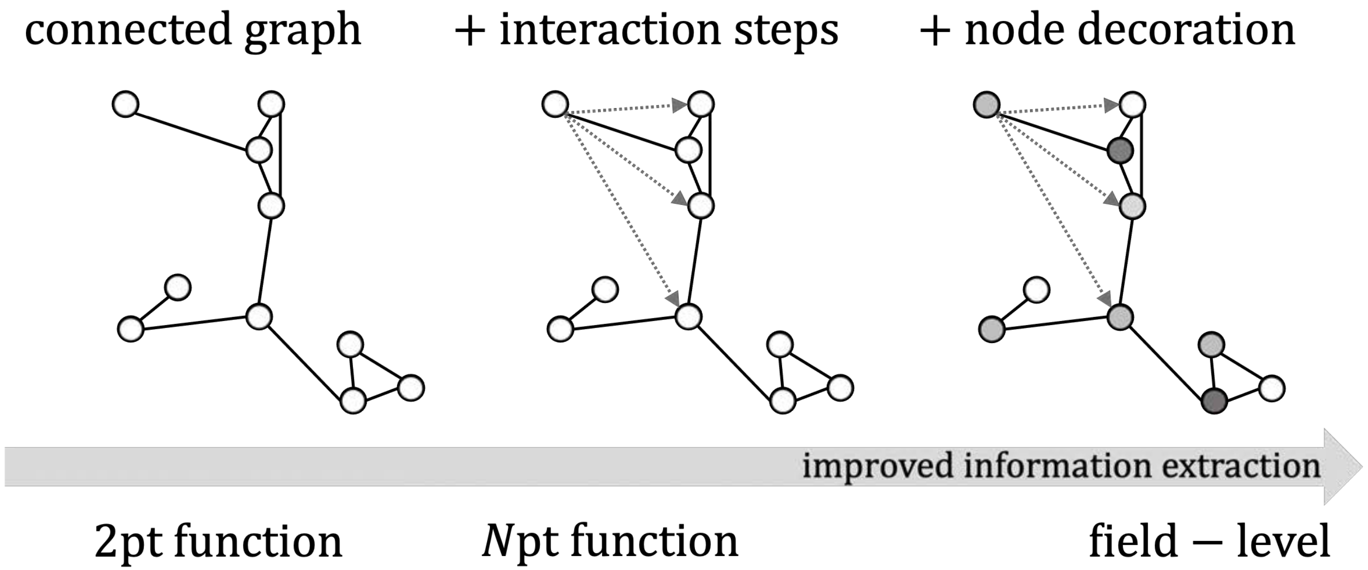



$\rightarrow F_{\alpha\beta}$

Makinen et al (2022) https://arxiv.org/abs/2207.05202

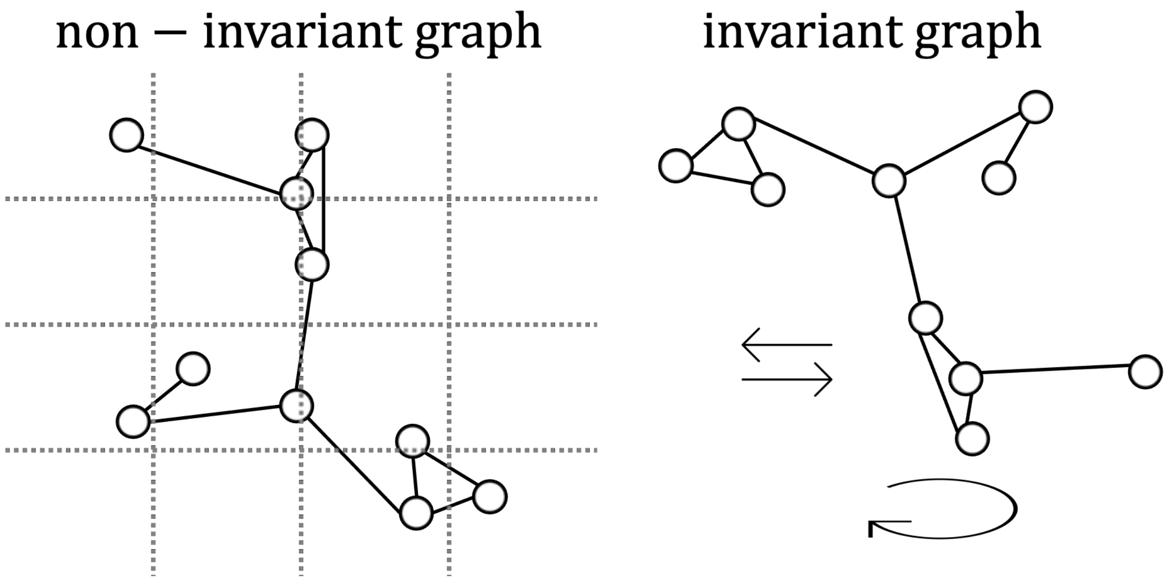
# Graphs: super modular

## Where is the information hiding ?



