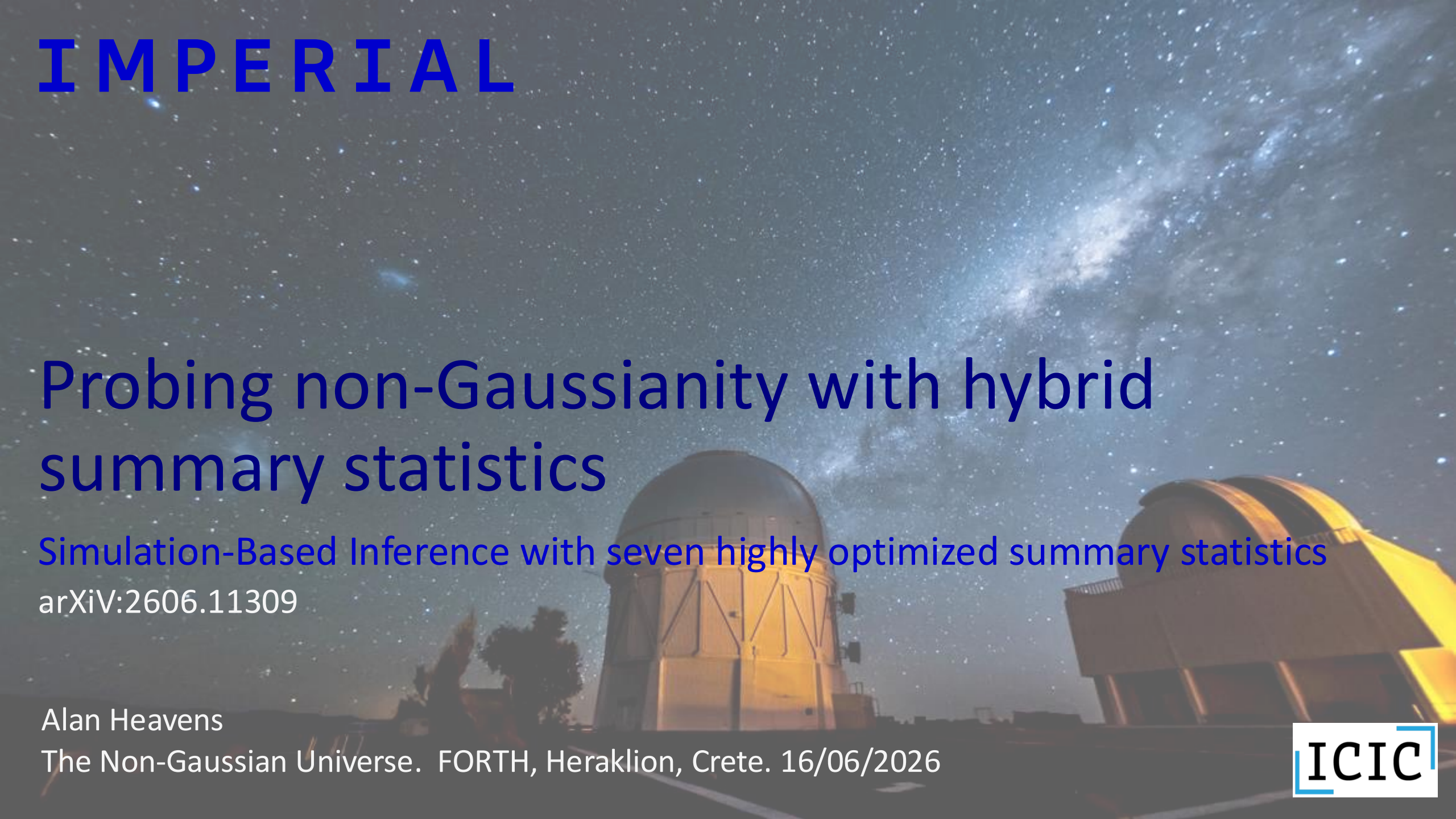


IMPERIAL

Probing non-Gaussianity with hybrid summary statistics

Simulation-Based Inference with seven highly optimized summary statistics

arXiV:2606.11309

Alan Heavens

The Non-Gaussian Universe. FORTH, Heraklion, Crete. 16/06/2026

ICIC

The Team

Imperial College:
Lucas Makinen (now DAMTP), Natalia Porqueres (now CEA), Alan Heavens

UCL:
Josh Williamson, Niall Jeffrey (now King's College London)

with important contributions from Marco Gatti, Lorne Whiteway, Ben Wandelt, Judit Prat, Ofer Lahav and others

Lucas Makinen

Natalia Porqueres.

Josh Williamson

Niall Jeffrey

Williamson & Makinen et al, arXiV:2606.11309

Outline

1. Goal: more information!
2. Simulation-based inference
3. (Conditional) Mutual Information
4. Hybrid summary statistics: 7 Physics-based + AI-based data points
5. Results from DES

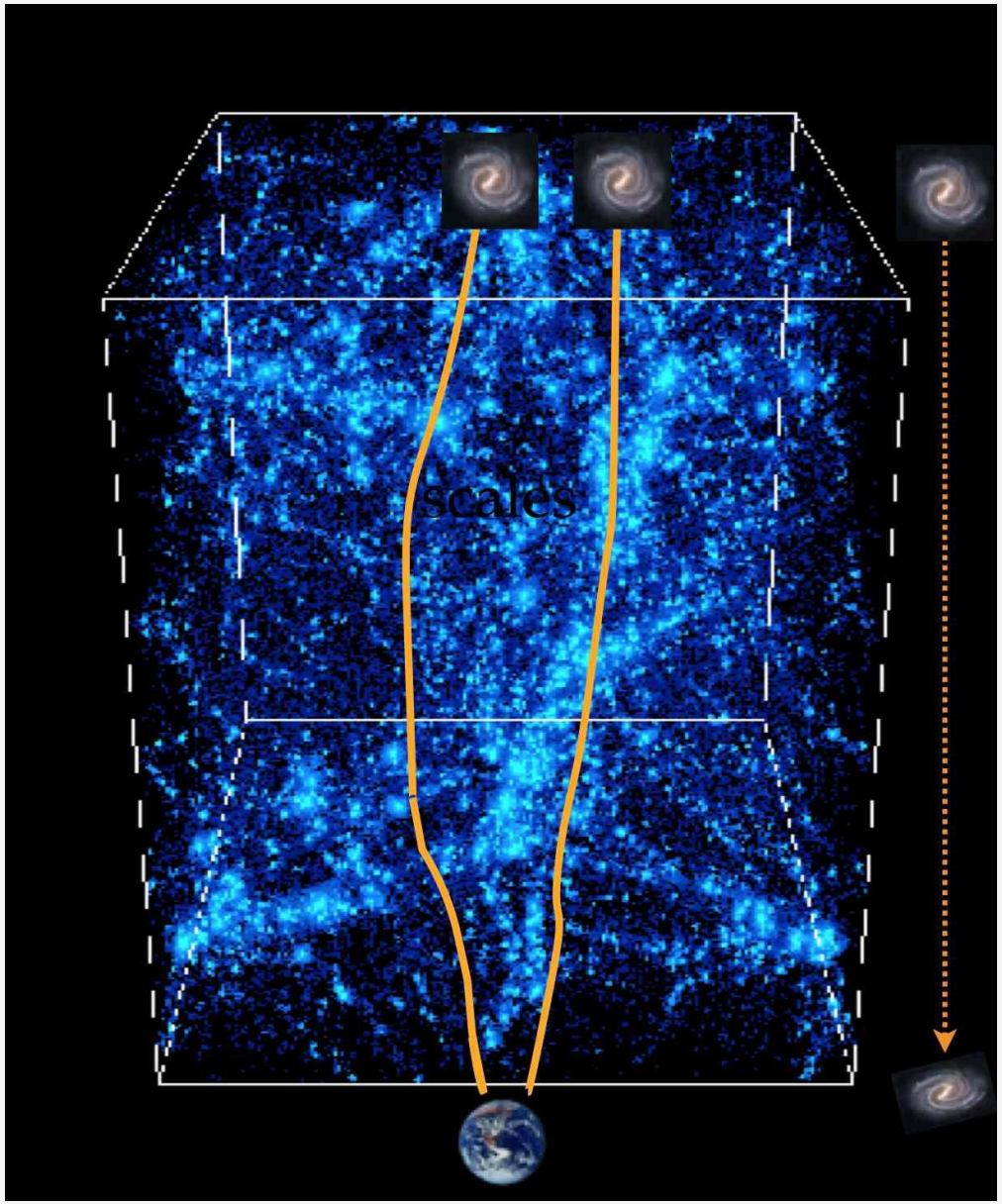
ICIC

Science goal

To extract more information than the power spectrum **in a principled way** for the cosmological parameters

$$S_8 \equiv \sigma_8 \left(\frac{\Omega_m}{0.3}\right)^{0.5}, \qquad \Omega_m\,, \qquad w = p/(\varrho\, c^2)$$

Background: Cosmic shear distorts the images of distant galaxies by ~1%, dependent on the **growth rate** of structure and sensitive to the **distance-redshift** relation, which depend principally on these parameters.



Credit: CEA-ASp

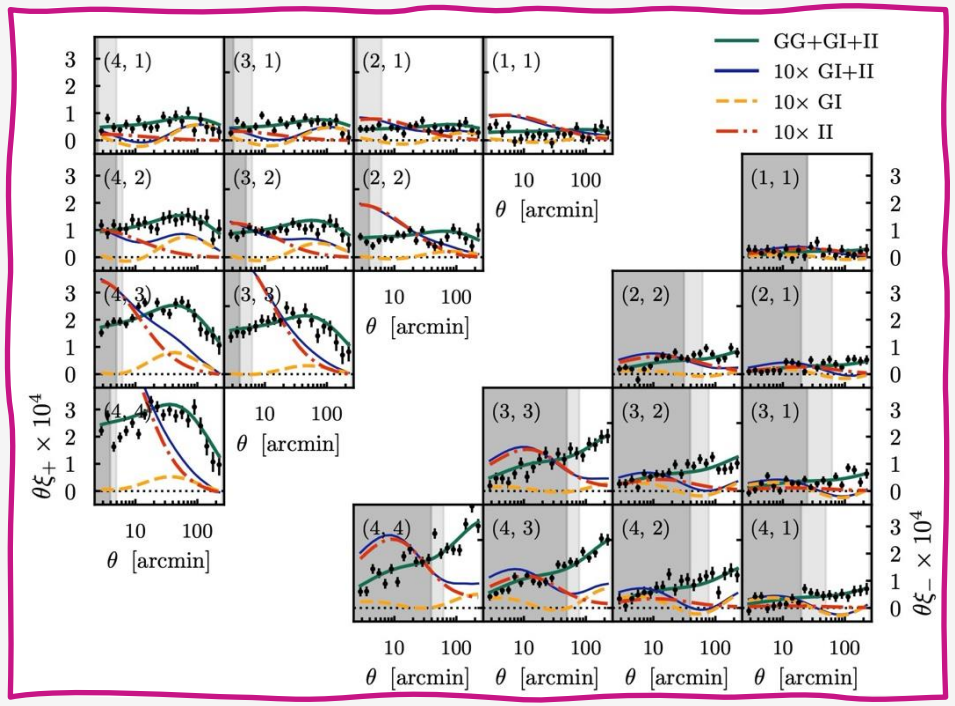
Why Simulation-Based Inference?

There is more information to extract, but how?

- SBI: simulation-based inference, AKA
- LFI: likelihood-free inference
- Implicit likelihood

Analysis challenges

- Fields are evolved to a non-Gaussian state
- Systematics and selection effects are complex
- Likelihood for any statistic may be hard or impossible to write down. Standard approaches:
  - **Approximate likelihood** (usually gaussian). Good?
  - **Challenge of covariance matrices** 4th order statistic
  - **Systematics** hard to include
  - **Information** not captured by the 2-point statistics – what else to include?

