COMPRESSION AND SIMULATION-BASED INFERENCE FOR NEXT-GENERATION COSMOLOGICAL SURVEYS

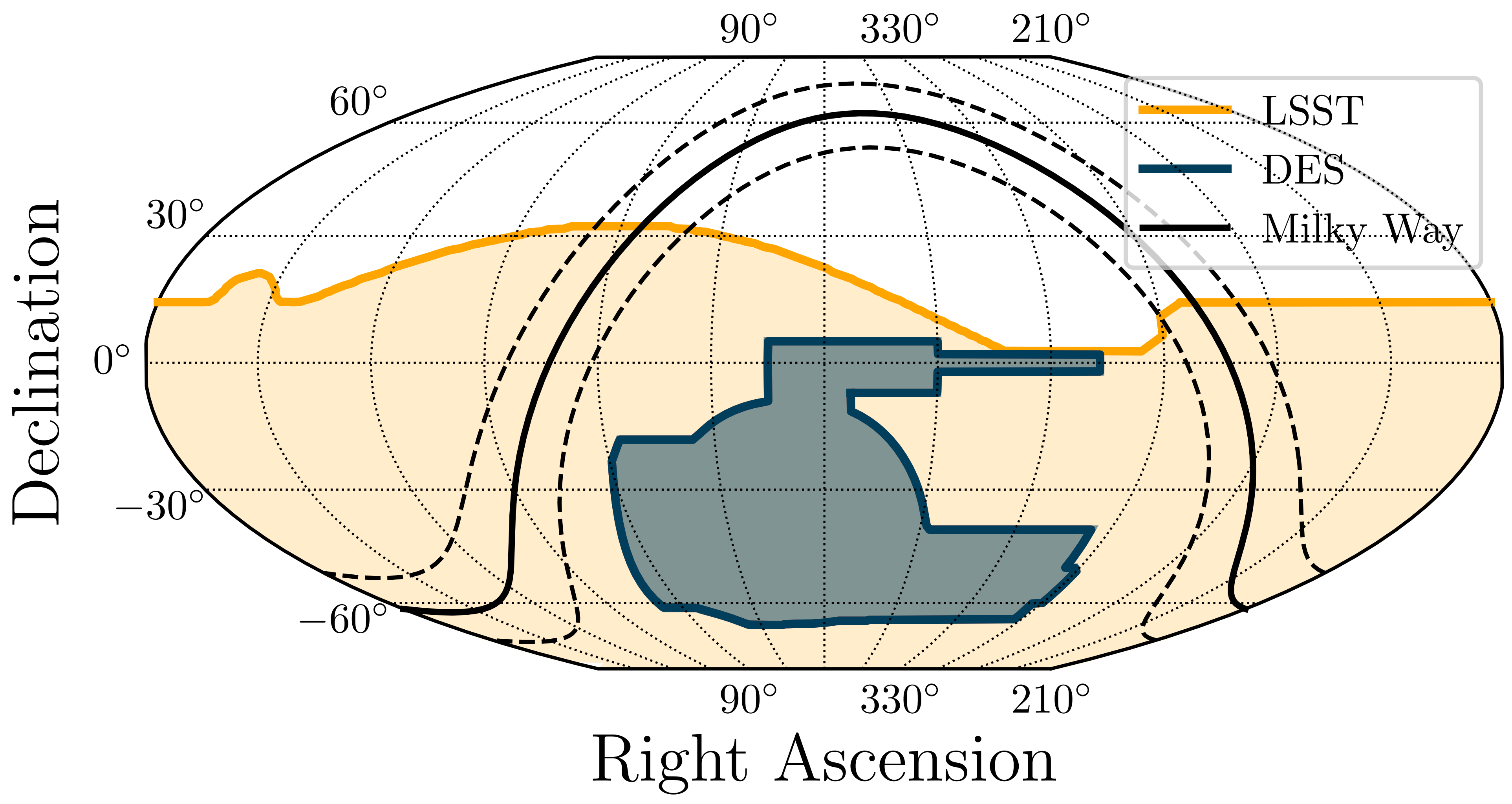
JUDIT PRAT MARTÍ
NORDITA FELLOW AT
STOCKHOLM UNIVERSITY AND KTH
SEPTEMBER 2025
AI FOR SCIENCE SYMPOSIUM

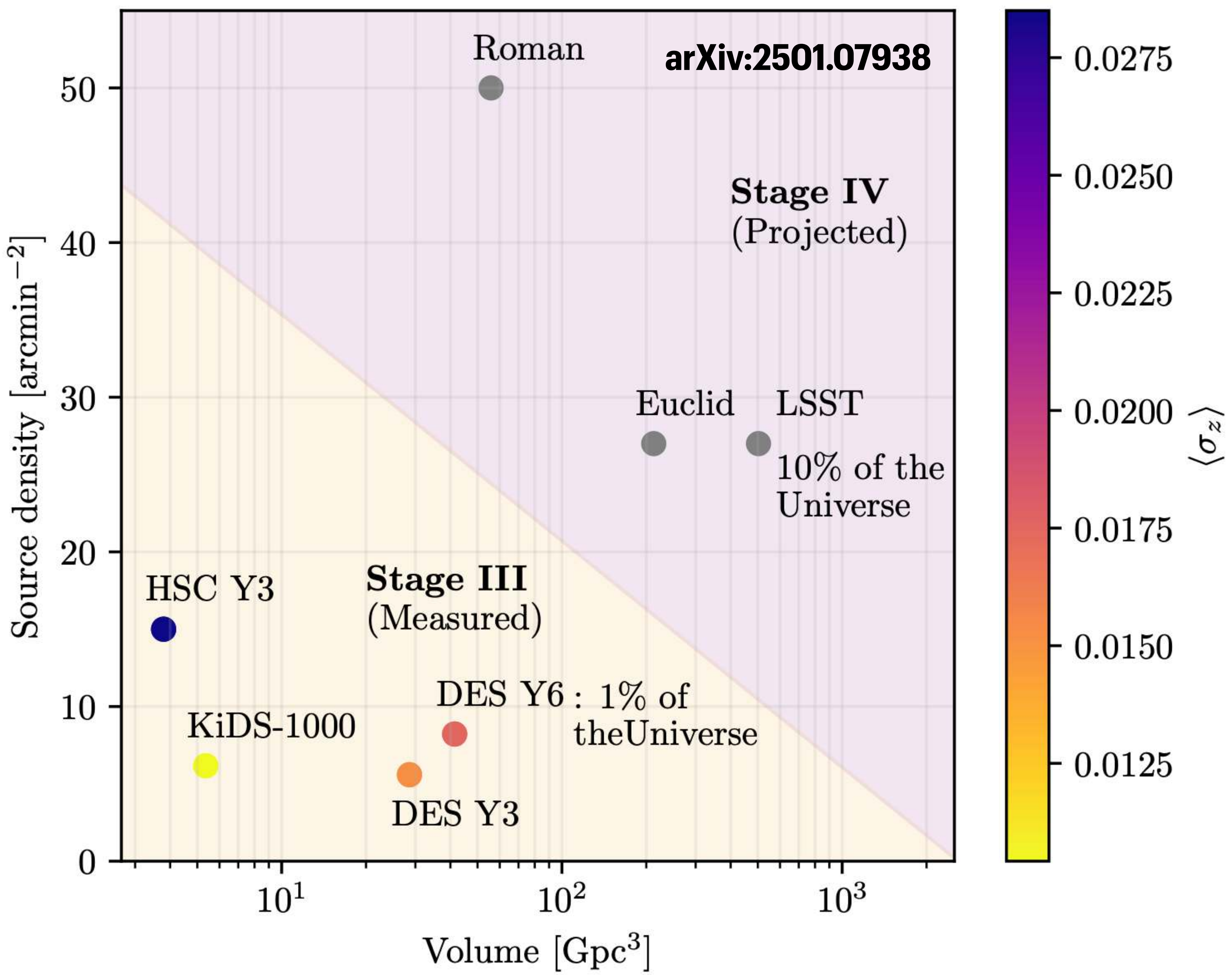
NORDITA
The Nordic Institute for
Theoretical Physics

Oscar Klein

AN EXPLOSION OF DATA FROM GALAXY SURVEYS

STARTING NOW!



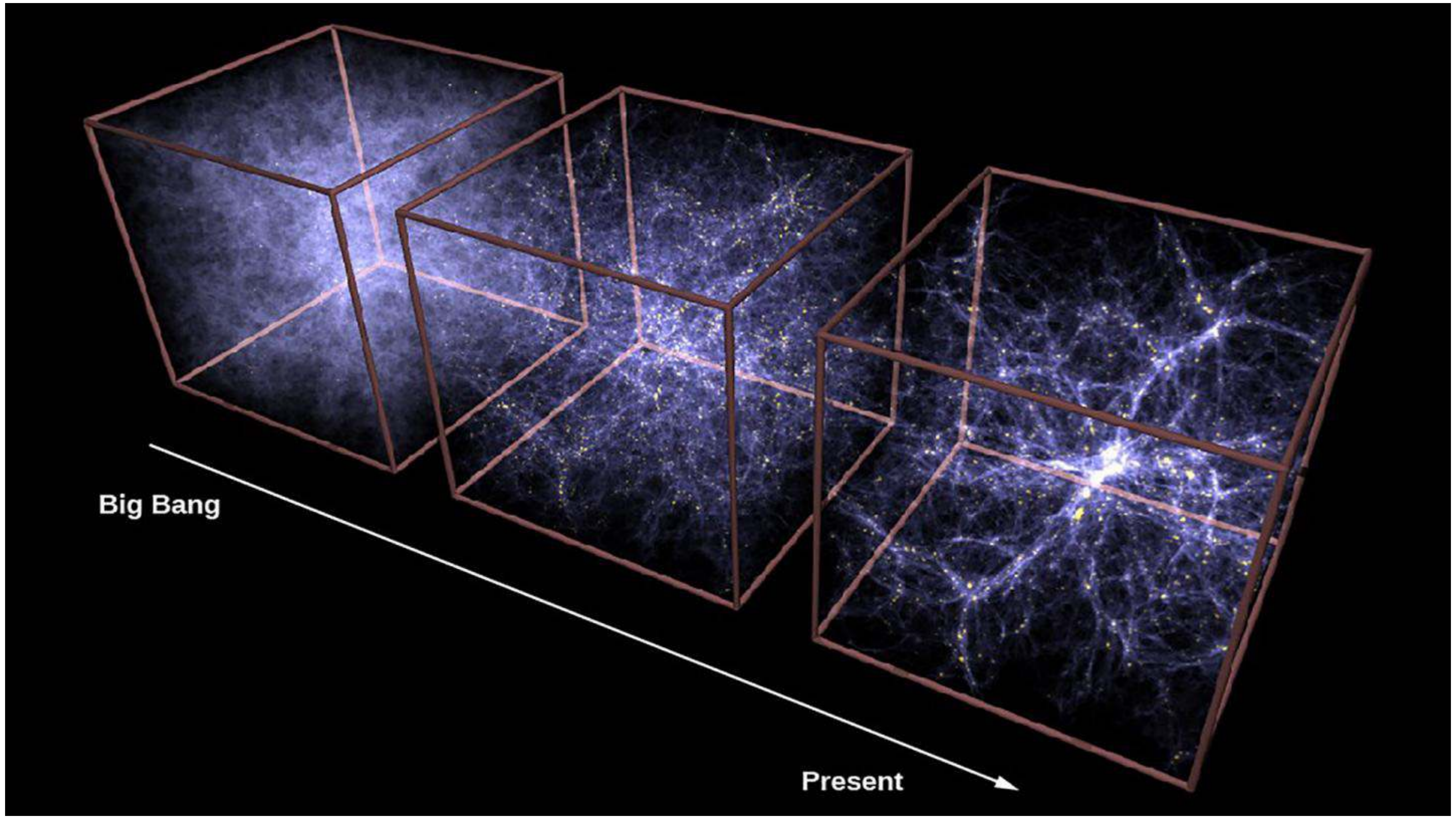


WHAT IS THE GOAL?

What is driving the accelerated expansion of the Universe? What is the nature of dark energy and dark matter?

Get *as much cosmological information* from the late-time Universe *that we can trust*.

The late-time Universe is non-Gaussian, as opposed to the early Universe. This information is harder to get, same tools do not work as well anymore.



WHERE DO WE STAND?

COSMOLOGY FROM WEAK GRAVITATIONAL LENSING DATA AND LARGE-SCALE STRUCTURE DATA

Standard analyses with theoretical modeling and two-point statistics:

- Captures Gaussian statistical information from cosmic surveys
- Well-established methodology with understood systematics, though modeling complexity increases as survey precision improves
- Currently the foundation for major cosmological results from large survey collaborations

Simulation based inference with higher-order statistics

- Extracts additional information while remaining closely tied to data (several analyses have already been successfully applied to real survey data)
- Promising approach but still not as mature as standard methods
- Can this become the new standard?

Using the full information content

- Represents the theoretical ideal for extracting maximum cosmological information
- Practical implementation is more difficult. Has been less applied to real data.
- Is the added complexity worth it? Roadblocks/ practical problems sometimes only become apparent when carrying analyses through to completion

Theoretically more information, but less applicable to data currently

More mature methods

WHERE DO WE STAND?

COSMOLOGY FROM WEAK GRAVITATIONAL LENSING DATA AND LARGE-SCALE STRUCTURE DATA

Standard analyses with theoretical modeling and two-point statistics:

- Captures Gaussian statistical information from cosmic surveys
- Well-established methodology with understood systematics, though modeling complexity increases as survey precision improves
- Currently the foundation for major cosmological results from large survey collaborations

Simulation based inference with higher-order statistics

- Extracts additional information while remaining closely tied to data (several analyses have already been successfully applied to real survey data)
- Promising approach but still not as mature as standard methods
- Can this become the new standard?

Using the full information content

- Represents the theoretical ideal for extracting maximum cosmological information
- Practical implementation is more difficult. Has been less applied to real data.
- Is the added complexity worth it? Roadblocks/practical problems sometimes only become apparent when carrying analyses through to completion

Theoretically more information, but less applicable to data currently →

← **More mature methods**

WHERE DO WE STAND?

COSMOLOGY FROM WEAK GRAVITATIONAL LENSING DATA AND LARGE-SCALE STRUCTURE DATA

Standard analyses with t modeling and two-point

- Captures Gaussian statistical inf cosmic surveys
- Well-established methodology w systematics, though modeling increases as survey precisio
- Currently the foundation for maj results from large survey col

Simulation based inference with higher-order statistics

- Extracts additional information while remaining closely tied to data (several analyses have already been successfully applied to real survey data)
- Promising approach but still not as mature as standard methods
- Can this become the new standard?

Using the full information content

- Represents the theoretical ideal for extracting maximum cosmological information
- Practical implementation is more difficult. Has been less applied to real data.
- Is the added complexity worth it? Roadblocks/ practical problems sometimes only become apparent when carrying analyses through to completion

Theoretically more information, but less applicable to data currently

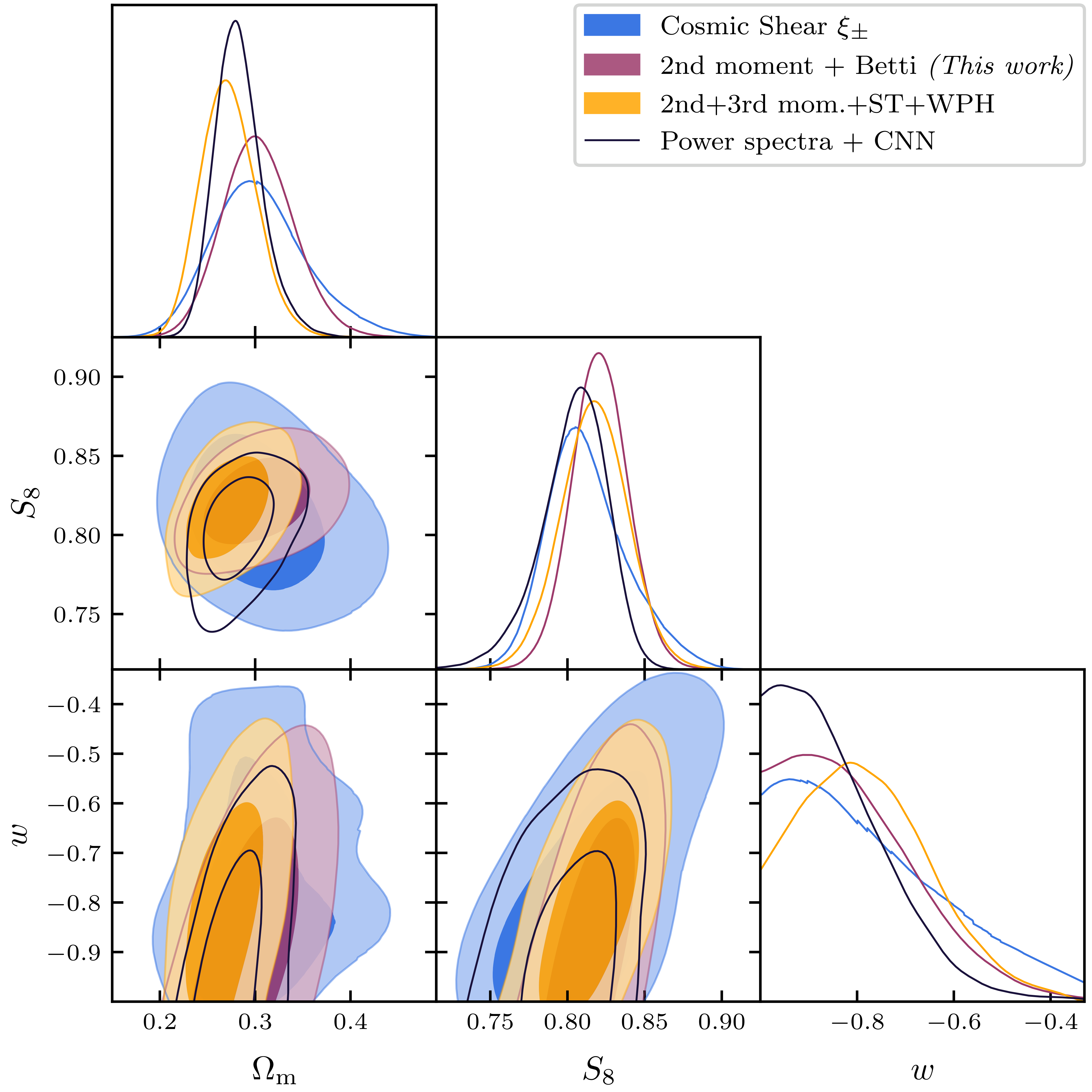
More mature methods

TALK BASED ON:
arXiv:2506.13439

DARK ENERGY SURVEY YEAR 3 RESULTS: wCDM COSMOLOGY FROM SIMULATION-BASED INFERENCE WITH PERSISTENT HOMOLOGY ON THE SPHERE

ABSTRACT
We present cosmological constraints from Dark Energy Survey Year 3 (DES Y3) weak lensing data using persistent homology, a topological data analysis technique that tracks how features like clusters and voids evolve across density thresholds. For the first time, we apply spherical persistent homology to galaxy survey data through the algorithm TopoS2, which is optimized for curved-sky analyses and HEALPix compatibility. Employing a simulation-based inference framework with the Gower Street simulation suite—specifically designed to mimic DES Y3 data properties—we extract topological summary statistics from convergence maps across multiple smoothing scales and redshift bins. After neural network compression of these statistics, we estimate the likelihood function and validate our analysis against baryonic feedback effects, finding minimal biases (under 0.3σ) in the $\Omega_m - S_8$ plane. Assuming the wCDM model, our combined Betti numbers and second moments analysis yields $S_8 = 0.821 \pm 0.018$ and $\Omega_m = 0.304 \pm 0.037$—constraints 70% tighter than those from cosmic shear two-point statistics in the same parameter plane. Our results demonstrate that topological methods provide a powerful and robust framework for extracting cosmological information, with our spherical methodology readily applicable to upcoming Stage IV wide-field galaxy surveys.