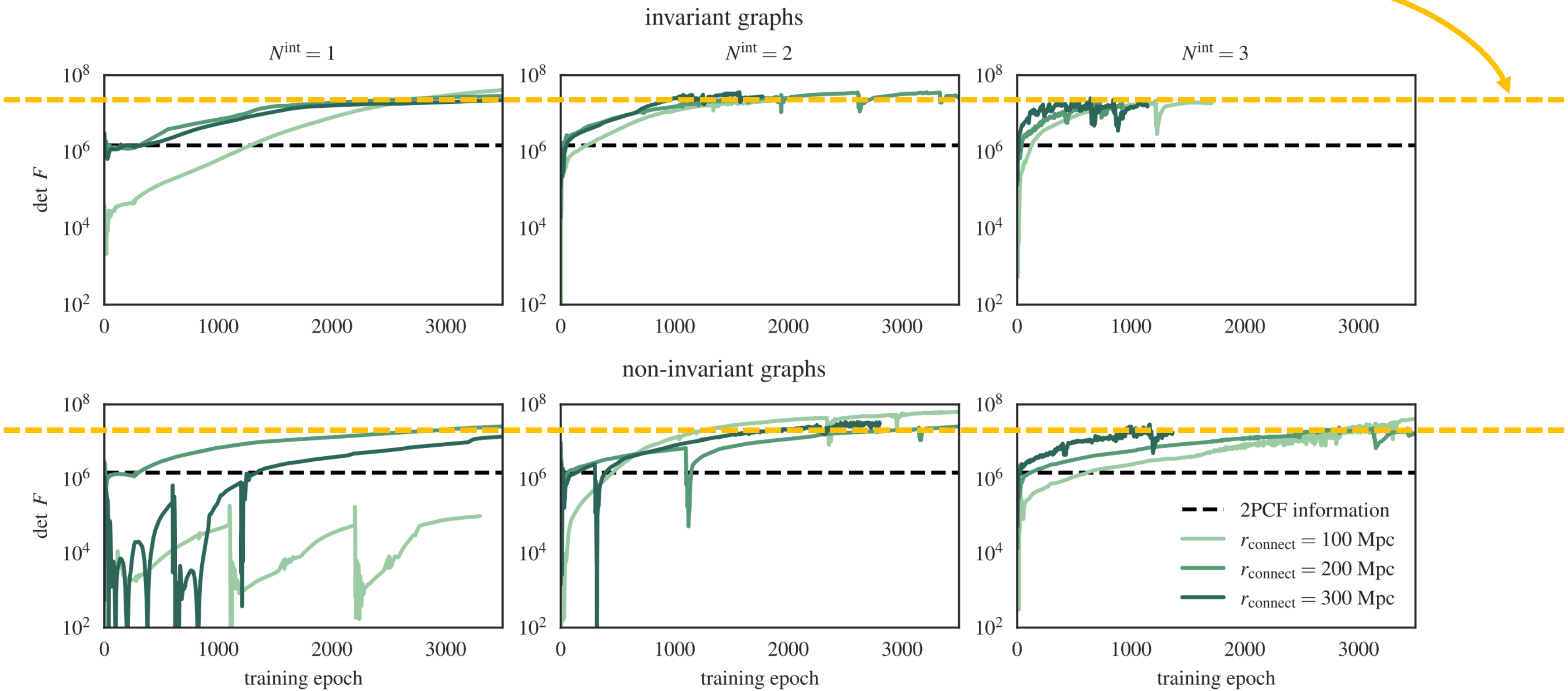

Makinen et al (2022) https://arxiv.org/abs/2207.05202

# Invariant vs non-invariant graphs



# Graphs: super modular

Information plateaus to the same level across graphs / network architectures