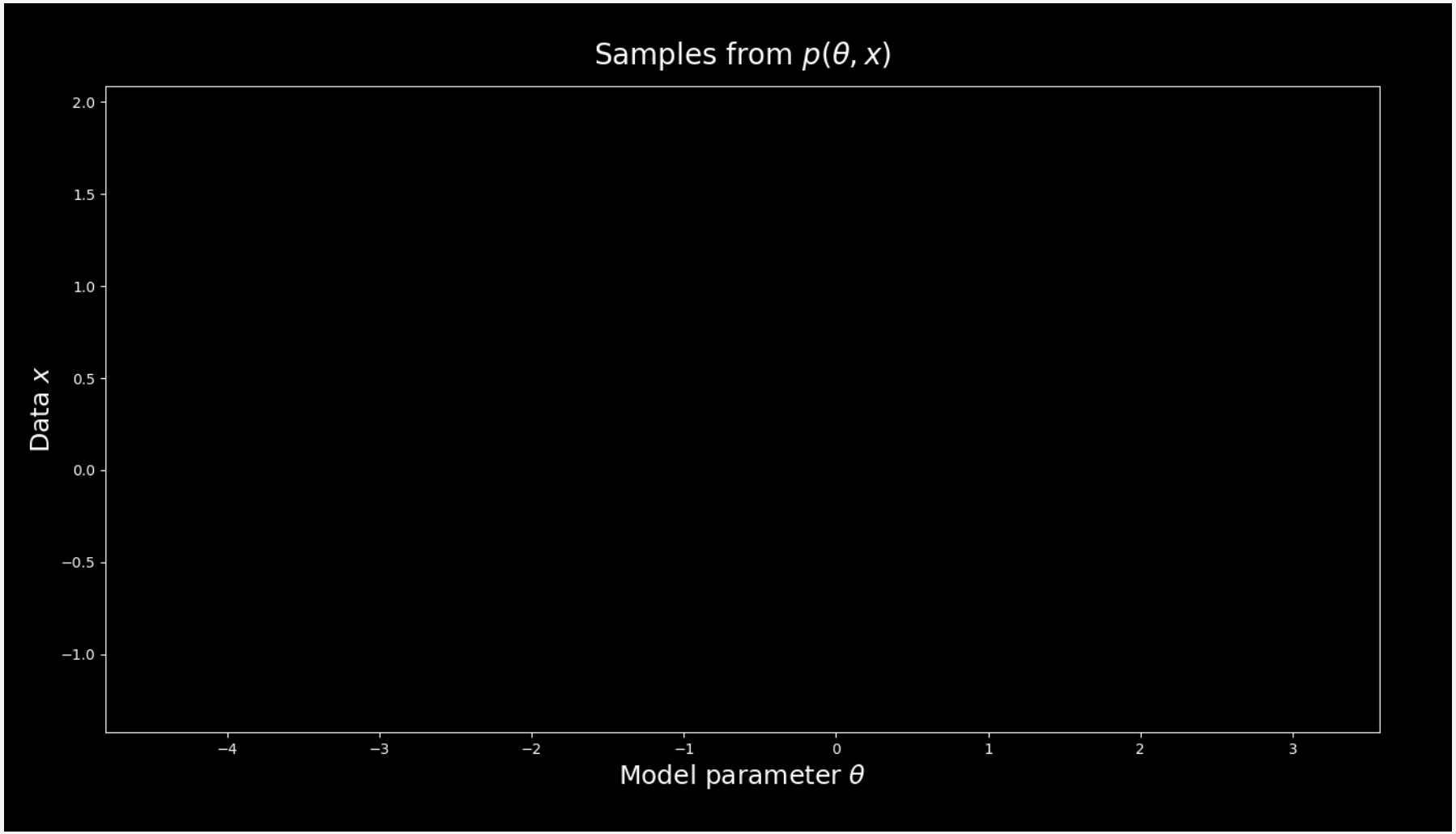


Secco et al. (2022). DES Y3

SBI (and Field-Level Inference) can handle these

- The likelihood function is not assumed (it is learned)
- Systematics and selection effects are (in principle) straightforward to include
- More cosmological information can be extracted

SBI in a nutshell



Challenges of SBI

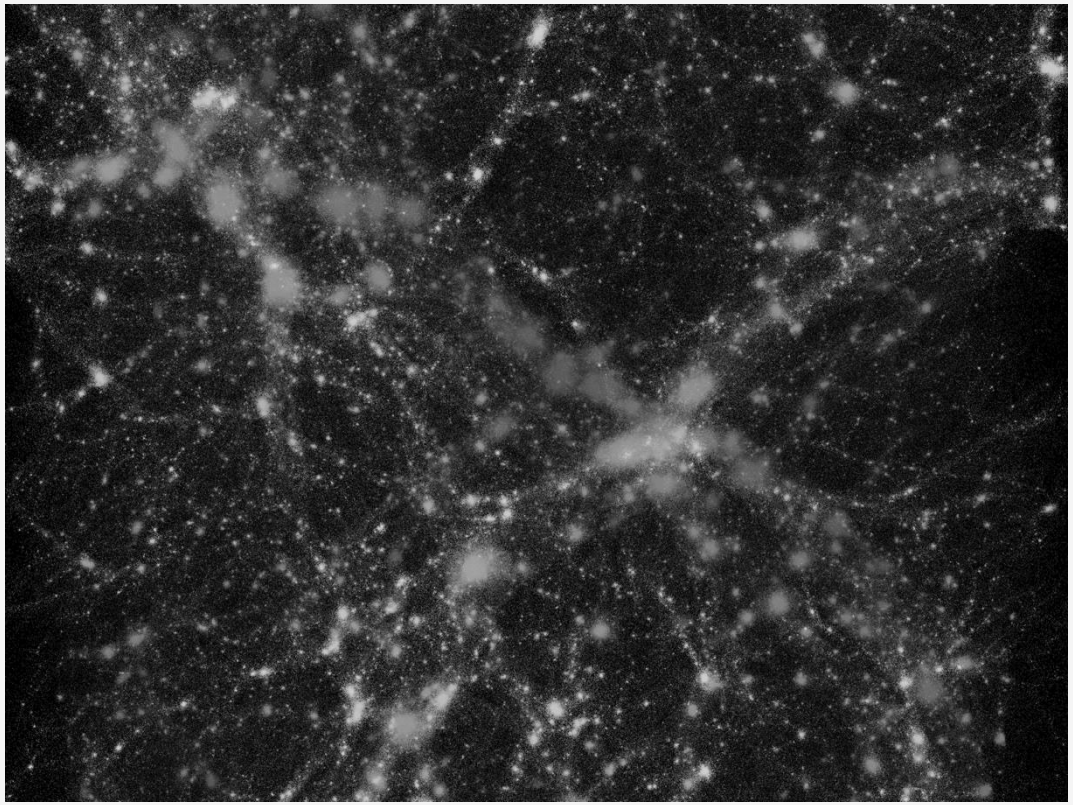
- Needs a good simulator
- Dimensionality

Concept: keep only those simulations that look like the real Universe

None do exactly

- Interpret ‘look like’ as same statistical properties
  - Which statistics?
  - The power spectra have too many points
- Needs *extreme* data compression

Key: find a few statistics that contain a lot of cosmology information and use these



Gower Street simulation

ICIC

Mutual Information

Find a handful of summary statistics that capture as much information about cosmological parameters as possible.

Not a crazy idea: analytic solutions exist for certain cases which compress data to one number per parameter without loss of Fisher information (e.g. MOPED)

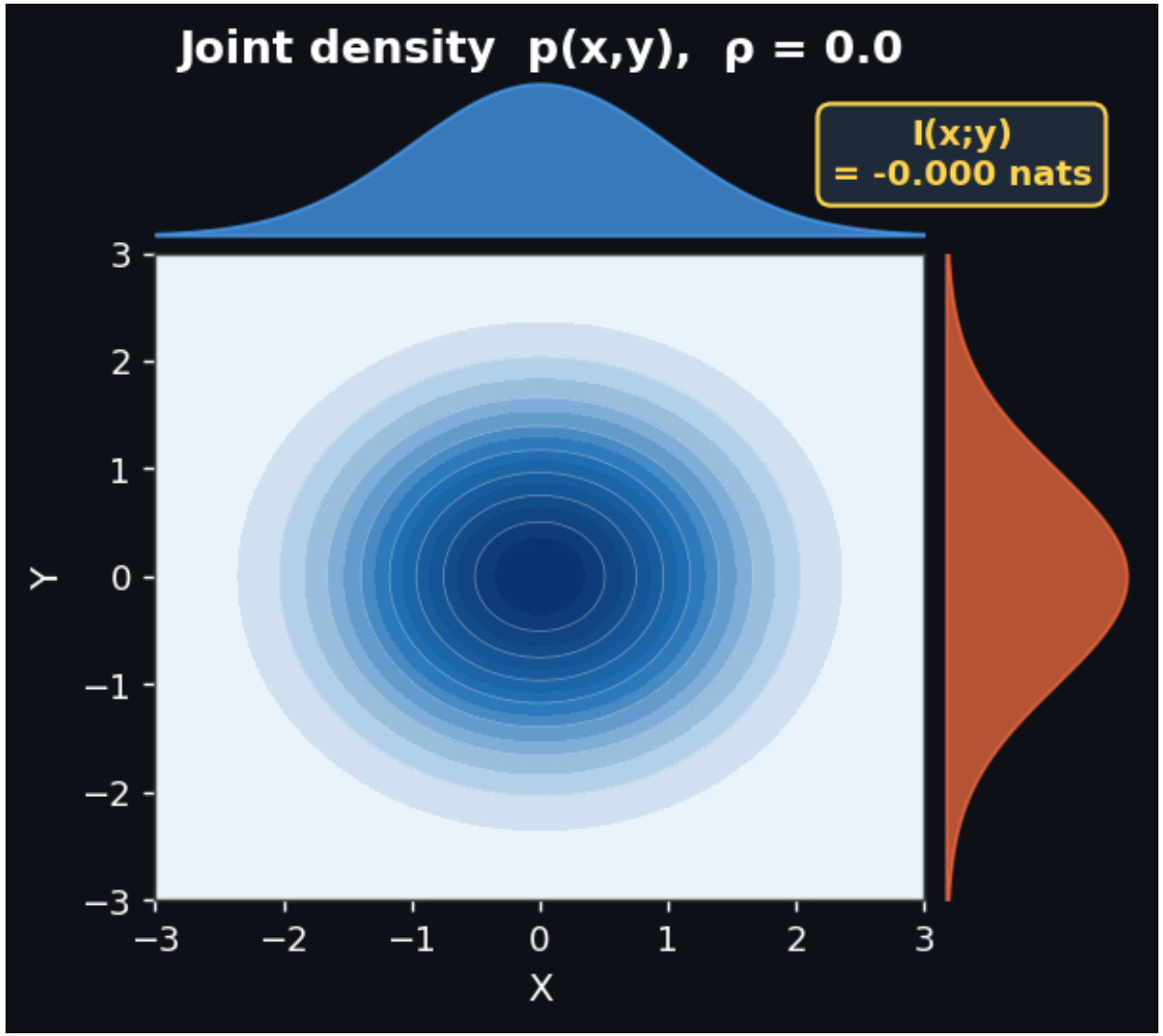
Maximise the Mutual Information:

$$\mathrm{I(X;\ Y)} = \int \mathrm{p(x,y)} \ln\left[\frac{\mathrm{p(x,y)}}{\mathrm{p(x)p(y)}}\right] \mathrm{dx\,dy}$$

