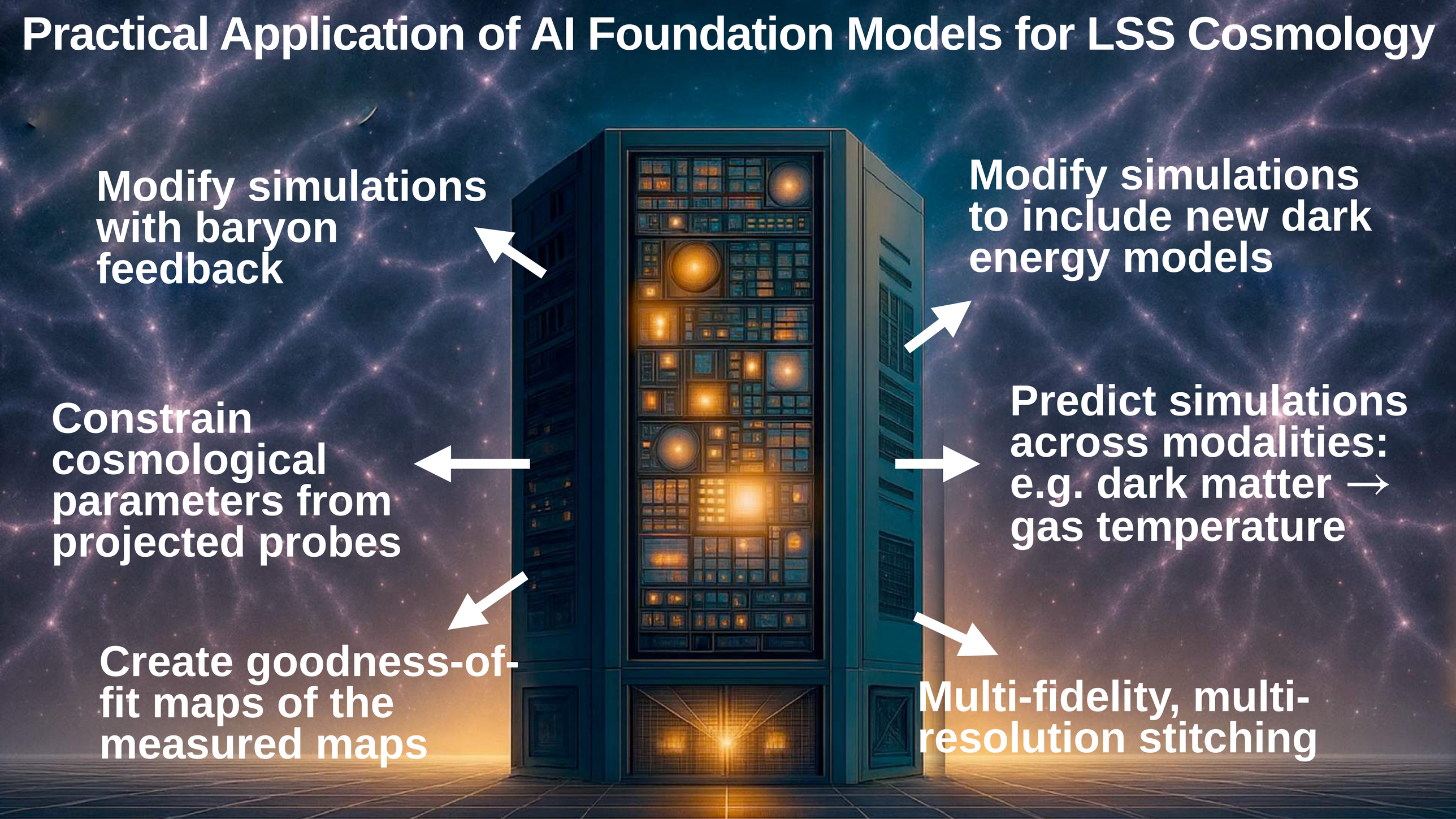
Towards AI Foundation Models for Large-Scale Structure Cosmology



Tomasz Kacprzak, University of Applied Sciences Northwestern Switzerland FHNW
AI+Astronomy, Zhejiang Lab, 19 Oct 2026

Practical Application of AI Foundation Models for LSS Cosmology



Simulating Cosmological Theory for Map-level AI Analysis



Large simulation volumes to avoid underestimating errors

SBI-ready theory modeling for astrophysical effects: galaxy bias, intrinsic alignment, baryon feedback

Large number of particles to resolve small-scale gravity

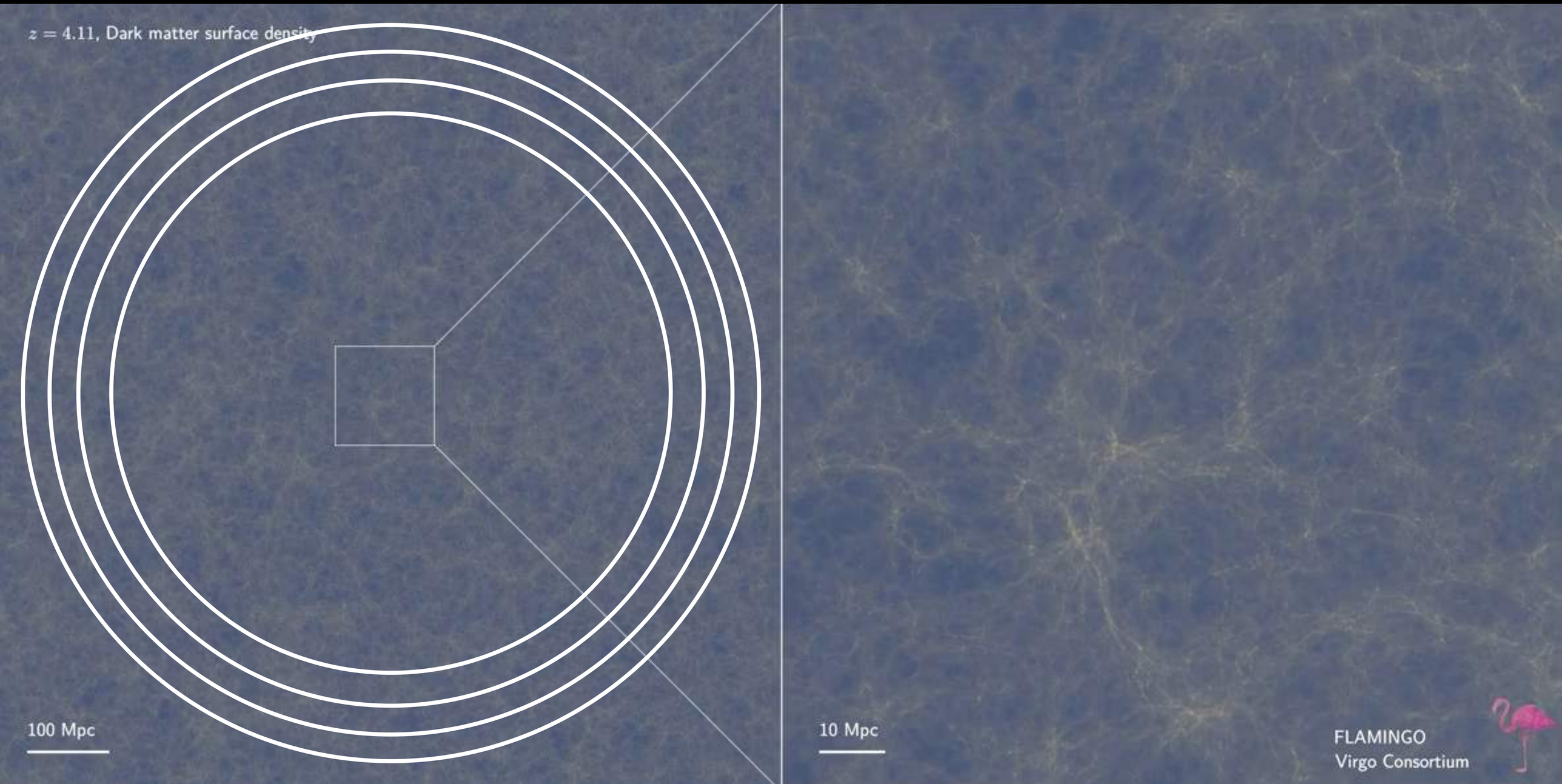
Efficient computing codes that fit the hardware (CPU/ GPU/NextGen)

Large number of simulations to cover the cosmology parameter space

Large production projects in supercomputing centers

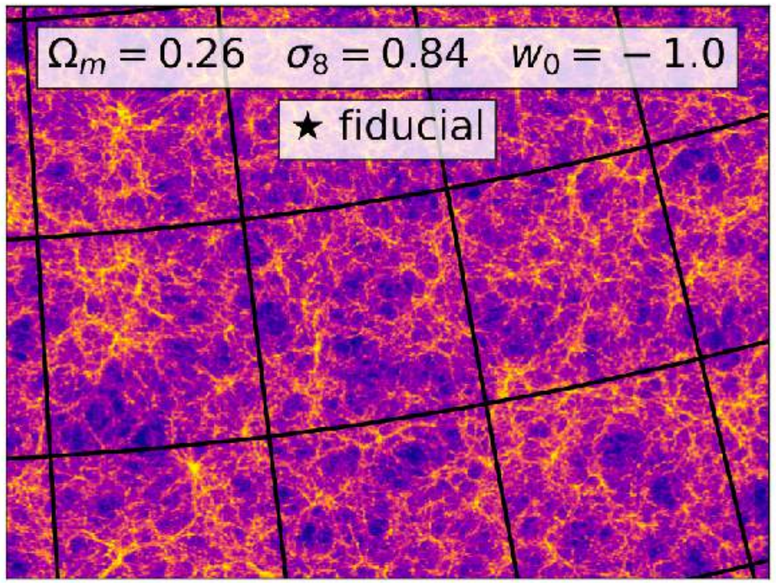
Large-scale matter distribution in the universe