Bringing analyses end to end on precursor data to identify challenges and mature methods for upcoming next-generation surveys…



MY COLLABORATORS

Marco Gatti
UChicago
(US)

Cyrille Doux
CNRS/IN2P3
(France)

Pratyush Pranav
Bennett
University (India)

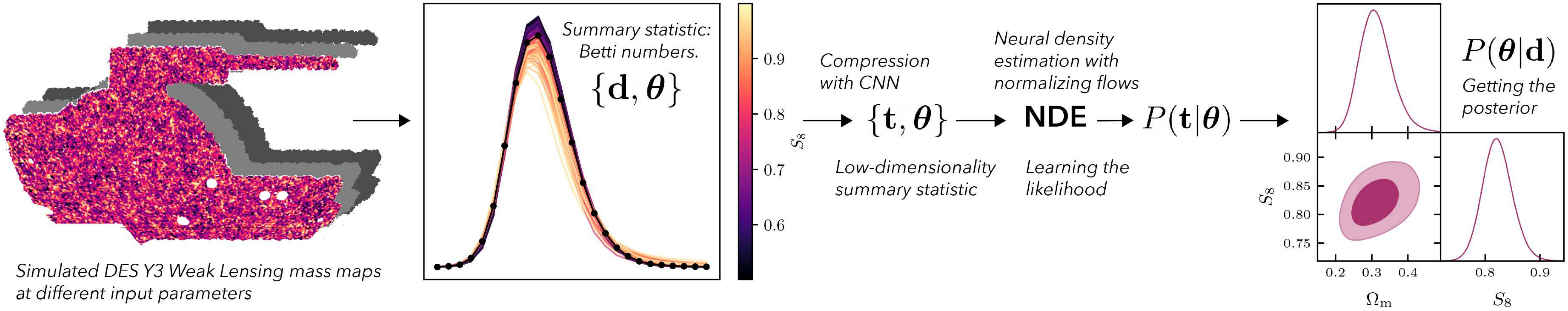
Chihway Chang
UChicago
(US)

Niall Jeffrey
UCL (UK)

arXiv:2506.13439

THE FRAMEWORK

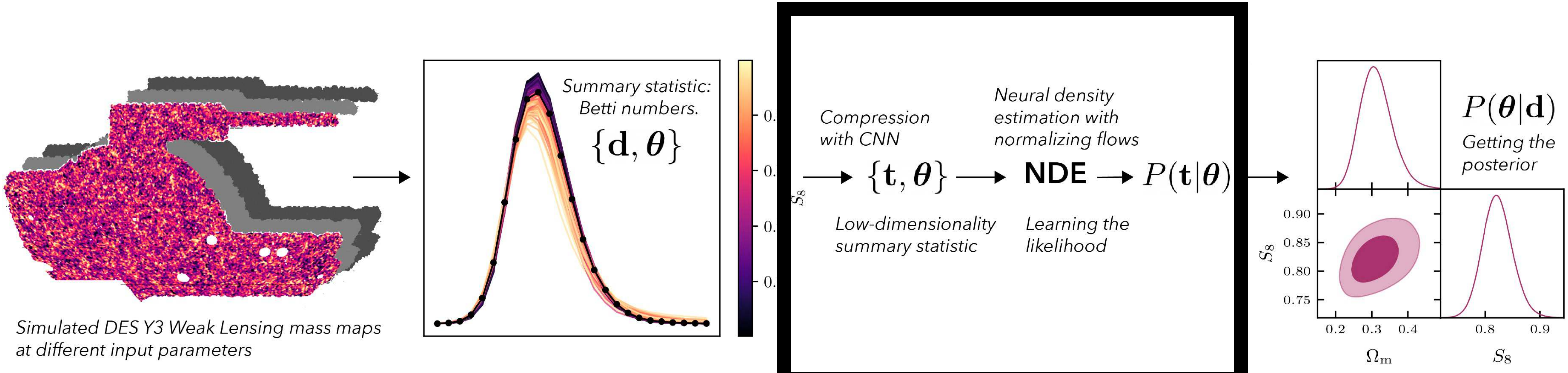
SIMULATION-BASED INFERENCE



arXiv:2506.13439

THE FRAMEWORK

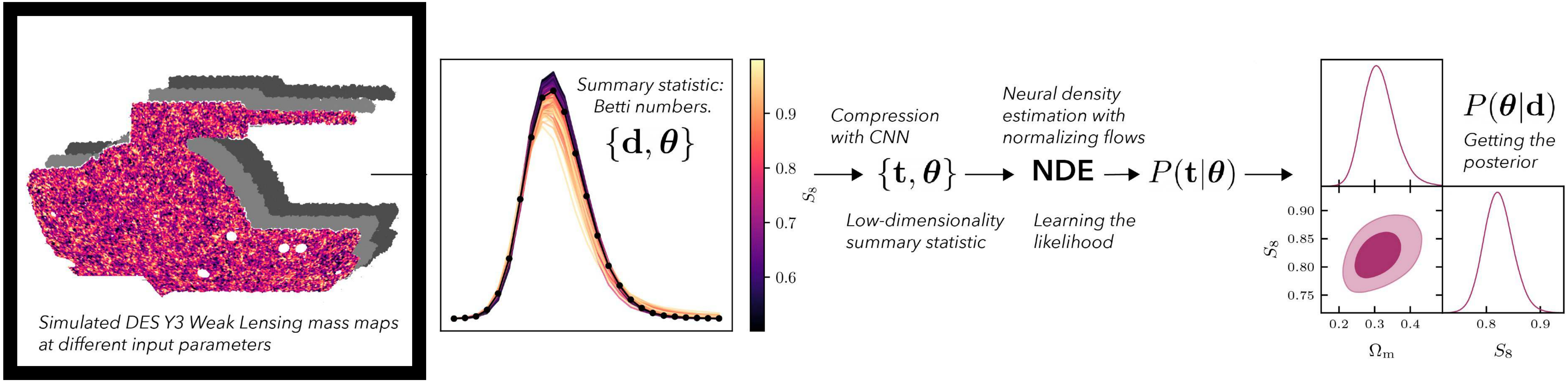
SIMULATION-BASED INFERENCE: NOT POSSIBLE WITHOUT MACHINE LEARNING!



arXiv:2506.13439

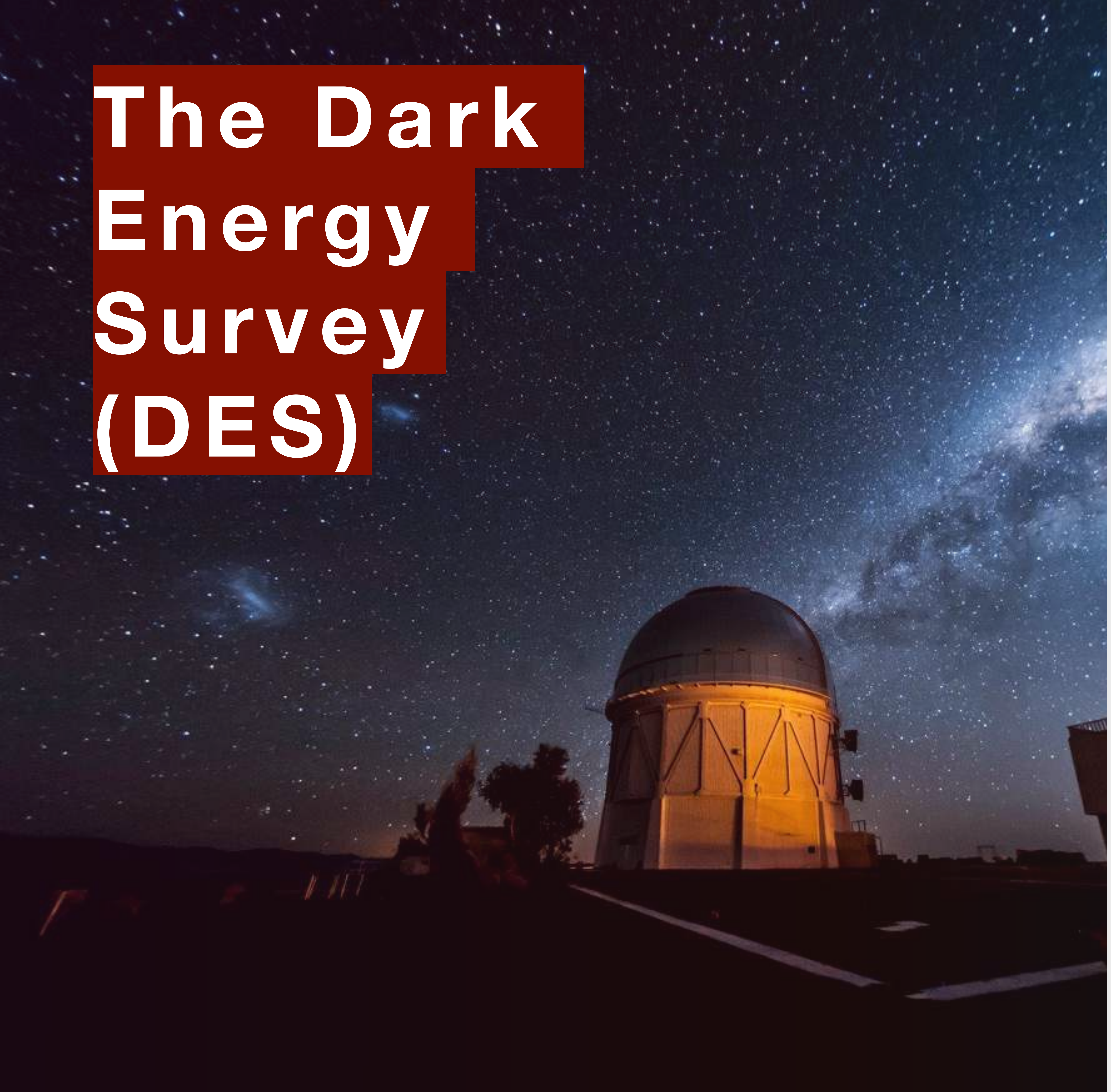
THE FRAMEWORK

SIMULATION-BASED INFERENCE

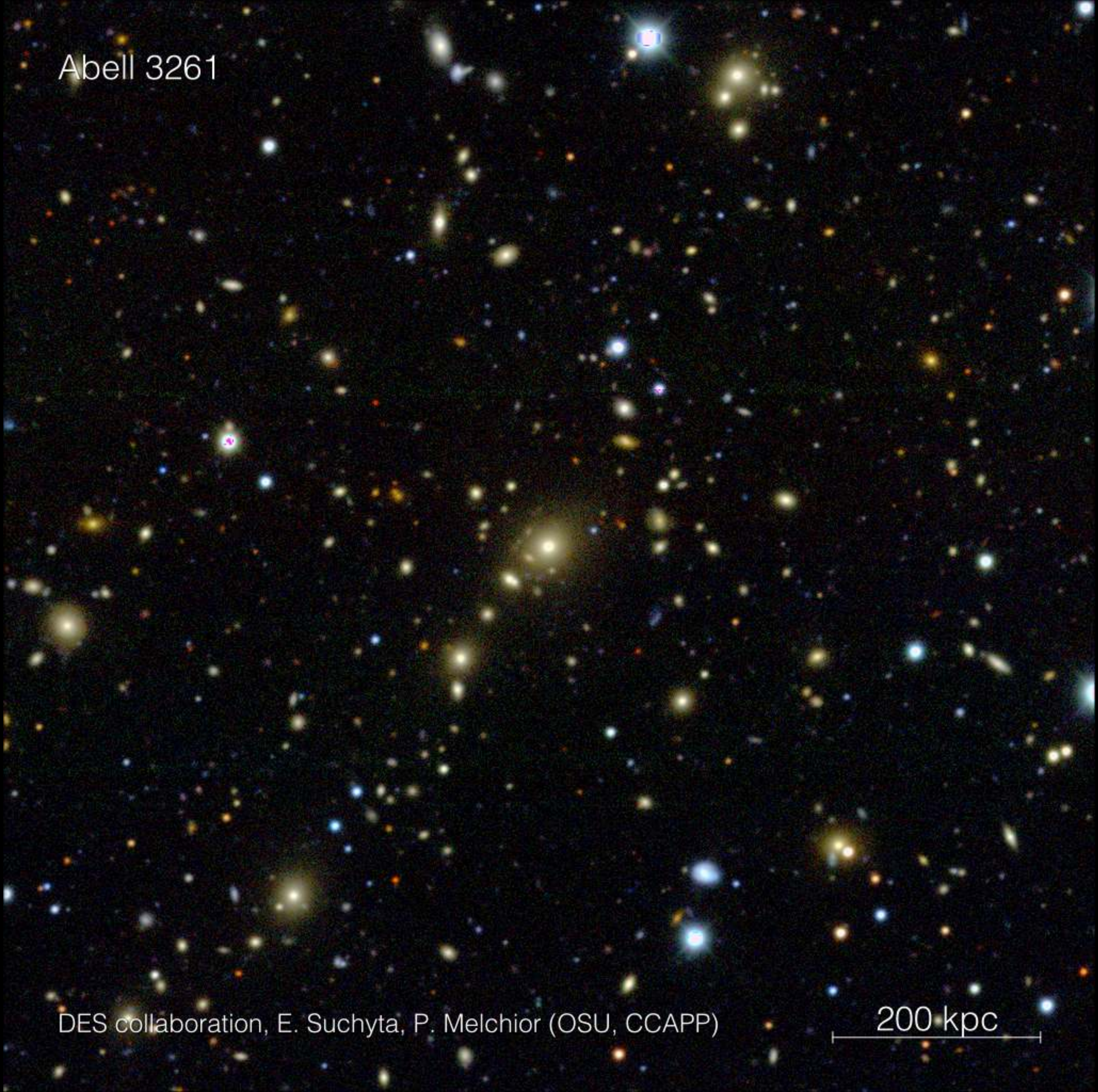


The data and simulations

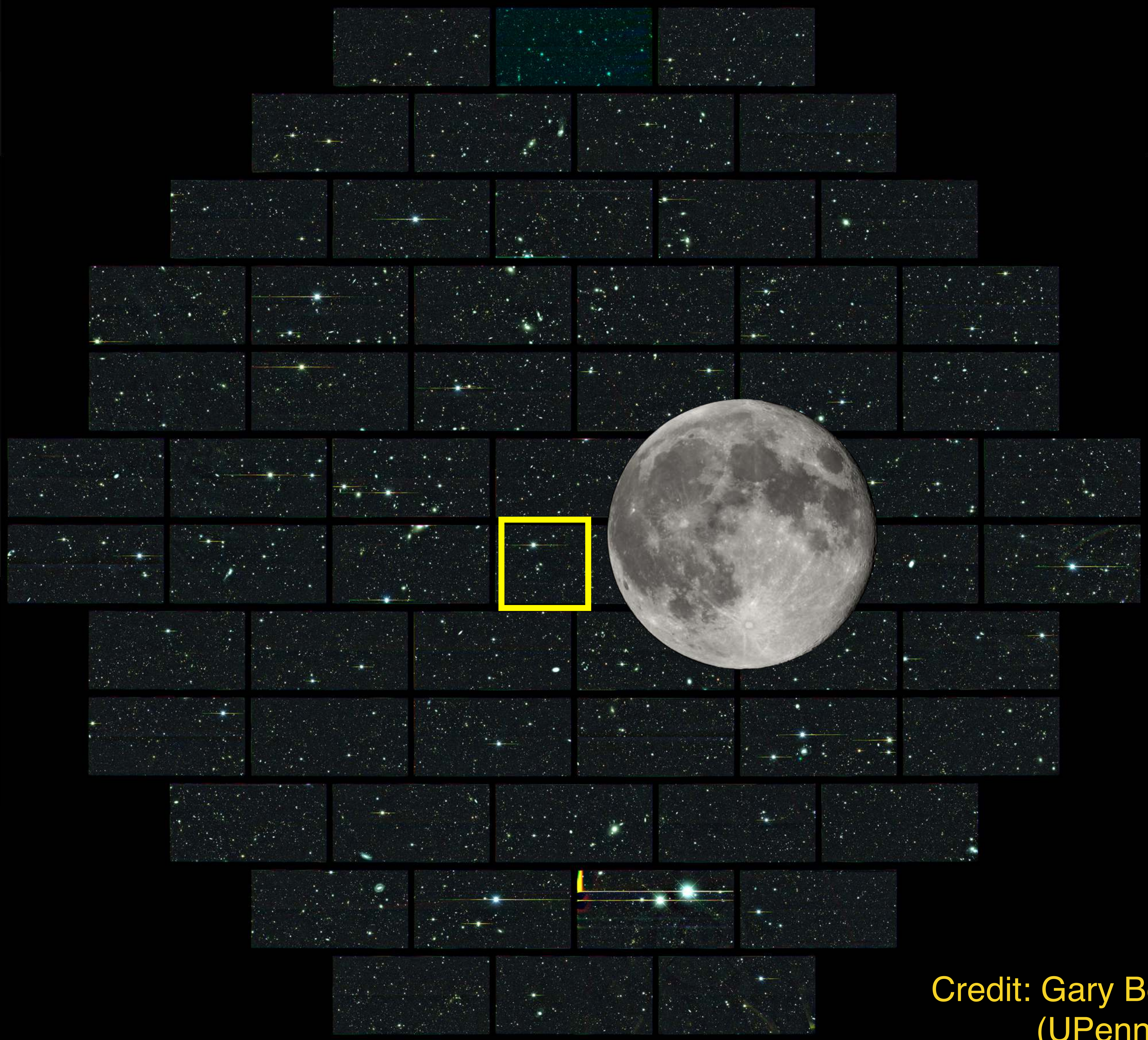
The Dark Energy Survey (DES)