Kullback-Leibler divergence:

$$\mathrm{D_{KL}}\left(p(\mathrm{x,y}) \,||\, p(\mathrm{x})p(\mathrm{y})\right) \geq 0$$

Case 1 — Independent Gaussian Variables (ρ = 0)



When ρ = 0, the joint distribution factorises completely:

$$p(x,y) = p(x) \cdot p(y)$$

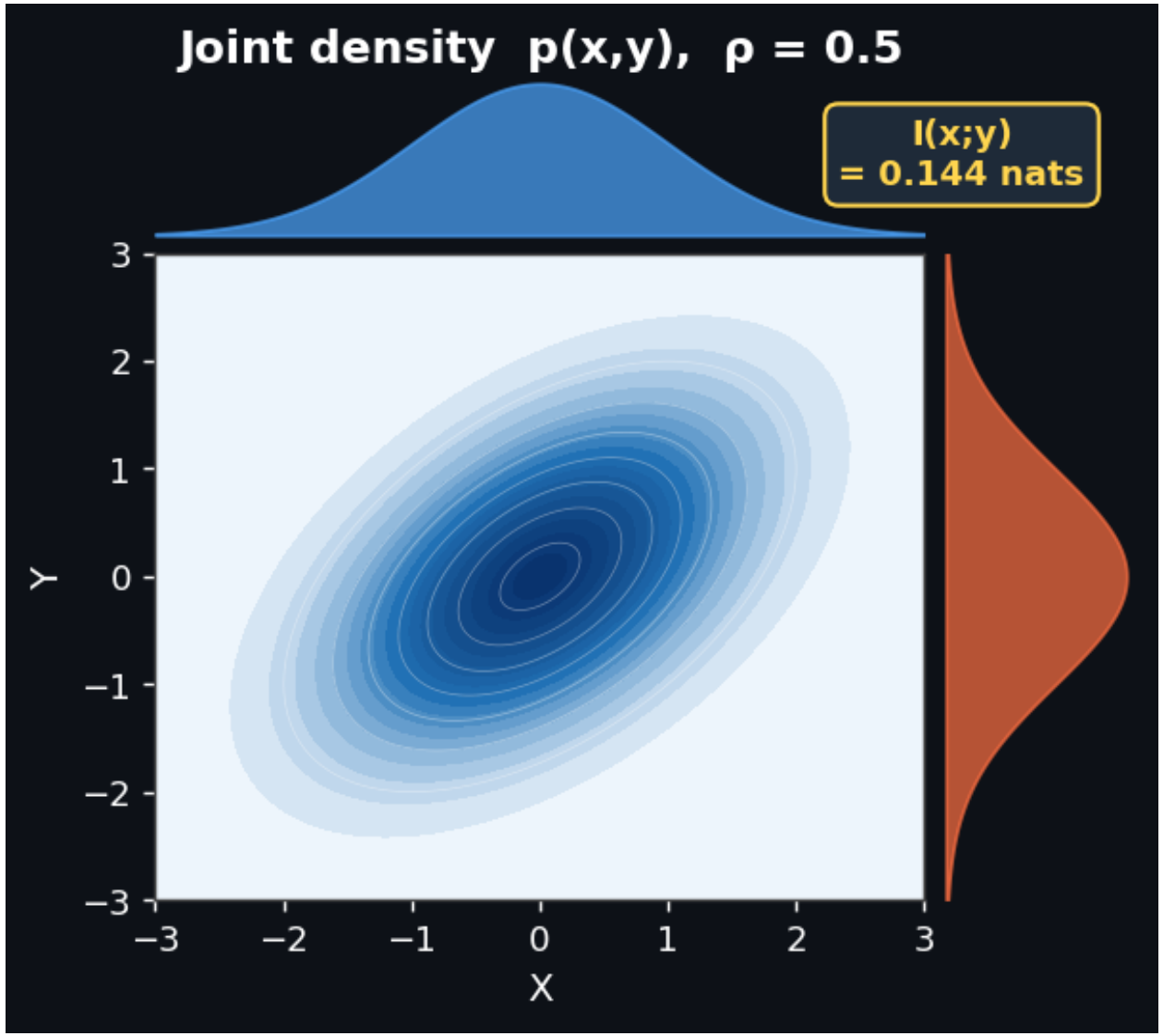
The contours are perfect circles in this case.

I(x;y) = 0 nats

Knowing Y tells you NOTHING about X.

ICIC

Case 2 — Weak Dependence (ρ = 0.5)



With ρ = 0.5, the ellipse tilts.

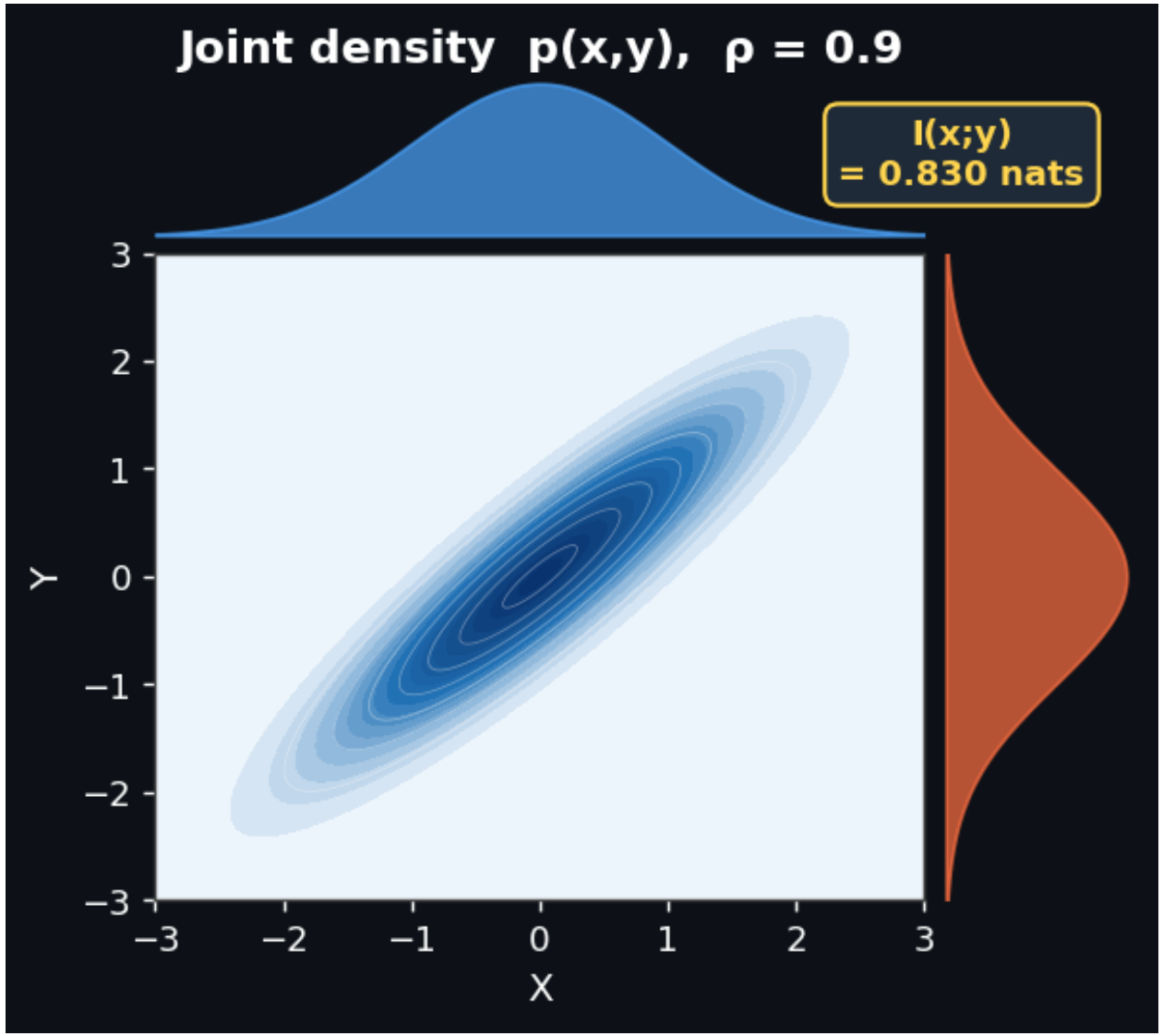
The joint distribution starts to diverge from p(x)·p(y).

I(x;y) ≈ 0.144 nats

Knowing Y gives moderate information about X.

ICIC

Case 3 — Strong Dependence (ρ = 0.9)



With ρ = 0.9, the ellipse is very elongated.

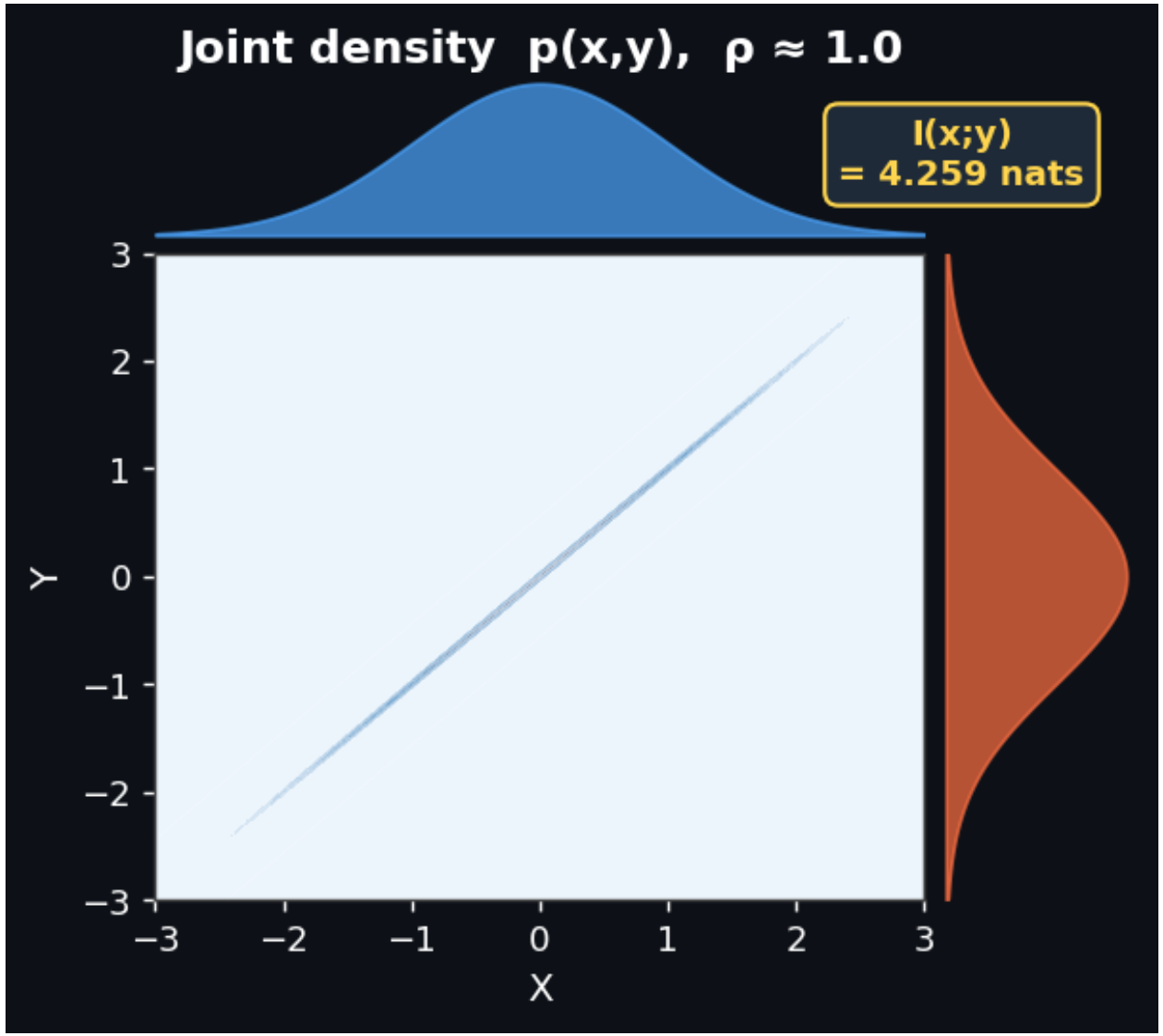
The joint distribution is far from the product of marginals.

