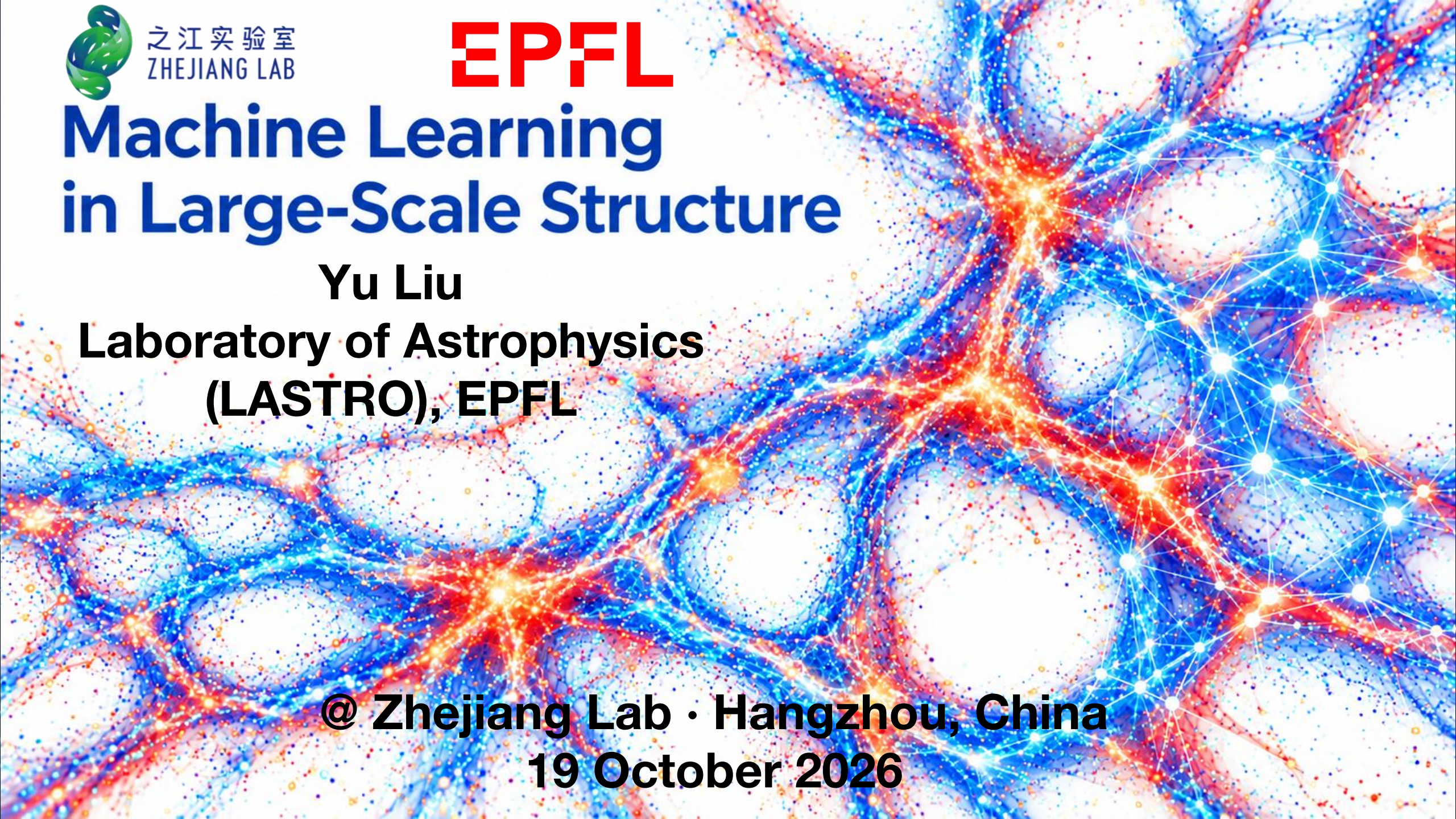
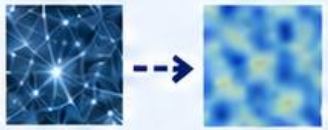
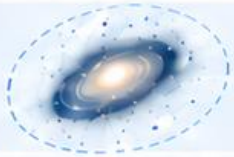
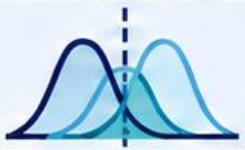

之江实验室
ZHEJIANG LAB
EPFL
Machine Learning
in Large-Scale Structure
Yu Liu
Laboratory of Astrophysics
(LASTRO), EPFL
@ Zhejiang Lab · Hangzhou, China
19 October 2026

Machine Learning in LSS

Overcoming key bottlenecks in cosmological analysis

SCIENTIFIC BOTTLENECKS	HOW MACHINE LEARNING CAN HELP
 **Expensive simulations and predictions**	Accelerate forward modelling with emulators, surrogate models and fast simulation methods.
 **Incomplete extraction of cosmological information**	Capture nonlinear and non-Gaussian information beyond conventional two-point statistics.
 **Nonlinear evolution and hidden cosmic fields**	Reconstruct initial conditions, continuous dark-matter density fields and velocity fields.
 **Complex astrophysical modelling**	Develop flexible models of galaxy–halo connections and baryonic effects, while accounting for their uncertainties.
 **Cosmological parameter inference**	Enable efficient simulation-based inference using learned data compression and neural likelihood or posterior estimation.
 **Large data volumes and complex survey effects**	Support scalable classification, redshift estimation, calibration and observational-systematics modelling.

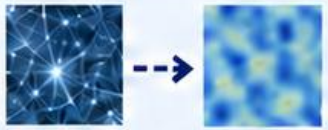
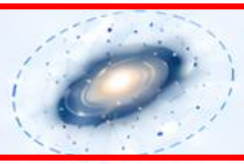
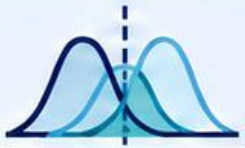


SCIENTIFIC GOAL

Extract more reliable cosmological information at an affordable computational cost.

Machine Learning in LSS

Overcoming key bottlenecks in cosmological analysis

SCIENTIFIC BOTTLENECKS	HOW MACHINE LEARNING CAN HELP
 **Expensive simulations and predictions**	Accelerate forward modelling with emulators, surrogate models and fast simulation methods.
 **Incomplete extraction of cosmological information**	Capture nonlinear and non-Gaussian information beyond conventional two-point statistics.
 **Nonlinear evolution and hidden cosmic fields**	Reconstruct initial conditions, continuous dark-matter density fields and velocity fields.
 **Complex astrophysical modelling**	Develop flexible models of galaxy–halo connections and baryonic effects, while accounting for their uncertainties.
 **Cosmological parameter inference**	Enable efficient simulation-based inference using learned data compression and neural likelihood or posterior estimation.
 **Large data volumes and complex survey effects**	Support scalable classification, redshift estimation, calibration and observational-systematics modelling.



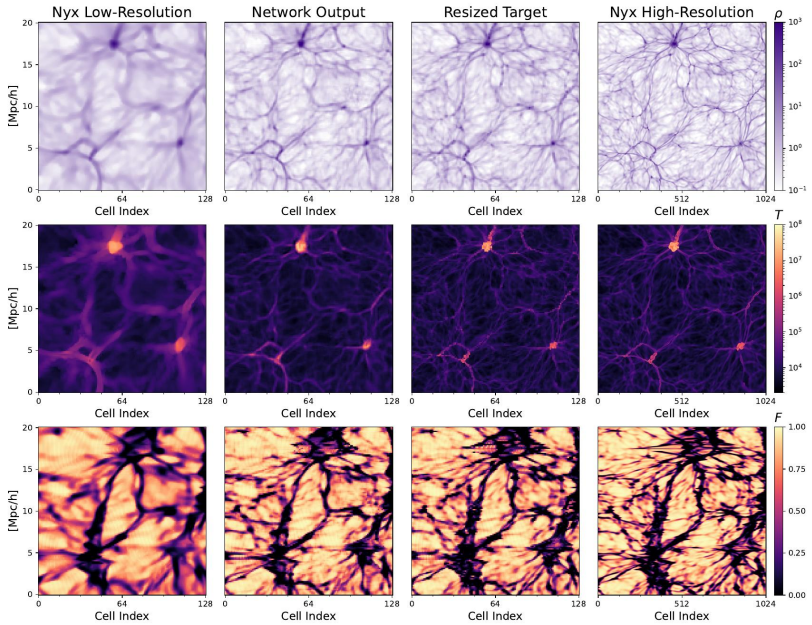
SCIENTIFIC GOAL

Extract more reliable cosmological information at an affordable computational cost.

Machine learning for cosmological simulations

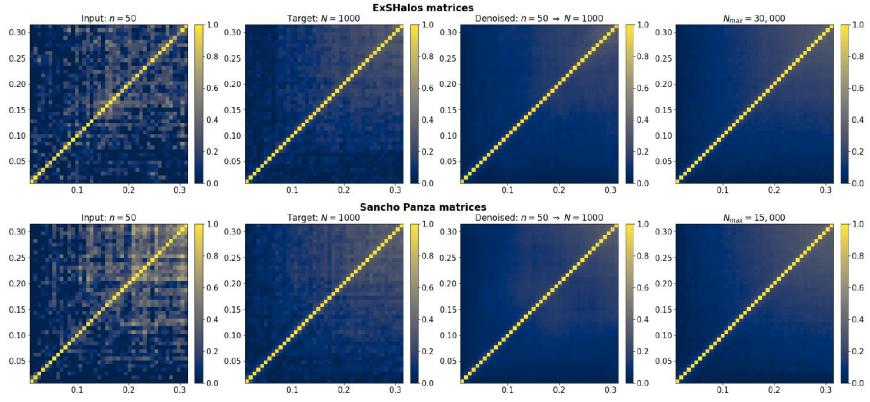
Resolution enhancement, fast physical models and covariance denoising