Matter distribution evolves under laws of gravity and expansion of the universe



Dark matter distributions carries information about cosmological parameters

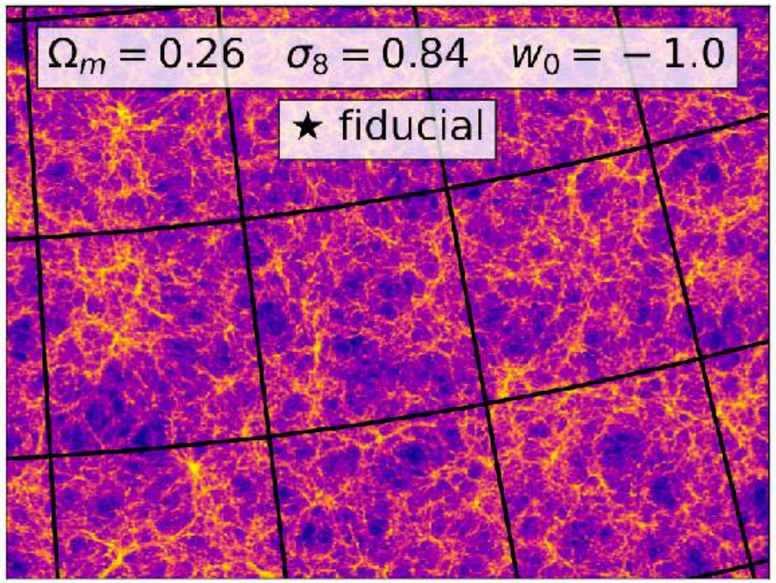
Dark energy density Ω_Λ, dark matter density Ω_m, such that $\Omega_\Lambda + \Omega_m = 1$



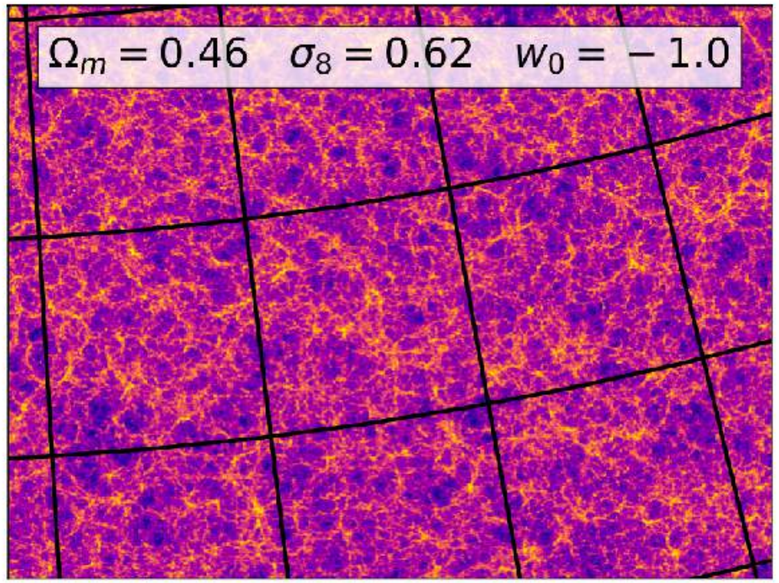
best-fitting today ★

Dark matter distributions carries information about cosmological parameters

Dark energy density Ω_Λ, dark matter density Ω_m, such that $\Omega_\Lambda + \Omega_m = 1$

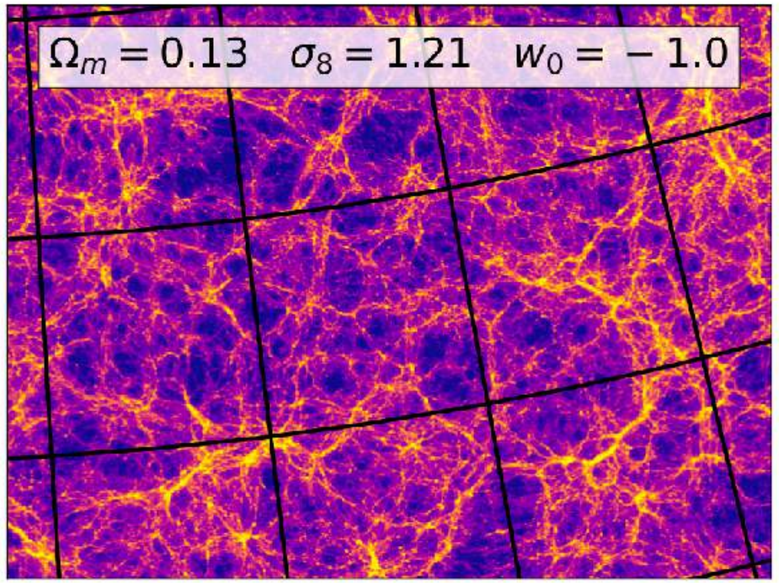


best-fitting today ★



high matter density

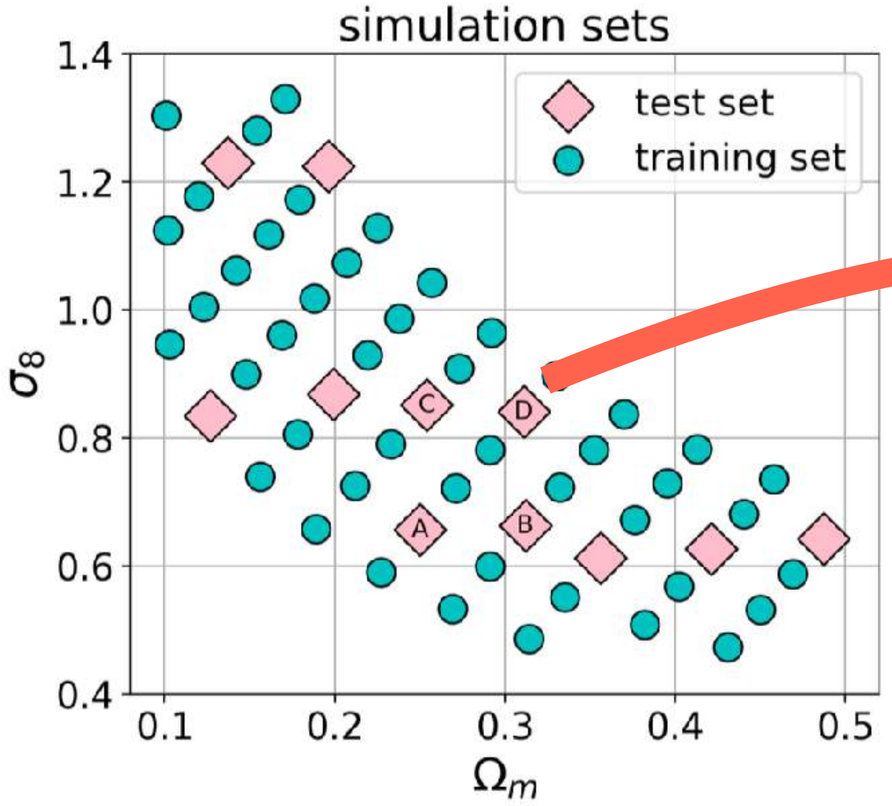
$\Omega_m \uparrow$, $\Omega_\Lambda \downarrow$

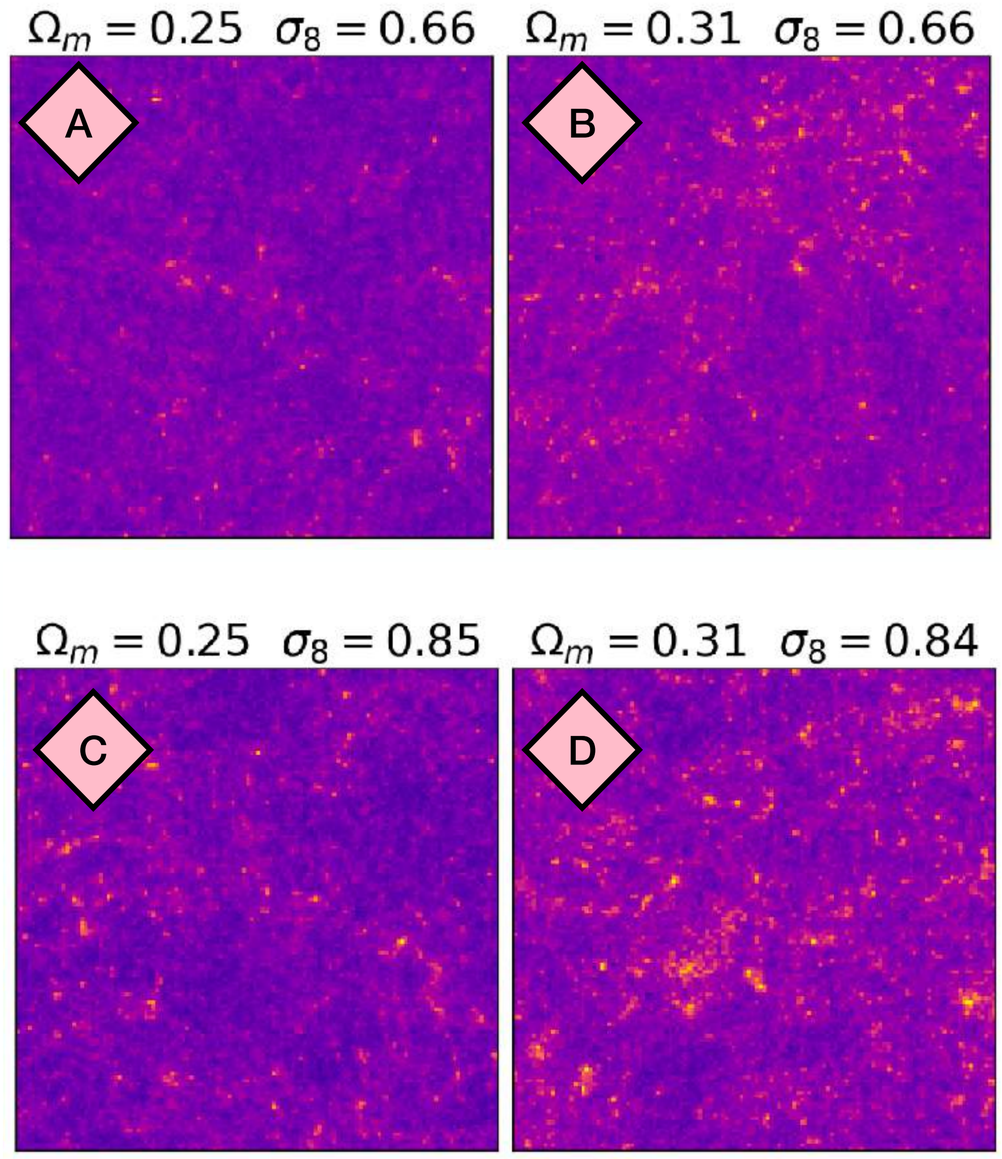


high dark energy

$\Omega_m \downarrow$, $\Omega_\Lambda \uparrow$

Generative AI for cosmological maps





Generative Adversarial Networks

Continuous Normalizing Flows trained with Flow Matching

Emulation of cosmological mass maps with conditional generative adversarial networks
Perraudin, +TK, et al. 2020, 2004.08139

$$\mathcal{L}_D = \mathbb{E}[D(x_g, c) - D(x, c)] \quad + \lambda \mathcal{R}_{\mathrm{GP}}$$

Networks: G generates maps. D assigns a scalar critic score. Minimize $\mathcal{L}_D$ over the critic weights and $\mathcal{L}_G$ over the generator weights, alternately.