I(x;y) ≈ 0.830 nats

Knowing Y gives a lot of information about X.

ICIC

Case 4 — Near-Perfect Dependence (ρ → 1)



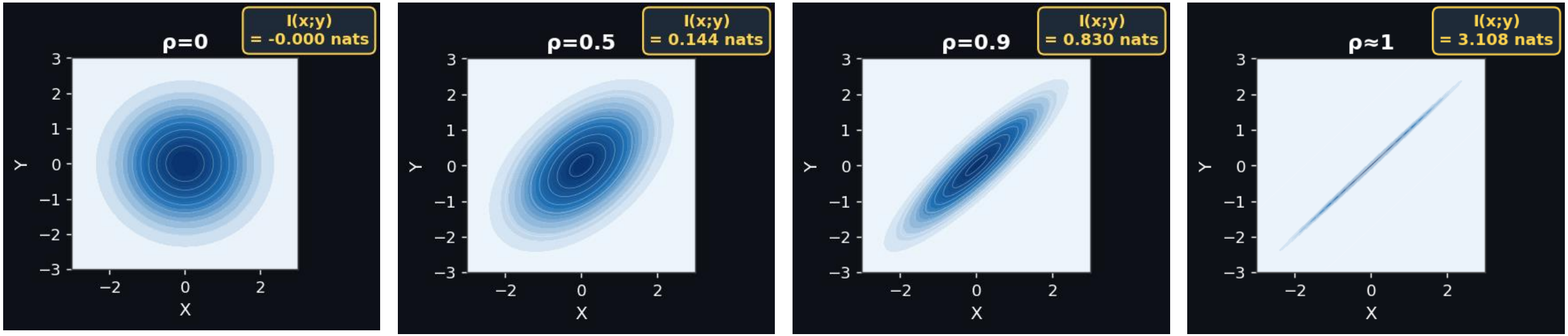
As ρ → 1, the distribution collapses onto the line Y = X.

All probability mass lies on a 1-D manifold.

I(x;y) → ∞ (≈ 4.3 nats here)

X is almost completely determined by Y.

Side-by-Side: How MI Grows with Dependence

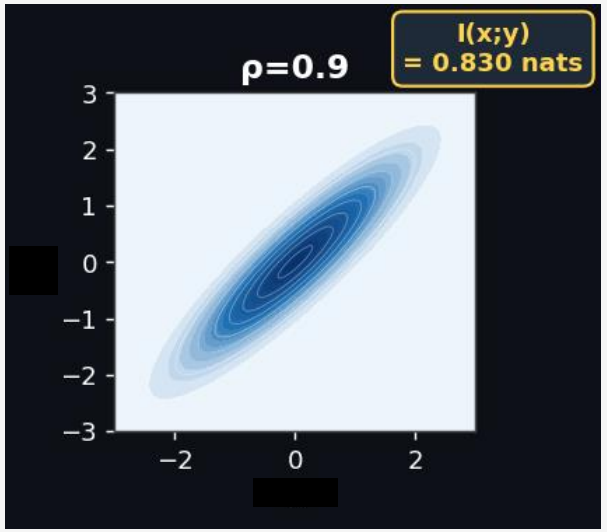


Mutual information is a highly nonlinear function of the distribution (cf correlation)

But MI is hard to estimate from samples.

Key: find summary statistics with large MI with the cosmological parameters

Basic idea:
Can we find a small set of summary statistics, $t_\psi(x)$, parametrized by ψ, that maximizes the MI with the cosmological parameters θ? i.e. maximize $I(\Theta; t_\psi(x))$



$t_\psi(x)$

θ

Can we just give the shear maps to a network? Not so easy.

Three technical slides

\begin{beach}

How can one do this? We can write the MI as

(VMIM: Jeffrey et al 2020)

$$\mathrm{I}(\Theta;\ \mathrm{T})\ \equiv\ \int \mathrm{p}(\theta,\mathrm{t})\ln\left(\frac{\mathrm{p}(\theta,\mathrm{t})}{\mathrm{p}(\theta)\mathrm{p}(\mathrm{t})}\right)\mathrm{d}\theta\ \mathrm{dt}\ =\ \mathrm{H}(\Theta)-\mathrm{H}(\Theta|\mathrm{T})$$

where the *differential entropy* and *conditional entropy* are

$$\mathrm{H}(\Theta)\ =\ -\int \mathrm{p}(\theta)\ln\mathrm{p}(\theta)\,\mathrm{d}\theta \text{ and } \mathrm{H}(\Theta|\mathrm{T})\ =\ -\int \mathrm{p}(\theta,\mathrm{t})\ln\mathrm{p}(\theta|\mathrm{t})\,\mathrm{d}\theta\ \mathrm{dt}.$$

The first term does not depend on our summary statistics t(x) so we can ignore it, and minimize $H(\Theta|T)$.

Vary t with a compression network: Data $\mathrm{x}\ \rightarrow\ \mathrm{t}_{\psi}(\mathrm{x})$ with network parameters ψ

ICIC

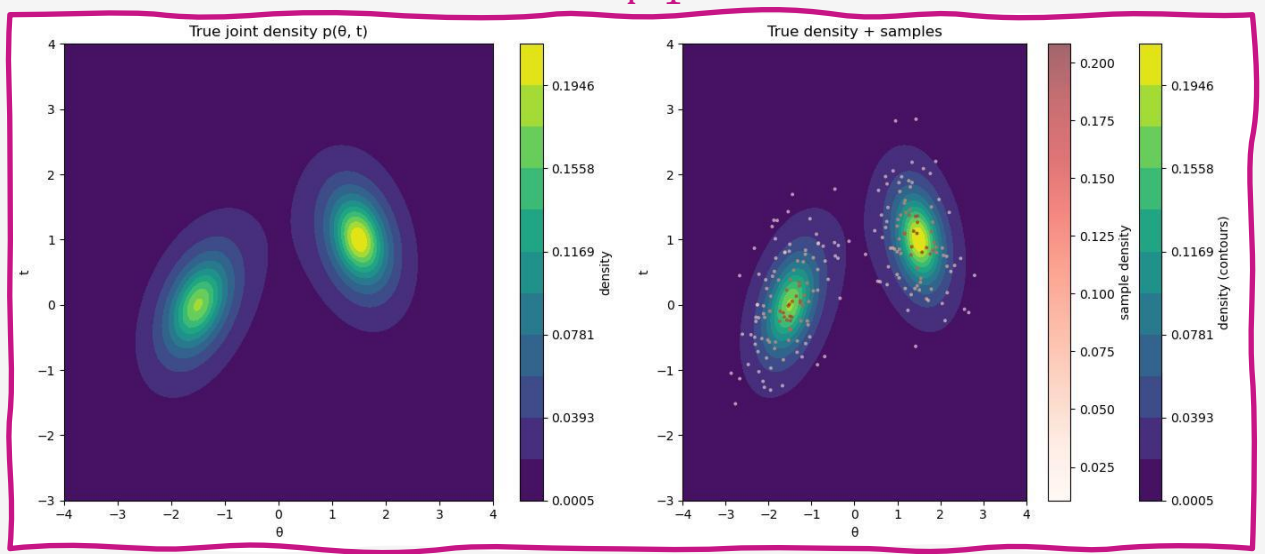
Compression network $\mathrm{t}_\psi(\mathrm{x})$ and NDE network

We approximate the probability distribution of the samples with a NDE, with network parameters φ: $\mathrm{p}(\theta|\mathrm{t}) \approx \mathrm{q}_\varphi(\theta|\mathrm{t})$ and we can write

$$\mathrm{H}(\Theta|\mathrm{T}) = -\mathbb{E}[\log \mathrm{q}_\varphi(\theta|\mathrm{t}_\psi)] - \mathbb{E}[\mathrm{D}_{\mathrm{KL}}(\mathrm{p}||\mathrm{q}_\varphi)]$$

$\mathrm{D}_{\mathrm{KL}} \geq 0$, we minimize the upper bound of H, or the lower (Barber-Agakov) bound of the MI. From the N samples of $\mathrm{p}(\theta, \mathrm{t})$ we have

$$\mathrm{H}(\theta|\mathrm{t}_\psi) \leq -\sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathrm{q}_\varphi(\theta_\mathrm{i}|\mathrm{t}_\psi(\mathrm{x}_\mathrm{i}))$$



Maximise w.r.t. φ and ψ simultaneously. q uses a relatively simple network, since optimizing for **t** is the hard part

Hybrid SBI (new): using physics knowledge to assist and make scaleable

(Makinen et al 2024)

Use both Physics-based + AI-based summary statistics

- **On large scales, power spectrum is highly informative.** Compress it to t_0
- **On small scales, there is more (nonlinear) information**

Use a neural network (CNN) to find summary statistics which contain *extra* information

Maximise the *Conditional* Mutual Information

$$I(\Theta;\ T_1|T_0) = \int p(\theta, t_1, t_0) \ln \left[\frac{p(\theta, t_1|t_0)}{p(\theta|t_0)\ p(t_1|\ t_0)}\right] d\theta\ dt_1\ dt_0$$

i.e. find extra information, given what we know already from the power spectrum

We can repeat this *sequentially*:

$I(\theta;\ T_k \ldots T_0) = I(\theta;\ T_0) + I(\theta;\ T_1|T_0) + \ldots + I(\theta;\ T_k|T_{k-1} \ldots T_0)$. \end{beach}

ICIC

Hybrid summaries

Starting point:

Power Spectrum (information from all scales) → 10 summaries

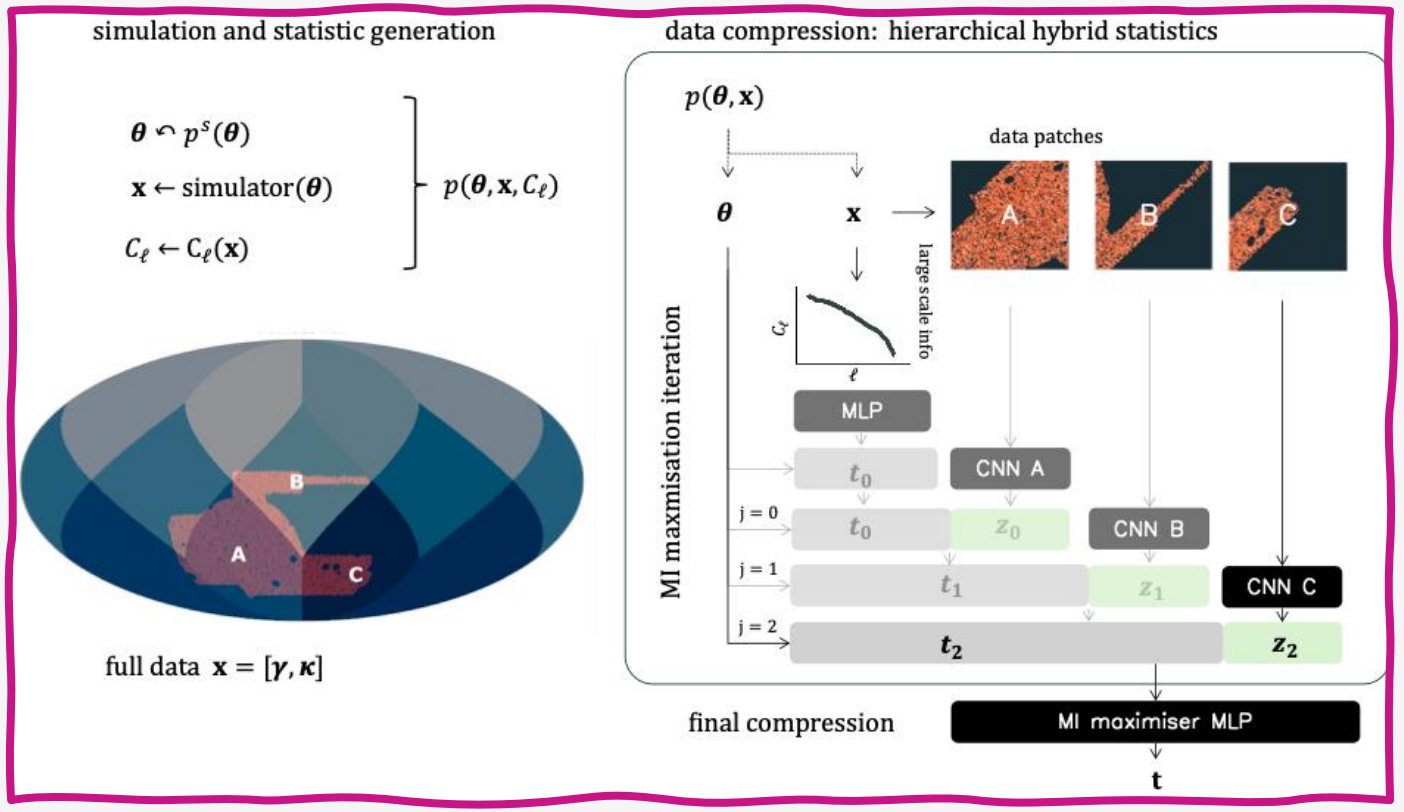
Compress CNN summaries from Patch A (small-scale information)

(Survey is too big to ingest at once)

Maximize the **conditional mutual information** → 4 extra summaries

Add Patch B and repeat

Final compression to 7 data points



Small technical point: we penalize large excursions in **t** to make it easier for the NDEs to characterize the distribution

Now fix the compression

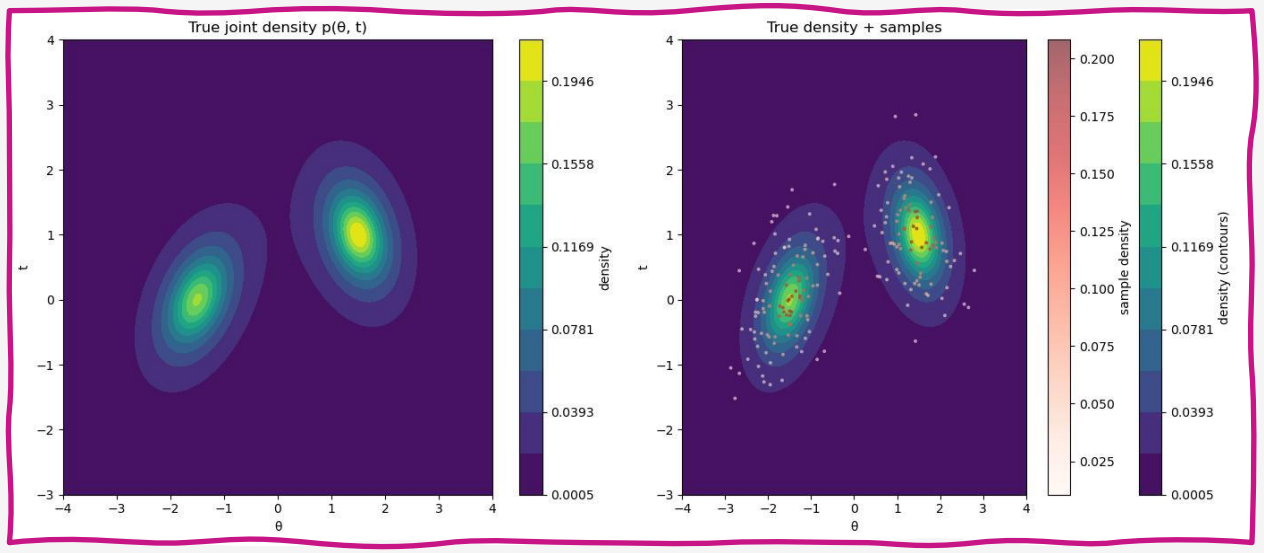
Forget about what has gone before; we have now 7 statistics (=functions of the data)

SBI just takes the 7 numbers as the data, and finds the 7 numbers for the simulated datasets

Now we do a more sophisticated job at estimating the joint distribution of t and θ

We use NLE, with 8 masked autoregressive flows (MAFs) to approximate $\mathrm{p(t|}\theta)$ $\boldsymbol{t_\psi(x)}$

It doesn't matter now how these functions were obtained



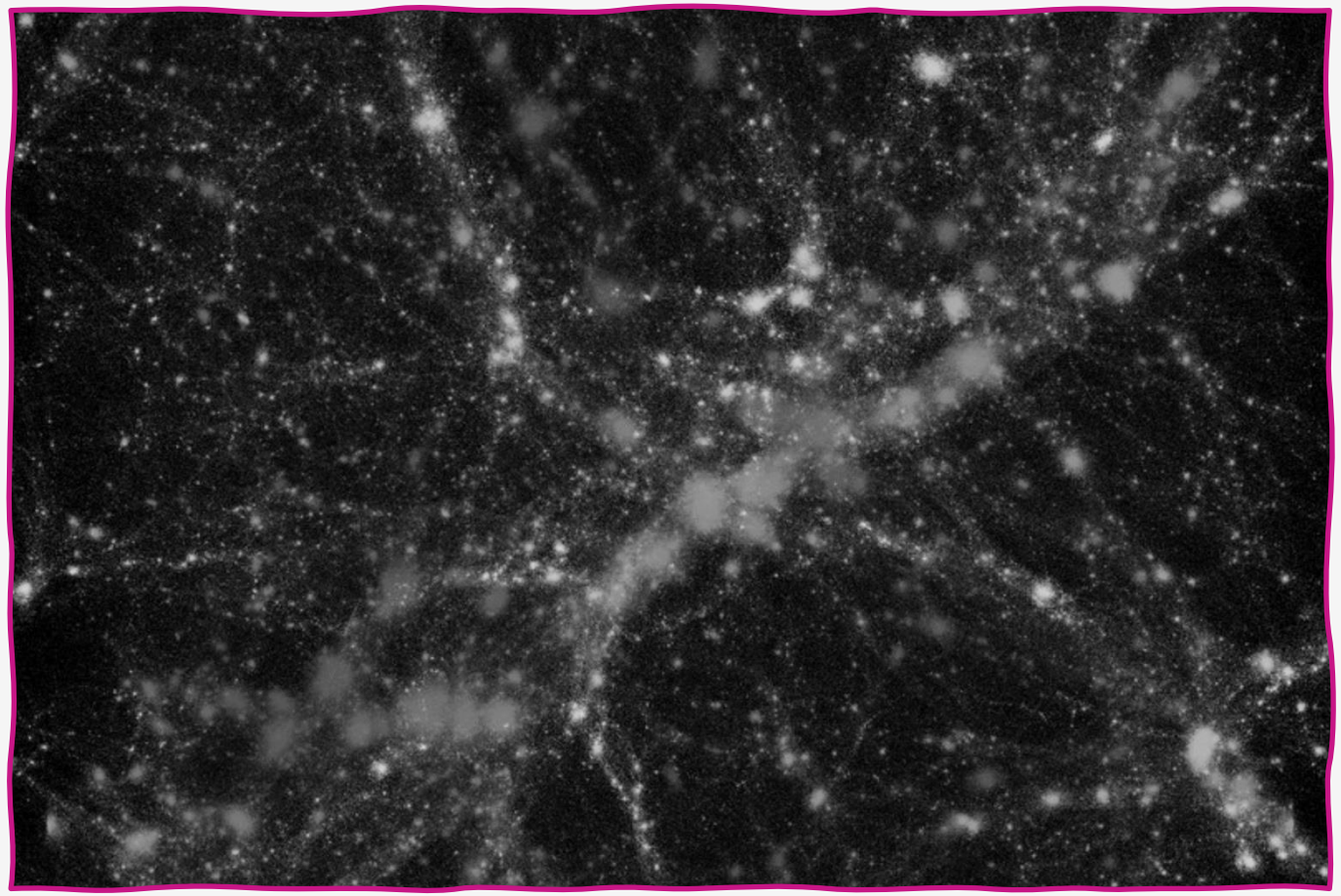
θ θ

Gower Street 1.0 simulations

wCDM

Gower Street simulations

- 1250 h^{-1} Mpc
- 1080^3 particles, N-body
- Systematics
  - NLA intrinsic alignments
  - Multiplicative biases
  - Source clustering
  - n(z) errors
  - h, m_ν, n_s, $\Omega_b h^2$
- marginalized
- 12,600 mock surveys

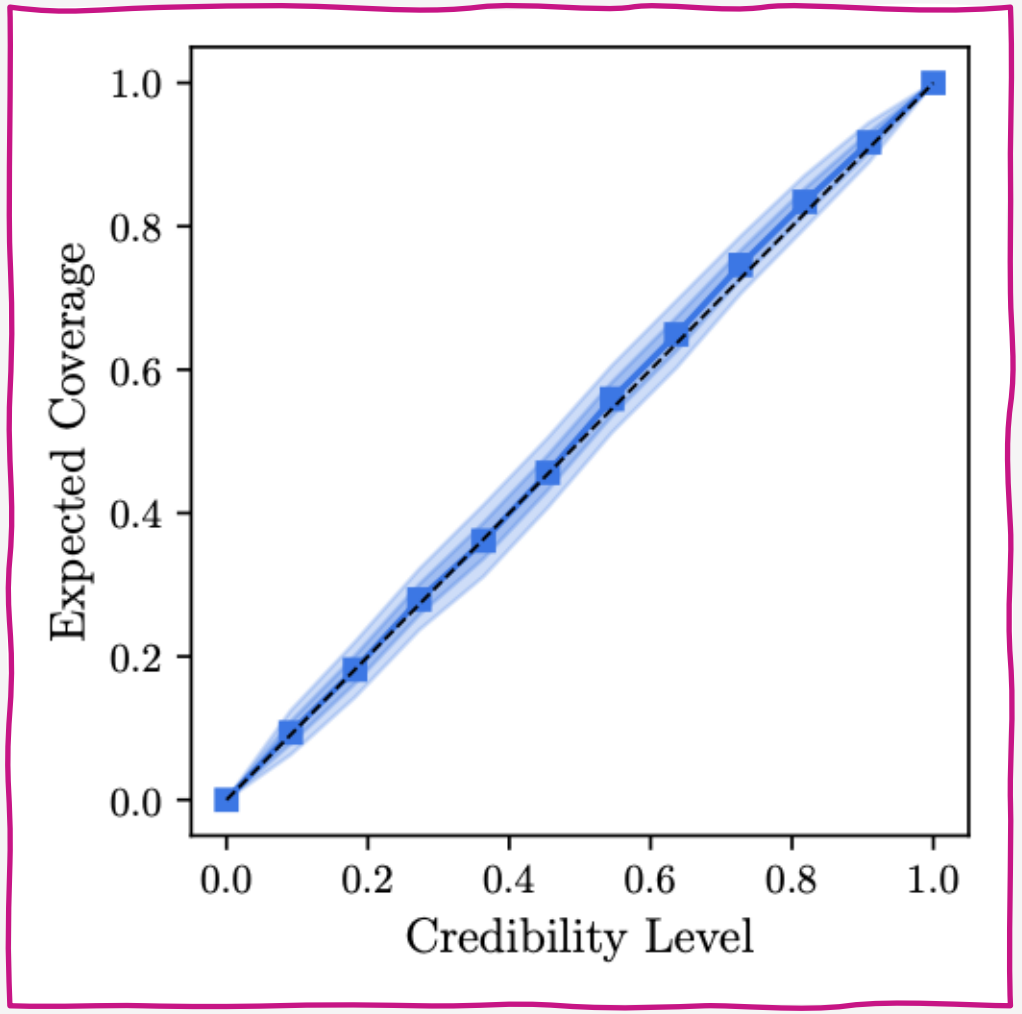


Jeffrey and Gatti et al. (2021)
Gatti et al. (2024)
Jeffrey et al. (2025)