GAN-enhanced gas and Lyα fields



Jacobus et al. (2023), Fig. 1 arXiv:2308.02637

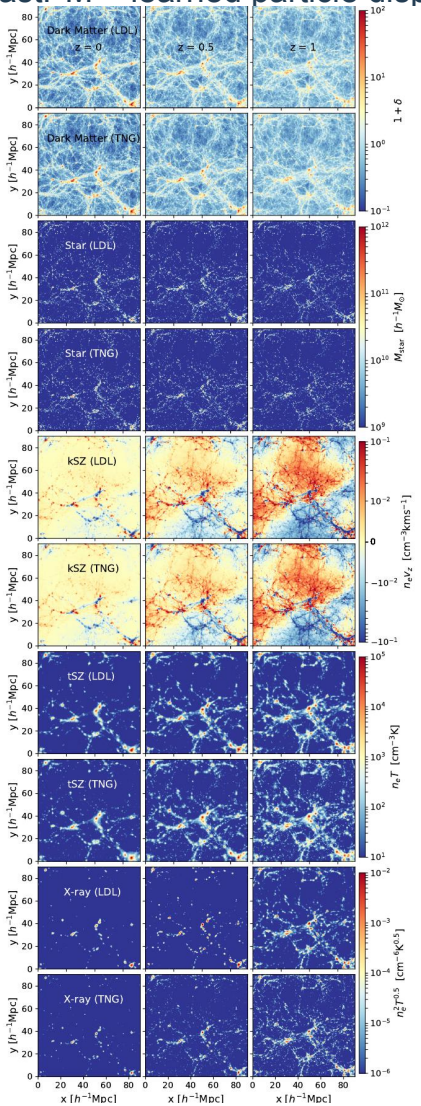
CNN covariance denoising



de Santi & Abramo (2022), Fig. 4 arXiv:2205.10881

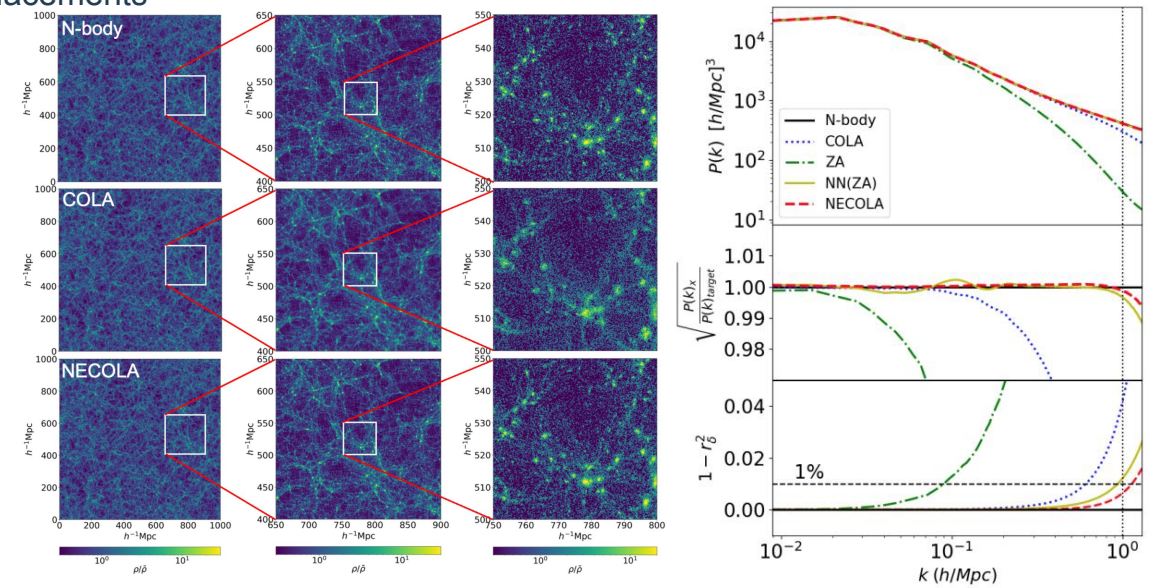
LDL: effective baryonic physics

FastPM + learned particle displacements



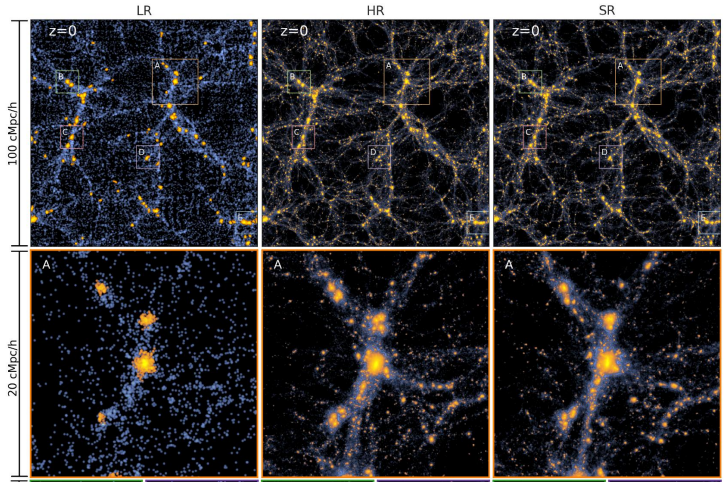
Dai & Seljak (2021), Fig. 1 arXiv:2010.02926

NECOLA: neural corrections to COLA



Kaushal et al. (2022), arXiv Fig. 1 arXiv:2111.02441

GAN-based particle super-resolution



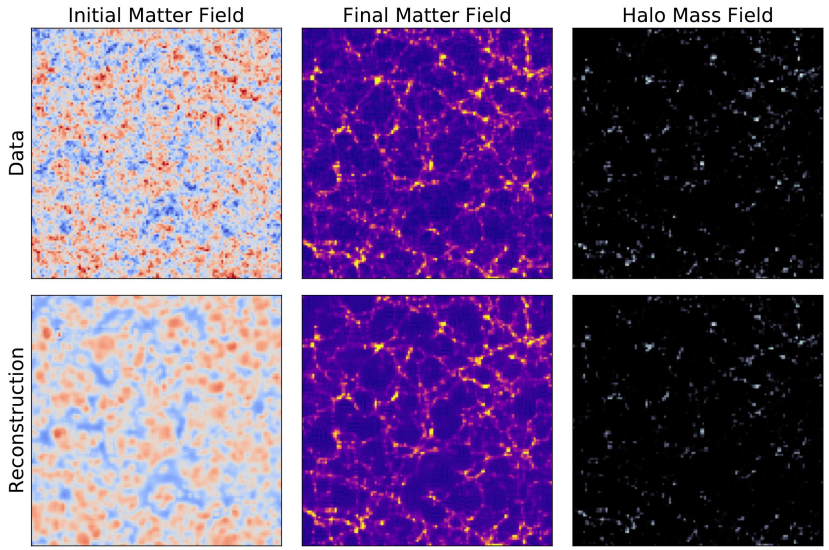
Li et al. (2021), Fig. 1 (upper rows) arXiv:2010.06608

ML-Assisted Reconstruction via Forward Modelling

Infer initial conditions by iteratively comparing forward predictions with simulated data

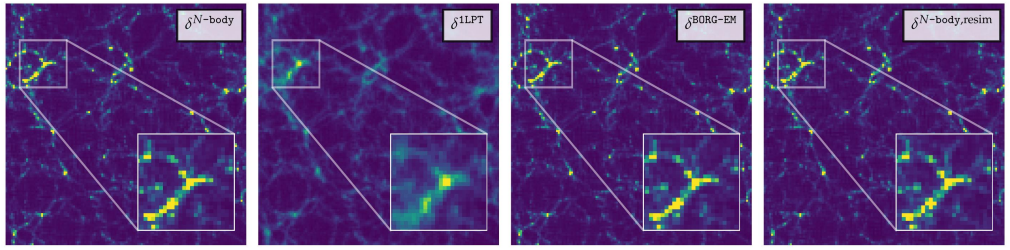
Learning the matter–halo connection

FastPM forward model + learned halo relation (MAP)



Modi, Feng & Seljak (2018), Fig. 9 arXiv:1805.02247

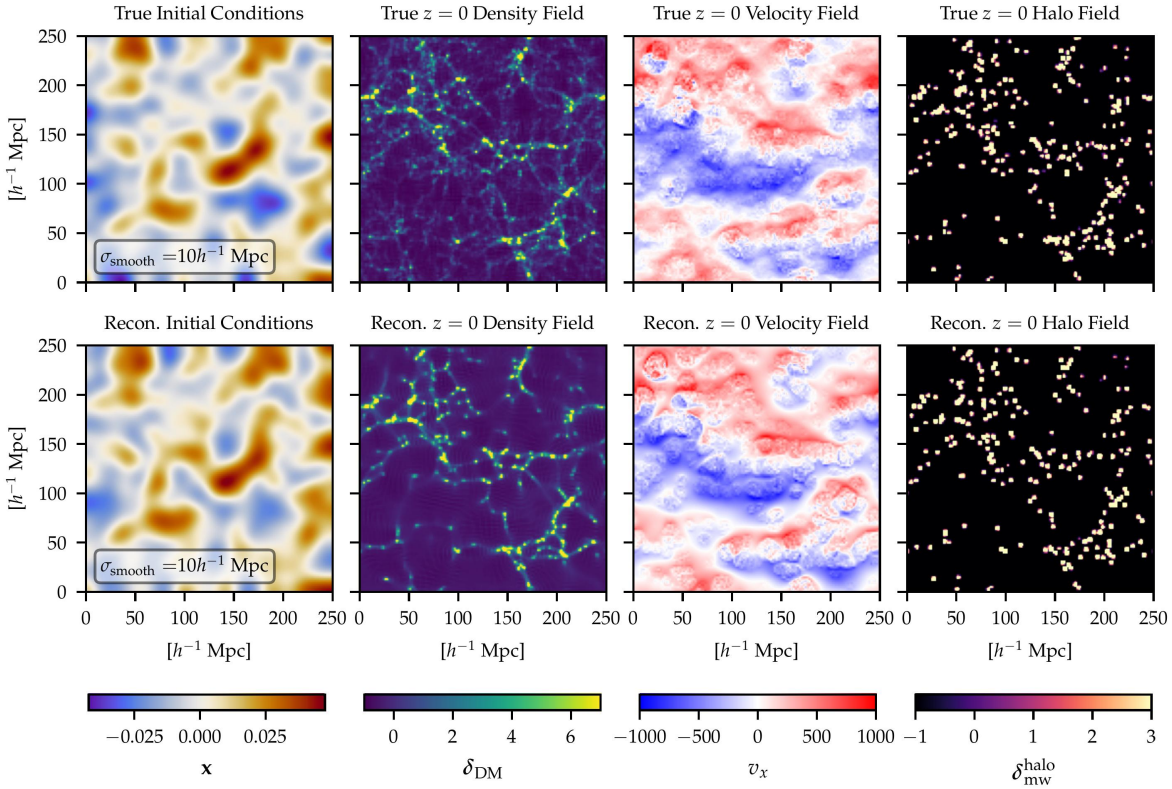
BORG-EM: gravity emulator + HMC



Doeser et al. (2024), Fig. 1 arXiv:2312.09271

LULO: learning to optimize initial conditions

Learned updates with full N-body simulations and halo finding in the forward model



Doeser, Ata & Jasche (2025), Fig. 4 arXiv:2502.13243

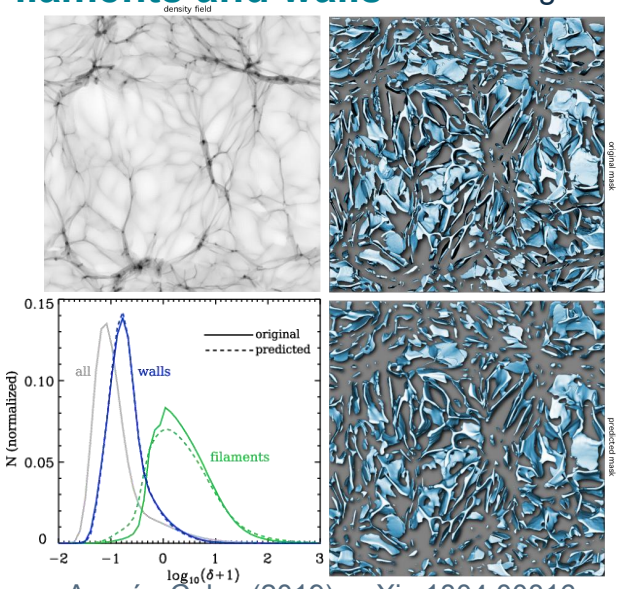
Related: CosmicRIM (Modi et al. 2021)

Learned MAP optimization arXiv:2104.12864

Machine learning for cosmic-structure identification

Cosmic-web classification and halo finding in simulations, with galaxy-group applications to survey data