Data and condition: x is a simulated training map. $c = (\Omega_m, \sigma_8)$ specifies its cosmology.

Generated map: $x_g = G(z, c)$, where $z \sim p_z$ is a

Increasing the Precision of Surrogate Models for Weak Lensing Mass Maps with Flow Matching
Li and Kacprzak 2026, 2605.23114

NEW PAPER!

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\|v(t, r_t, c) - u\|_2^2\right]$$

Network: v predicts a velocity at every map pixel. Minimize $\mathcal{L}_{\mathrm{FM}}$ over its weights.

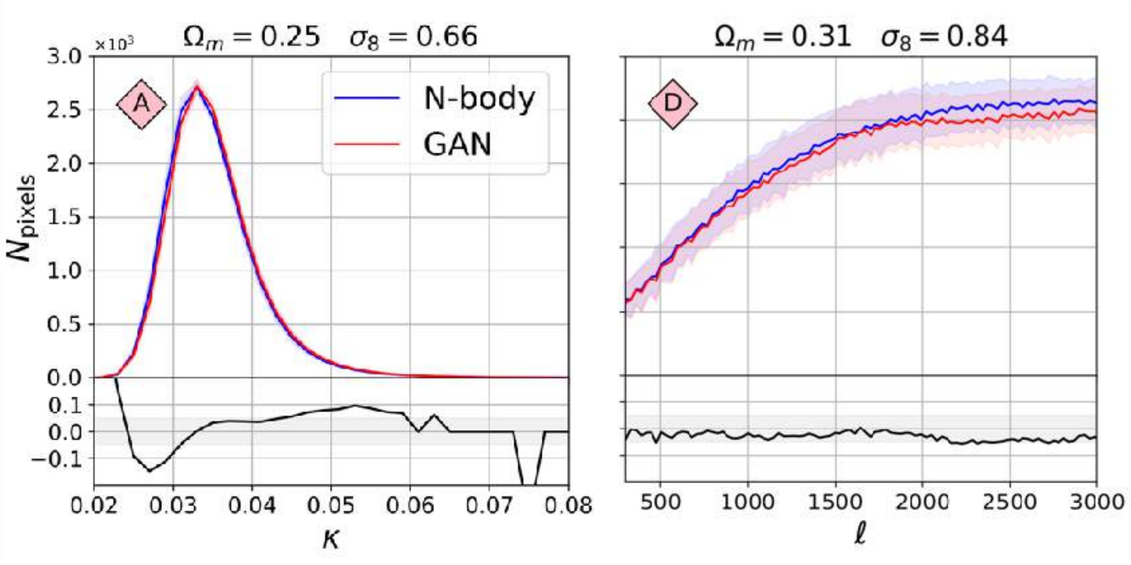
Data and condition: x is a log-transformed, globally standardized training map. $c = (\Omega_m, \sigma_8)$ specifies its cosmology.

Training input: $r_t = t\, r + s_t(c) z$, where $r = x - \mu(c)$ is the residual map and $\mu(c)$ is the ensemble mean at fixed cosmology.

Generative Adversarial Networks

Continuous Normalizing Flows trained with Flow Matching

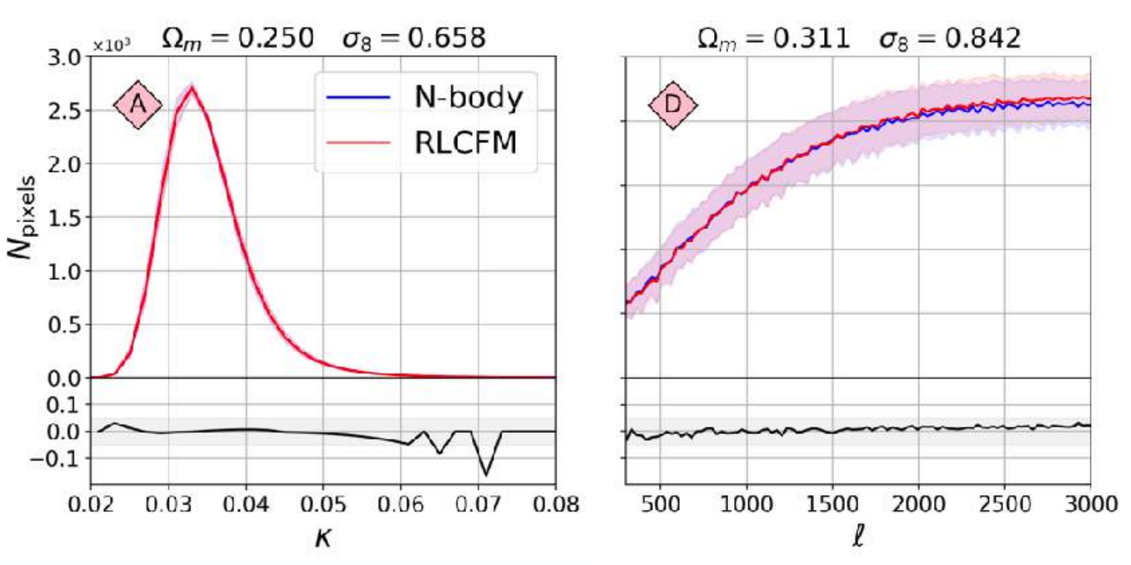
Emulation of cosmological mass maps with conditional generative adversarial networks
Perraudin, +TK, et al. 2020, 2004.08139



Map pixel histograms

Map angular power spectrum

Increasing the Precision of Surrogate Models for Weak Lensing Mass Maps with Flow Matching
Li and Kacprzak 2026, 2605.23114