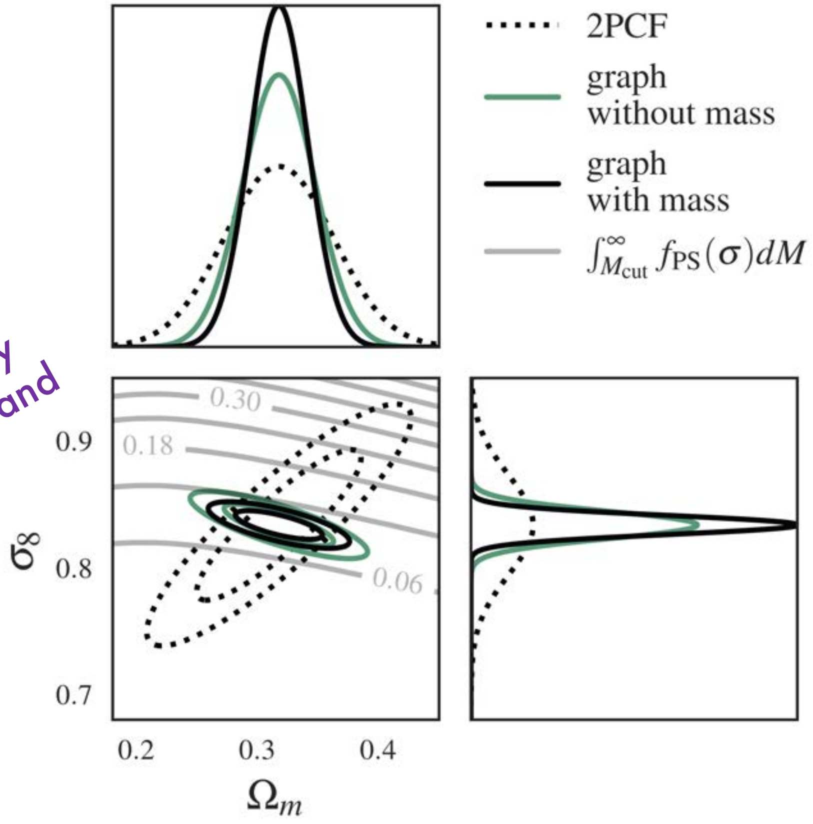



Makinen et al (2022) https://arxiv.org/abs/2207.05202

# Graphs: super modular

**Where is the information hiding ?**

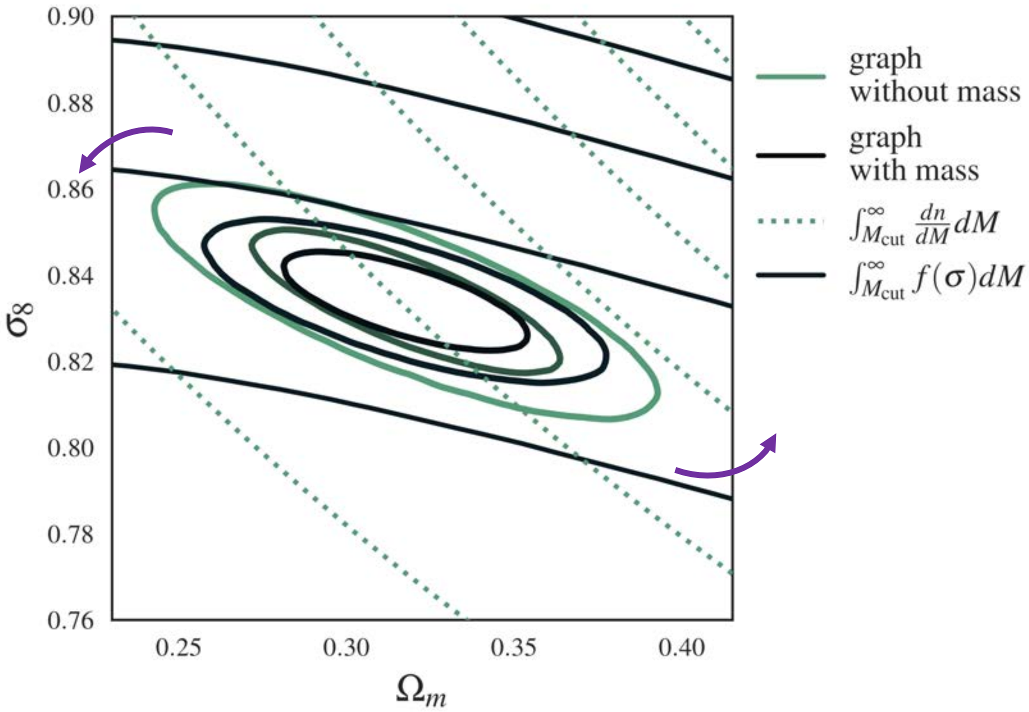
Network automatically combines clustering and mass information !