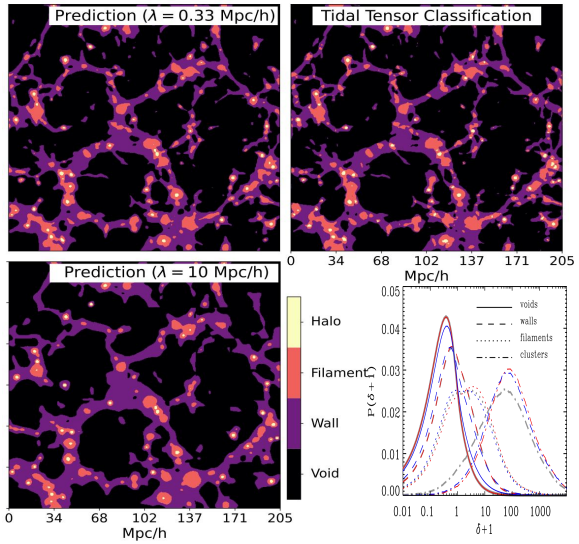
Filaments and walls U-Net segmentation



Aragón-Calvo (2019), arXiv:1804.00816

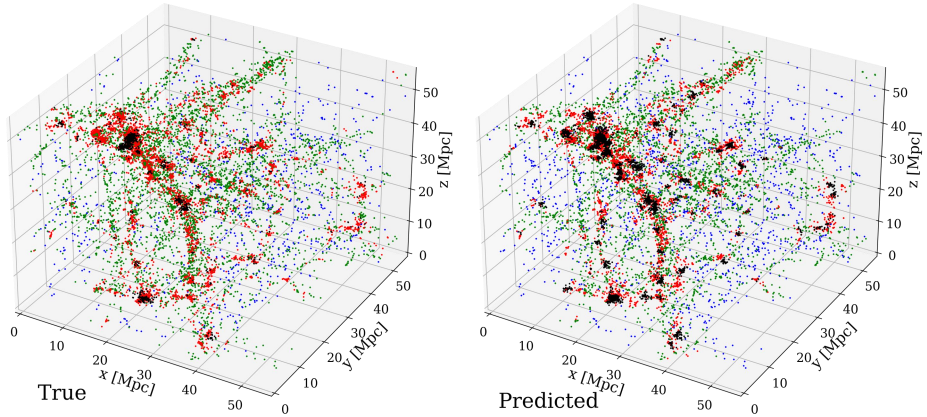
Voids and web environments

DeepVoid: U-Net + curriculum learning



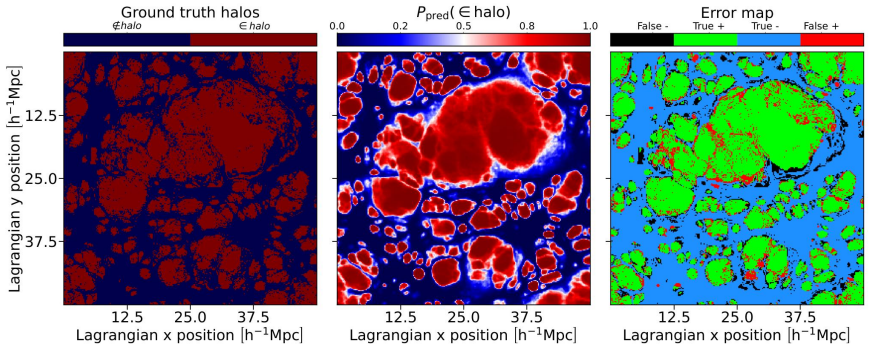
Kumagai et al., arXiv:2504.21134

Galaxy environments 3D CNN classification



Inoue et al. (2022), arXiv:2207.09072

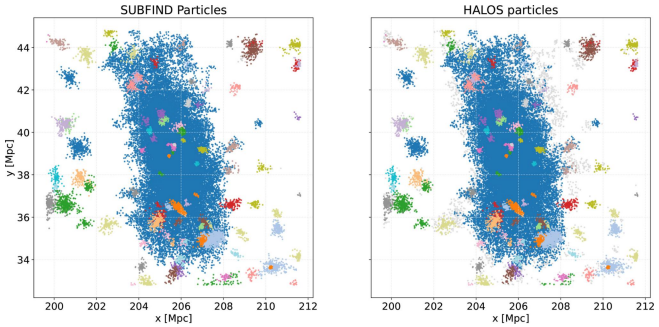
Protohalo membership V-Net from initial conditions



López-Cano et al. (2024), arXiv:2311.12110

Subhalo identification

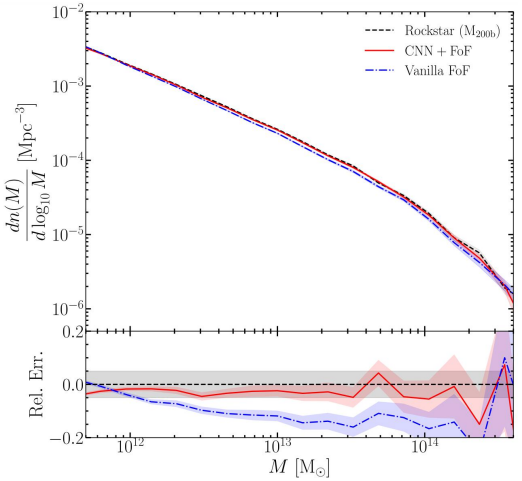
HALOS: MLP + HDBSCAN



Spampinato et al. (2026), A&C 56, 101114

Halo catalogues

CNN + FoF

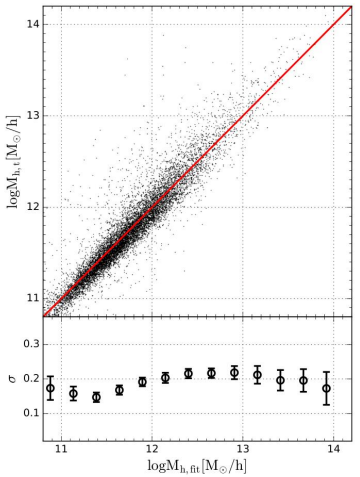


arXiv:2602.21246

Galaxy groups

FoF + random forests

zCOSMOS mock validation

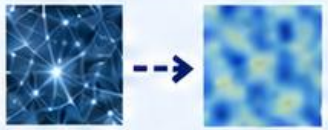
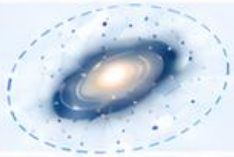
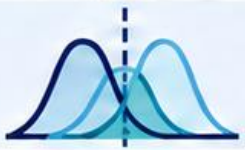


arXiv:2006.05426

Related: YOLO-CL (SDSS clusters), arXiv:2301.09657. GAT+ (simulated galaxy environments), arXiv:2512.05909.

Machine Learning in LSS

Overcoming key bottlenecks in cosmological analysis

SCIENTIFIC BOTTLENECKS	HOW MACHINE LEARNING CAN HELP
 **Expensive simulations and predictions**	Accelerate forward modelling with emulators, surrogate models and fast simulation methods.
 **Incomplete extraction of cosmological information**	Capture nonlinear and non-Gaussian information beyond conventional two-point statistics.
 **Nonlinear evolution and hidden cosmic fields**	Reconstruct initial conditions, continuous dark-matter density fields and velocity fields.
 **Complex astrophysical modelling**	Develop flexible models of galaxy–halo connections and baryonic effects, while accounting for their uncertainties.
 **Cosmological parameter inference**	Enable efficient simulation-based inference using learned data compression and neural likelihood or posterior estimation.
 **Large data volumes and complex survey effects**	Support scalable classification, redshift estimation, calibration and observational-systematics modelling.



SCIENTIFIC GOAL

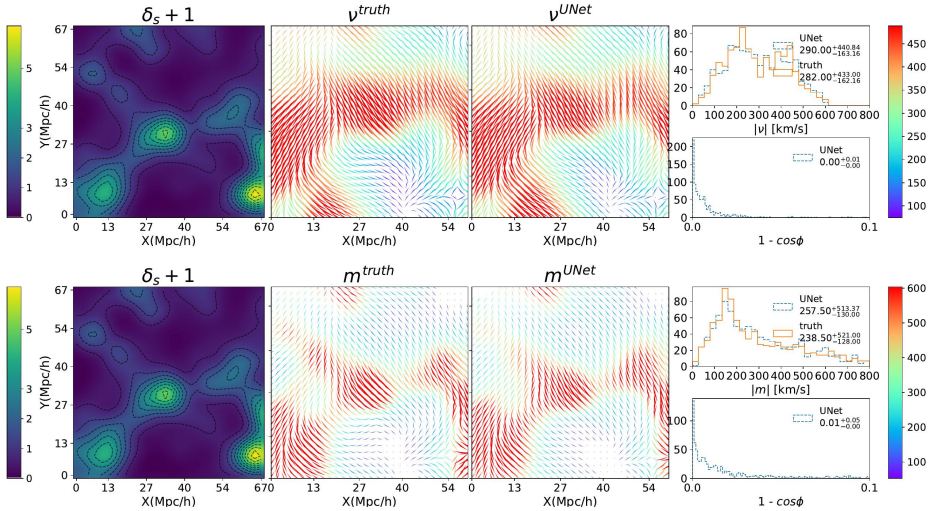
Extract more reliable cosmological information at an affordable computational cost.

Machine learning for cosmic-field reconstruction

Supervised reconstruction of late-time density and velocity fields from different tracers

UNet: dark-matter velocity and momentum

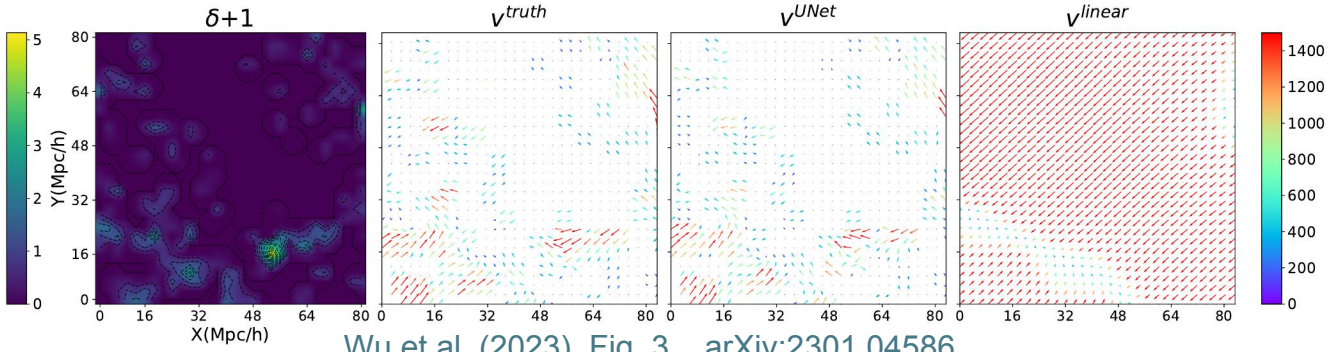
Velocity (top) and momentum (bottom) from sparse halos



Xiao et al. (2024), Fig. 4 arXiv:2411.11280

UNet: halo peculiar velocities

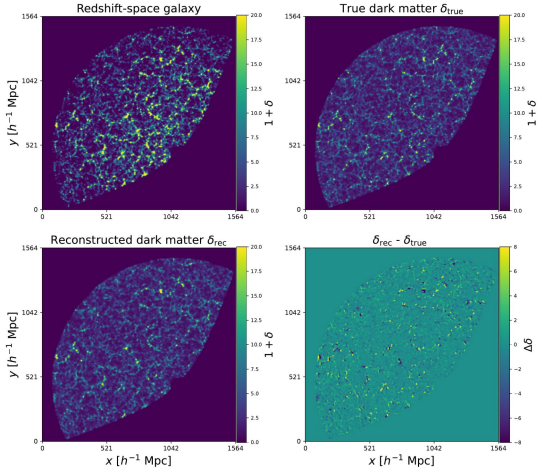
Nonlinear flow patterns beyond the linear prediction