ICIC

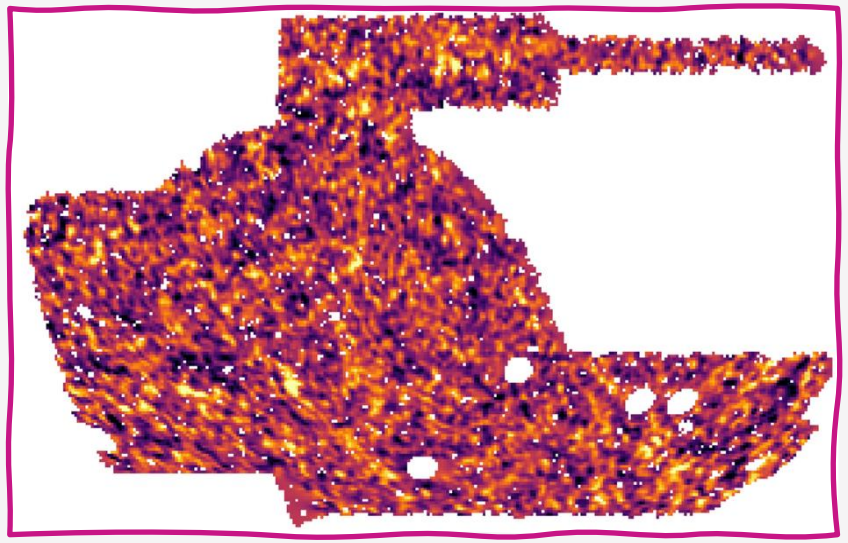
Coverage tests on simulations

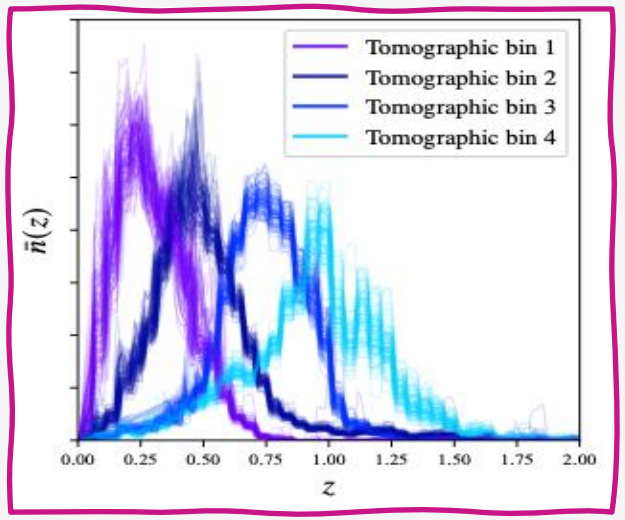
How reliable are the credible regions?



Dark Energy Survey (DES) Year 3 data

- 5000 square degrees
- 4 tomographic bins
- 100 million galaxies
- Convergence maps (Kaiser-Squires)

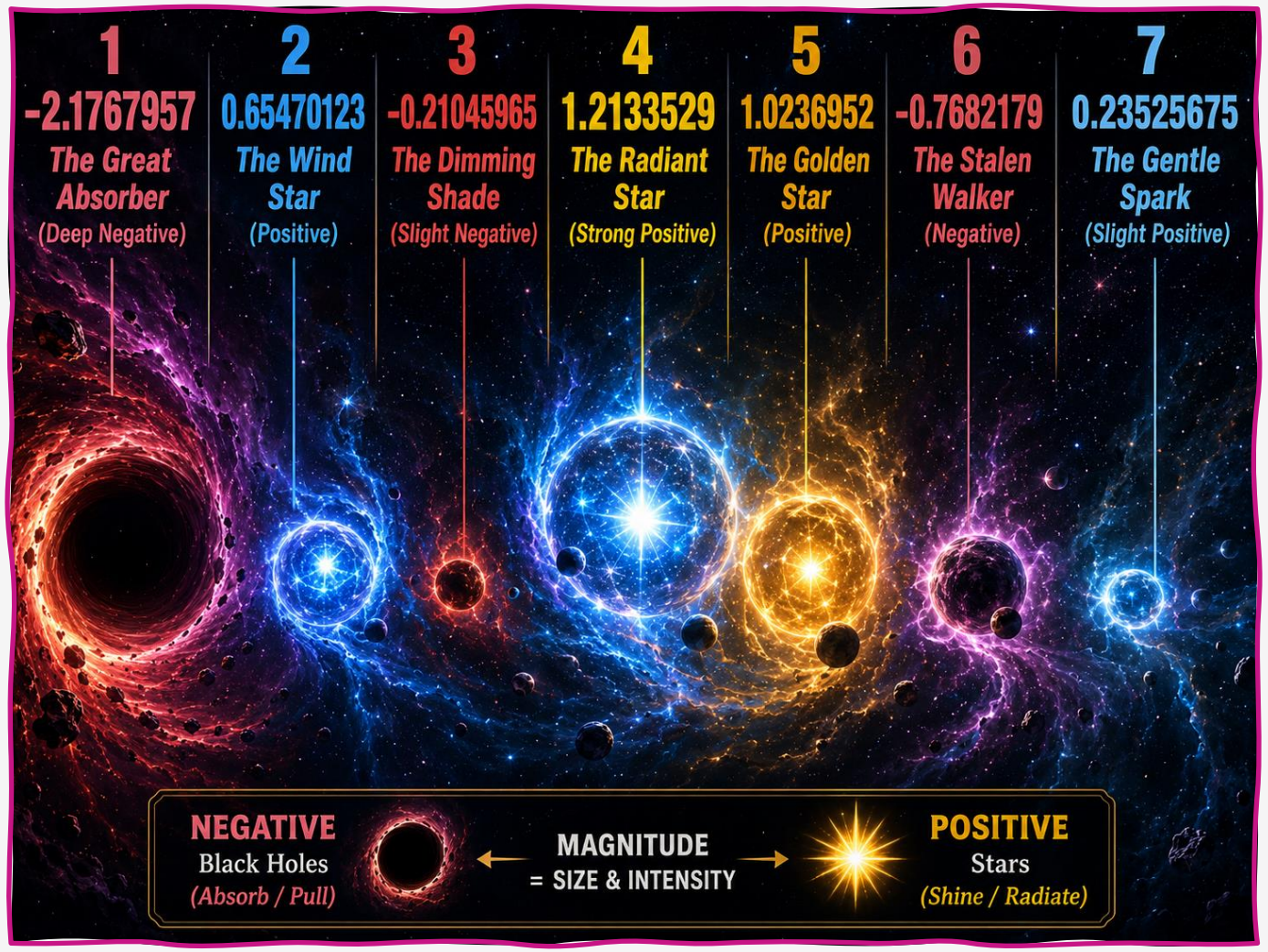




DES Collaboration (2016)
Shape catalogue Huff and Mandelbaum (2017)
Convergence maps Jeffrey et al (2021)

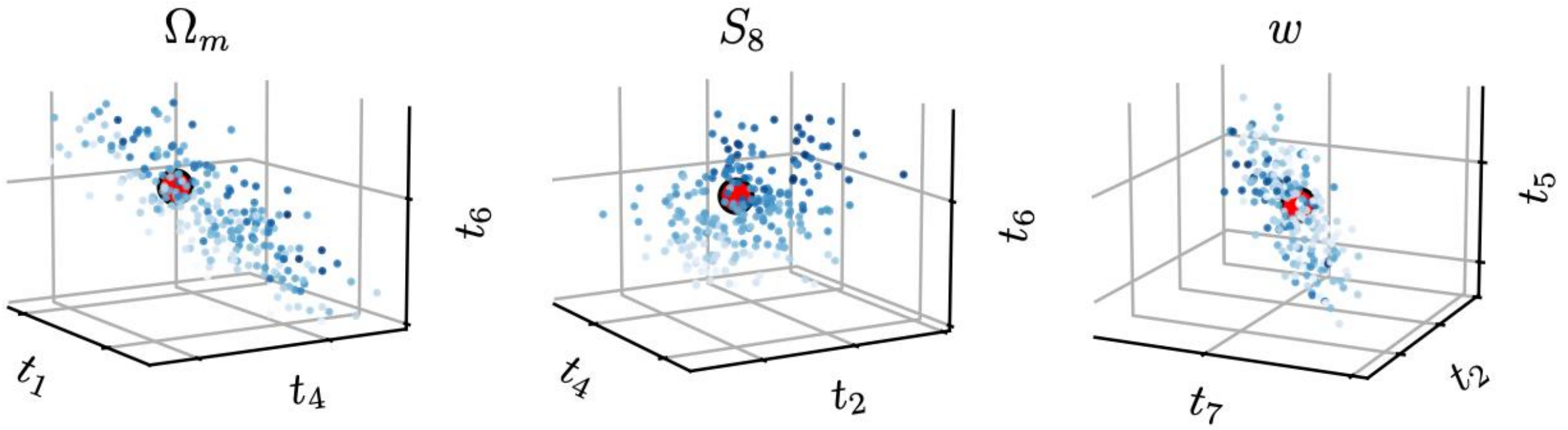
ICIC

The seven numbers $t_\psi(x_{DES})$



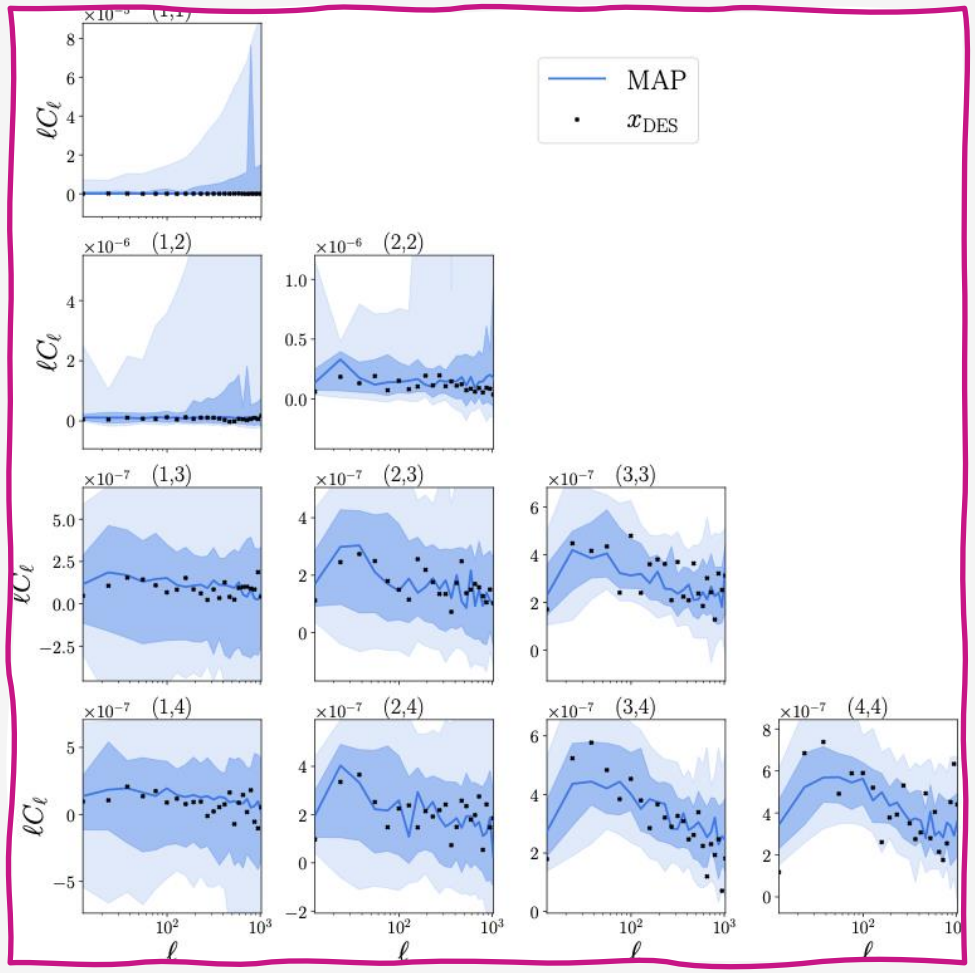
How similar are the simulations to the real Universe?