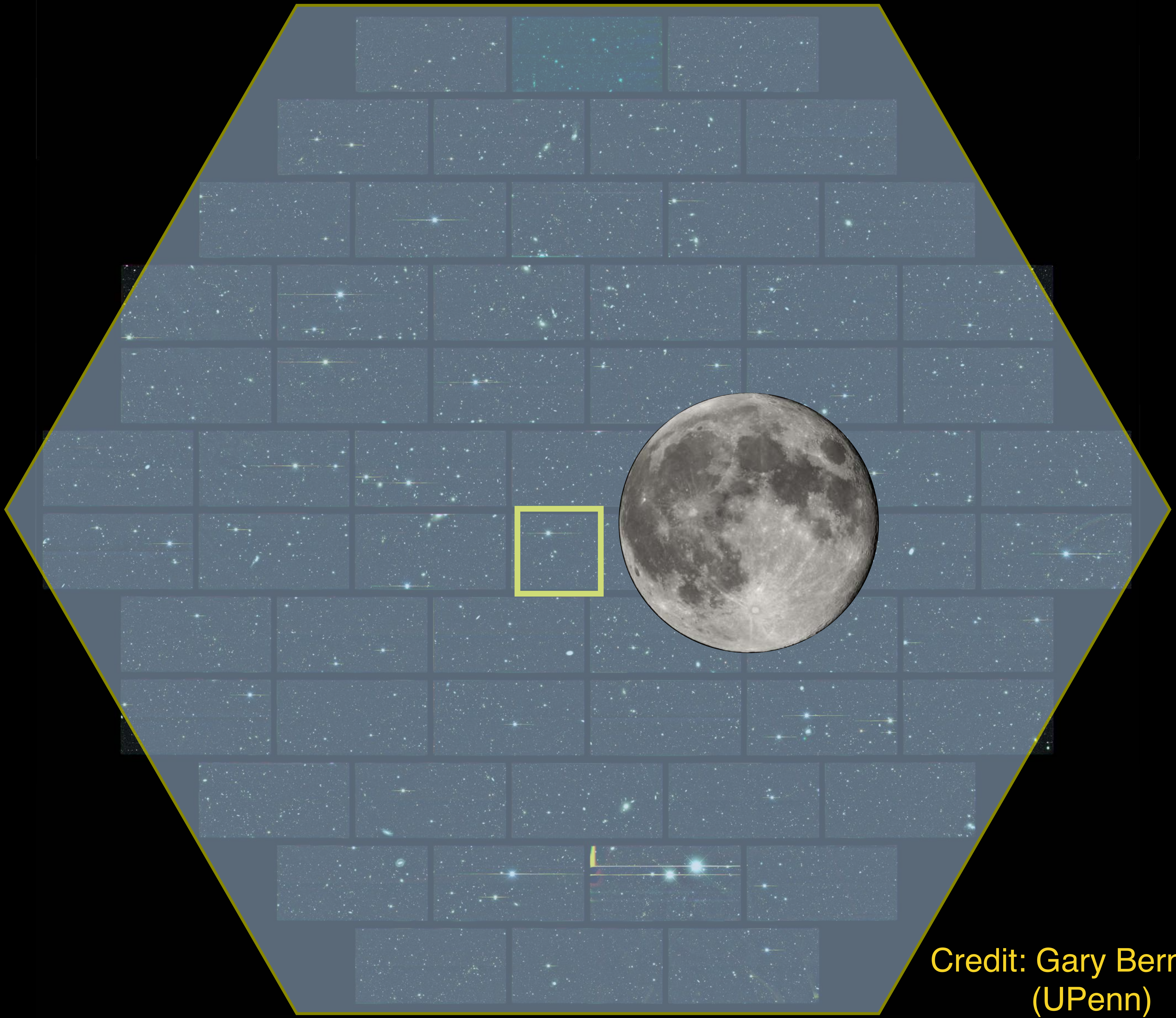
- Imaging galaxy survey.
- 5000 sq. deg. after 6 years (2013-2019)
- 570-Megapixel digital camera, DECam, mounted on the Blanco 4-meter telescope at Cerro Tololo Inter-American Observatory (Chile).


Abell 3261
DES collaboration, E. Suchyta, P. Melchior (OSU, CCAPP)
200 kpc

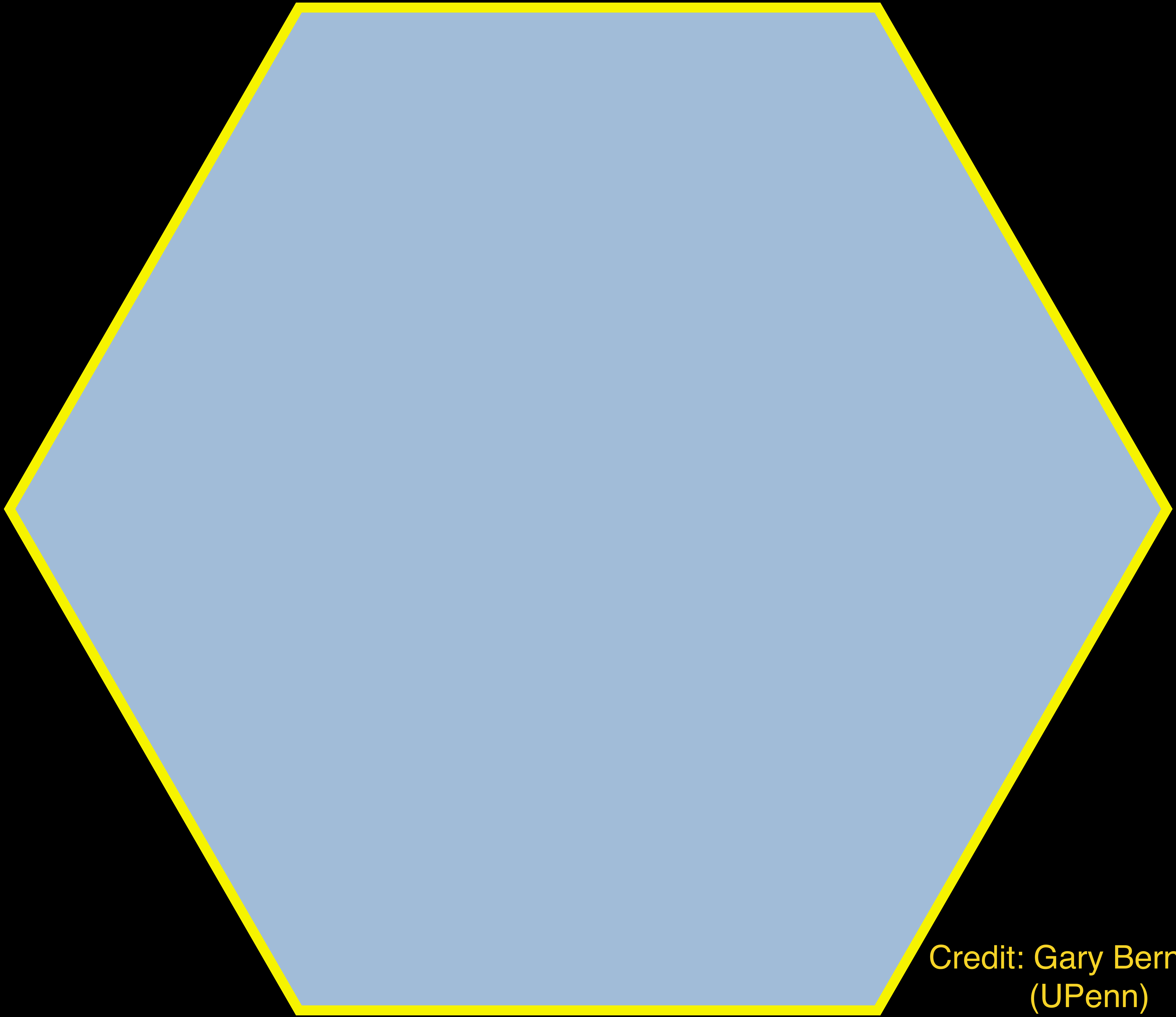
Credit: Gary Bernstein
(UPenn)



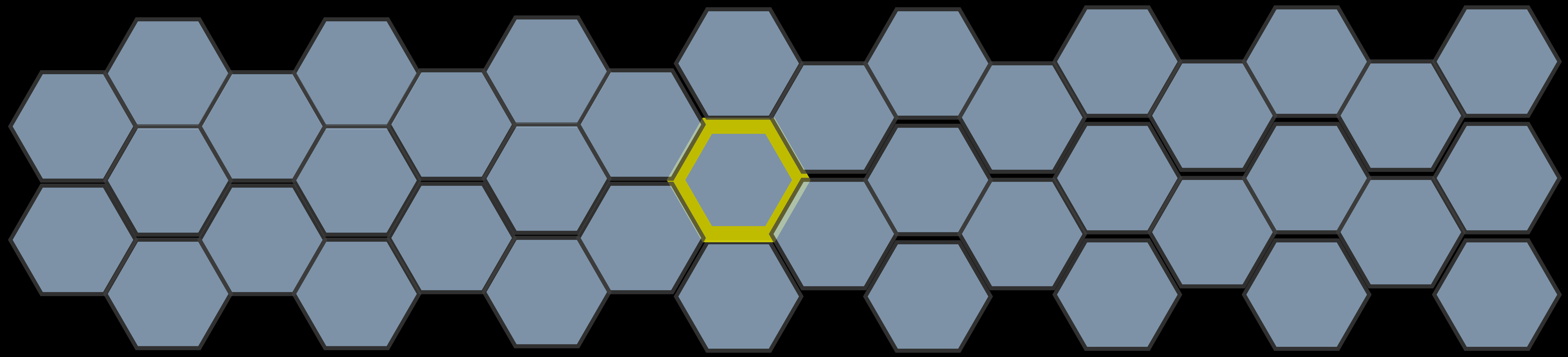
Credit: Gary Bernstein
(UPenn)



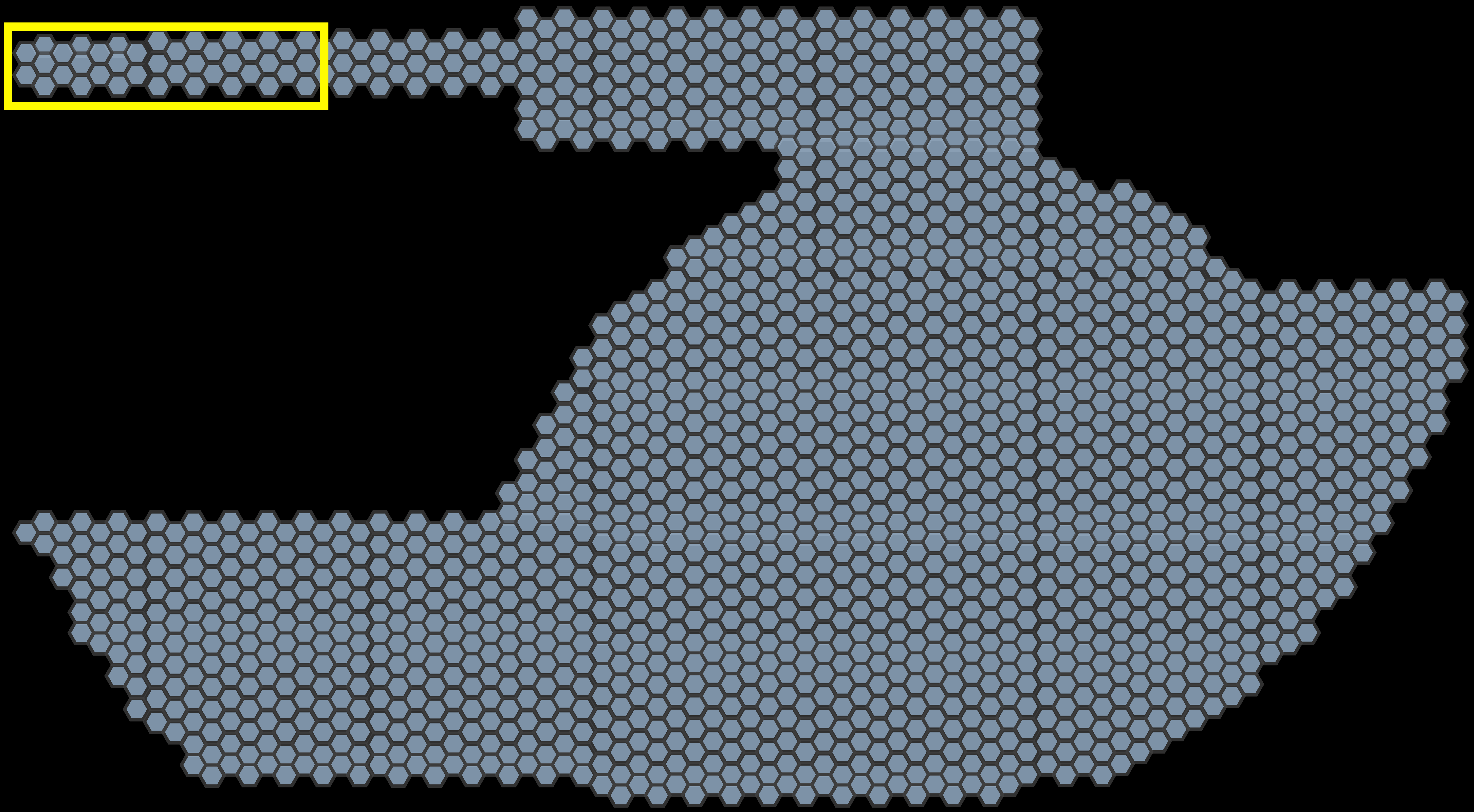
Credit: Gary Bernstein
(UPenn)



Credit: Gary Bernstein
(UPenn)

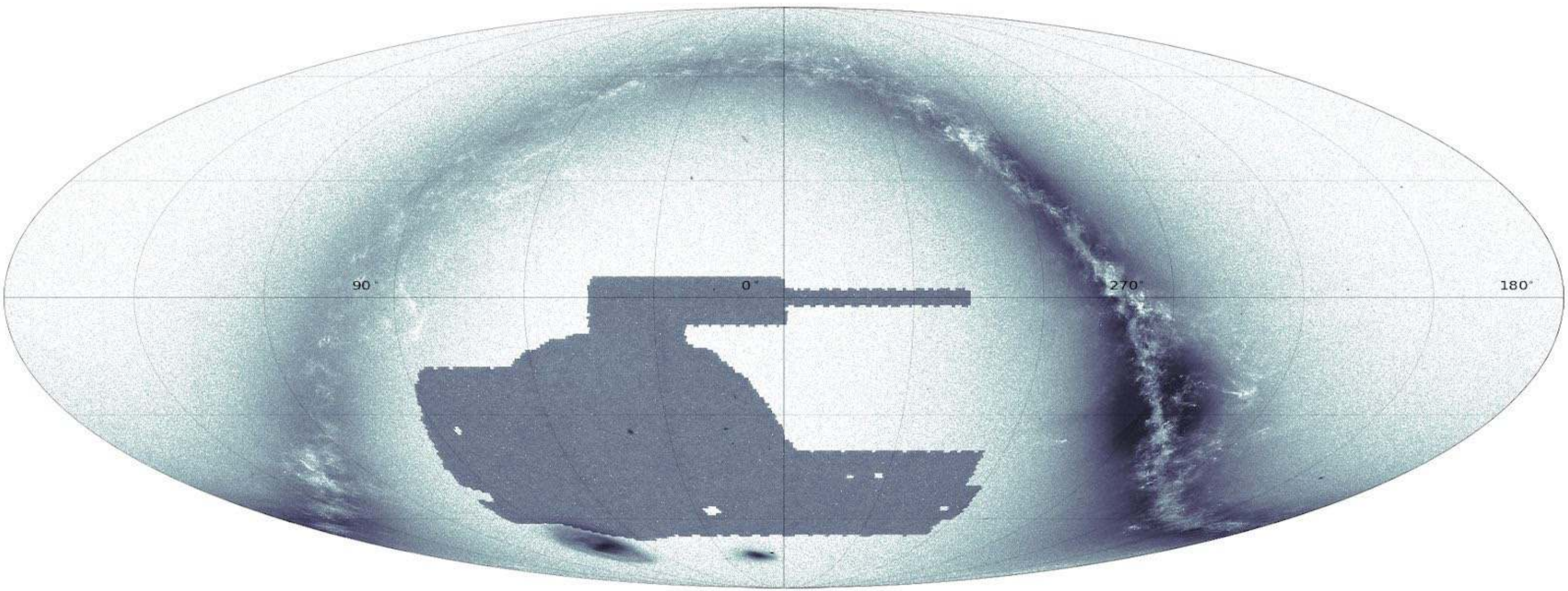


Credit: Gary Bernstein
(UPenn)