Wu et al. (2023), Fig. 3 arXiv:2301.04586

DarkAI: 3D U-Net for matter density

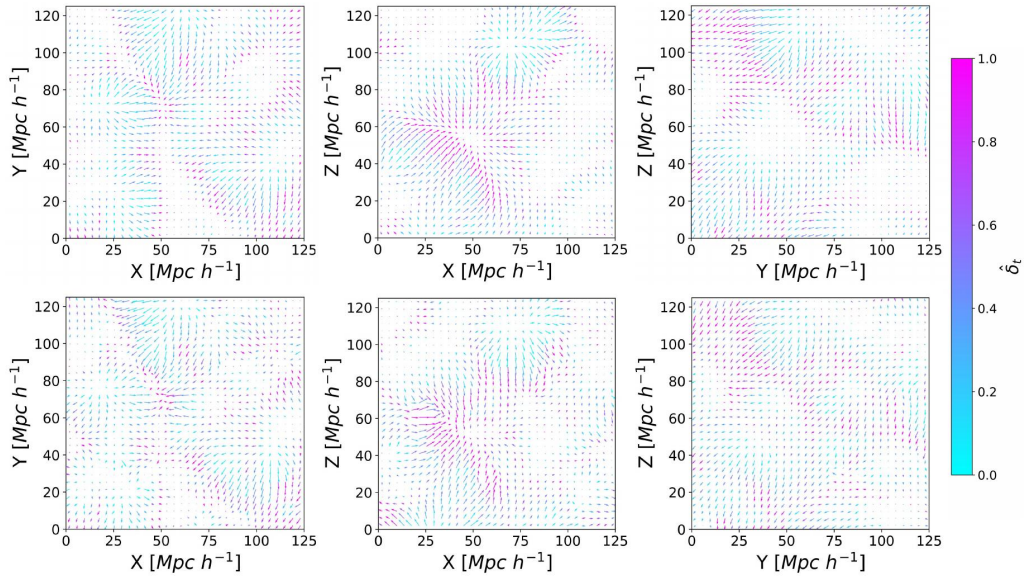
Supervised regression using DESI-like galaxy mocks with survey selection and RSD



Shi et al. (2025), Fig. 7
arXiv:2501.12621

Earlier DarkAI: Wang et al. (2024)
arXiv:2305.11431

V-net: density and velocity fields



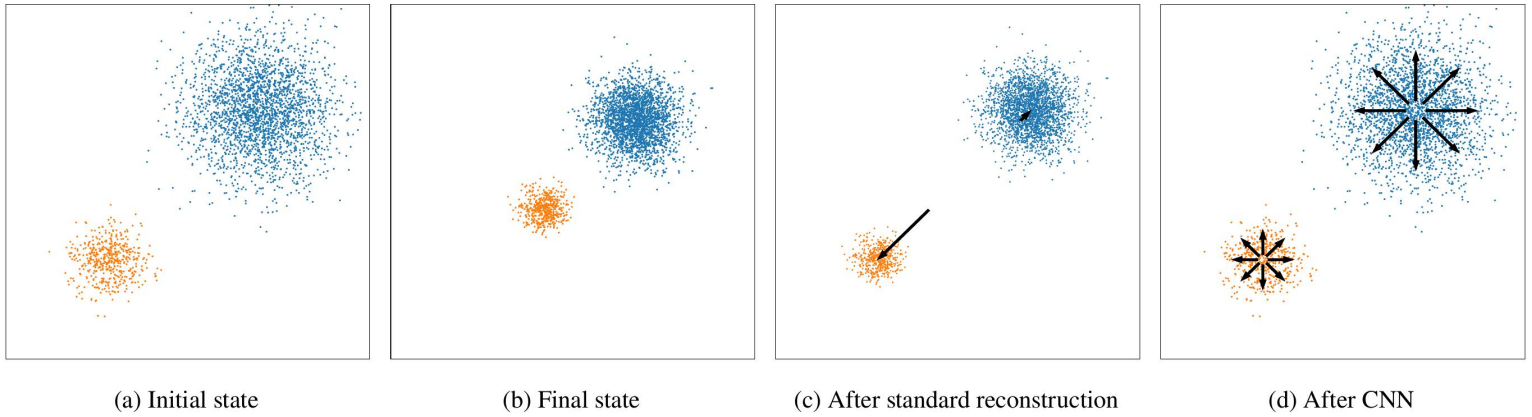
Qin et al. (2023), Fig. 13 (velocity panels) arXiv:2302.02087

Reconstructed (top) and true (bottom) velocities

Machine learning for initial-condition reconstruction

Standard reconstruction handles bulk flows. CNNs learn the remaining nonlinear mapping.

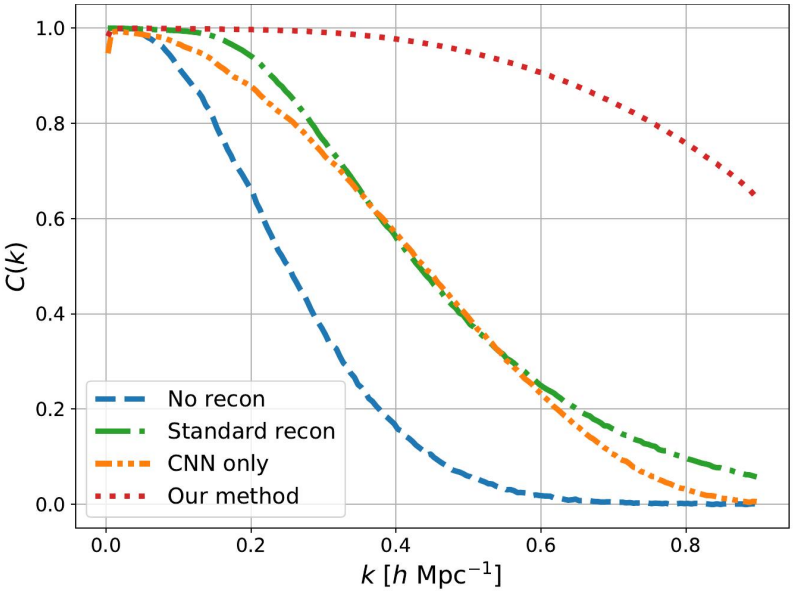
Standard reconstruction + CNN



(a) Initial state (b) Final state (c) After standard reconstruction (d) After CNN

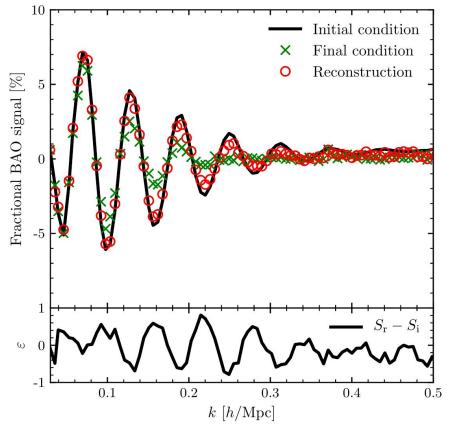
Shallue & Eisenstein (2023), Fig. 1 arXiv:2207.12511

Recovery of initial modes



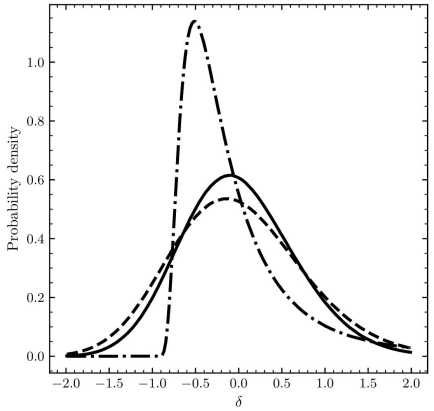
Same study, Fig. 2 (left), real-space matter

BAO contrast recovery



Mao et al. (2021), Fig. 6

A more Gaussian one-point PDF



Mao et al. (2021), Fig. 3

Dashed: initial
Solid: reconstructed
Dash-dot: final

A Gaussian-looking PDF alone does not prove a Gaussian random field.

Related developments

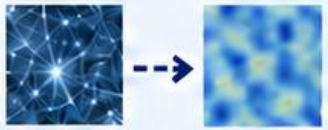
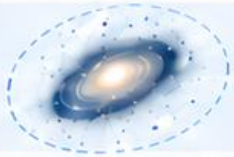
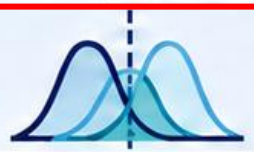
Chen et al. (2023): robustness tests
arXiv:2306.10538

Parker et al. (2025): biased tracers
arXiv:2504.01092

Zang et al. (2026): BOSS DR12
arXiv:2609.15943

Machine Learning in LSS

Overcoming key bottlenecks in cosmological analysis

SCIENTIFIC BOTTLENECKS	HOW MACHINE LEARNING CAN HELP
 **Expensive simulations and predictions**	Accelerate forward modelling with emulators, surrogate models and fast simulation methods.
 **Incomplete extraction of cosmological information**	Capture nonlinear and non-Gaussian information beyond conventional two-point statistics.
 **Nonlinear evolution and hidden cosmic fields**	Reconstruct initial conditions, continuous dark-matter density fields and velocity fields.
 **Complex astrophysical modelling**	Develop flexible models of galaxy–halo connections and baryonic effects, while accounting for their uncertainties.
 **Cosmological parameter inference**	Enable efficient simulation-based inference using learned data compression and neural likelihood or posterior estimation.
 **Large data volumes and complex survey effects**	Support scalable classification, redshift estimation, calibration and observational-systematics modelling.



SCIENTIFIC GOAL

Extract more reliable cosmological information at an affordable computational cost.

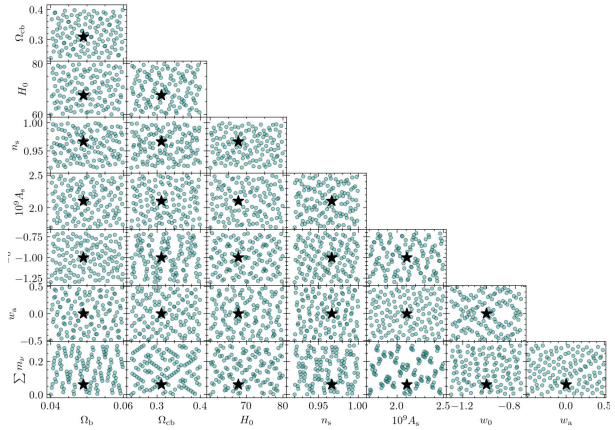
Cosmological emulators

Fast theory predictions for cosmological parameter inference

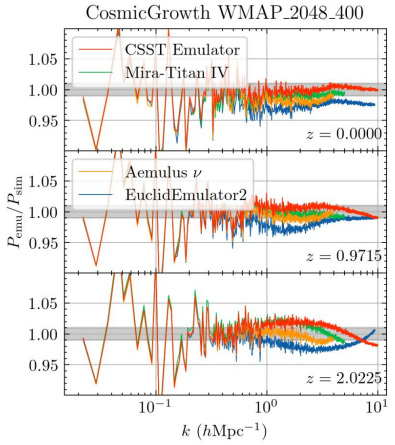
Other related works, e.g., arXiv:2109.04984, arXiv:1508.02654, arXiv:2105.01081, arXiv:1705.03388, arXiv:1804.05867, arXiv:2306.03144, arXiv:2206.04594