DES summaries sit in the simulation space



Posterior checks, and baryon physics check

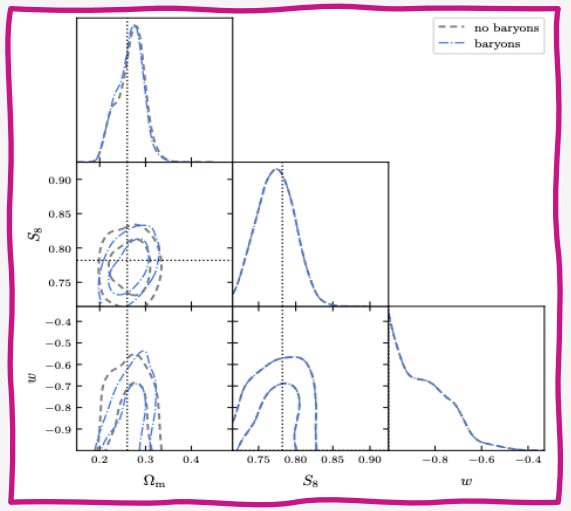
Do the models actually account for the data?



Baryons/gravity model

Results robust to changes of the forward model:

Using CosmoGridV1 simulations + baryonification, parameters shift by $< 0.3\sigma$



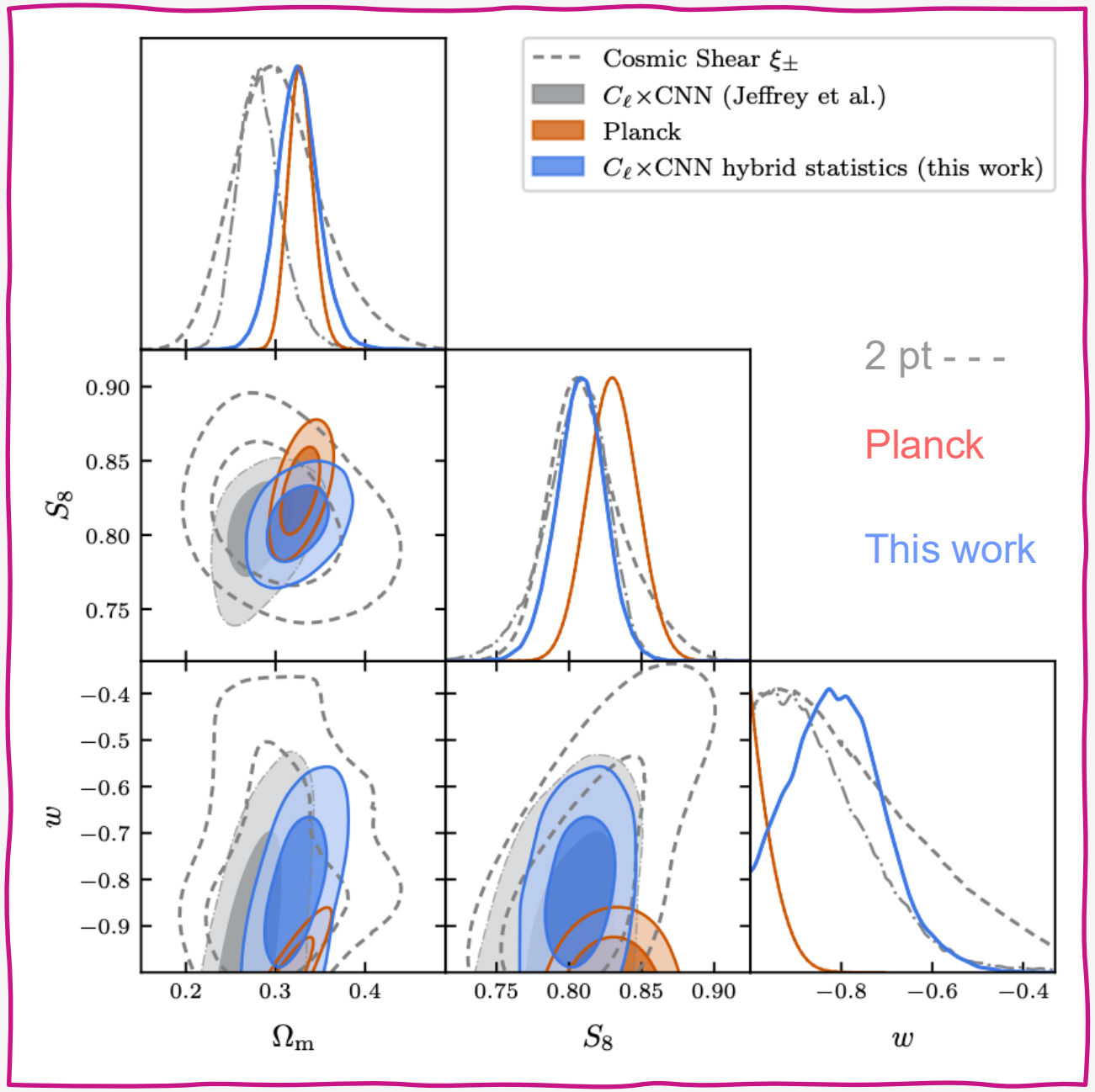
Results

Results

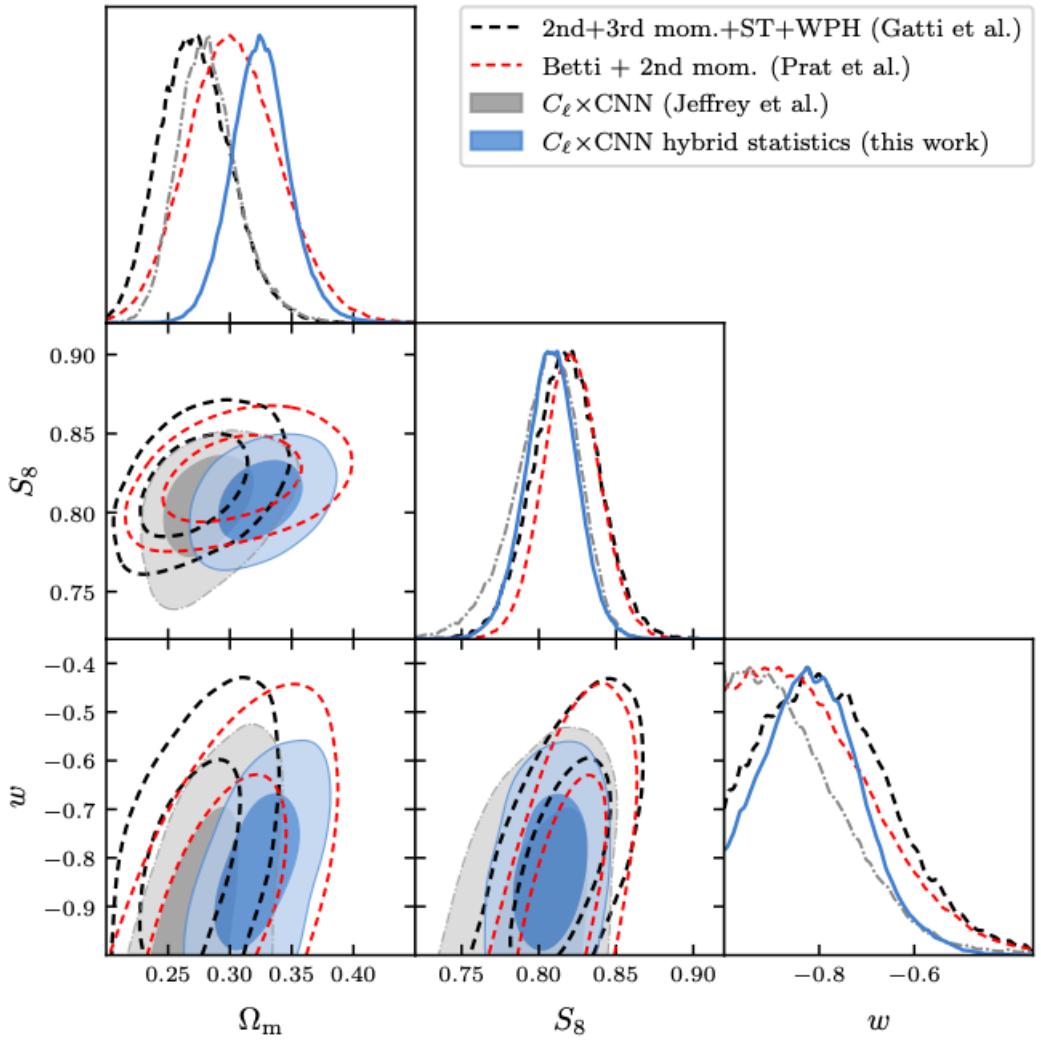
3 parameters

$$S_8 = 0.808 \pm 0.017$$
$$\Omega_m = 0.325 \pm 0.024$$
$$w < -0.77$$

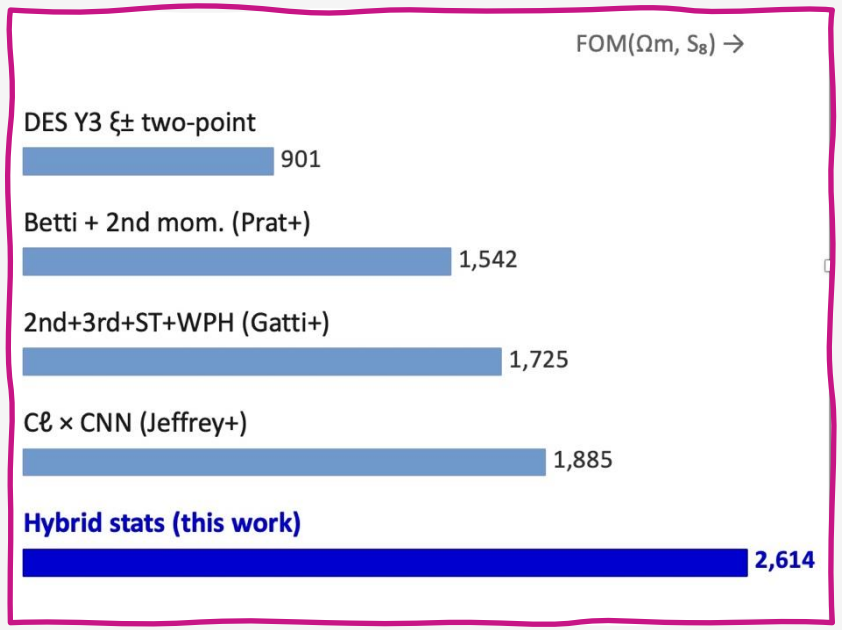
Precise agreement with Planck in both S_8 and Ω_m



Comparison with state-of-the-art



- **DES-Y3 analyses: 2pt:** FOM = 900 (Secco et al.)