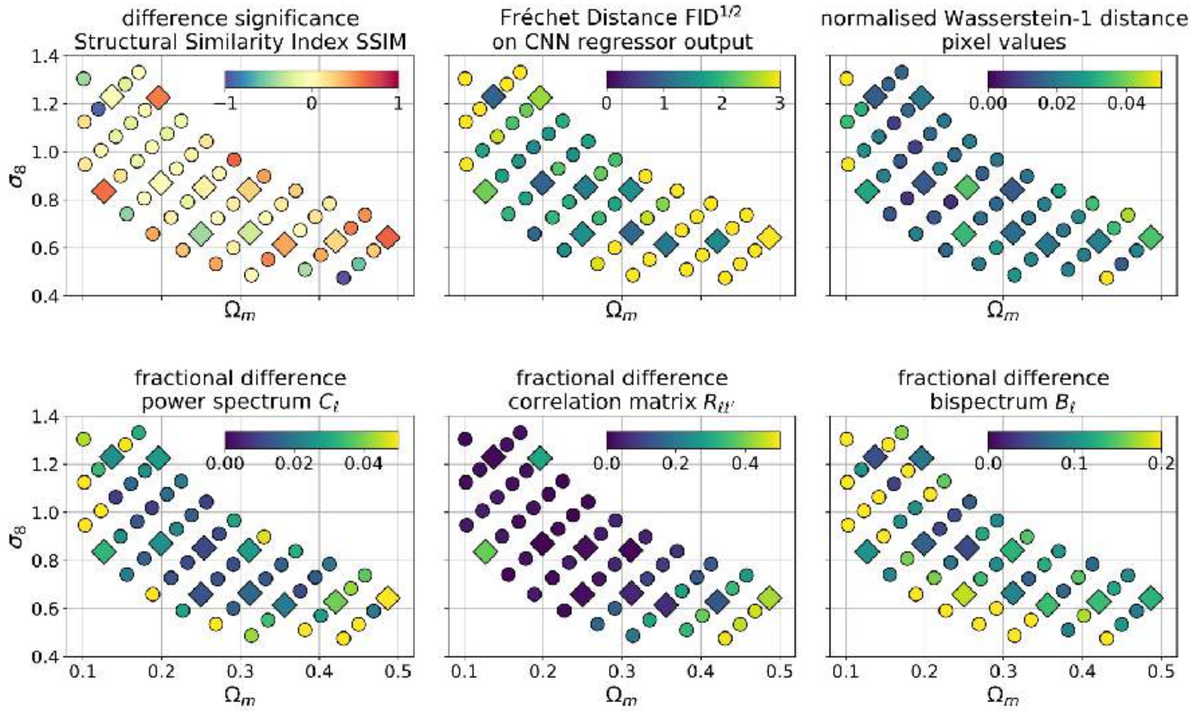
Map pixel histograms

Map angular power spectrum

Generative Adversarial Networks

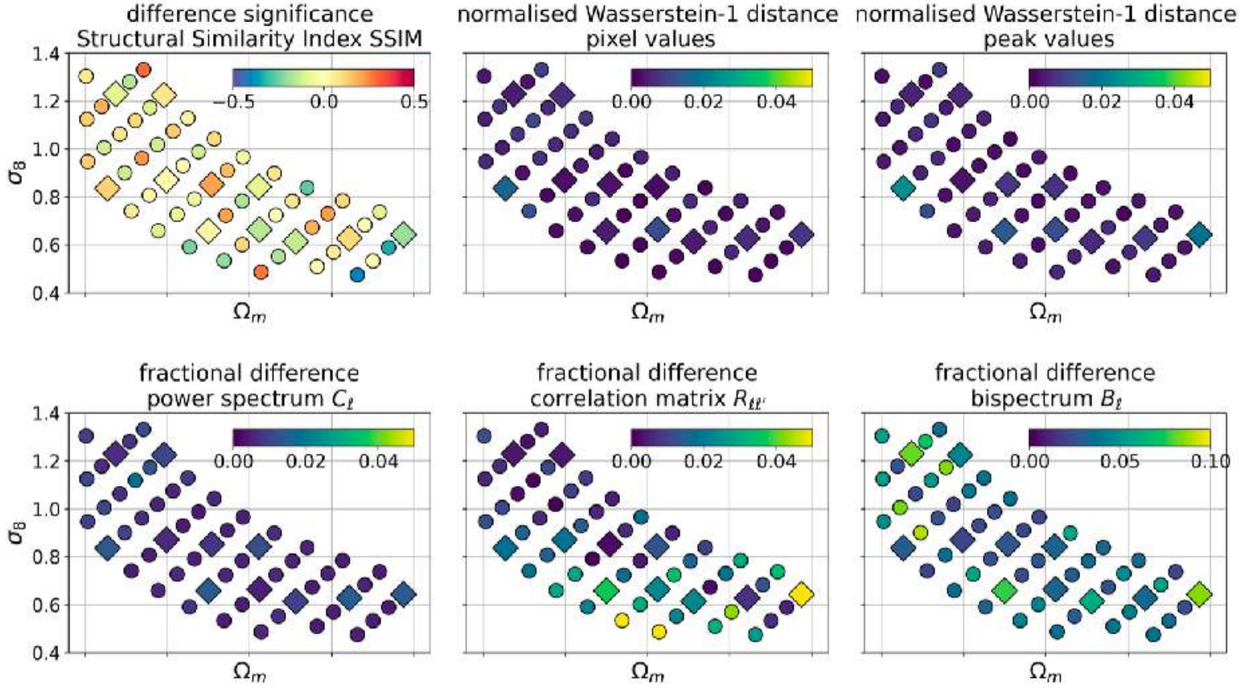
Continuous Normalizing Flows trained with Flow Matching

Emulation of cosmological mass maps with conditional generative adversarial networks
Perraudin, +TK, et al. 2020, 2004.08139



5-10% accuracy on summary statistics

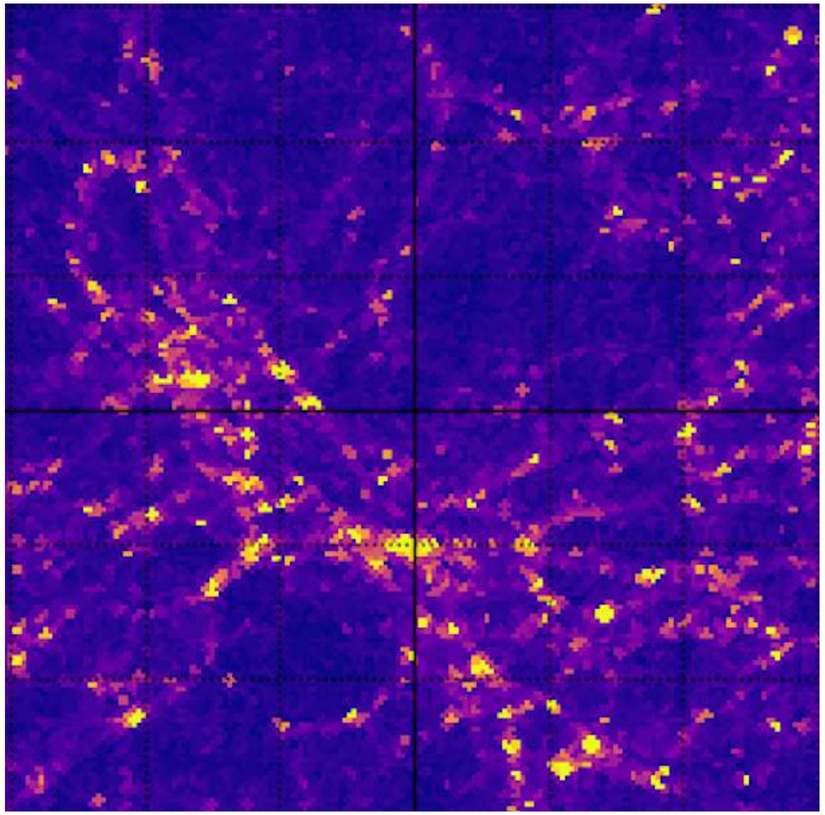
Increasing the Precision of Surrogate Models for Weak Lensing Mass Maps with Flow Matching
Li and Kacprzak 2026, 2605.23114