Nonlinear matter power spectra

CSST Emulator PCA + Gaussian processes



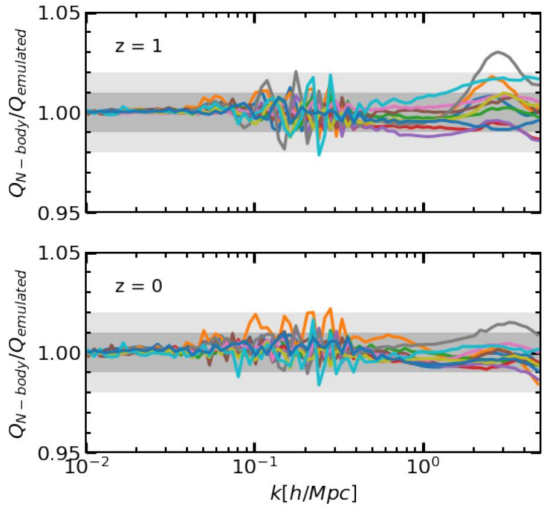
129 training cosmologies



External validation

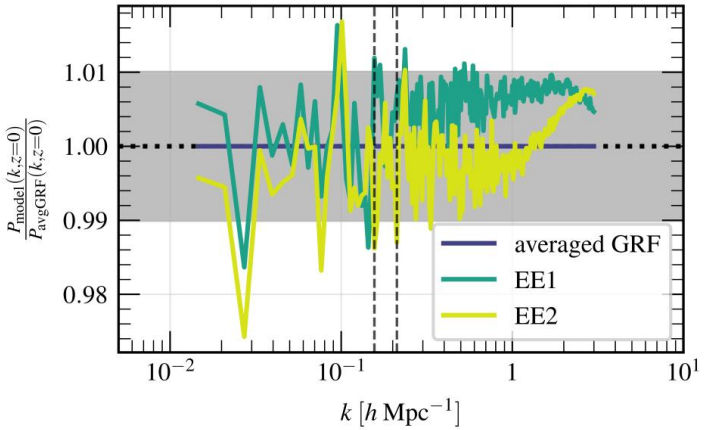
Chen et al. (2025) arXiv:2502.11160

BACCO Rescaling + PCA + GPR



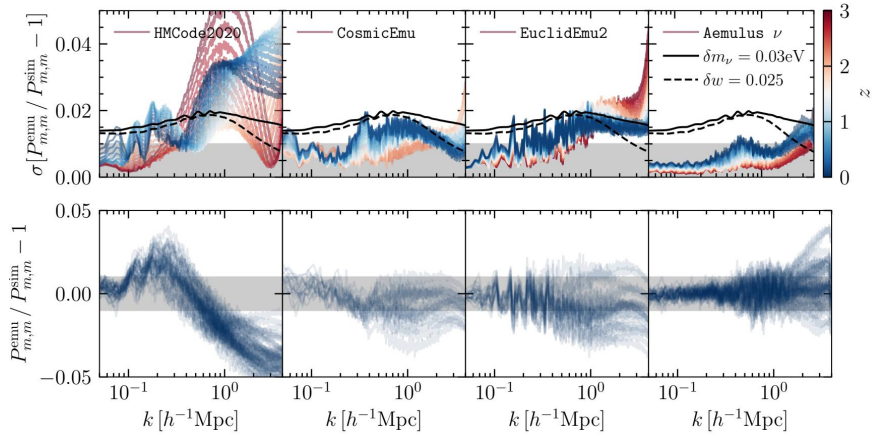
Angulo et al. (2021) arXiv:2004.06245

EuclidEmulator2 PCA + polynomial chaos



Knabenhans et al. (2021) arXiv:2010.11288

Aemulus ν PCA + polynomial chaos

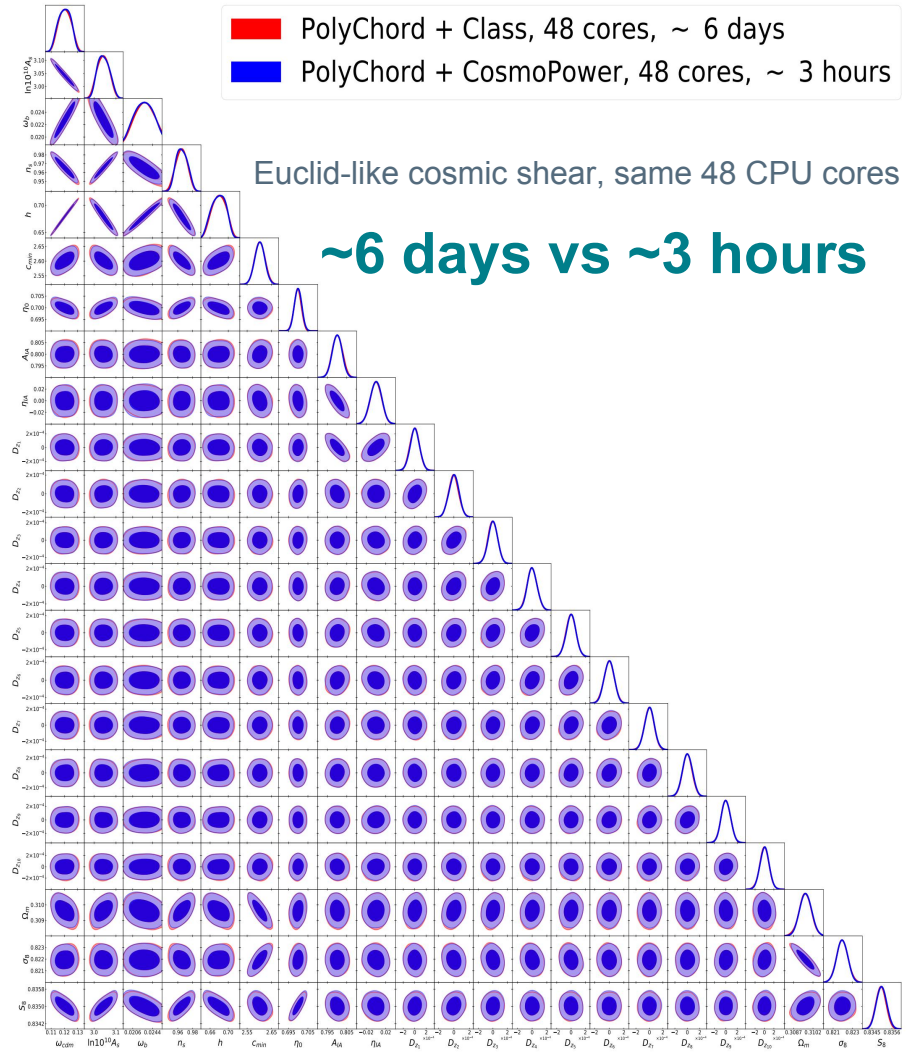


DeRose et al. (2023) arXiv:2303.09762

Reliable inference requires validated emulator accuracy and reliable training simulations.

Accelerated Bayesian inference

CosmoPower Neural-network power spectra



Euclid-like cosmic shear, same 48 CPU cores

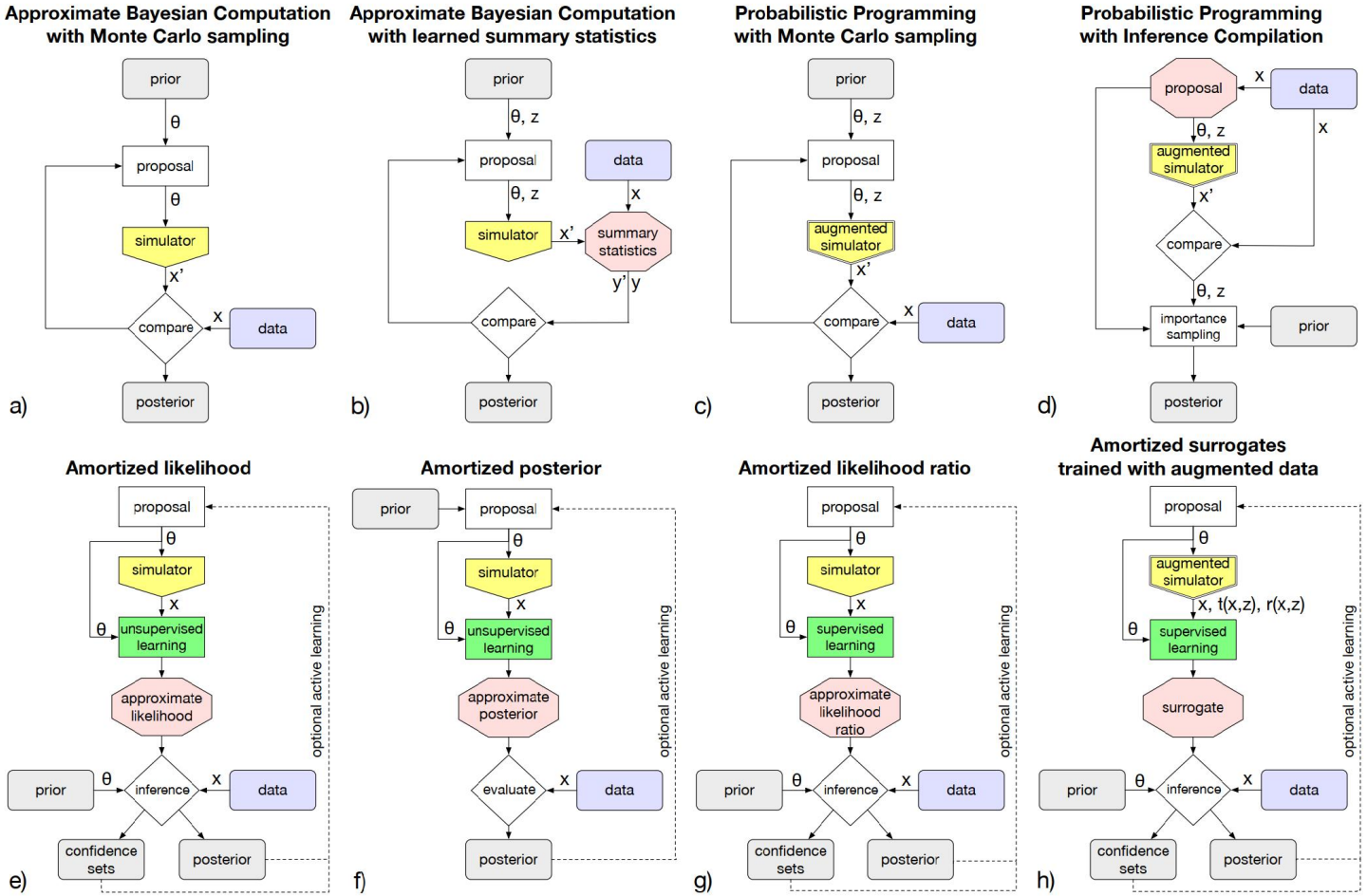
~6 days vs ~3 hours

Spurio Mancini et al. (2022) arXiv:2106.03846

Learned data compression and neural SBI

Compression retains parameter-sensitive features. Neural SBI learns the probability model for inference.

Overview of different approaches to simulation-based inference



Learned compression

Summarize a field or statistic with a compact feature vector s = f(x), optimized for the parameters of interest.