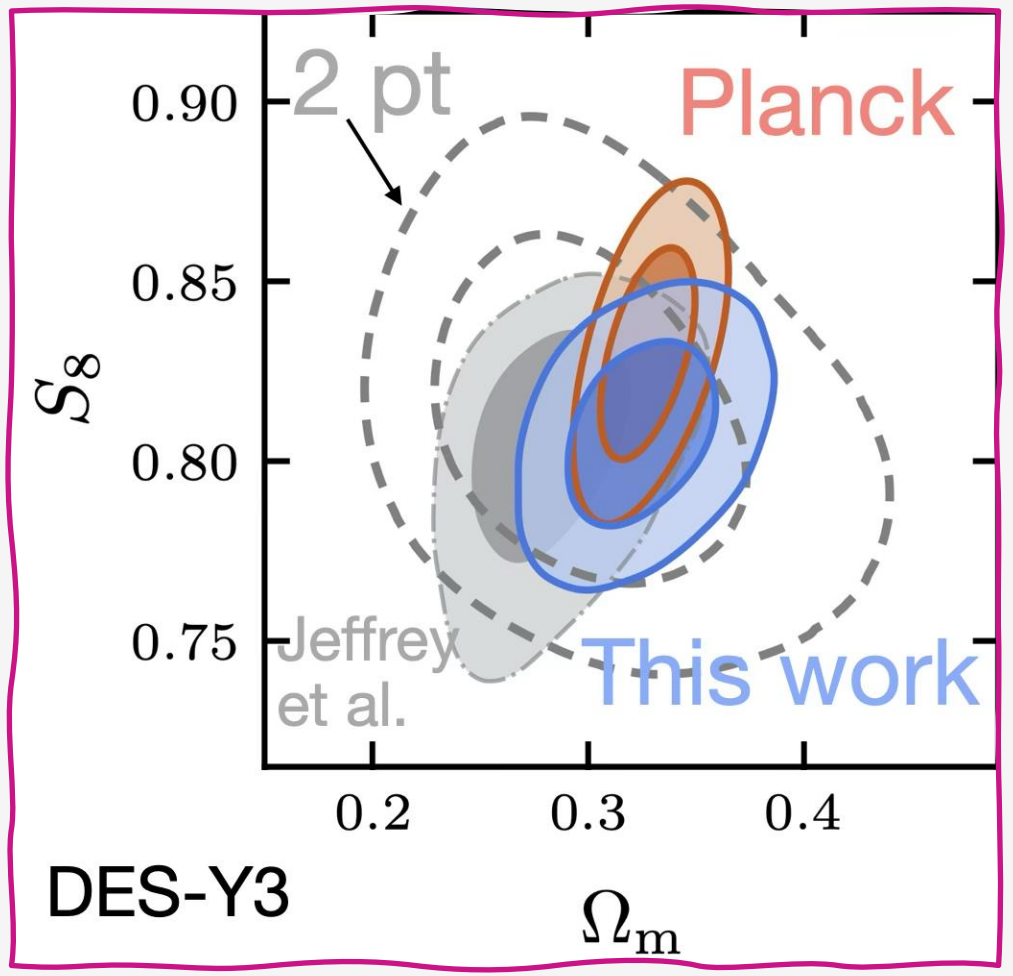
- **DES-Y6 cosmic shear only:**
  - Abbott et al. (2026): 2 pt. FOM = 2100 (LCDM)

Hybrid SBI: FOM = 2600

Summary

- **DES-Y3 weak lensing with hybrid SBI**

- **Most precise cosmology from cosmic shear alone**
- **Precise agreement with Planck in both** S_8 and Ω_m
- 100 Million galaxies compressed to only **7 data points**
- **Scalable to Euclid, LSST**
- Applicable to other probes
- Hybrid mix of power spectrum and CNN summaries; information-theory maximization of information content;
- **Not a black box** – SBI does not care how the summaries were obtained. Principled Bayesian result
- Williamson & Makinen et al. arxiv:2606.11309 today



$$S_8 = 0.808 \pm 0.017$$
$$\Omega_m = 0.325 \pm 0.024$$
$$w < -0.77$$

Backup slides

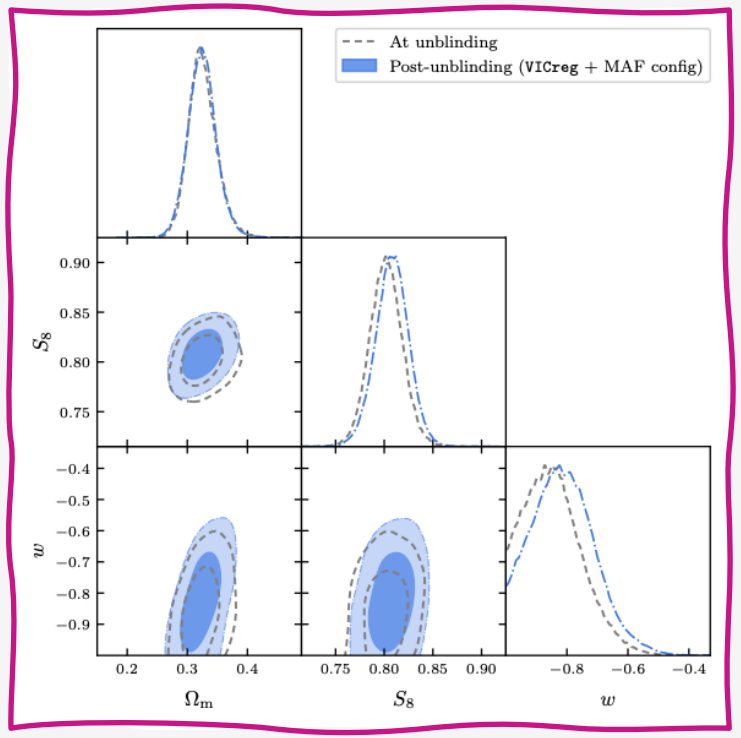
Simulation range

Parameter	Simulation Sampling Distribution
Ω_m	ActiveLearning$[0.15, 0.49]$
σ_8	ActiveLearning$[0.5, 1.0]$
h	$\mathcal{N}(0.7022, 0.0245)$
n_s	$\mathcal{N}(0.9649, 0.0063)$
$\Omega_b h^2$	$\mathcal{N}(0.02237, 0.00015)$
w	$\mathcal{U}[-1, -\frac{1}{3}]$
$\ln(m_\nu)$	$\mathcal{U}[\ln(0.06), \ln(0.14)]$

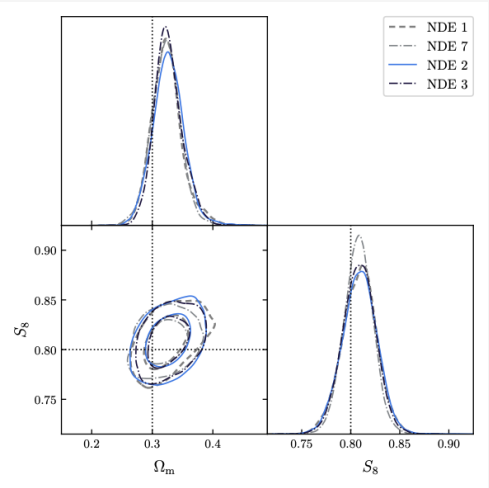
Priors

Parameter	Prior
Ω_m	$\mathcal{U}[0.15, 0.52]$
S_8	$\mathcal{U}[0.5, 1.0]$
w	$\mathcal{U}[-1, -\frac{1}{3}]$
n_s	$\mathcal{N}(0.9649, 0.0063)$
h	$\mathcal{N}(0.7022, 0.0245)$
$\Omega_b h^2$	$\mathcal{N}(0.02237, 0.00015)$
$\ln(m_\nu)$	$\mathcal{U}[\ln(0.06), \ln(0.14)]$
A_{IA}	$\mathcal{U}[-3, 3]$
η_{IA}	$\mathcal{U}[-5, 5]$
m_1	$\mathcal{N}(-0.0063, 0.0091)$
m_2	$\mathcal{N}(-0.0198, 0.0078)$
m_3	$\mathcal{N}(-0.0241, 0.0076)$
m_4	$\mathcal{N}(-0.0369, 0.0076)$
$\bar{n}_i(z)$	$p_{\text{HYPERRANK}}(\bar{n}_i(z) \mid x_{\text{phot}})$

Post-unblinding changes

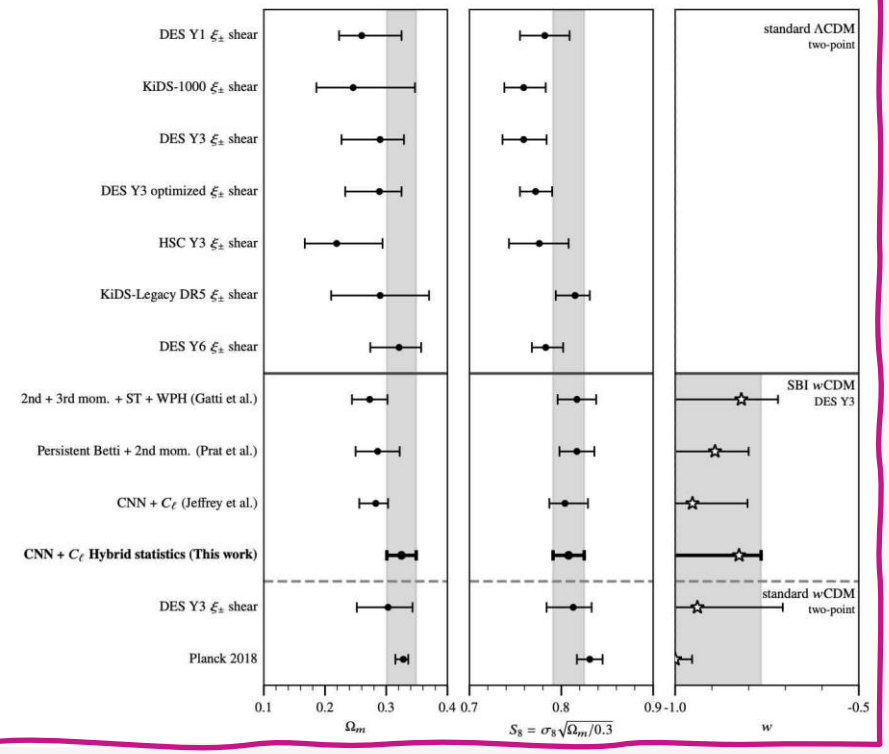


Different NDEs



Backup slides

Results comparison



	DES Y3 $\xi_\pm$ likelihood*	$C_\ell\times$ CNN (Jeffrey et al. 2025)	**Hybrid statistics (this work)**
Ω_m	$0.303^{+0.040}_{-0.051}$	$0.283^{+0.020}_{-0.027}$	$\mathbf{0.325 \pm 0.024}$
S_8	$0.813^{+0.020}_{-0.029}$	$0.804^{+0.025}_{-0.017}$	$\mathbf{0.808 \pm 0.017}$
w	< -0.707	< -0.804	$\mathbf{< -0.766}$
MAP(w)	-0.940	-0.953	$\mathbf{-0.826}$
FOM(Ω_m, w)	138	384	**470**
FOM(S_8, w)	391	389	**592**
FOM(Ω_m, S_8)	901	1,885	**2,614**
FOM(Ω_m, S_8, w)	10,344	18,430	**29,444**

*reanalysed