1% accuracy on summary statistics

Modify Standard Simulations to Include Extended Dark Energy/Matter Models

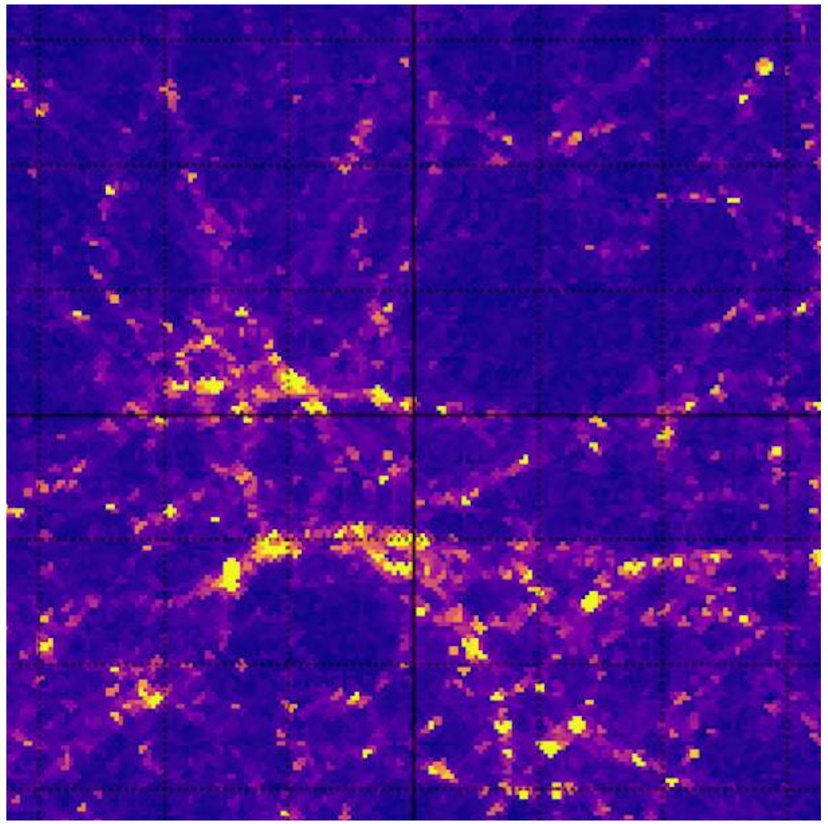
$w_a = 0$



Standard model, no time-evolution of dark energy equation of state

AI modification

$w_a = +0.15$

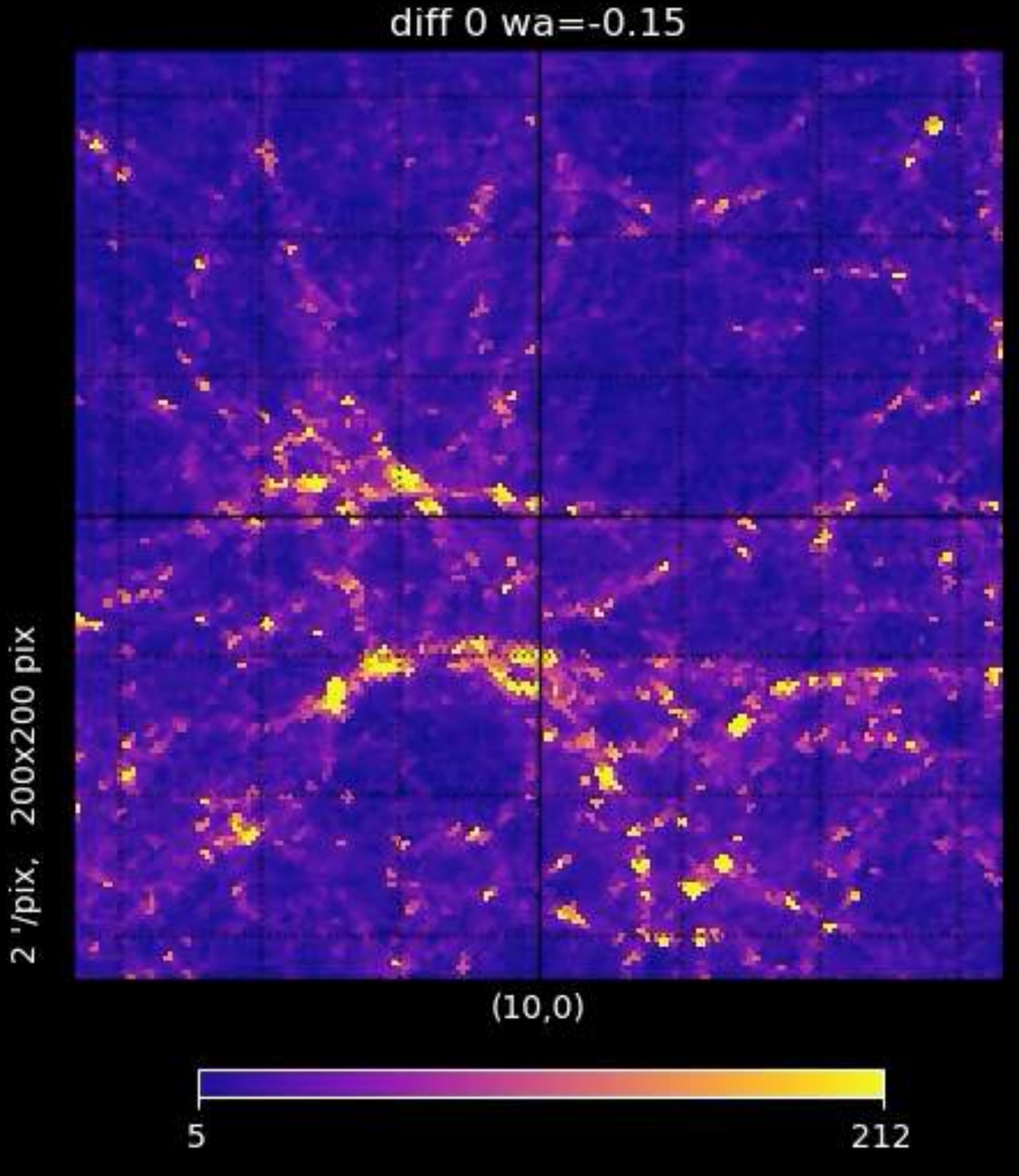


Extended model, time-evolving dark energy

$w(t) = w_0 + w_a(1 - a_t)$

Rapid testing of dark energy models on map-level

- “Exotic physics” simulations are quite expensive as we do as it is risky to invest engineering efforts to optimize their codes
- Solution: use generative AI model!
- Train a foundation model on standard cosmology simulations, e.g. CosmoGridV1
- Create a small number of “exotic physics” simulations with matched initial conditions
- Train a foundation model adapter to learn how to convert the “standard” to “exotic”
- Deploy adapter on the full grid
- Repeat for many “exotic” models, detect some with sufficient statistical significance
- In case of a detection, run full simulation grid with expensive simulations, if needed
- MSc project Noor Schepers - completed!



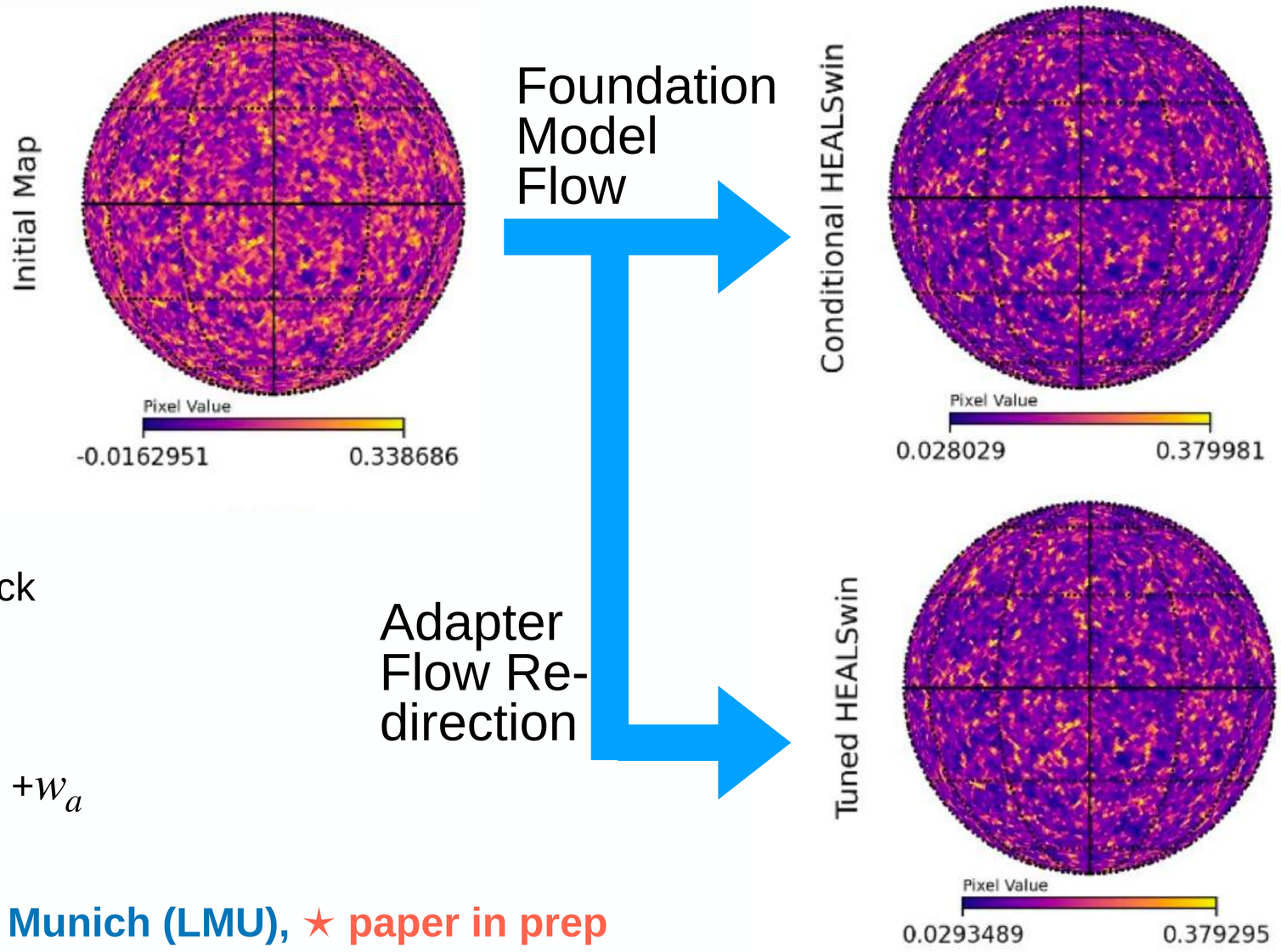
Towards the foundation model for cosmological large-scale structure lightcones

Step 1: Train the foundation models using abundant standard physics simulations

- HealSwin Transformer, U-net architecture
- 44m trainable parameters
- 17500 dark matter maps
- Start from phase-locked gaussian field
- Condition on standard cosmology parameters

Step 2: Train the adapter to generate extended physics maps

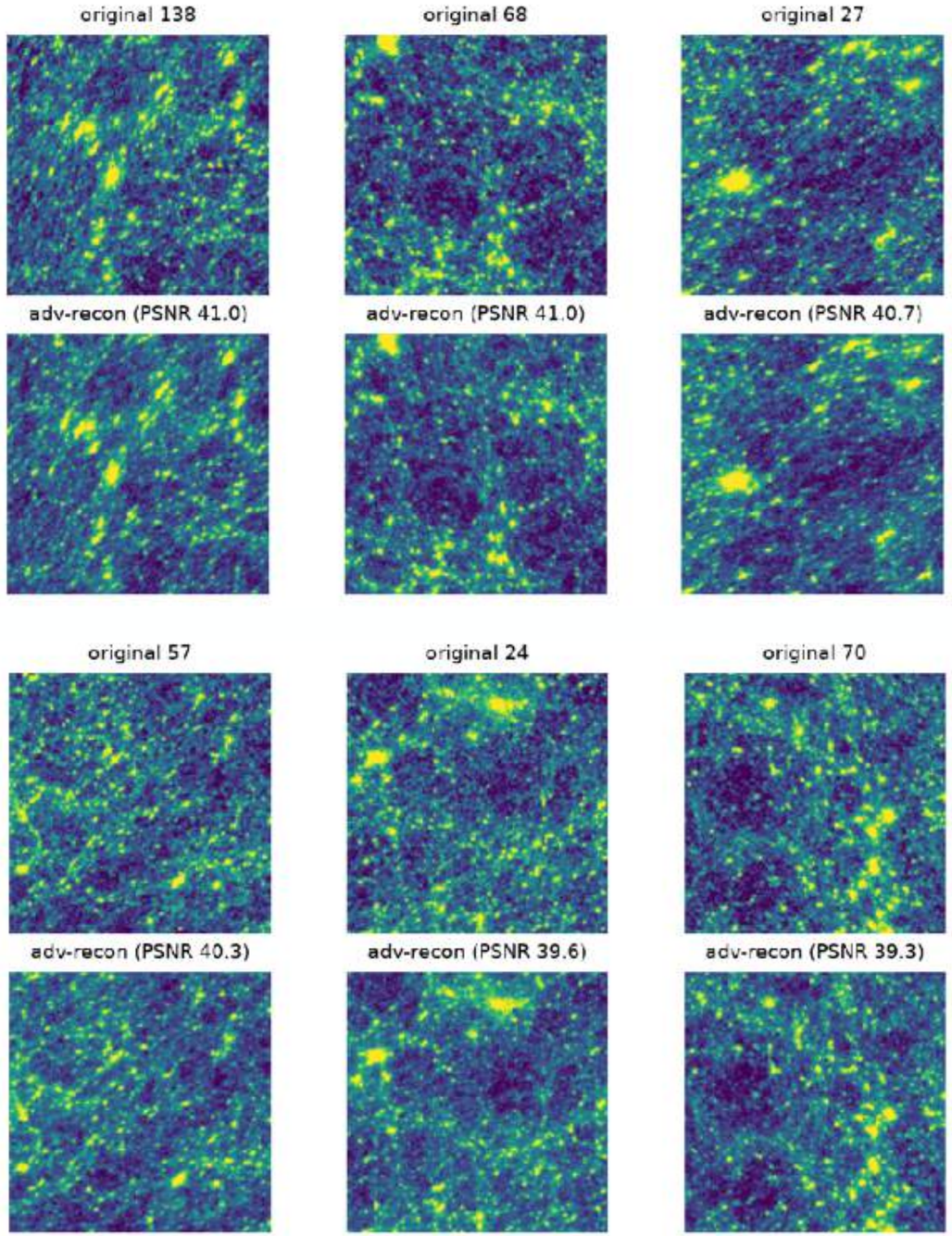
- Bottleneck adapter after each transformer block
- 2m trainable parameters
- 493 dark matter maps
- Start from phase-locked gaussian field
- Condition on standard cosmology parameters +w_a