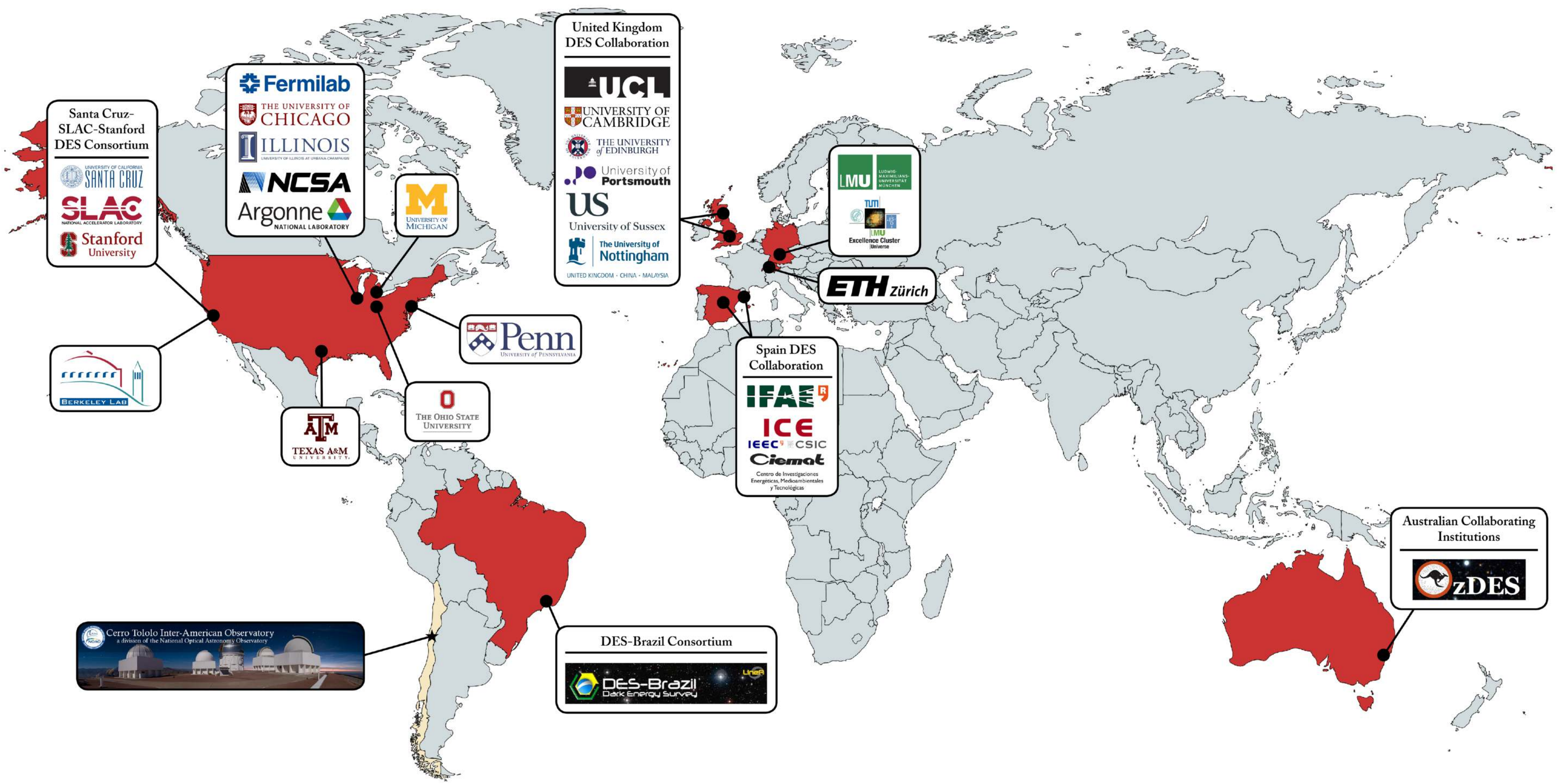


Credit: Gary Bernstein
(UPenn)

The Dark Energy Survey (DES)



Dark Energy Survey Collaboration

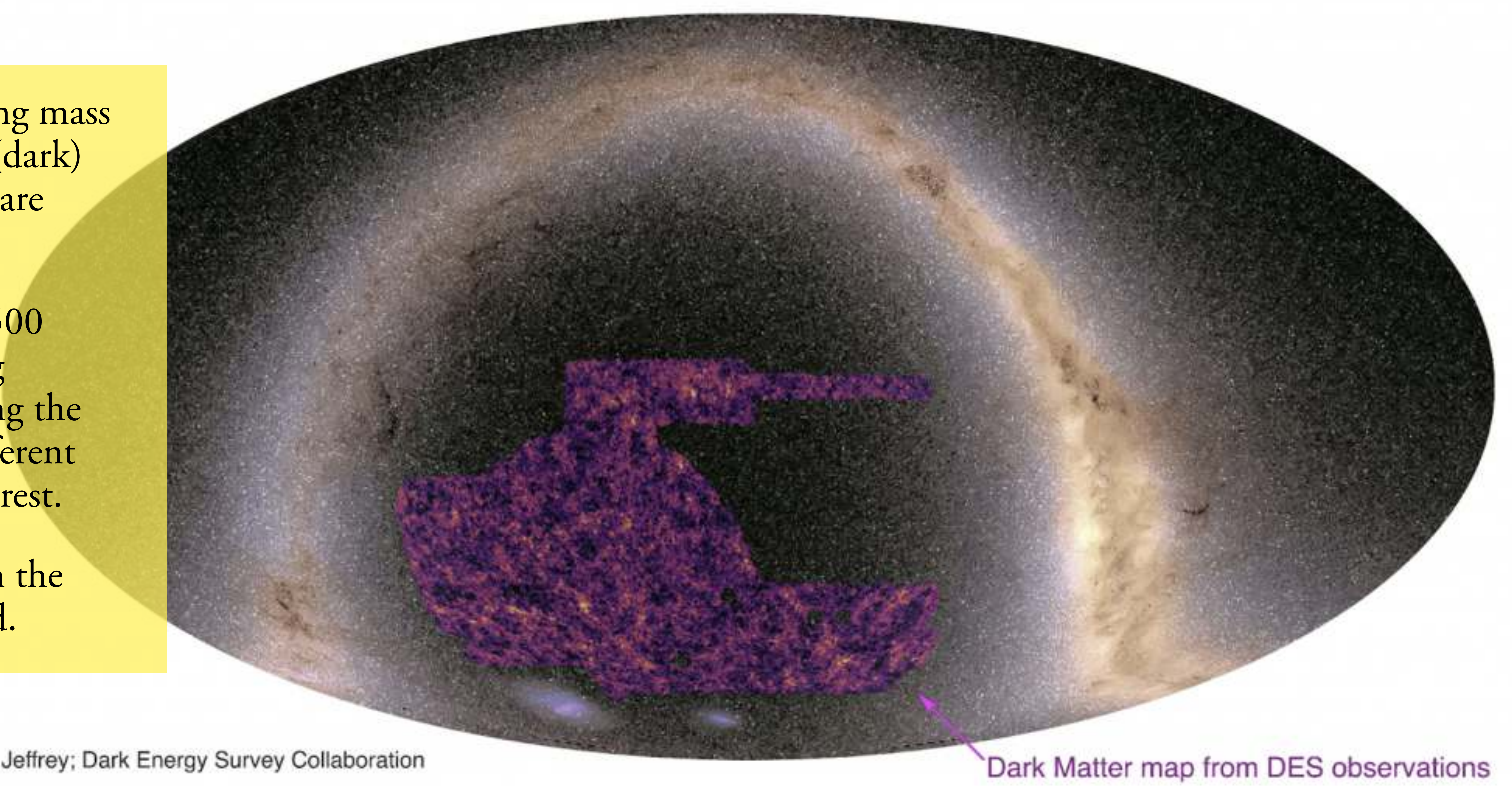


THE DATA AND SIMS: WEAK LENSING MASS MAPS

Our data are weak lensing mass maps, which probe the (dark) matter distribution and are sensitive to cosmology.

We have produced ~12500 *convergence* weak lensing simulated maps matching the data properties with different input parameters of interest.

These are scalar maps on the sphere on a healPIX grid.

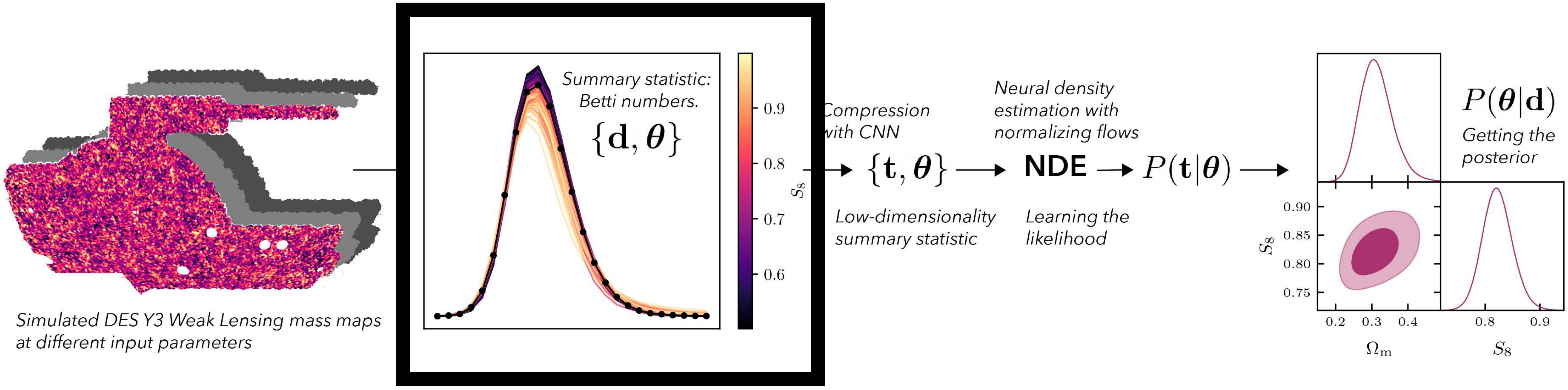


N. Jeffrey; Dark Energy Survey Collaboration

arXiv:2506.13439

THE FRAMEWORK

SIMULATION-BASED INFERENCE

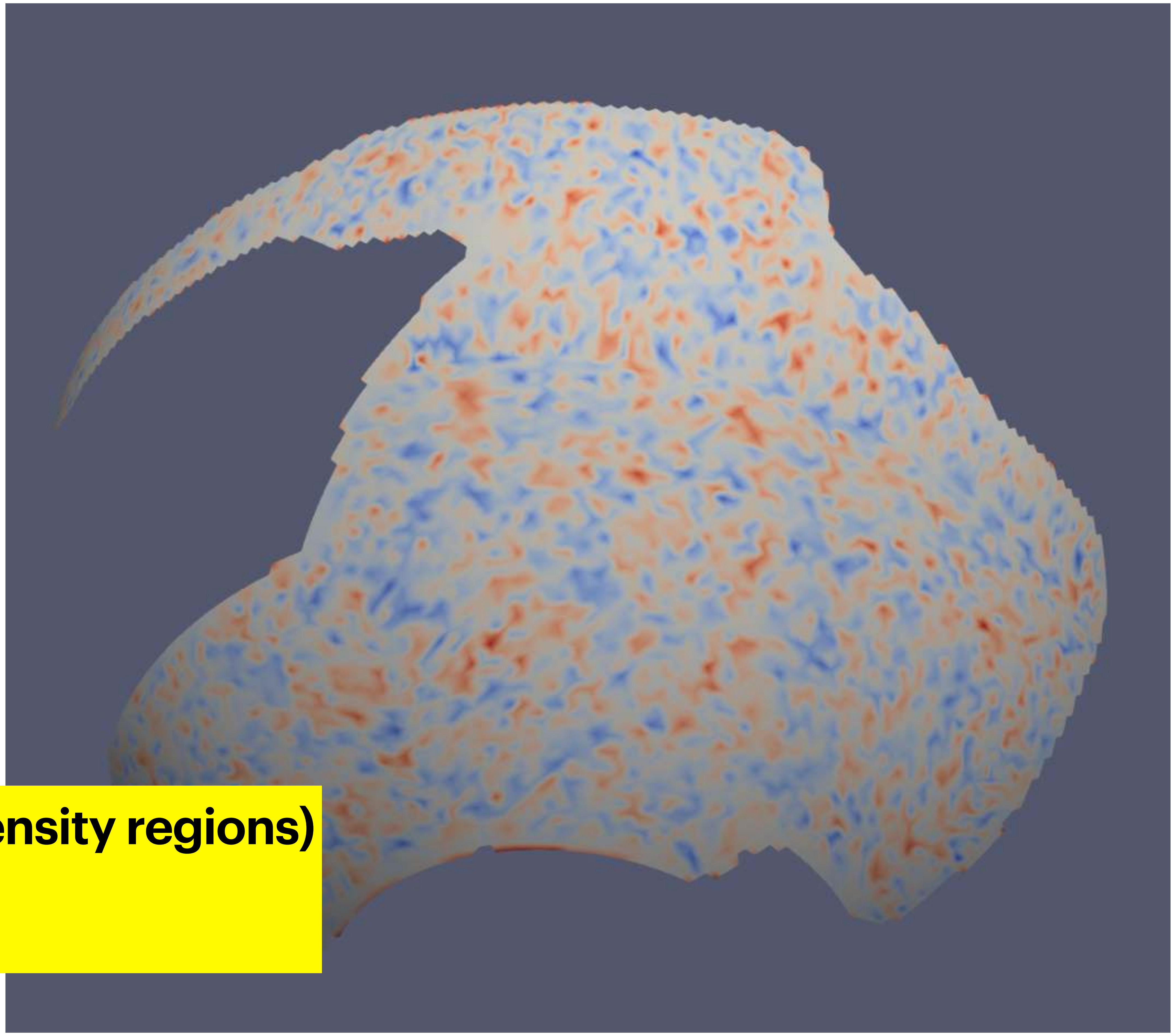


The summary statistic: Persistent homology

WHAT IS PERSISTENT HOMOLOGY?

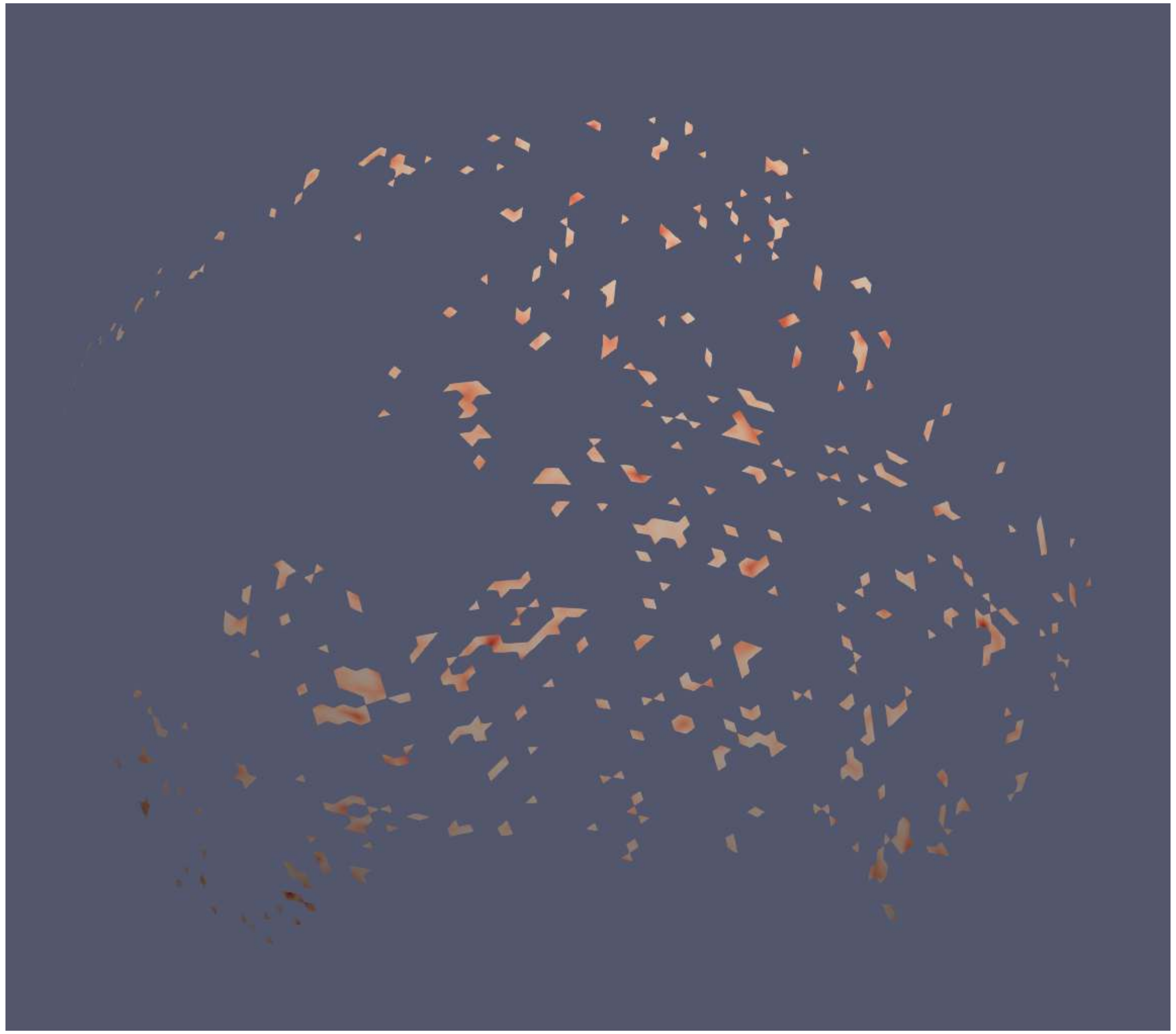
Persistent homology is a mathematical technique from topological data analysis that systematically **tracks how topological features appear and disappear** as the analysis threshold is systematically varied across the entire range of convergence values in the weak lensing mass maps.