REVIEW CONTRIBUTION

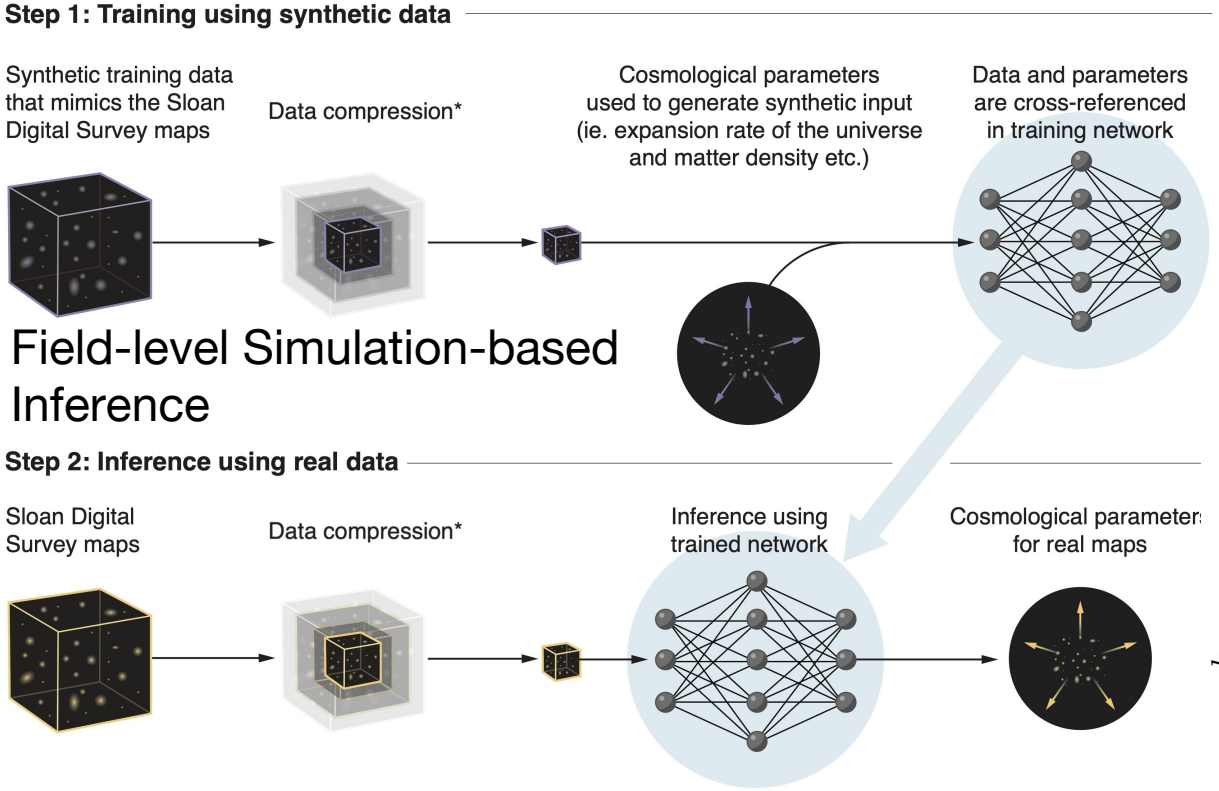
Connects simulation, learned summaries and probabilistic inference in a common methodological framework.

SimBIG uses neural posterior estimation.

SBI avoids evaluating the simulator likelihood directly. Its validity still depends on the simulator.

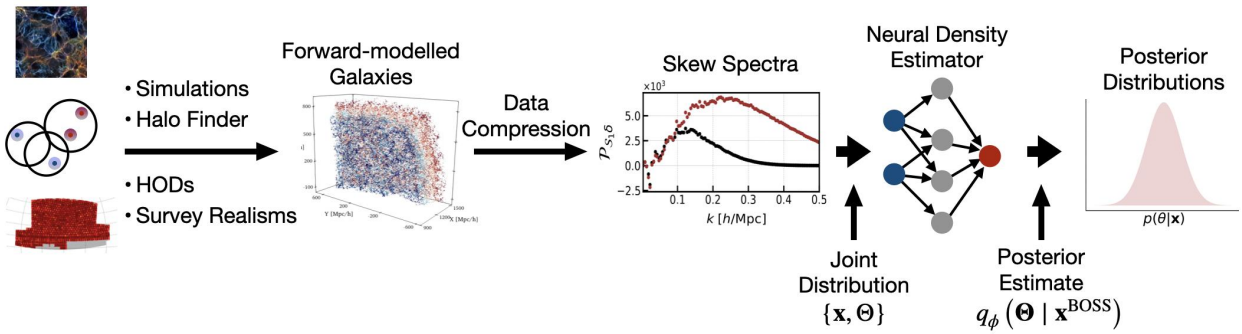
Cranmer, Brehmer & Louppe (2020), PNAS 117, 30055–30062. Selected workflows. Published version: Fig. 1(E–G).
DOI: 10.1073/pnas.1912789117. Original figure 3(e–g) from arXiv:1911.01429v3. Selected panels.

SimBIG: simulation-based cosmological inference



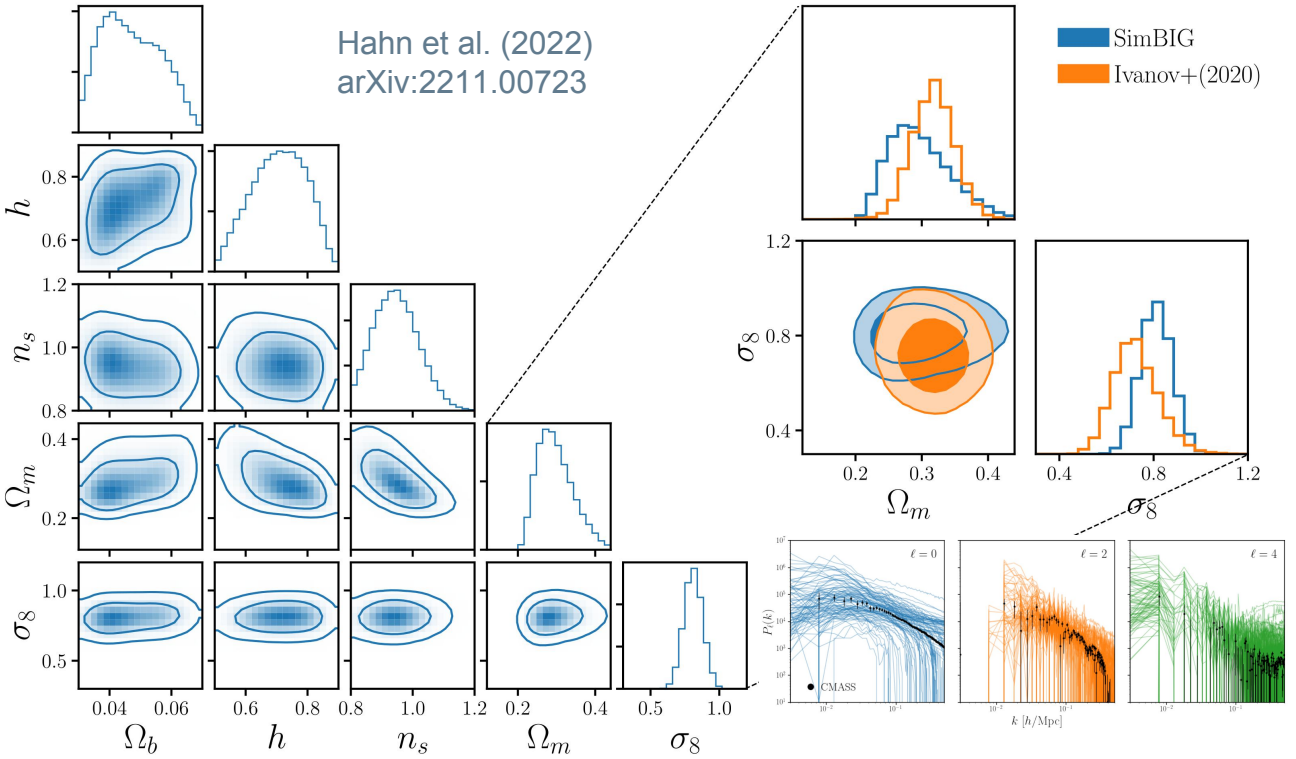
Lemos et al. (2023), Fig. 1 arXiv:2310.15256

Schematic illustrating the various elements of the SimBIG forward modelling pipeline



Schematic view of the SimBIG pipeline

Hou et al. (2024), Fig. 1 arXiv:2401.15074

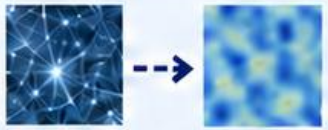
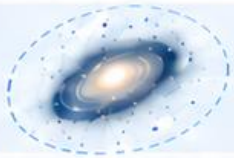
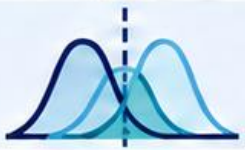


Other related works

arXiv:2211.00660, arXiv:2507.03086, arXiv:2404.04228, arXiv:1903.00007

Machine Learning in LSS

Overcoming key bottlenecks in cosmological analysis

SCIENTIFIC BOTTLENECKS	HOW MACHINE LEARNING CAN HELP
 **Expensive simulations and predictions**	Accelerate forward modelling with emulators, surrogate models and fast simulation methods.
 **Incomplete extraction of cosmological information**	Capture nonlinear and non-Gaussian information beyond conventional two-point statistics.
 **Nonlinear evolution and hidden cosmic fields**	Reconstruct initial conditions, continuous dark-matter density fields and velocity fields.
 **Complex astrophysical modelling**	Develop flexible models of galaxy–halo connections and baryonic effects, while accounting for their uncertainties.
 **Cosmological parameter inference**	Enable efficient simulation-based inference using learned data compression and neural likelihood or posterior estimation.
 **Large data volumes and complex survey effects**	Support scalable classification, redshift estimation, calibration and observational-systematics modelling.



SCIENTIFIC GOAL

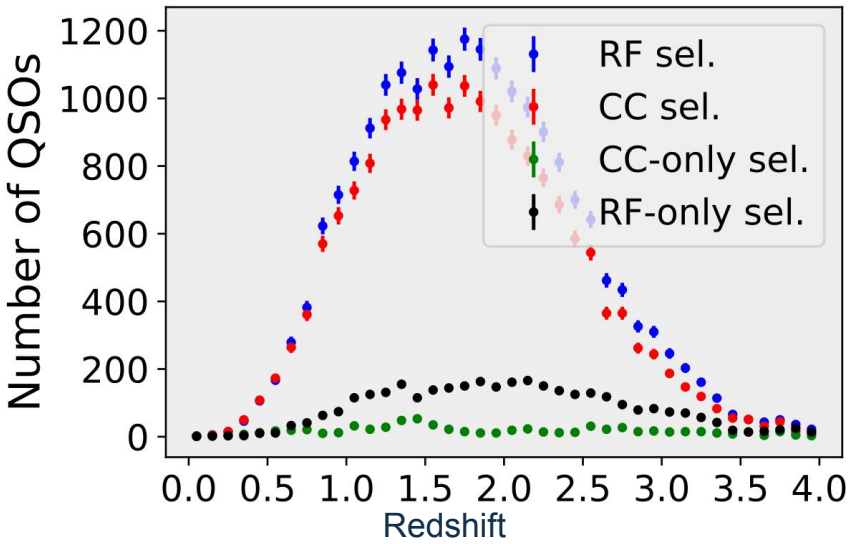
Extract more reliable cosmological information at an affordable computational cost.

Machine learning in galaxy-survey analysis

Target selection, photometric redshifts and imaging-systematics mitigation

Quasar target selection

Random forest on optical–IR photometry
DESI quasar targeting



+15%

confirmed quasars overall

+21% at z > 2.1

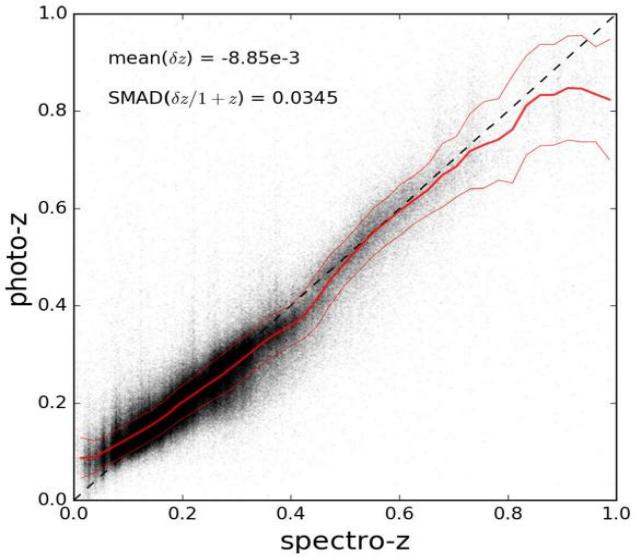
Versus colour cuts at matched target density (approximately 310 targets per deg²).