MSc thesis Noor Schepers 2026 University of Munich (LMU), ★ paper in prep

A good tokenizer will be extremely useful

- CosmoTok project led by Shulei Ni at the Zhejiang Lab
- Using CosmoGridV1 data
- Encoder/decoder based on transformer embeddings,
- Working on independent 128×128 pixel patches created from healpix maps, 16×16 grid, 256 tokens per parch
- Finite Scalar Quantizer (FSQ)
- Training with adversarial model loss
- Very good reconstruction of images
- Good reconstruction of power spectra
- Compression is a difficult challenge as the maps are more like “complicated noise” than “structured image”
- Can enable flow matching in latent space
- Collaboration in progress



Simulating Cosmological Theory for Map-level AI Analysis



Large simulation volumes to avoid underestimating errors

SBI-ready theory modeling for astrophysical effects: galaxy bias, intrinsic alignment, baryon feedback

Large number of particles to resolve small-scale gravity

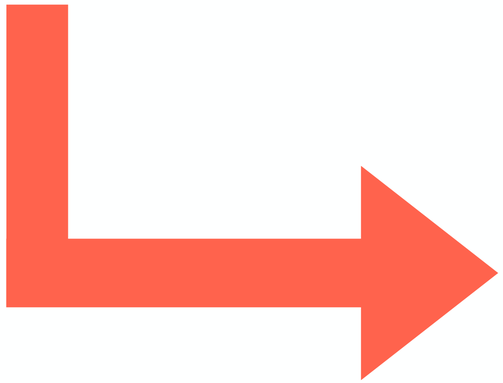
Efficient computing codes that fit the hardware (CPU/ GPU/NextGen)

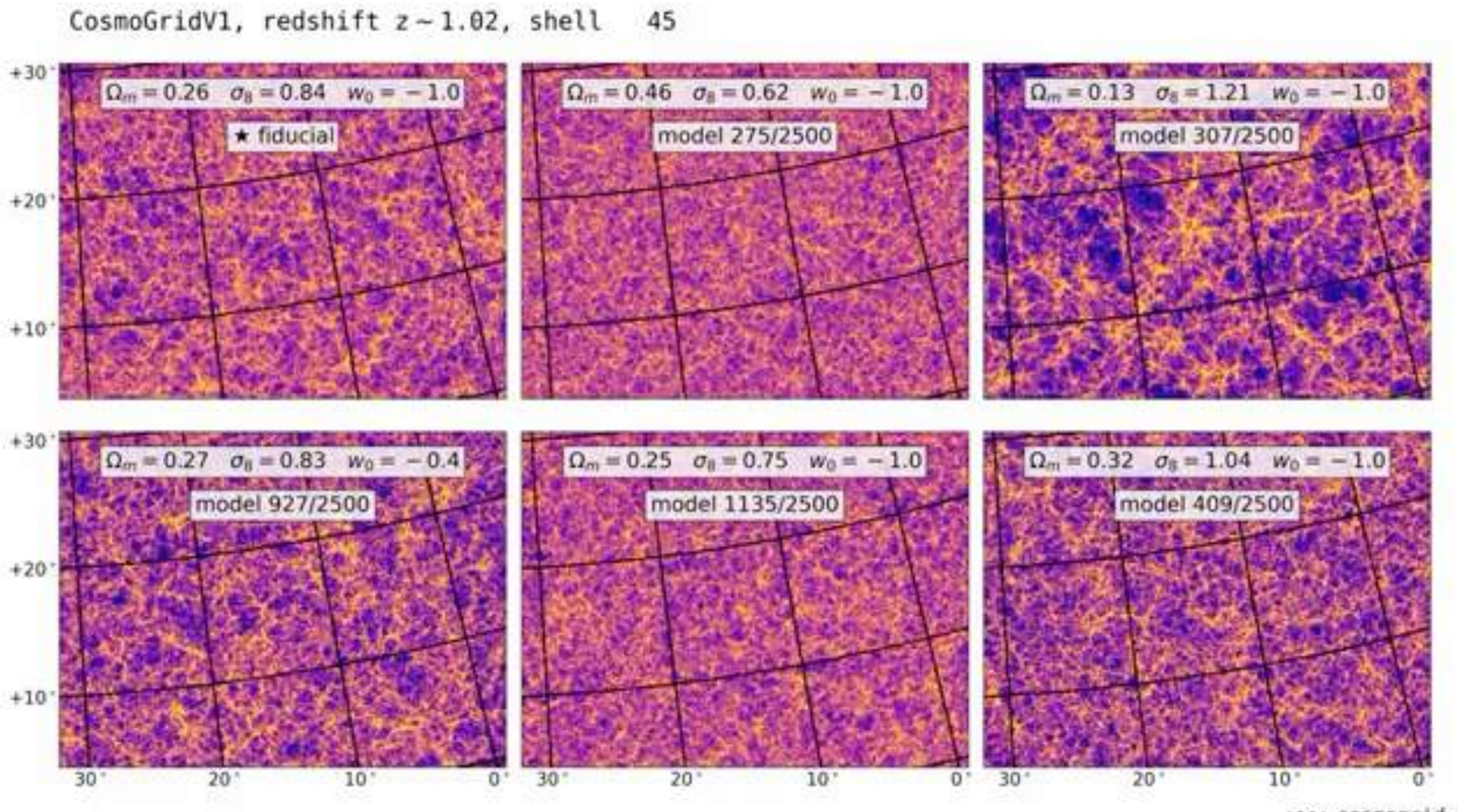
Large number of simulations to cover the cosmology parameter space