- **Connected components (high density regions)**
- **Voids (low density regions)**



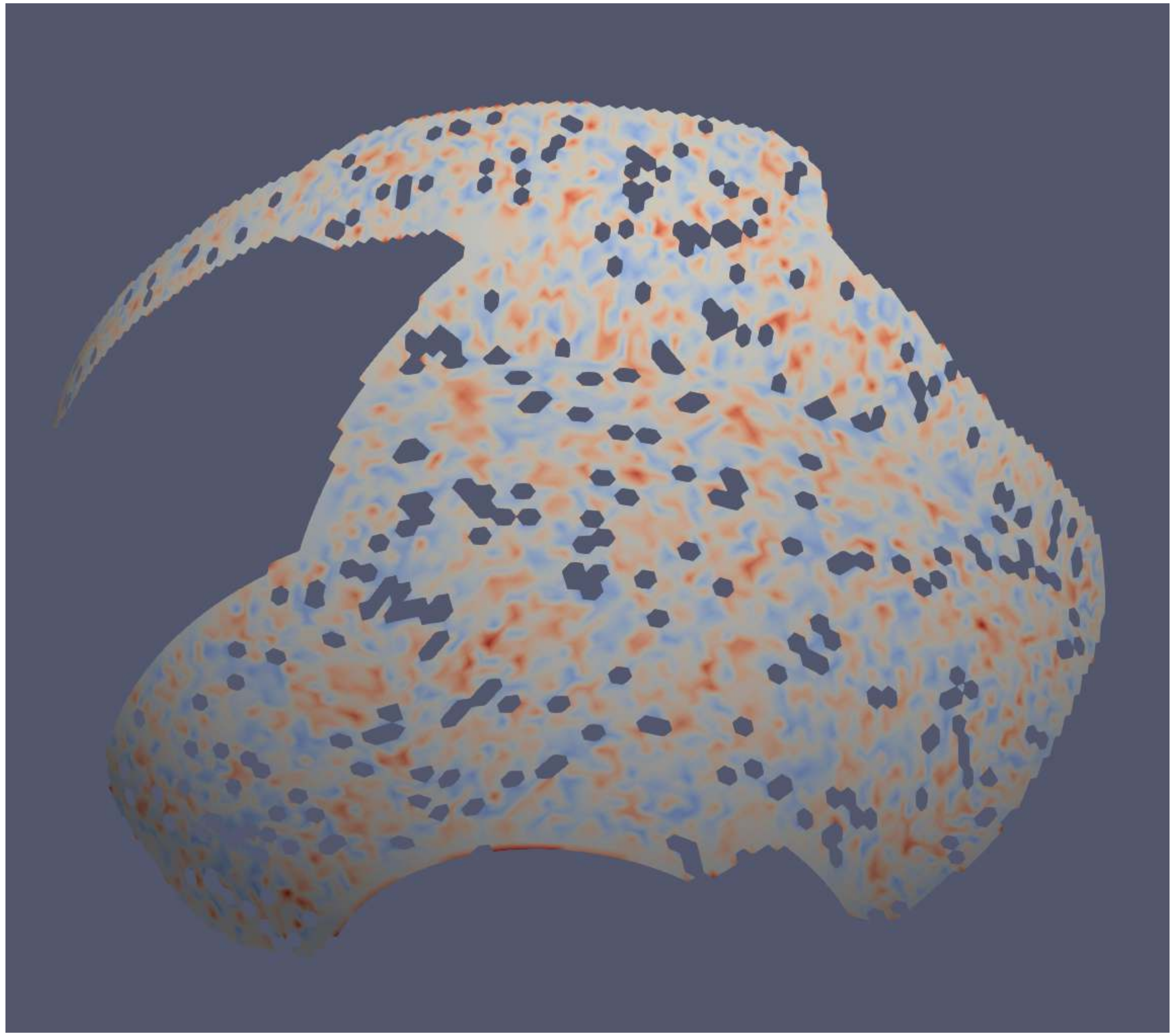
WHAT IS PERSISTENT HOMOLOGY?

Persistent homology is a mathematical technique from topological data analysis that systematically tracks how topological features—specifically **clusters** (***connected high-density regions***) and voids (holes encircled by matter)—appear and disappear as the analysis threshold is systematically varied across the entire range of convergence values in the weak lensing mass maps.



WHAT IS PERSISTENT HOMOLOGY?

Persistent homology is a mathematical technique from topological data analysis that systematically tracks how topological features—specifically clusters (connected high-density regions) and ***voids (holes encircled by matter)***—appear and disappear as the analysis threshold is systematically varied across the entire range of convergence values in the weak lensing mass maps.

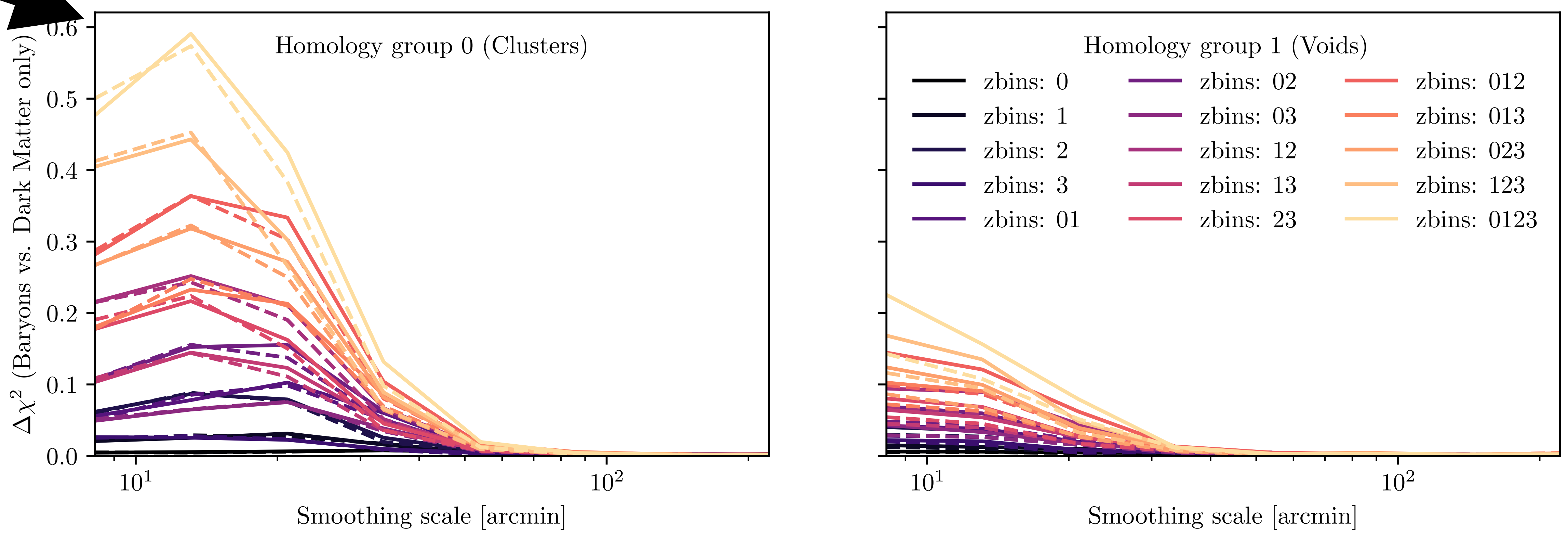


WHY PERSISTENT HOMOLOGY?

Multi-scale topological information that captures higher-order information: Unlike other statistics that probe features at fixed thresholds, persistent homology tracks how topological structures (clusters and voids) evolve across all density thresholds simultaneously.

Physically interpretable features: Different parts of the data vector are directly connected to different physical entities clusters (connected components) and voids (holes), which have different systematics, and thus can be controlled in a "manual way". This contrasts with other methods (such as map-level compressions or statistics that do not distinguish between different physical structures) where separating such systematics is more challenging.

Spherical and directly applicable to native data structure, thus scalable for larger surveys: Other analyses have assumed flat geometry (with some complications). This is the first persistent homology analysis with spherical geometry on the native grid of the data HealPIX (using TopoS2), making it easier to handle the data and avoid removing parts of the data.

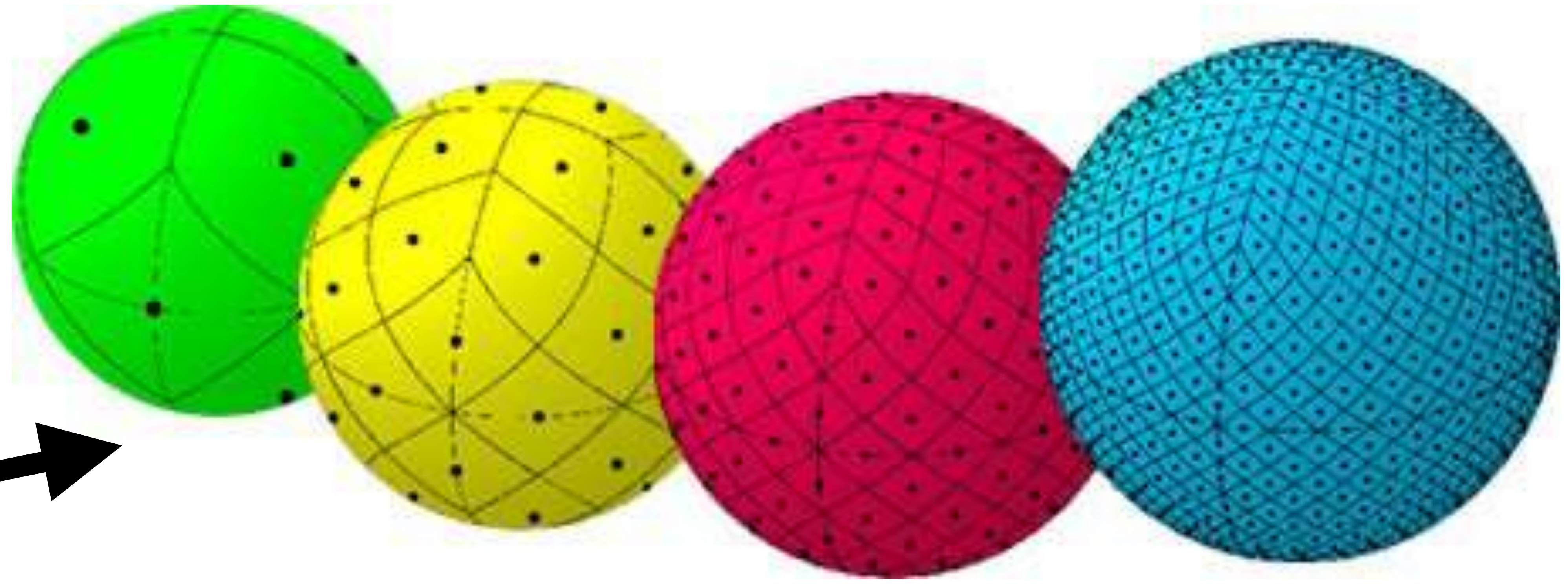


WHY PERSISTENT HOMOLOGY?

Multi-scale topological information: Unlike other statistics that probe features at fixed thresholds, persistent homology tracks how topological structures (clusters and voids) evolve across all density thresholds simultaneously.

Physically interpretable features: Different parts of the data vector are directly connected to different physical entities clusters (connected components) and voids (holes), which have different systematics, and thus can be controlled in a “manual way”. This contrasts with other methods (such as map-level compressions or statistics that do not distinguish between different physical structures) where separating such systematics is more challenging.

Spherical and directly applicable to native data structure, thus scalable for larger surveys: Other analyses have assumed flat geometry (with some complications). This is the first persistent homology analysis with spherical geometry on the native grid of the data HealPIX (using TopoS2), making it easier to handle the data and avoid removing parts of the data.

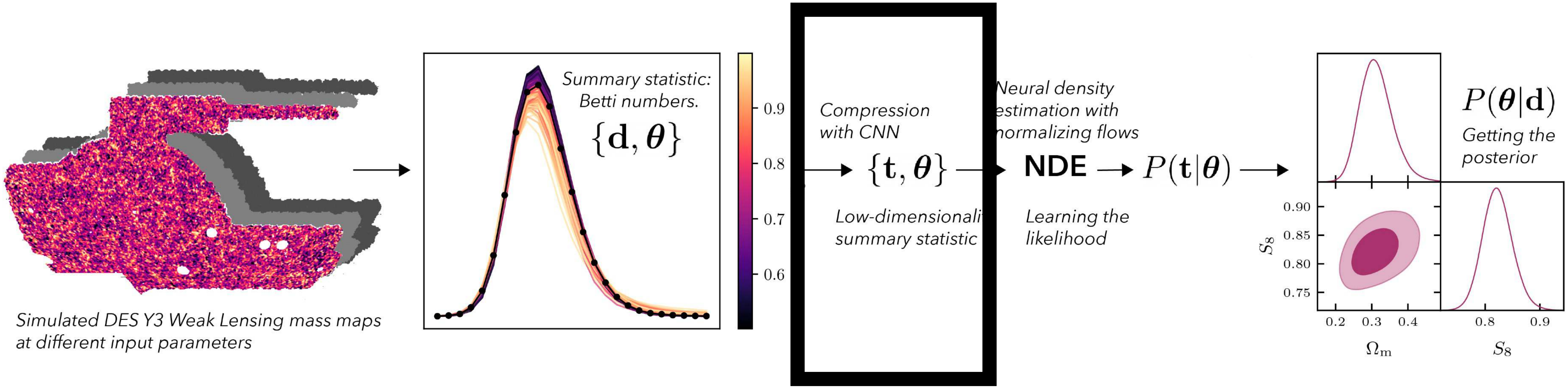


HealPIX grid

arXiv:2506.13439

THE FRAMEWORK

SIMULATION-BASED INFERENCE



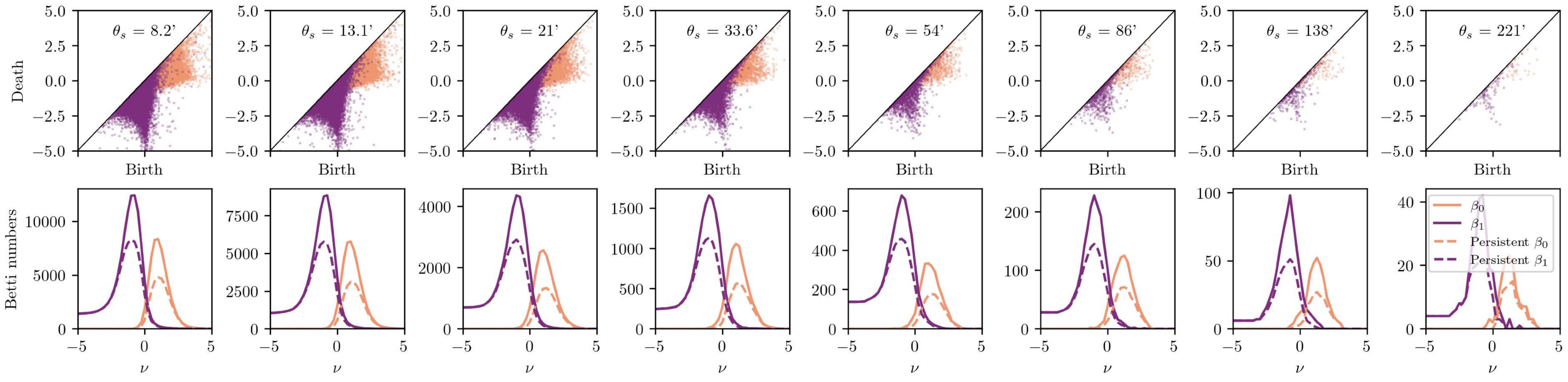
The compression

We use separate neural networks for each target parameter.

We train with mean-squared error (MSE) loss.

THE BETTI NUMBERS DATA VECTOR

FOR DES Y3 DATA

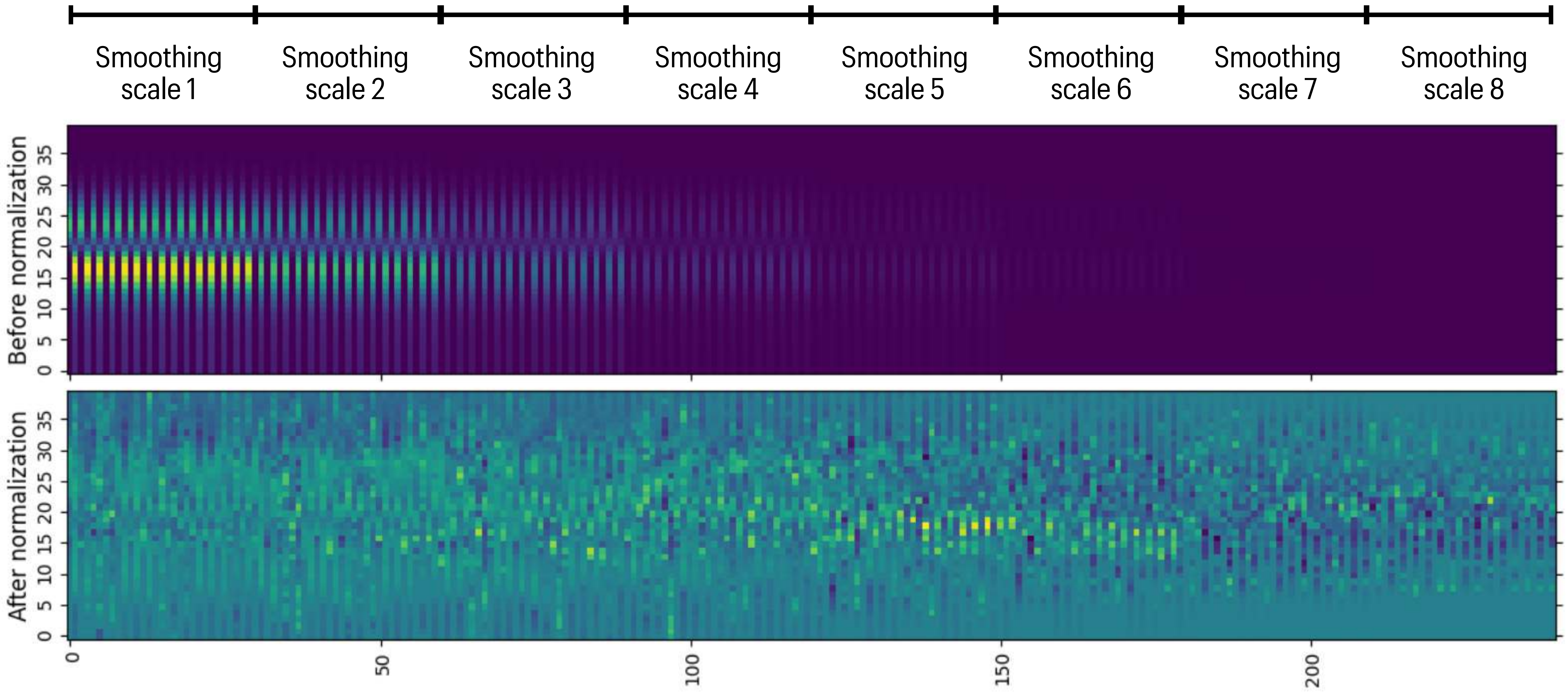


DATA IN ITS ORIGINAL FORM

THE CHALLENGE:

Betti numbers data vector has 9,600 points= 40 points per Betti curve × 8 smoothing scales × 15 redshift bins × 2 homology groups

We have this for ~10000 simulations in the training set.