Chaussidon et al. (2023), arXiv:2208.08511

Related: arXiv:1011.6392 (XDQSO)
arXiv:2010.11280 (DESI random forest)

Photometric redshifts

ANNz2 neural ensembles on ugri photometry
KiDS DR3



39.2 million

sources in the public catalogue
20.5 million in the fiducial sample

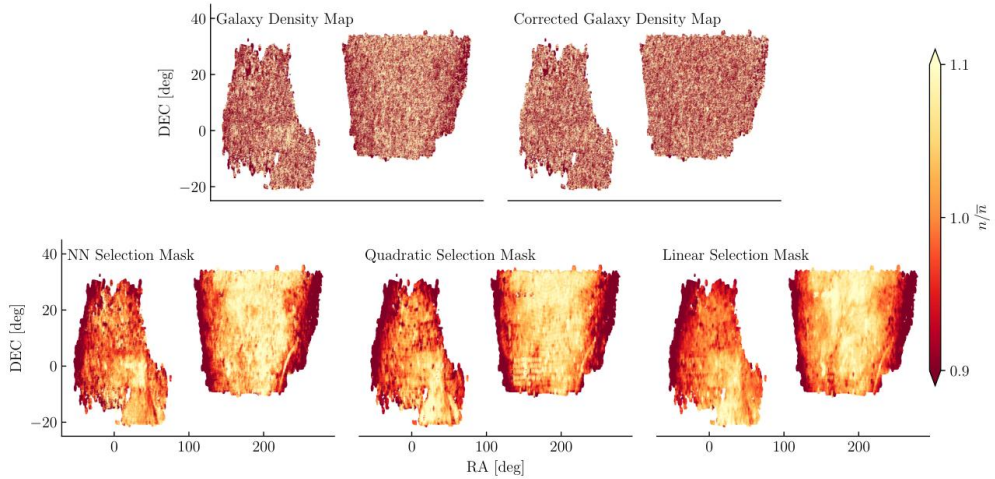
Training overlap in this plot. Limited coverage restricts extrapolation (photo-z ≲ 0.9, r ≲ 23.5).

Bilicki et al. (2018), arXiv:1709.04205

Related: arXiv:1604.03593 (GPz)
arXiv:2410.19390 (CLAP)

Imaging systematics

Neural correction weights from imaging maps
DECaLS DR7 emission-line-galaxy targets



Observed and corrected density maps with learned selection masks

≈12× lower

residual χ²: 9567 to 767
Density trends within the mock footprint

Residual systematics remain. Mock tests must check for cosmological signal loss.

Rezaie et al. (2020), arXiv:1907.11355

Related: arXiv:2108.03640 (DESI QSOs)
arXiv:2411.12024 (DESI ELGs)

Future plans with machine learning

Building on my work in reconstruction and non-Gaussian statistics

MY CURRENT WORK	PLANNED AI EXTENSION
Biased-tracer reconstruction	Learn corrections to improve recovery of initial conditions and BAO information.
DTFE Minkowski functionals	Build morphology emulators across cosmology and nuisance parameters.
Optimized void statistics	Build emulators for optimized void observables and characterize their statistical uncertainty.

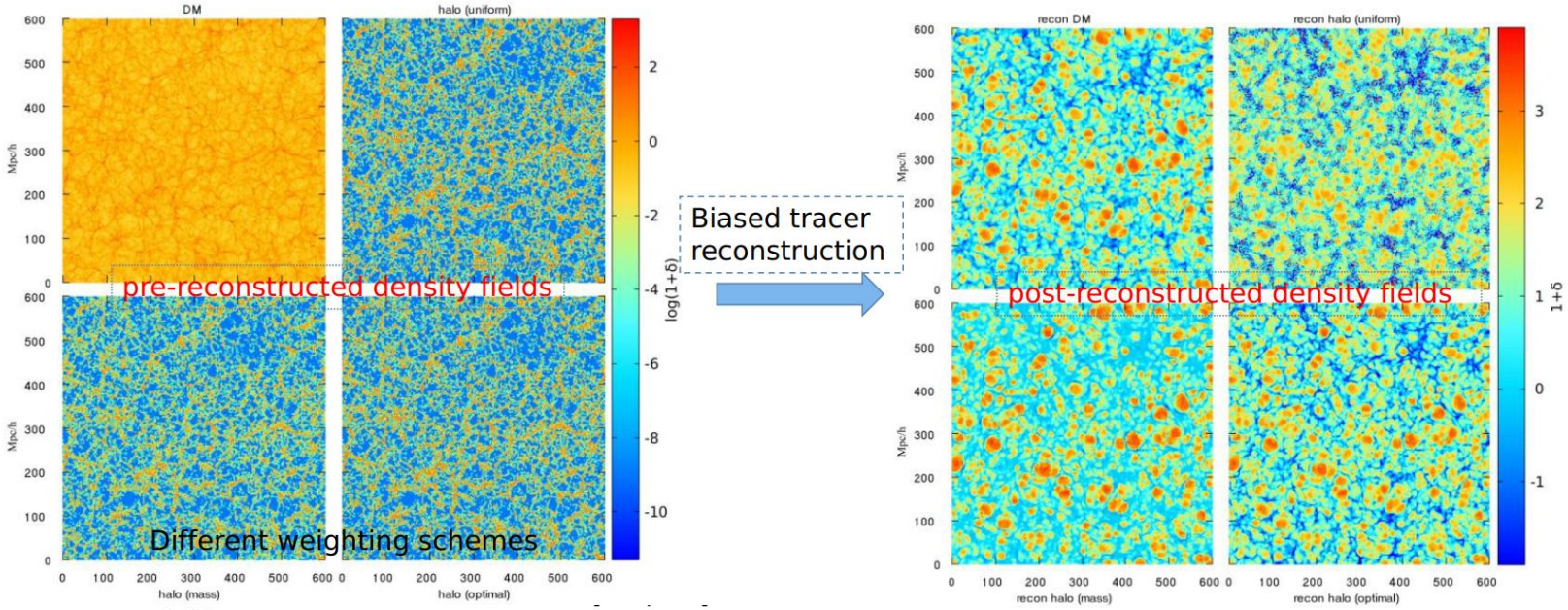
Shared goal: cosmological inference with SBI

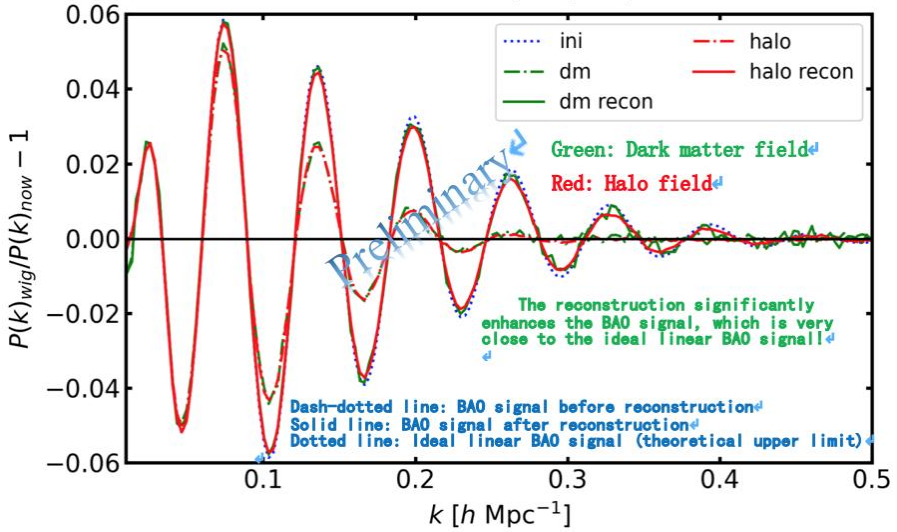
Model survey effects and uncertainties. Validate on independent mocks before applying the pipeline to galaxy survey data.