Makinen et al (2022) https://arxiv.org/abs/2207.05202

# What's being learned ?

adding mass pushes us towards the HALO MASS FUNCTION



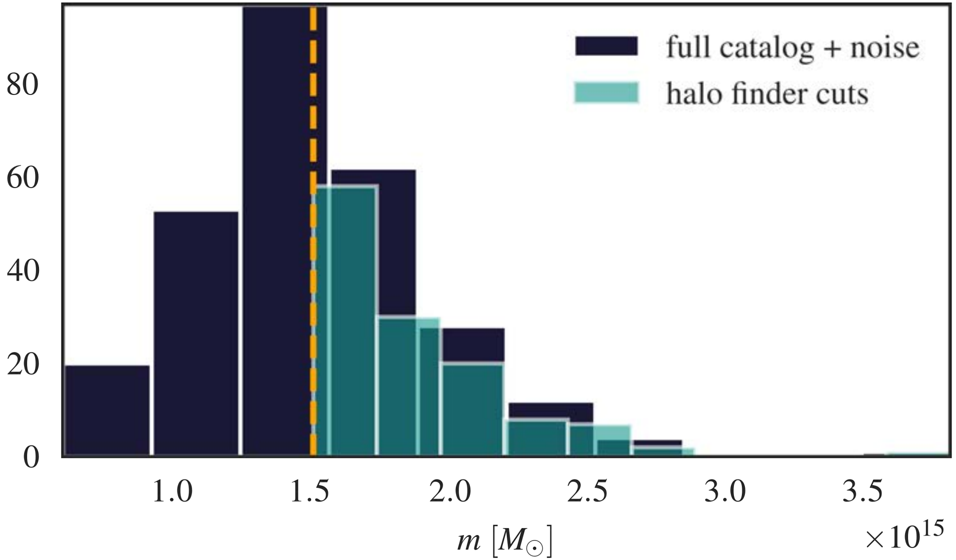
# What's being learned ?