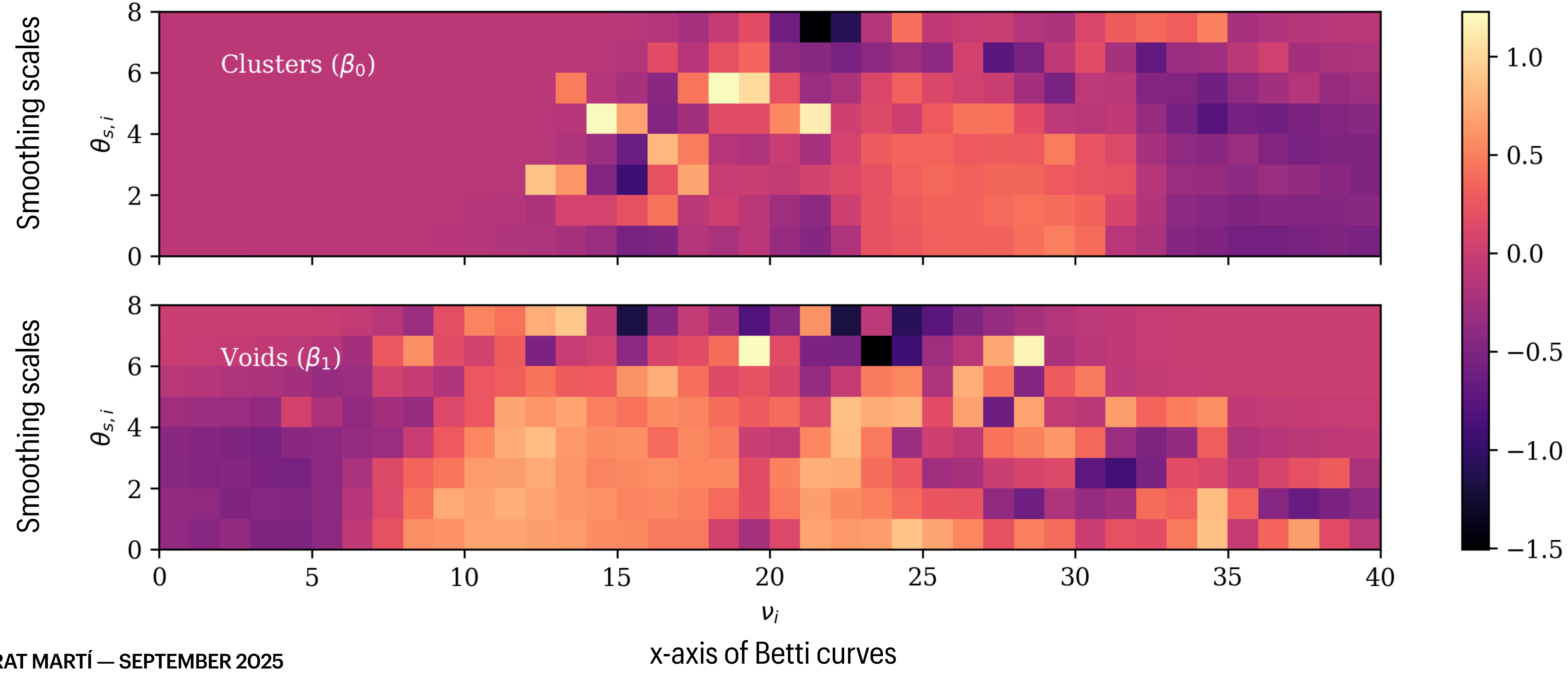


We first tried a fully connected layer as it had been done in analyses of other statistics, but it didn't work well...

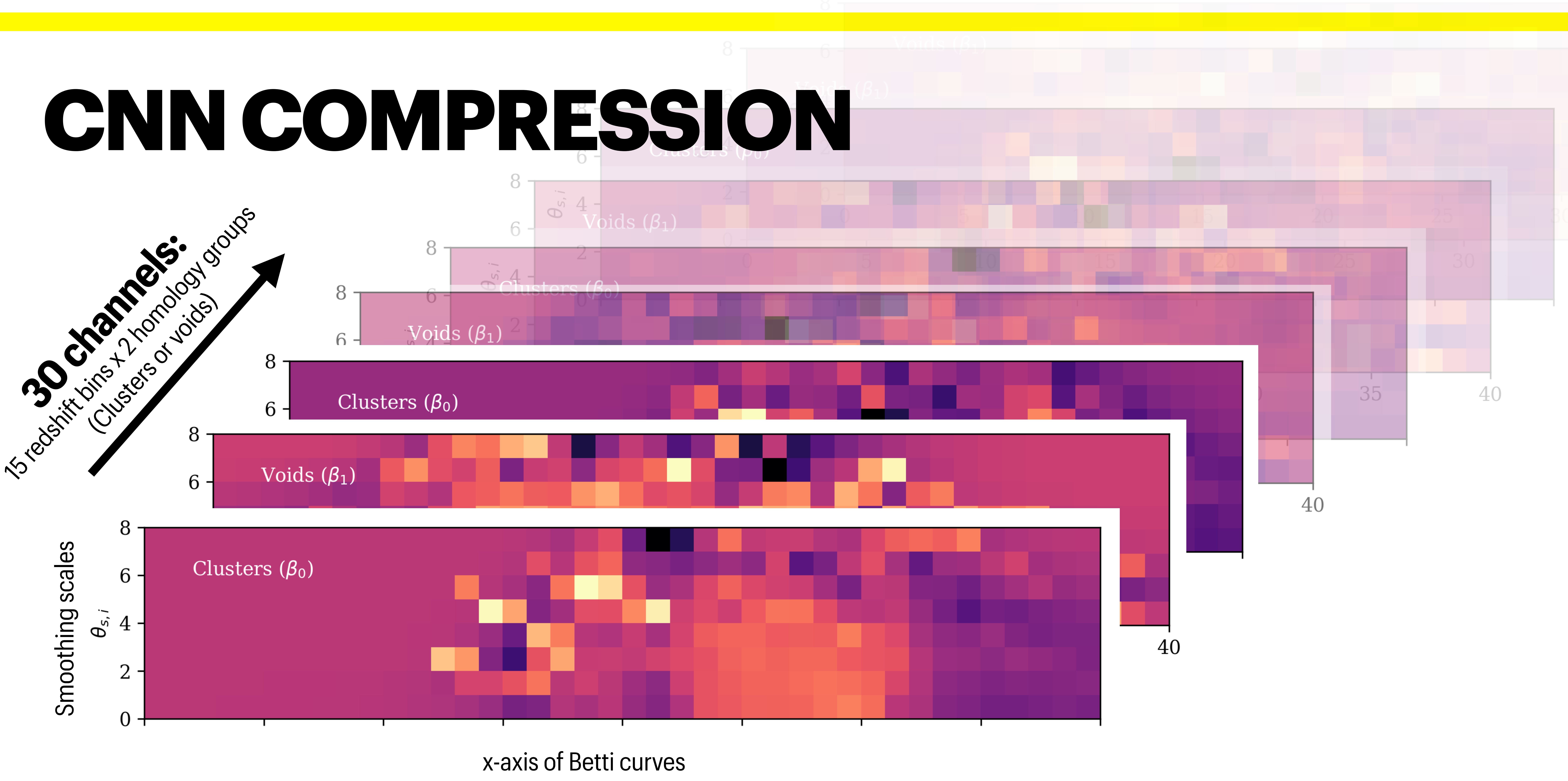
On persistent homology, previous analyses were using a rudimentary manual method to compress the data, so we had to innovate..

CNN COMPRESSION

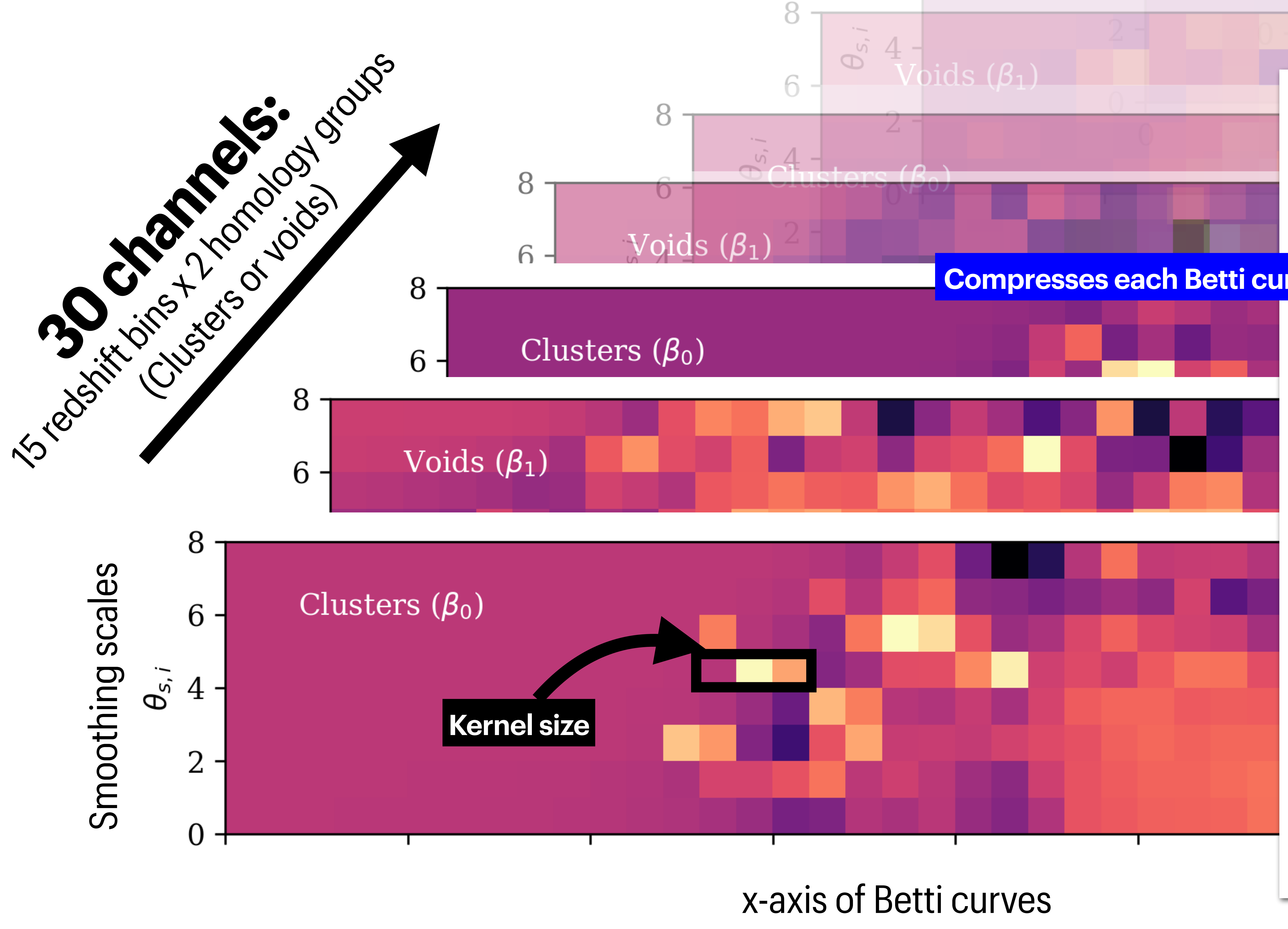
THE SOLUTION: Specialized CNN architecture designed to efficiently capture the correlations between smoothing scales and redshift bins.



CNN COMPRESSION

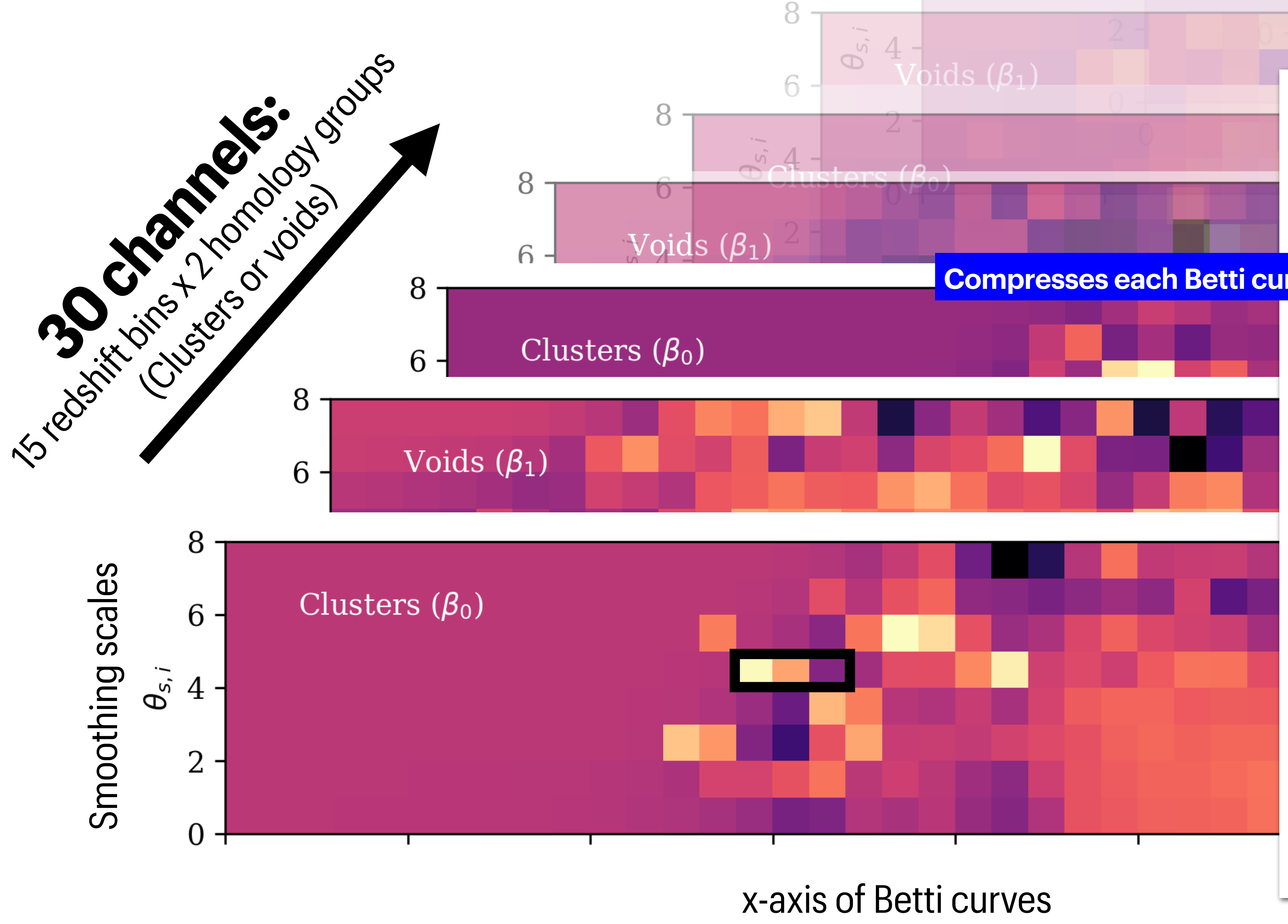


CNN COMPRESSION



Layer Type	Output Shape	Kernel Size	Parameters
Input	(40, 240)	-	0
Reshape	(40, 8, 30)	-	0
SeparableConv2D	(40, 8, 32)	(3, 1)	1,082
BatchNormalization	(40, 8, 32)	-	128
SeparableConv2D	(40, 8, 32)	(5, 3)	1,536
BatchNormalization	(40, 8, 32)	-	128
AveragePooling2D	(20, 8, 32)	(2, 1)	0
Conv2D	(20, 8, 16)	(3, 3)	4,624
BatchNormalization	(20, 8, 16)	-	64
AveragePooling2D	(10, 4, 16)	(2, 2)	0
Conv2D	(10, 4, 8)	(3, 3)	1,160
Flatten	(320)	-	0
Dropout (0.3)	(320)	-	0
Dense	(32)	-	10,272
Dense	(1)	-	33
		Total Parameters:	19,027