Nonlinear biased tracer reconstruction

Our work: Nonlinear biased tracer reconstruction with mass information

OUR WORK





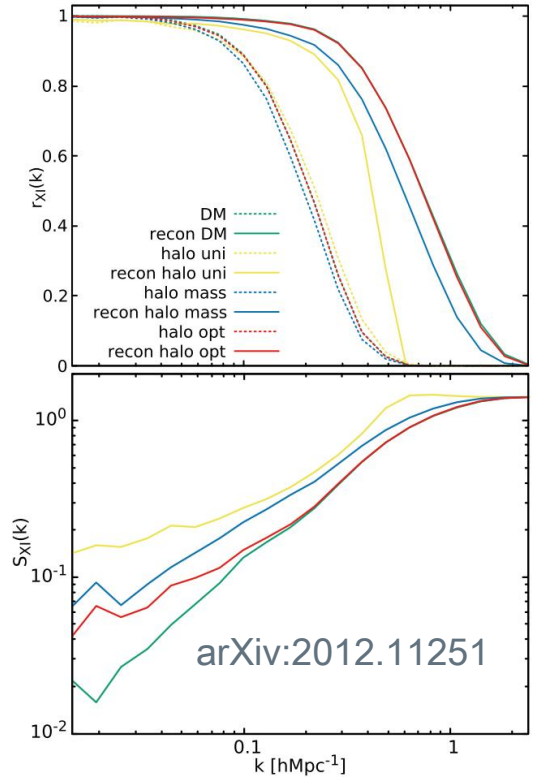


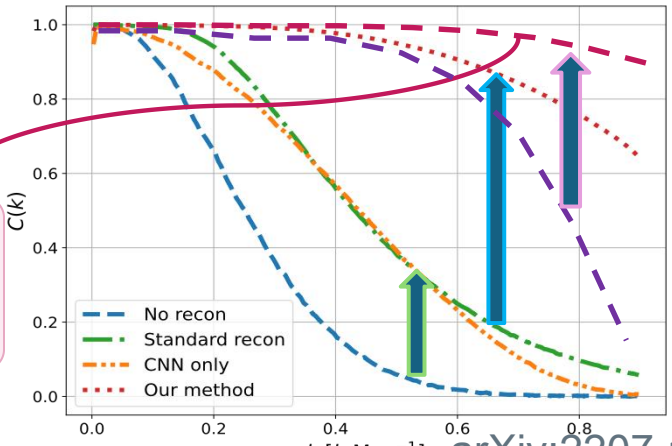
Schematic curve

Schematic hypothesis
biased tracer
reconstruction + ML

Correlation with the initial conditions

New method substantially outperforms standard reconstruction





Recovering linear BAO information from biased tracers with high number density

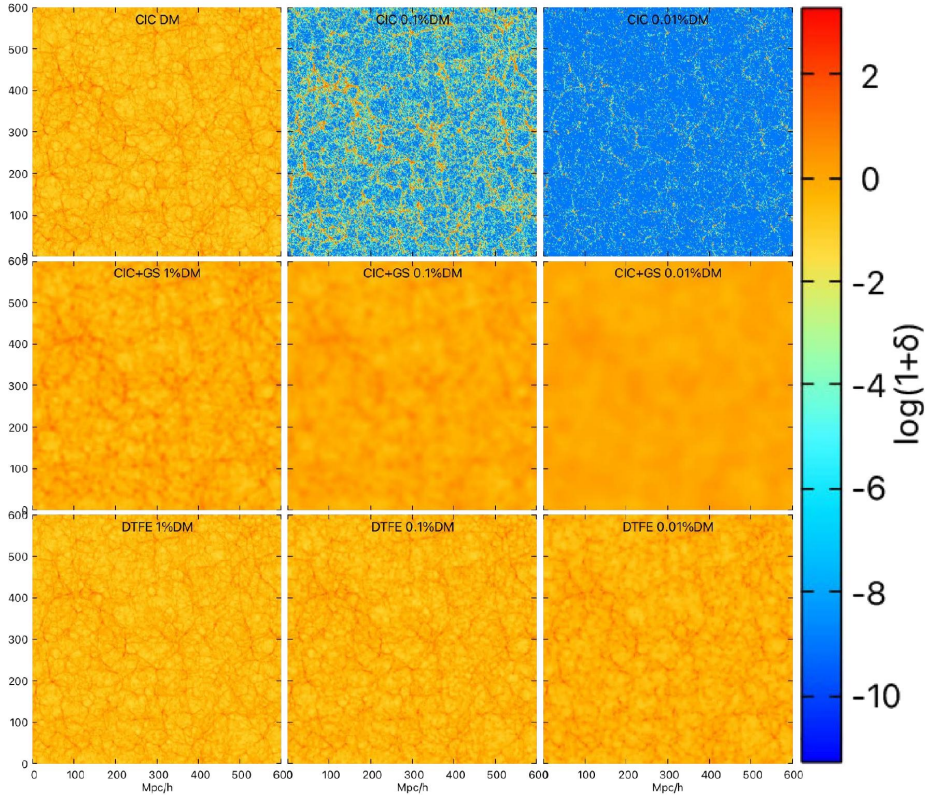
arXiv:2207.12511

DTFE-enhanced Minkowski functionals

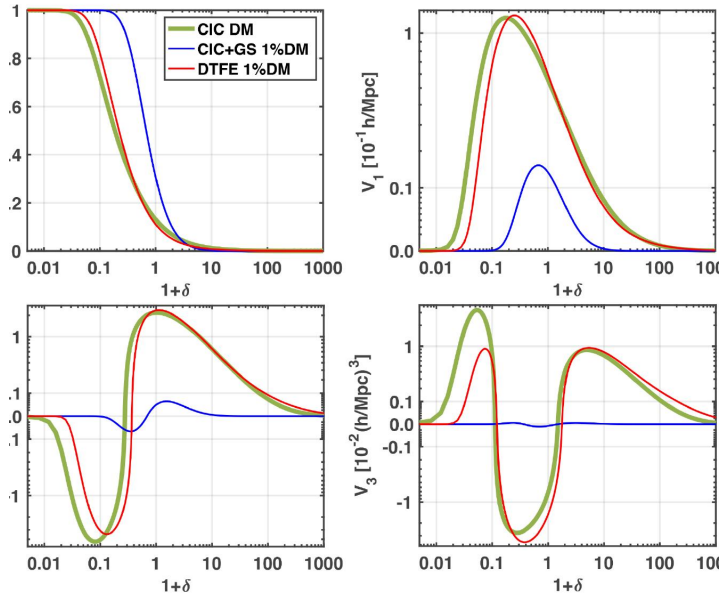
OUR WORK

Adaptive tessellation retains small-scale structure that fixed Gaussian smoothing suppresses

Morphology recovery



Minkowski functionals

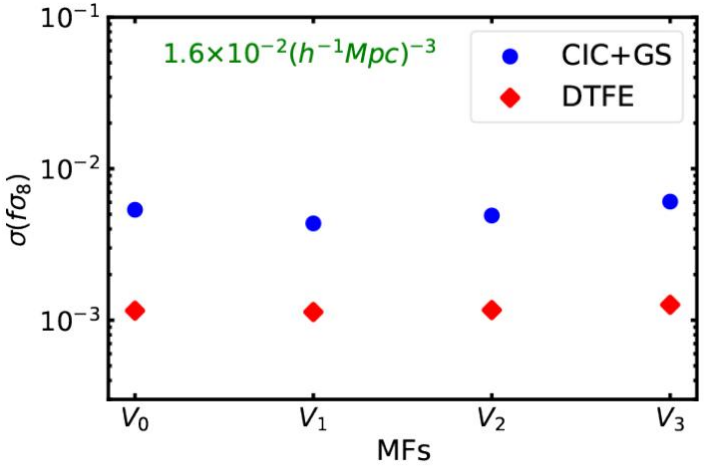


DTFE closely recovers the noise-free MF curves. Gaussian smoothing suppresses the amplitudes of V_1 , V_2 , and V_3 (i.e., the abundance of structrual elements).

3–5×

tighter constraints on the structure-growth rate than CIC with Gaussian smoothing

Forecast uncertainty on $f\sigma_8$



Key result

DTFE exposes substantially more morphological information in sparse-tracer fields.



Next step

Train an emulator for DTFE MFs and integrate it into an SBI pipeline for survey data.

Liu et al. (2024), The Astrophysical Journal Supplement Series 273, 33, arXiv:2405.0890

Optimized void statistics

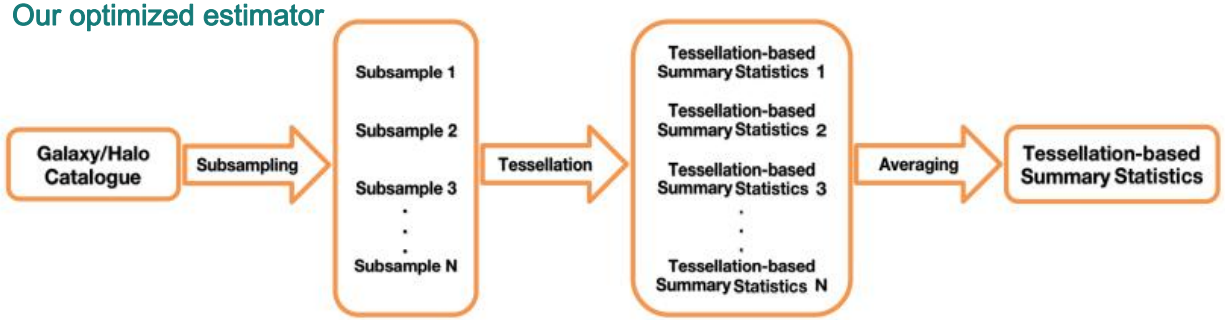
OUR WORK

Subsampling and averaging reduce the statistical uncertainties of tessellation-based statistics



The Traditional Scheme

Our optimized estimator



The Optimized Scheme

Tessellation instability can dominate the statistical scatter of void observables.

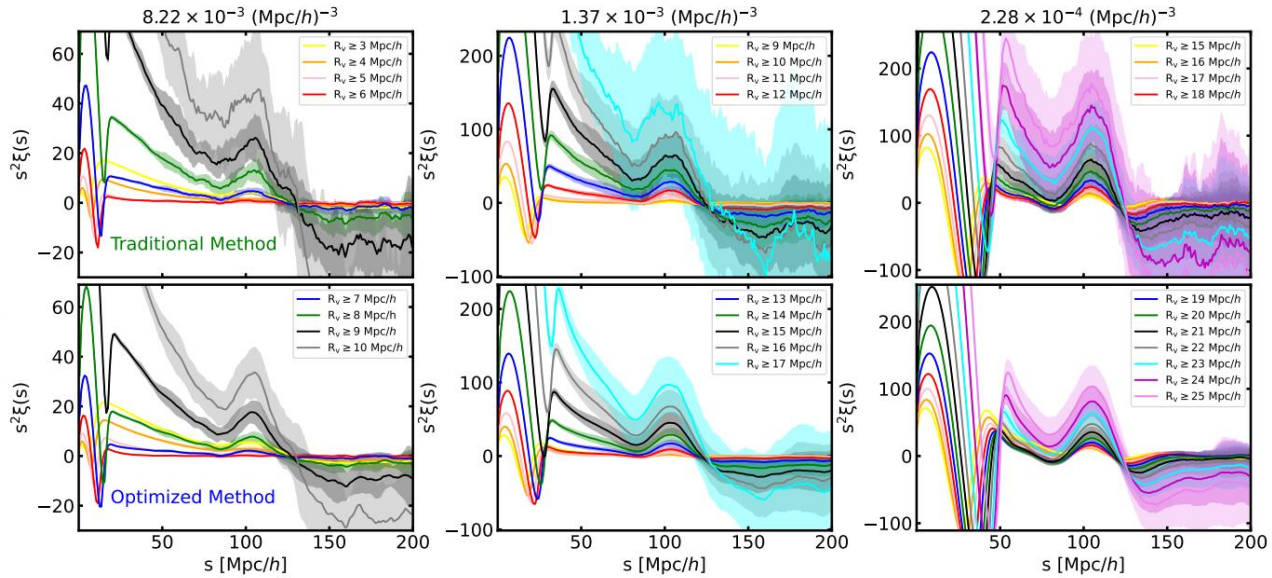
Key result Subsampling and averaging substantially reduce the statistical uncertainty of tessellation-based void statistics.

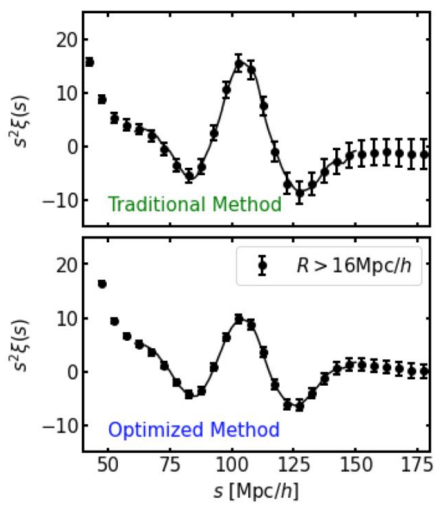


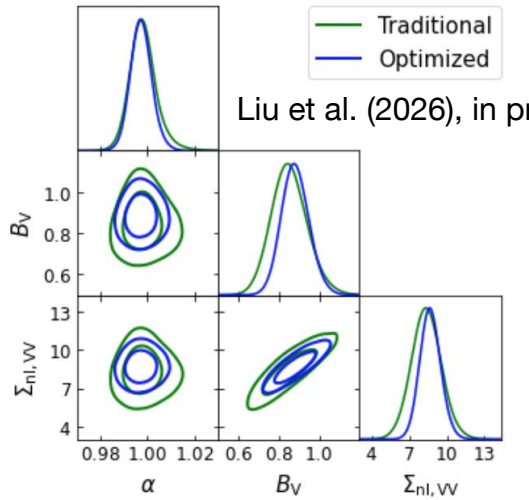
Next step Train an emulator for the optimized void statistics, then build an SBI pipeline for observational applications.

Liu et al. (2026), Optimization of Tessellation-based Statistics: Void Statistics, arXiv:2606.21864

Evidence from void statistics







Liu et al. (2026), in process

The optimized estimator produces smaller error bars and tighter posterior contours while preserving the physical signal.

What does machine learning learn?

Approximations to complex functions and probability distributions,
learned from simulations or observations.

FUNCTIONS & REPRESENTATIONS

Predict observables and reconstruct fields
Compress high-dimensional data

PROBABILITY MODELS

Likelihoods, posteriors and likelihood ratios
Conditional realizations

These roles overlap: reconstruction can itself be probabilistic.

Scientific credibility requires two levels of validation

Numerical fidelity

Accuracy on independent validation data

Scientific validity

Physical realism and calibrated uncertainties

Thank you!

Questions & Discussion

E-mail： y.liu@epfl.ch