fixing catalogue length removes cardinality feature – network can no longer learn number or mass density !

| catalogue $N^v$ | graph assembly | $\ln\det F$ | epistemic | **aleatoric** |
|---|---|---|---|---|
| fixed | without mass | | $5.03 \pm 0.47$ | $\mathbf{5.98 \pm 1.06}$ |
| | with mass | | $12.43 \pm 1.44$ | $\mathbf{12.39 \pm 0.22}$ |
| | 2PCF | 9.74 | | |
| variable | without mass | | $17.89 \pm 0.33$ | $\mathbf{17.66 \pm 0.27}$ |
| | with mass | | $17.40 \pm 0.57$ | $\mathbf{17.85 \pm 0.12}$ |
| | 2PCF | 14.19 | | |

# Adding Noise

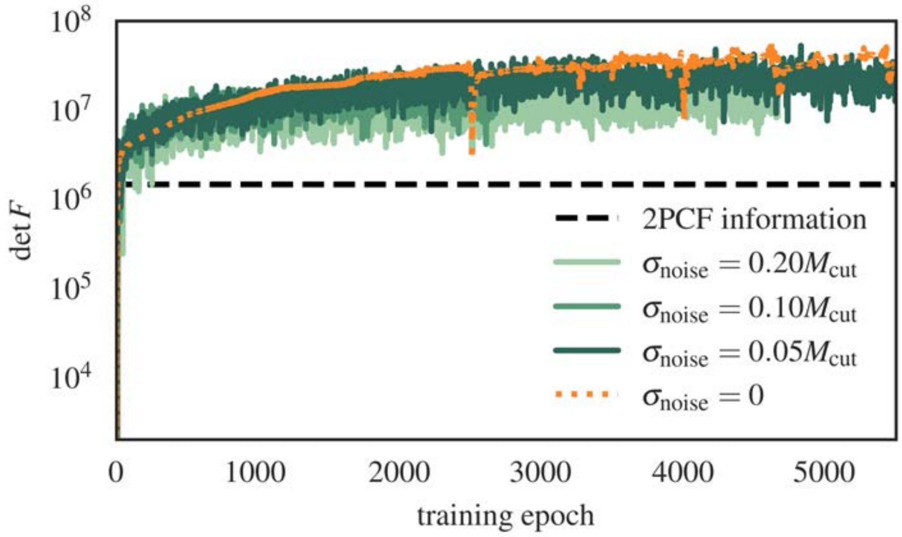
Forward-simulate noise and catalogue cuts !




Makinen et al (2022) https://arxiv.org/abs/2207.05202

# Adding Noise

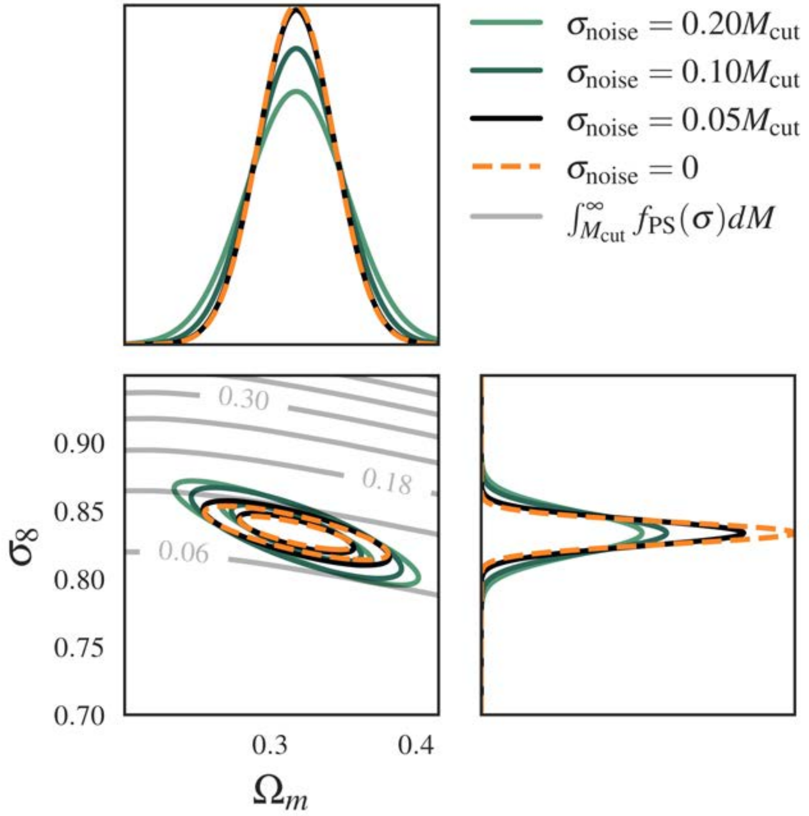
forward-simulate noise and catalogue cuts !