Large production projects in supercomputing centers

Simulating Theory Prediction for SBI and AI Training



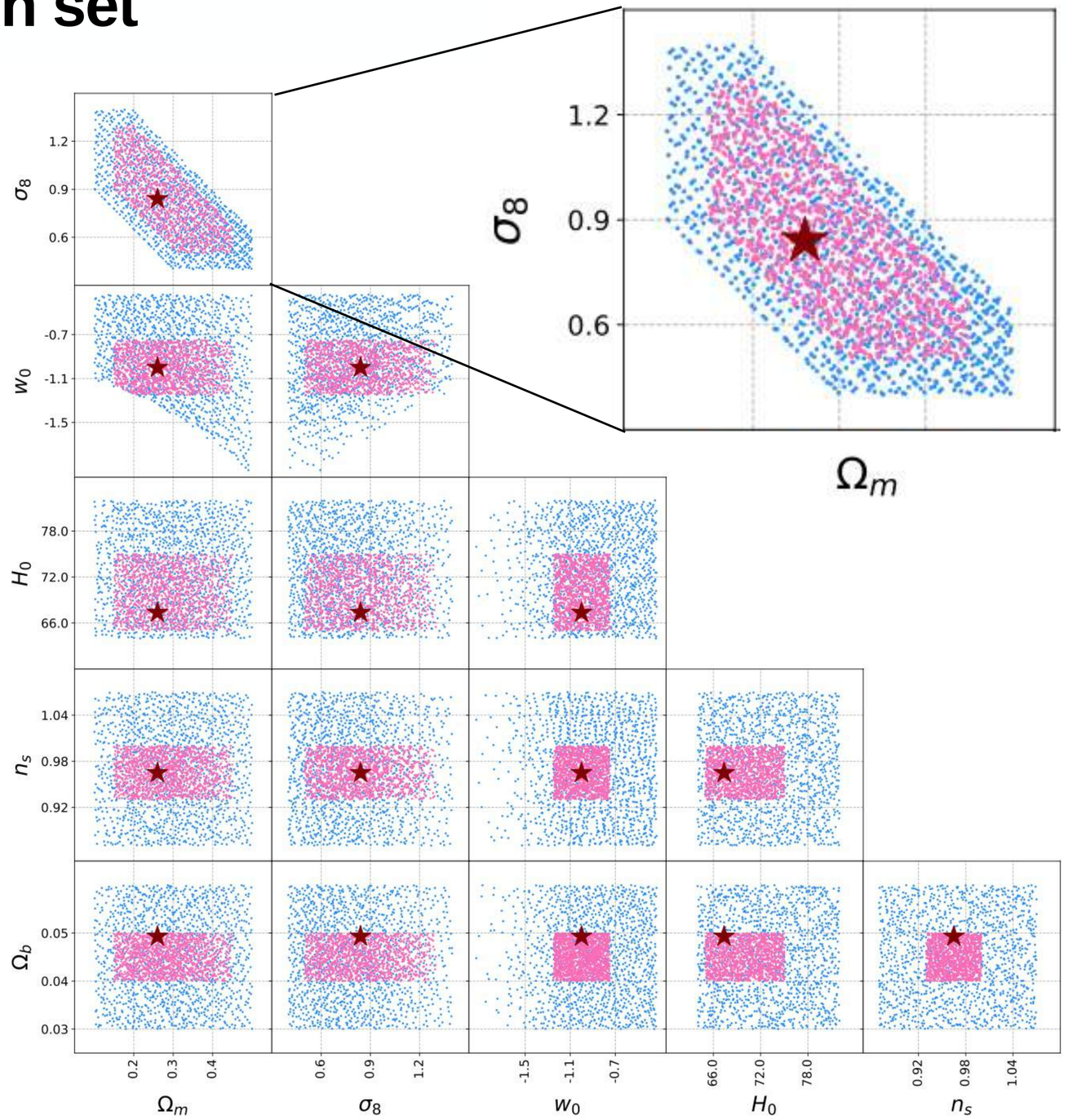




The public CosmoGridV1 simulation set

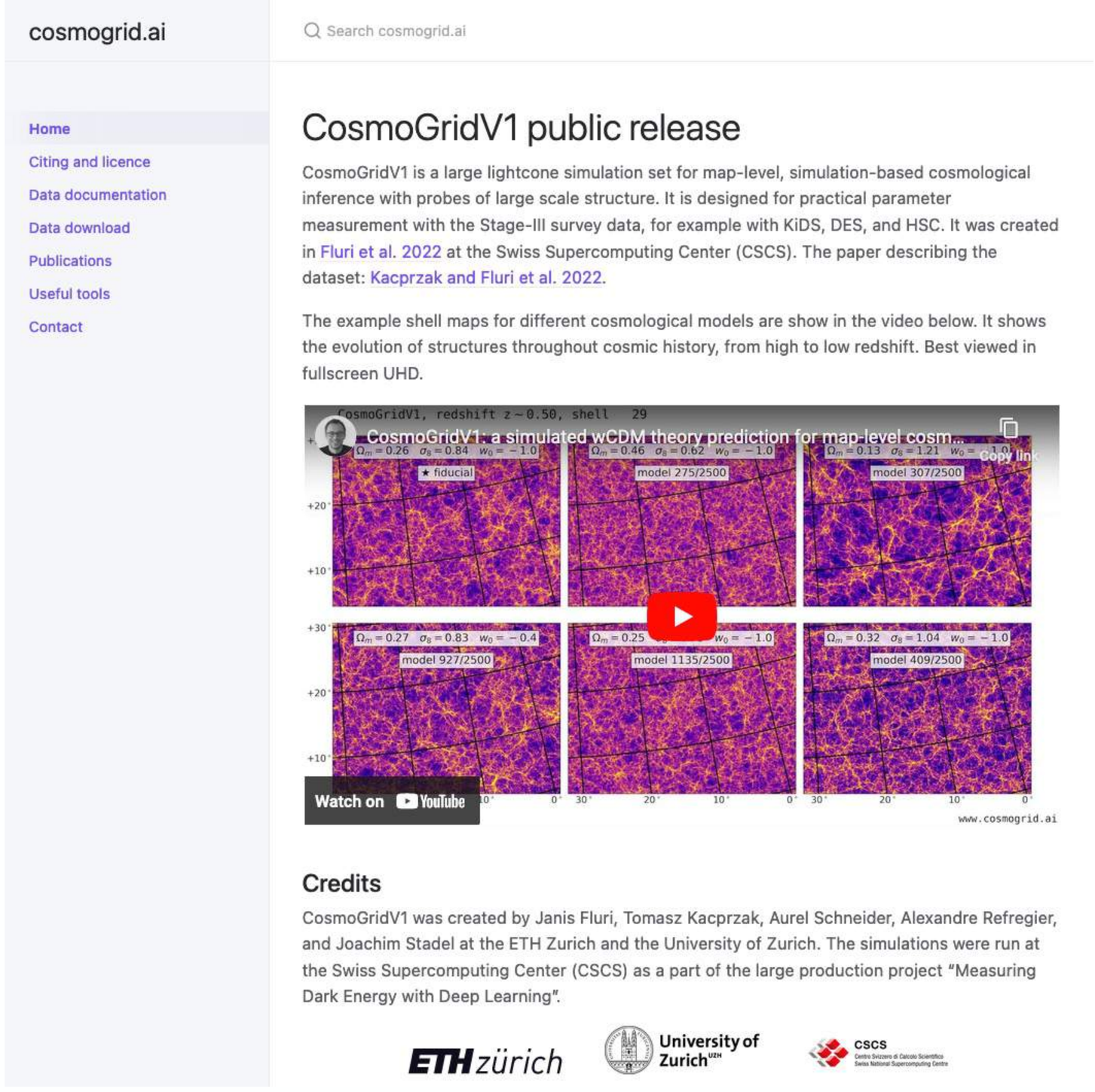
TK, Fluri, Schneider, Refregier, Stadel 2209.04662

- 2500 unique cosmologies in wCDM (5 cosmological parameters)
- 7 unique initial conditions simulations per cosmology
- Total ~20’000 independent N-body simulations
- Medium-resolution maps, time resolution: 70 maps starting from 11 billion years ago
- Benchmark simulations available with bigger boxes, more particles and more shells
- CSCS Large Production Project, 2020-21
- 750’000 GPU node hours (Piz Daint)
- 300TB of raw simulation data, CC-BY licence
- Map-level baryonification



The CosmoGridV1 Data

- → Data available at **www.cosmogrid.ai**
- → Full data documentation available
- → Fast transfer with Globus
- → Creative Commons BY licence
- → CosmoGridV1.1 in preparation to include higher order fields (intrinsic alignments, quadratic bias, magnification, source clustering)
- → Projection codes available on Github, but still work in progress: github.com/CosmoGridCollab/cosmogridv11



SBI for Stage IV surveys

Stage III:

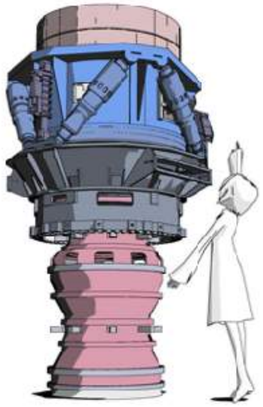
- Area: 1300 - 5000 deg^2
- Number of galaxies: 21-600 m, ~6 gal/arcmin^2
- i-band magnitude limit: ~24
- Max redshift: z=1.2
- Error on Ω_m 5%, on w_0 20%

Stage IV:

- Area: 14000 - 18000 deg^2
- Number of galaxies: 20 b, ~30 gal/arcmin^2
- i-band magnitude limit: ~26
- Max redshift: z=3
- Error on Ω_m 1%, on w_0 5%

DARK ENERGY SURVEY

Stage III



Hyper Suprime Cam

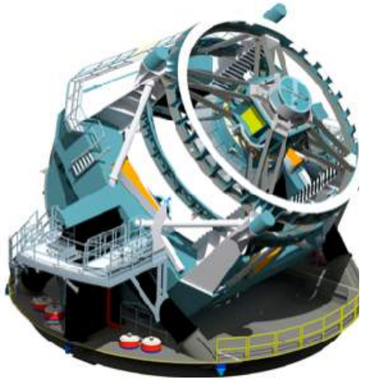
KiDS

Kilo Degree Survey

Euclid



Stage IV



Rubin/LSST

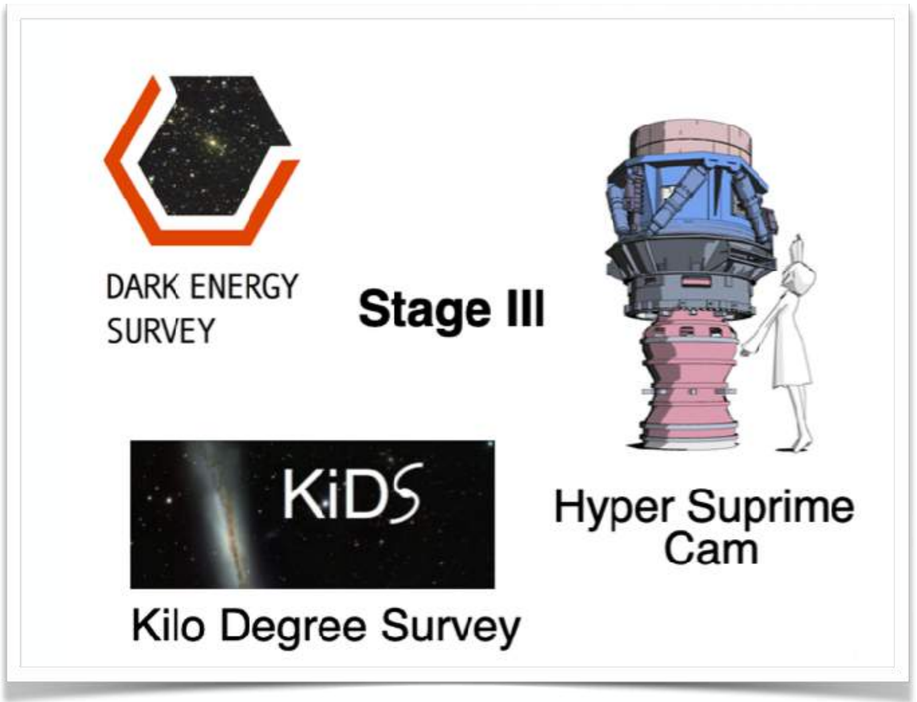


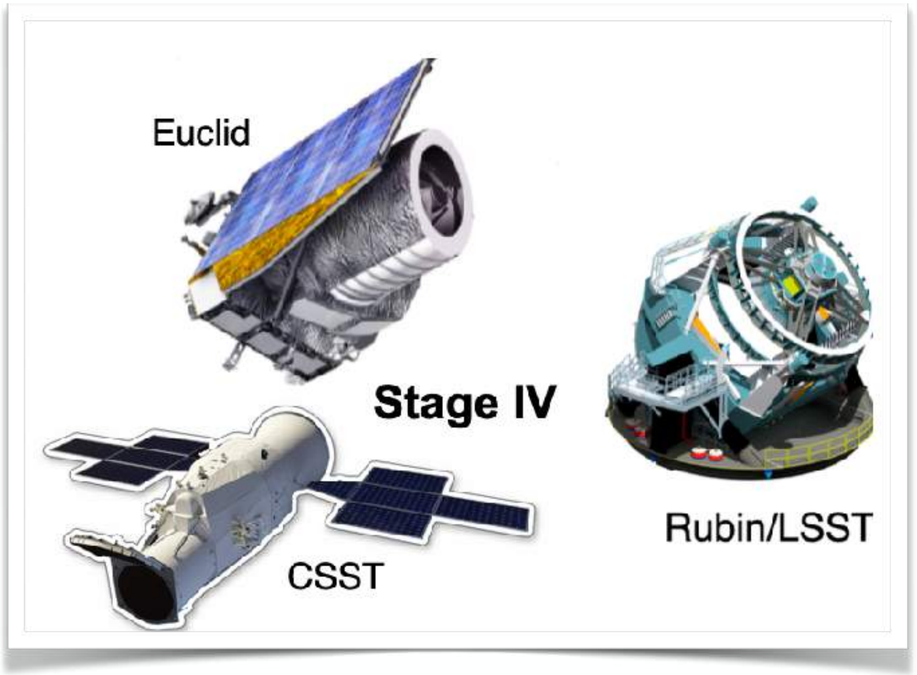
CSST

SBI for Stage IV surveys

Key improvements required:

1. Increased number of particles for lower lensing map noise
2. Increased number of shells for more precise ray-tracing
3. Realistic galaxy properties for understanding survey selection function (luminosity functions, colours, sizes)
4. Spatially varying redshift distributions
5. Halo shapes for intrinsic alignment models
6. Improved modelling for galaxy clustering, implement multiple models in literature:
    1. Lower mass halos for better SHAM models
    2. Store more projected field information for EPT/LPT/ EFT clustering models
7. **More to be investigated!**





The CosmoGridV2 Project

- CosmoGridV1 was built for Stage-3 surveys requirements
- SBI theory prediction for Stage-4 surveys, meeting their precision requirements
- Dark matter -only, skeleton for adding LSS tracers
- Ready for LSS combined probes
- Enough simulations to capture cosmic variance on the map level for human and AI features
- Collaboration of several groups, including ETH Zurich, Uni Zurich, Uni Manchester, Uni Newcastle, LMU Observatory
- Open source and open data project, useful for all LSS surveys: Euclid, LSST, CSST
- Possibly several proposals for different HPC centres
- **Participation welcome!** We are publishing a CosmoGridV2 newsletter, I can add you to the mailing list if interested

Weak lensing

- Lightcone Healpix shells
- Baryon feedback using baryonification models or AI-based “painting”
- Enough data products to implement multiple intrinsic alignment methods: Tidal Alignment Tidal Torque (TATT, based on EPT), halo model IA

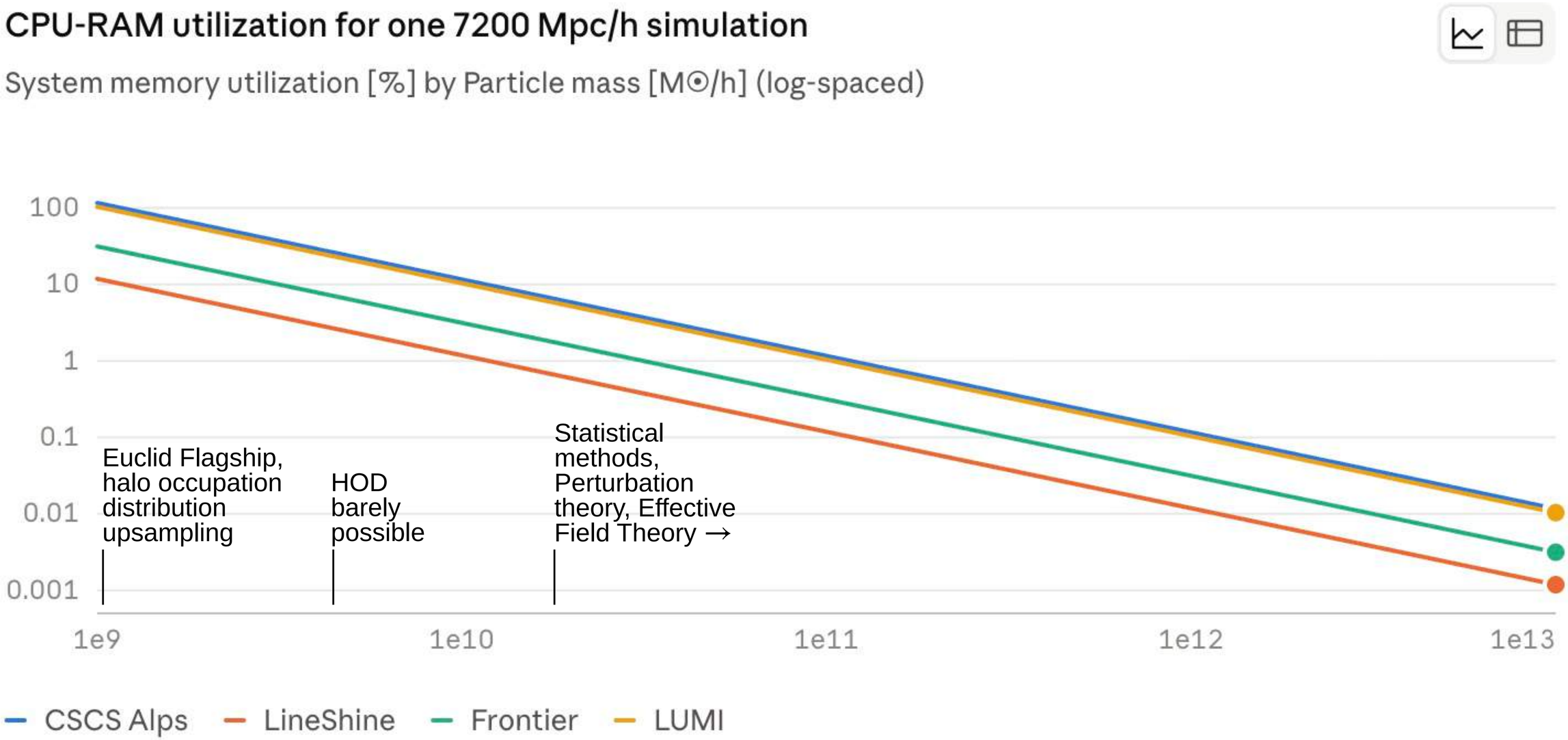
Galaxy clustering

- Halo substructure catalogs
- Enough data products to implement multiple galaxy-halo connection models: HOD, SHAM, EPT/LPT, non-local bias, possibly EFT
- Galaxy properties (luminosity functions, colors, etc) added later on-the-fly, marginalise in joint inference with cosmology

Additional probes

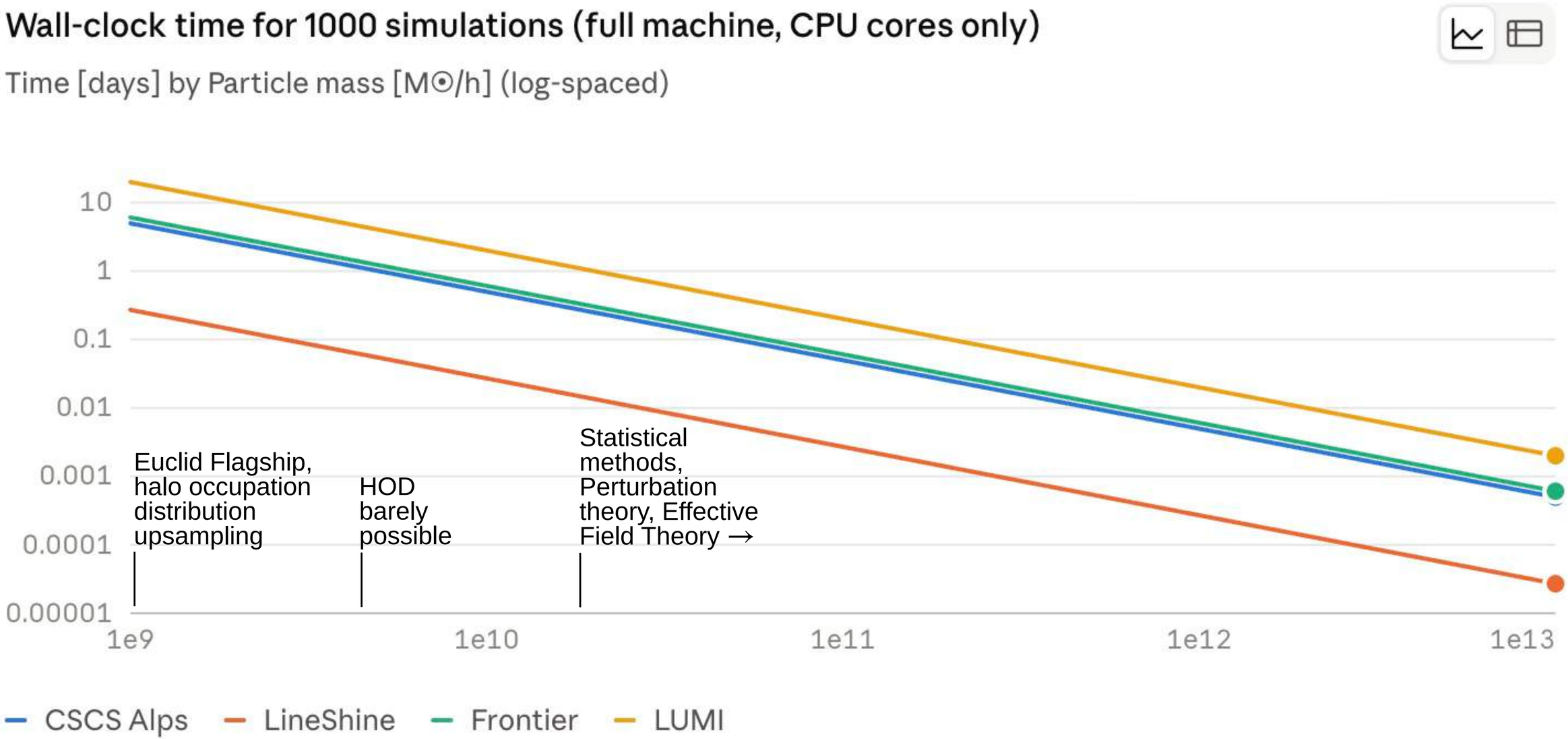
- tSZ/kSZ, probe combination with CMB
- X-ray intensities, combination with eRosita

What will we need?



Assuming 48 bytes per particle, but we will probably need 64…

What will we need?



Assuming 1000 cosmologies, but we will probably need at least 10000…

The way forward →

- → Moving towards **Computational Cosmology with AI Foundation Models** enrichment
- → Reaching the **information floor** of cosmological datasets with AI- based parameter inference
- → Building large simulations in a **collaborative way**, publishing data sets to the community
- → Using large supercomputing facilities to produce open-data simulated theory prediction
- → Capitalising on **latest advances in AI** in practical cosmological measurements