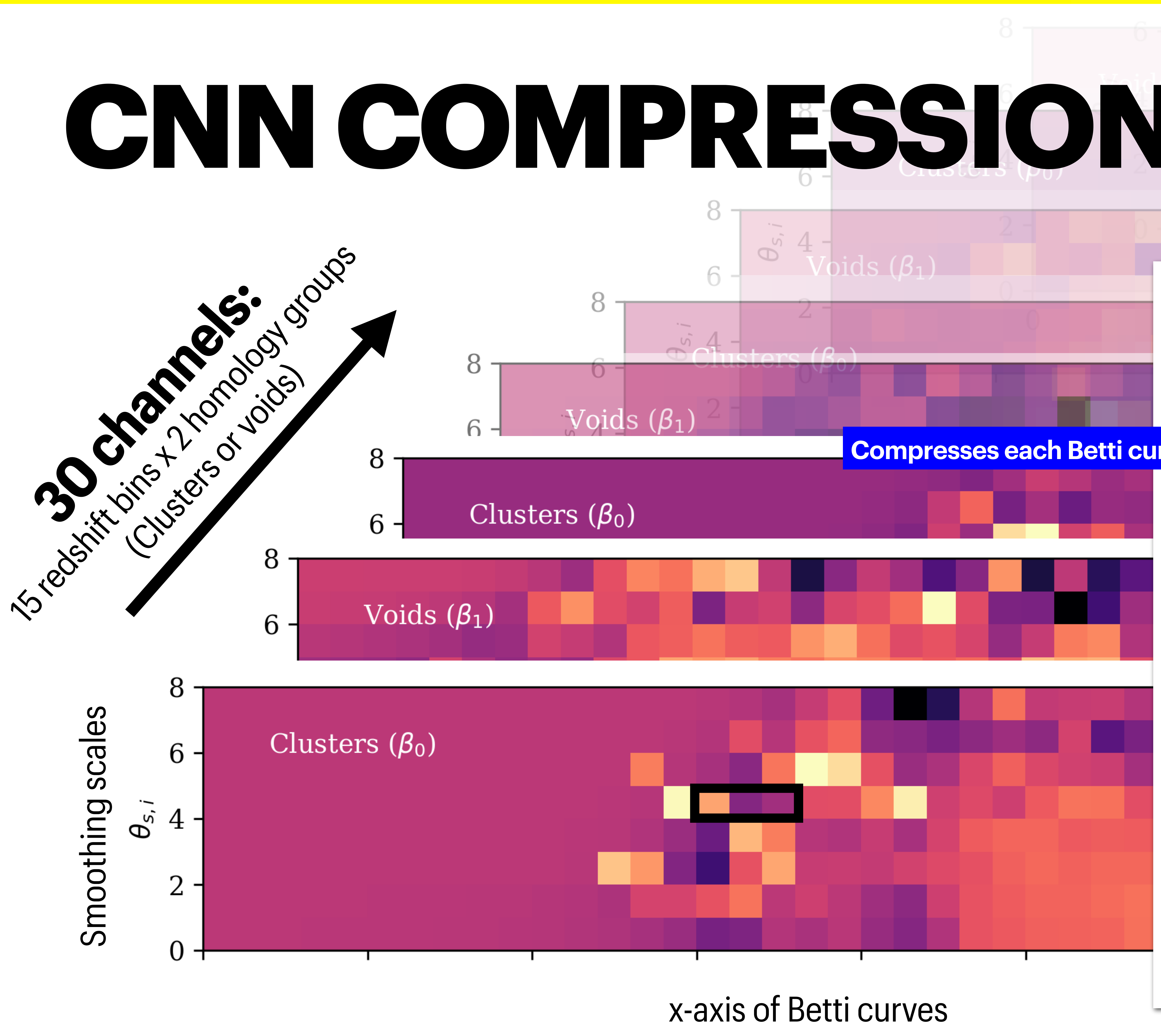
CNN COMPRESSION



Compresses each Betti curve:

Layer Type	Output Shape	Kernel Size	Parameters
Input	(40, 240)	-	0
Reshape	(40, 8, 30)	-	0
SeparableConv2D	(40, 8, 32)	(3, 1)	1,082
BatchNormalization	(40, 8, 32)	-	128
SeparableConv2D	(40, 8, 32)	(5, 3)	1,536
BatchNormalization	(40, 8, 32)	-	128
AveragePooling2D	(20, 8, 32)	(2, 1)	0
Conv2D	(20, 8, 16)	(3, 3)	4,624
BatchNormalization	(20, 8, 16)	-	64
AveragePooling2D	(10, 4, 16)	(2, 2)	0
Conv2D	(10, 4, 8)	(3, 3)	1,160
Flatten	(320)	-	0
Dropout (0.3)	(320)	-	0
Dense	(32)	-	10,272
Dense	(1)	-	33
		Total Parameters:	19,027

CNN COMPRESSION



Compresses each Betti curve:

Layer Type	Output Shape	Kernel Size	Parameters
Input	(40, 240)	-	0
Reshape	(40, 8, 30)	-	0
SeparableConv2D	(40, 8, 32)	(3, 1)	1,082
BatchNormalization	(40, 8, 32)	-	128
SeparableConv2D	(40, 8, 32)	(5, 3)	1,536
BatchNormalization	(40, 8, 32)	-	128
AveragePooling2D	(20, 8, 32)	(2, 1)	0
Conv2D	(20, 8, 16)	(3, 3)	4,624
BatchNormalization	(20, 8, 16)	-	64
AveragePooling2D	(10, 4, 16)	(2, 2)	0
Conv2D	(10, 4, 8)	(3, 3)	1,160
Flatten	(320)	-	0
Dropout (0.3)	(320)	-	0
Dense	(32)	-	10,272
Dense	(1)	-	33
		Total Parameters:	19,027

SEPARABLE CONV 2D OR *DEPTHWISE-SEPARABLE*

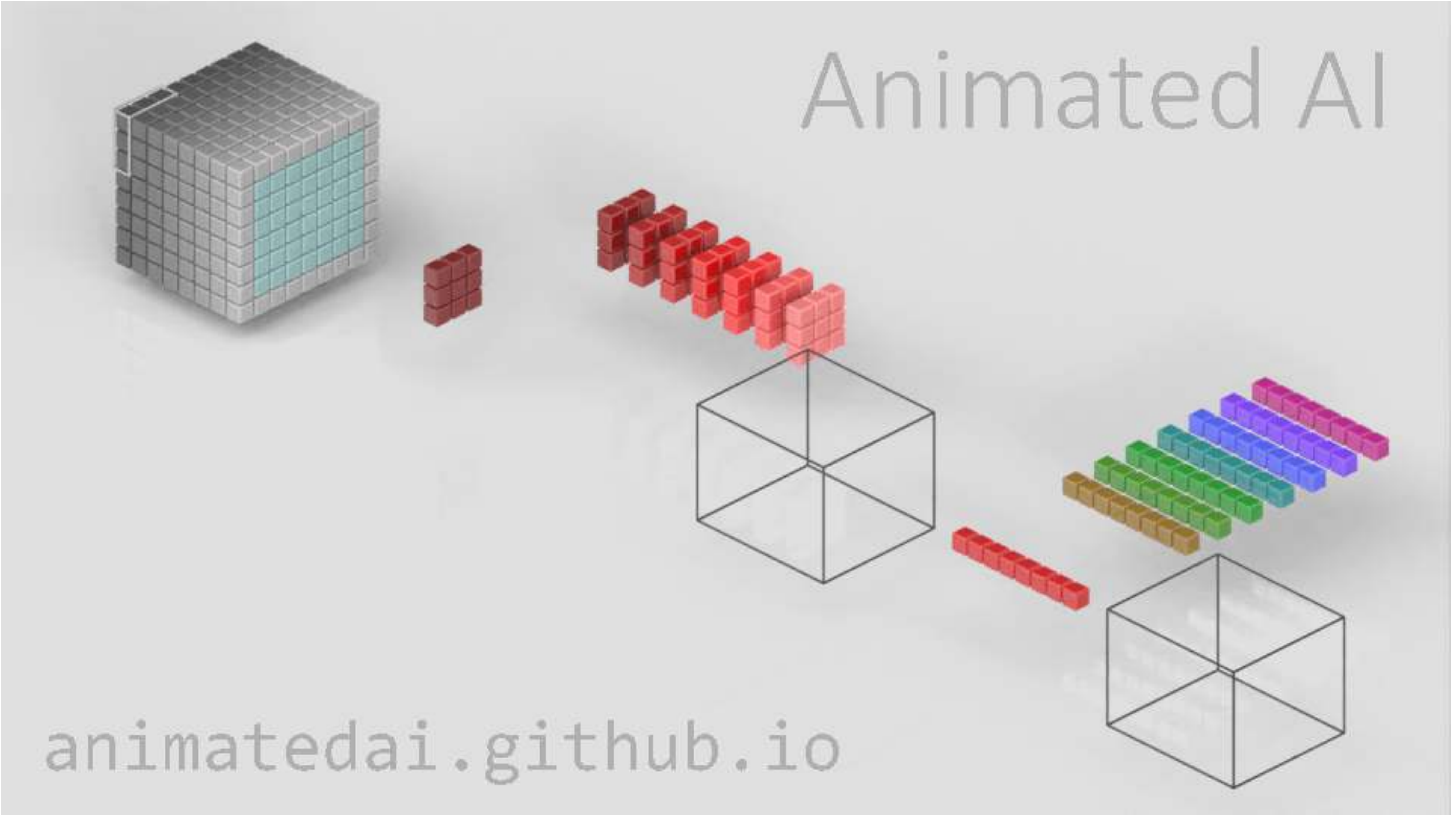
WHAT IS IT?

An efficient approximation that reduces computational cost and model parameters by separating the filtering and combination steps:

- **Standard convolutions** perform filtering and feature combination simultaneously.

- **Depthwise-separable** convolutions first use a depthwise convolution that acts separately on channels and then a pointwise convolution to combine the channels.

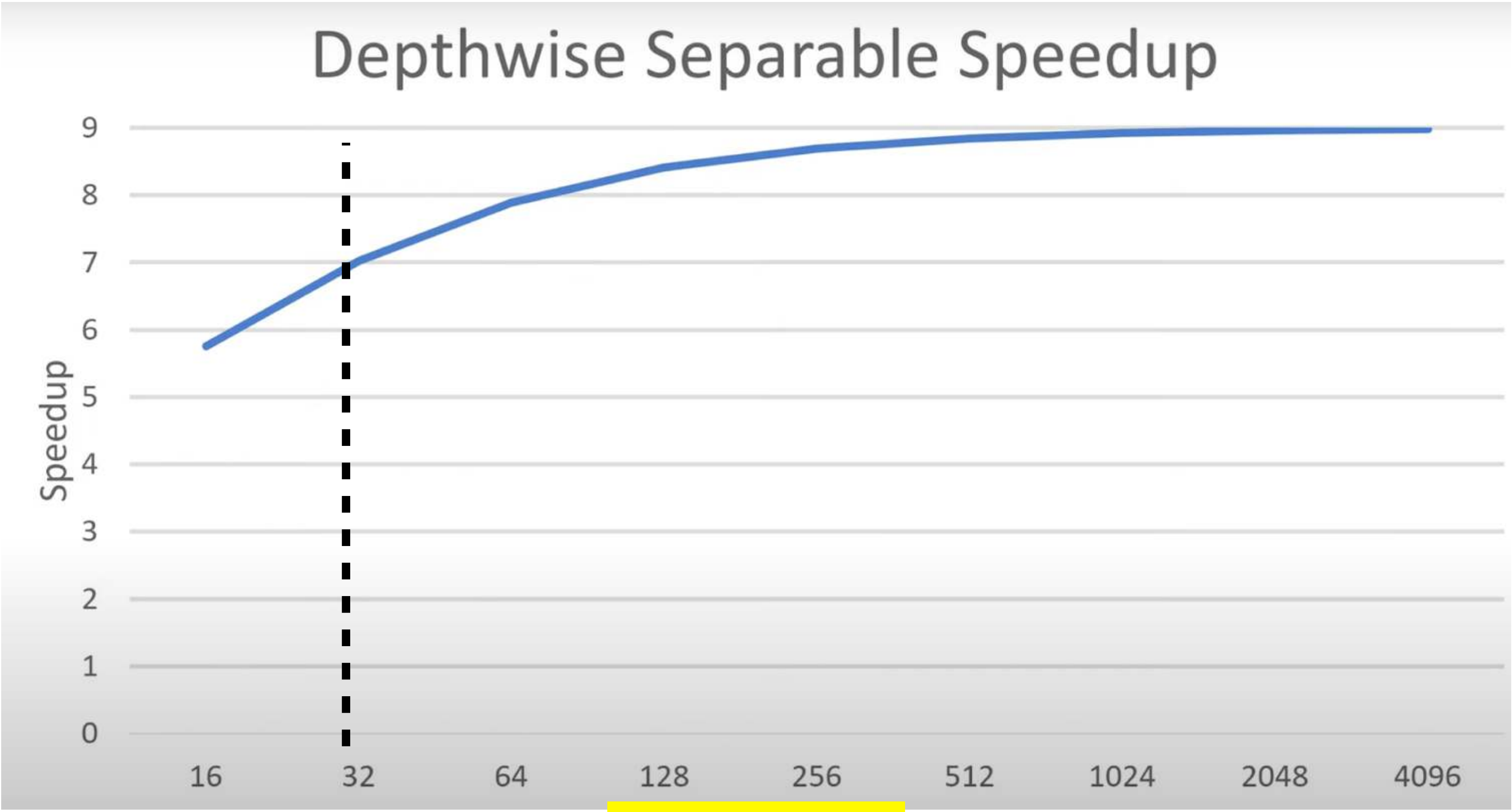
This approach allows the network to initially learn features specific to each redshift bin configuration and topological structure (clusters or voids) separately.



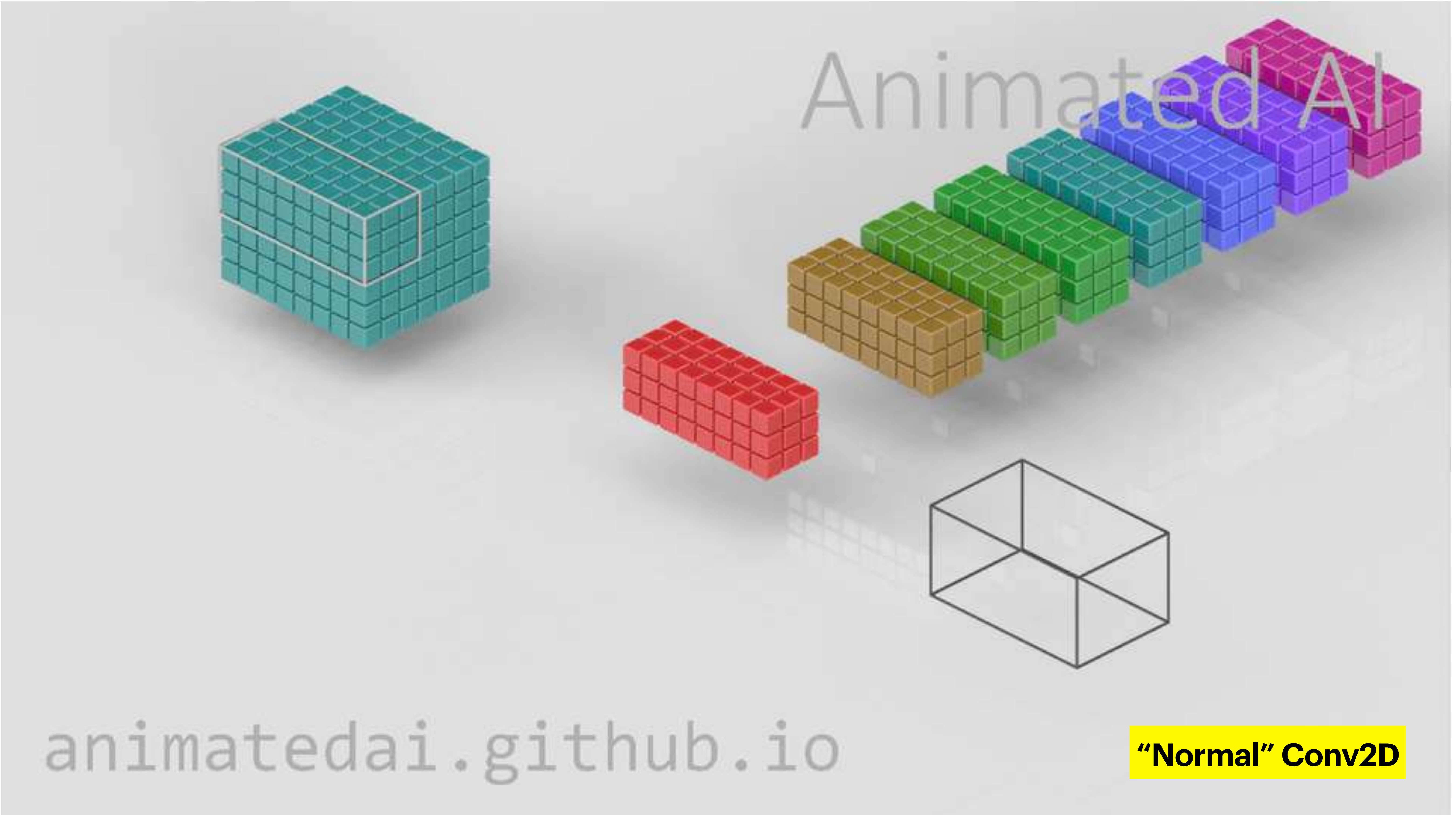
HOW DO THEY COMPARE TO NORMAL CONVULTIONS?