Makinen et al (2022) https://arxiv.org/abs/2207.05202

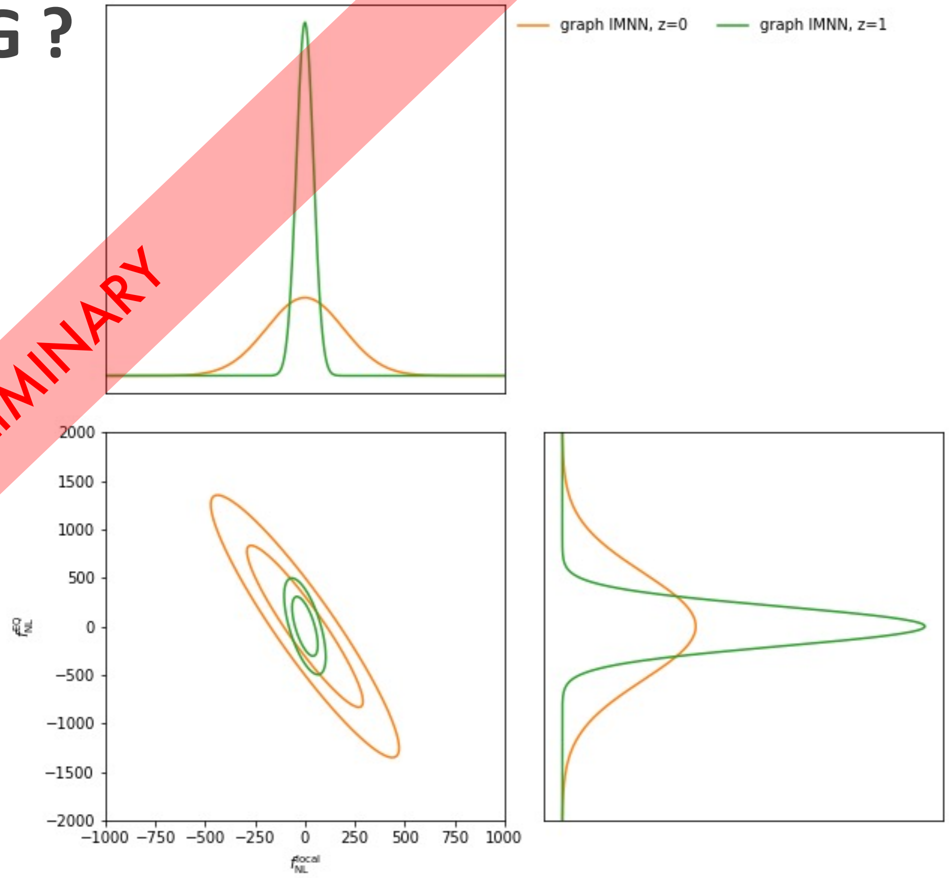
# Adding Noise

forward-simulate noise and catalogue cuts !




Makinen et al (2022) https://arxiv.org/abs/2207.05202

# are graphs sensitive to pNG ?




Makinen et al (in prep)

# TAKEAWAYS

- Cosmology is just an optimization problem !
- IMNNs can help find asymptotically lossless

  statistics automatically
- Using neural networks that respect symmetries and data structures improves compression

# Get the code !

Browser-based inference tutorial: https://bit.ly/cosmicGraphsColab

Blog: https://tlmakinen.github.io/blog/2022/09/12/cosmicgraphs

Github: https://github.com/tlmakinen/cosmicGraphs

# THANKS !

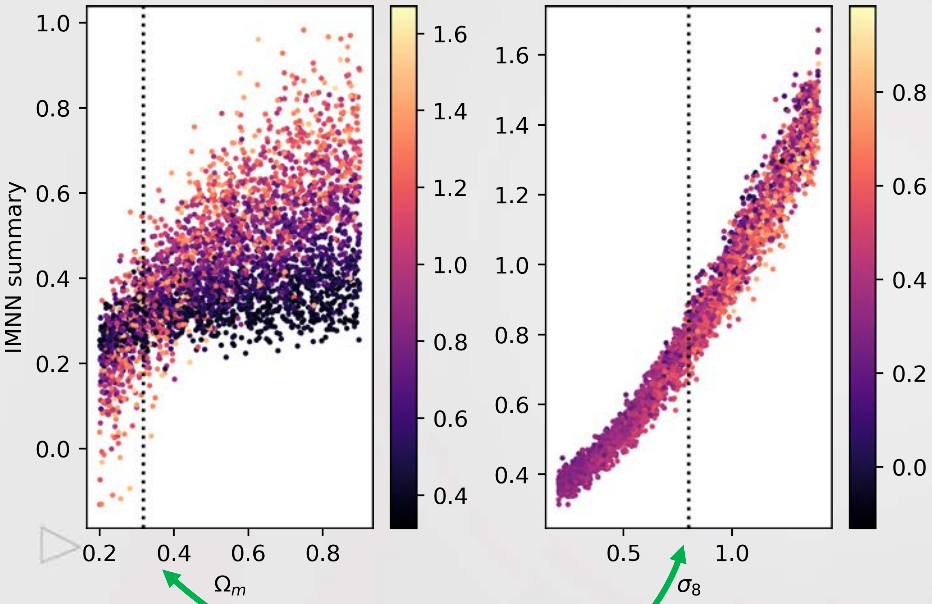
https://tlmakinen.github.io/

https://github.com/tlmakinen

@LucasMakinen

# Steps to SBI with IMNNs

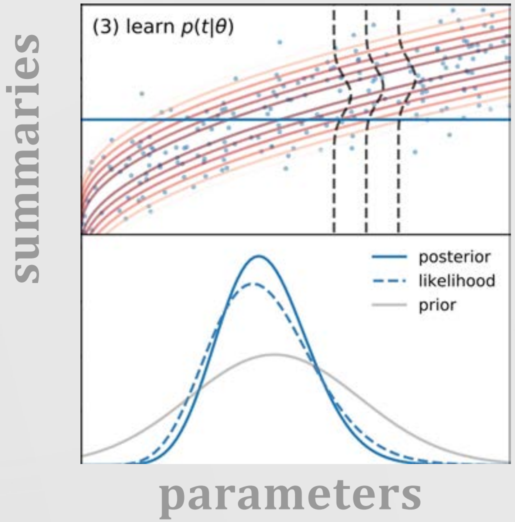
1. Train IMNN at fiducial model
2. Simulate over a prior (one that contains your Fisher)
3. Feed simulations through IMNN funnel
4. Perform density estimation in compression space




IMNN trained here

# Neural Density Estimation

**Goal**: parameterize the posterior $p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ with Conditional Masked Autoregressive Flows




Alsing et al (2018): https://arxiv.org/abs/1903.00007