THEY ARE MORE EFFICIENT

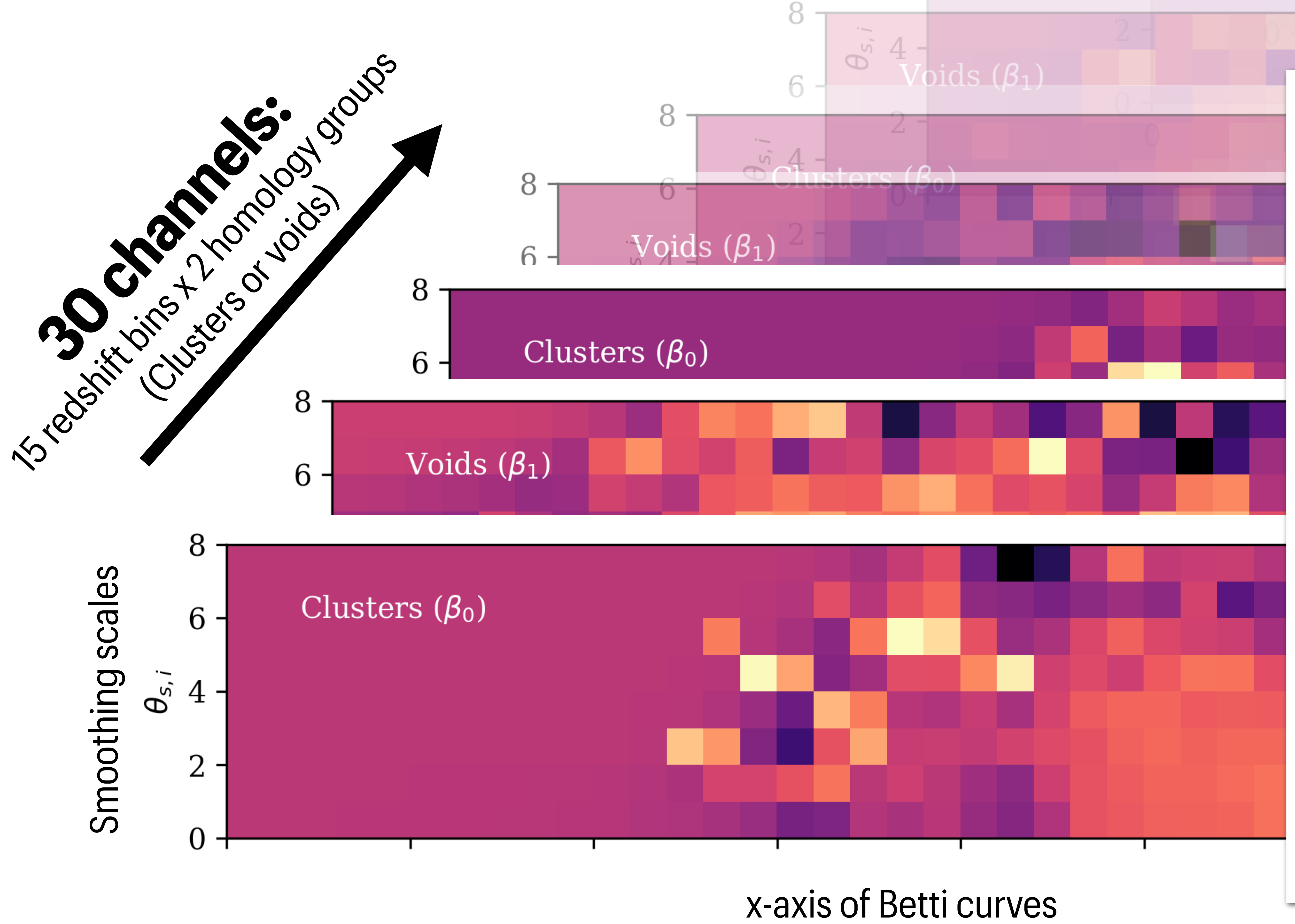
This approach significantly reduces the parameter count compared to standard convolutions.



# of channels

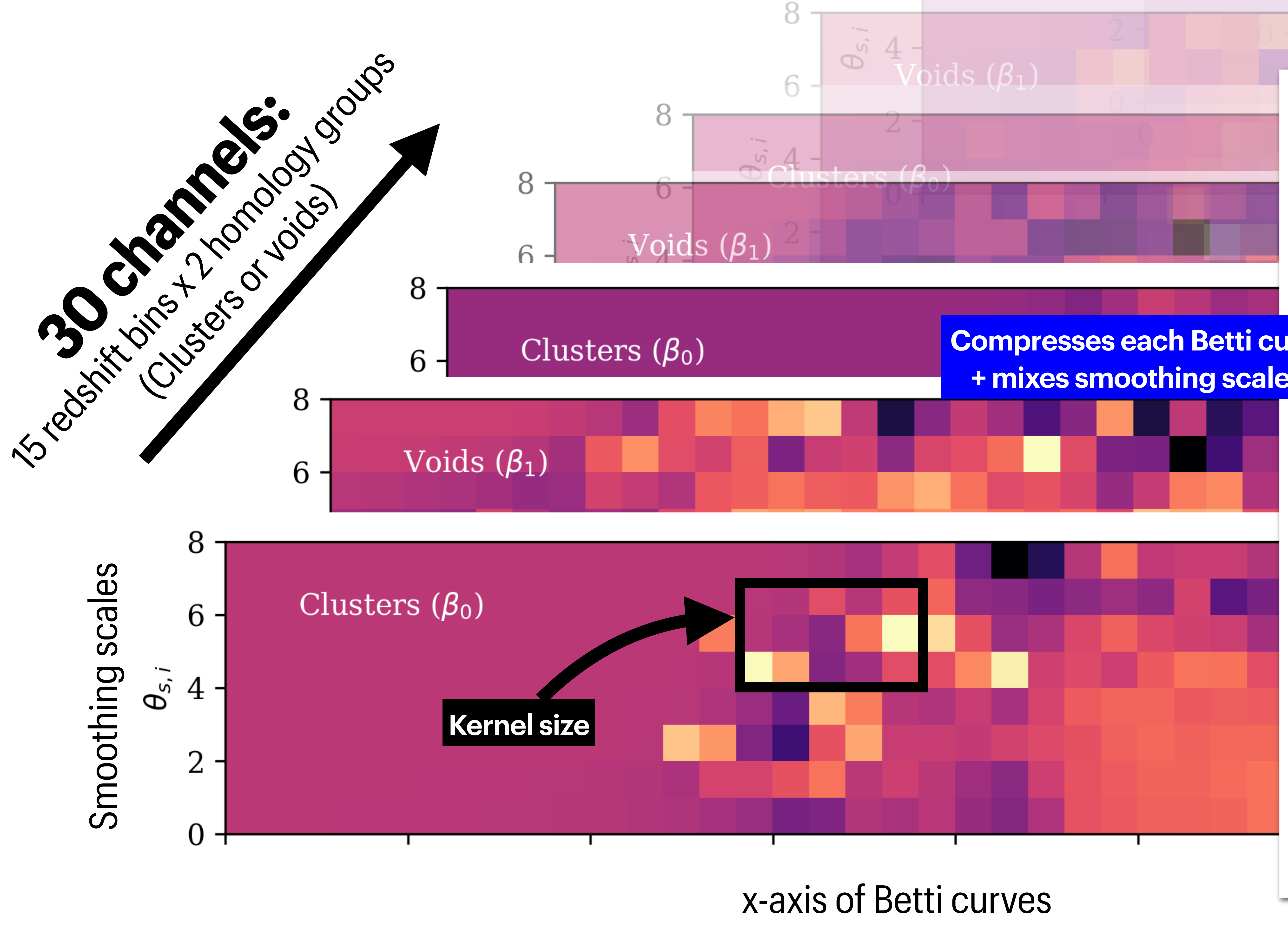


CNN COMPRESSION



Layer Type	Output Shape	Kernel Size	Parameters
Input	(40, 240)	-	0
Reshape	(40, 8, 30)	-	0
SeparableConv2D	(40, 8, 32)	(3, 1)	1,082
BatchNormalization	(40, 8, 32)	-	128
SeparableConv2D	(40, 8, 32)	(5, 3)	1,536
BatchNormalization	(40, 8, 32)	-	128
AveragePooling2D	(20, 8, 32)	(2, 1)	0
Conv2D	(20, 8, 16)	(3, 3)	4,624
BatchNormalization	(20, 8, 16)	-	64
AveragePooling2D	(10, 4, 16)	(2, 2)	0
Conv2D	(10, 4, 8)	(3, 3)	1,160
Flatten	(320)	-	0
Dropout (0.3)	(320)	-	0
Dense	(32)	-	10,272
Dense	(1)	-	33
		Total Parameters:	19,027

CNN COMPRESSION



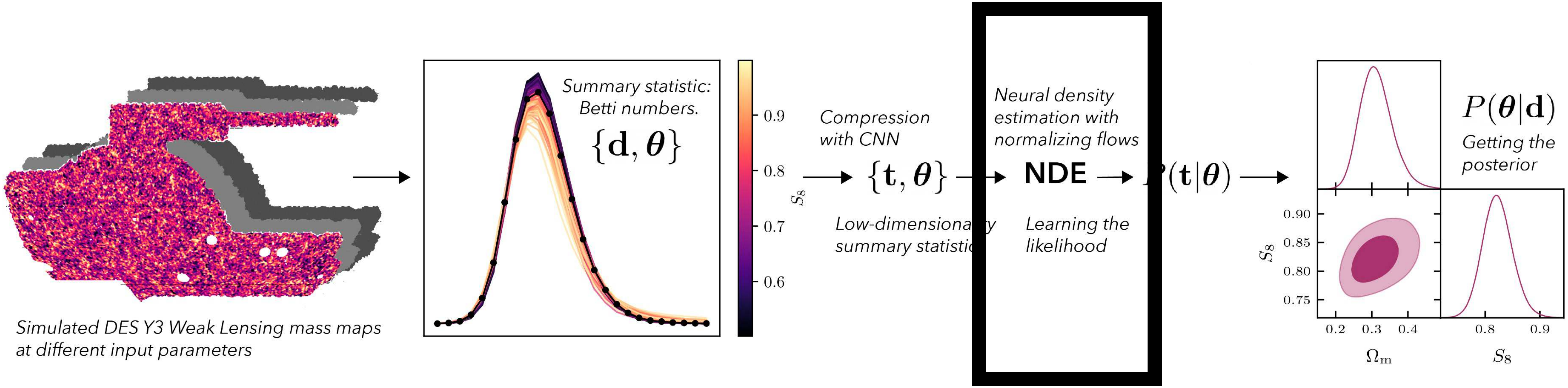
Compresses each Betti curve + mixes smoothing scales:

Layer Type	Output Shape	Kernel Size	Parameters
Input	(40, 240)	-	0
Reshape	(40, 8, 30)	-	0
SeparableConv2D	(40, 8, 32)	(3, 1)	1,082
BatchNormalization	(40, 8, 32)	-	128
SeparableConv2D	(40, 8, 32)	(5, 3)	1,536
BatchNormalization	(40, 8, 32)	-	128
AveragePooling2D	(20, 8, 32)	(2, 1)	0
Conv2D	(20, 8, 16)	(3, 3)	4,624
BatchNormalization	(20, 8, 16)	-	64
AveragePooling2D	(10, 4, 16)	(2, 2)	0
Conv2D	(10, 4, 8)	(3, 3)	1,160
Flatten	(320)	-	0
Dropout (0.3)	(320)	-	0
Dense	(32)	-	10,272
Dense	(1)	-	33
		Total Parameters:	19,027

arXiv:2506.13439

THE FRAMEWORK

SIMULATION-BASED INFERENCE



The Neural Density Estimation

NEURAL LIKELIHOOD ESTIMATION

From the pairs of compressed statistics and input parameters $\{t_i, \theta_i\}$, we learn a density q that approximates p, such that $p(t|\theta) \approx q(t|\theta)$. For this purpose, we use the pyDELFI package[2] (Alsing et al. 2019) to estimate $p(t|\theta)$ with an ensemble of neural density estimators (NDEs). NDEs use neural networks to parameterize densities, including conditional probability densities. An NDE estimates $q(t|\theta, \varphi)$ by varying the neural network parameters φ (e.g., weights and biases) to minimize the loss function

$$U(\varphi) = -\sum_{i=1}^{N} \log q(t_i|\theta_i; \varphi) \tag{2}$$

over N mock data points t_i. This loss function minimizes the Kullback-Leibler divergence (Kullback & Leibler 1951) between the true conditional density $p(t|\theta)$ and our approximation $q(t|\theta; \varphi)$. The KL divergence is defined as:

$$D_{\mathrm{KL}}(p(t|\theta)||q(t|\theta; \varphi)) = \int p(t|\theta) \log \left(\frac{p(t|\theta)}{q(t|\theta; \varphi)} \right) dt, \tag{3}$$

which can be expanded as:

$$D_{\mathrm{KL}} = \int p(t|\theta) \log p(t|\theta) dt - \int p(t|\theta) \log q(t|\theta; \varphi) dt. \tag{4}$$

pyDELFI package: https://github.com/justinalsing/pydelfi

An important advantage of neural likelihood estimation (as opposed to learning the posterior directly) is that the learned function is independent of the proposal distribution used to generate the training samples.

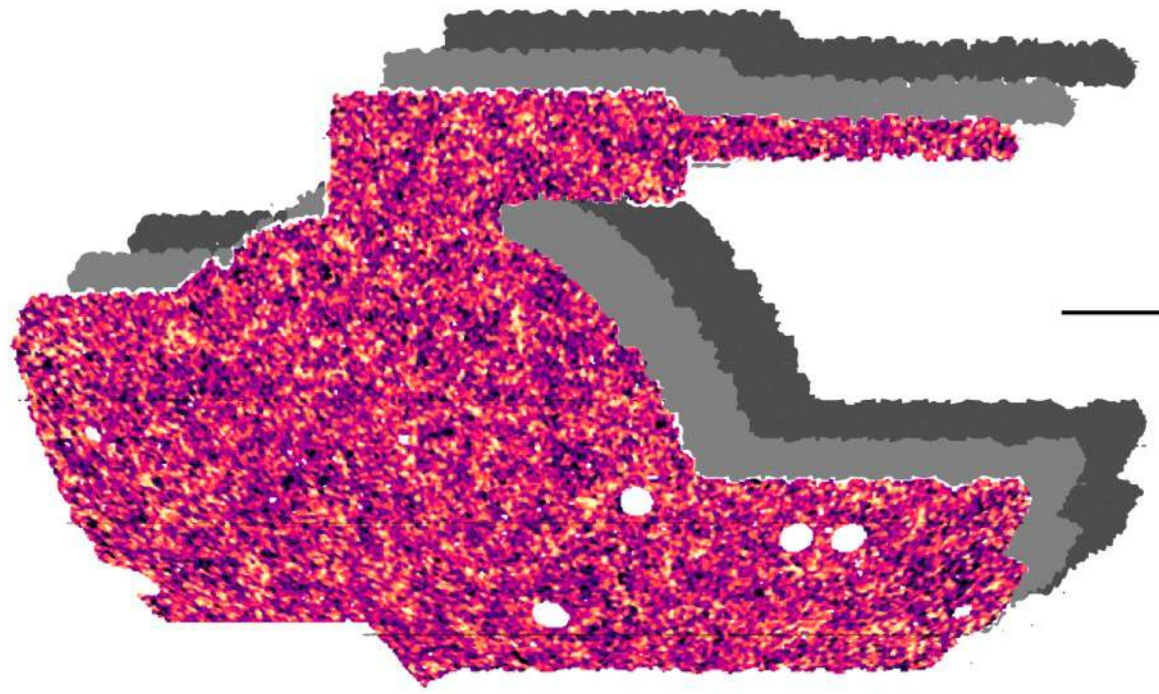
arXiv:2506.13439

THE FRAMEWORK

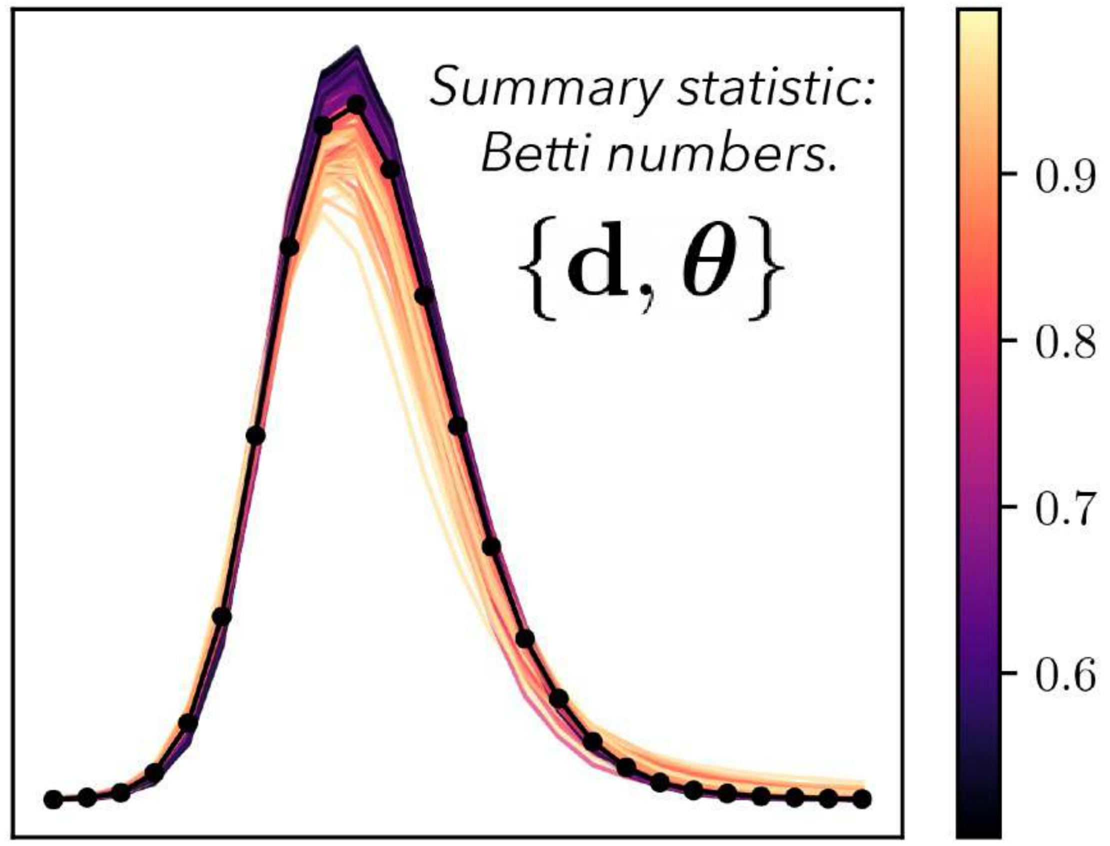
SIMULATION-BASED INFERENCE

Regular Bayesian inference with MCMC

$$p(\theta|t_O) = \frac{p(t_O|\theta)\, p(\theta)}{p(t_O)}$$



Simulated DES Y3 Weak Lensing mass maps at different input parameters



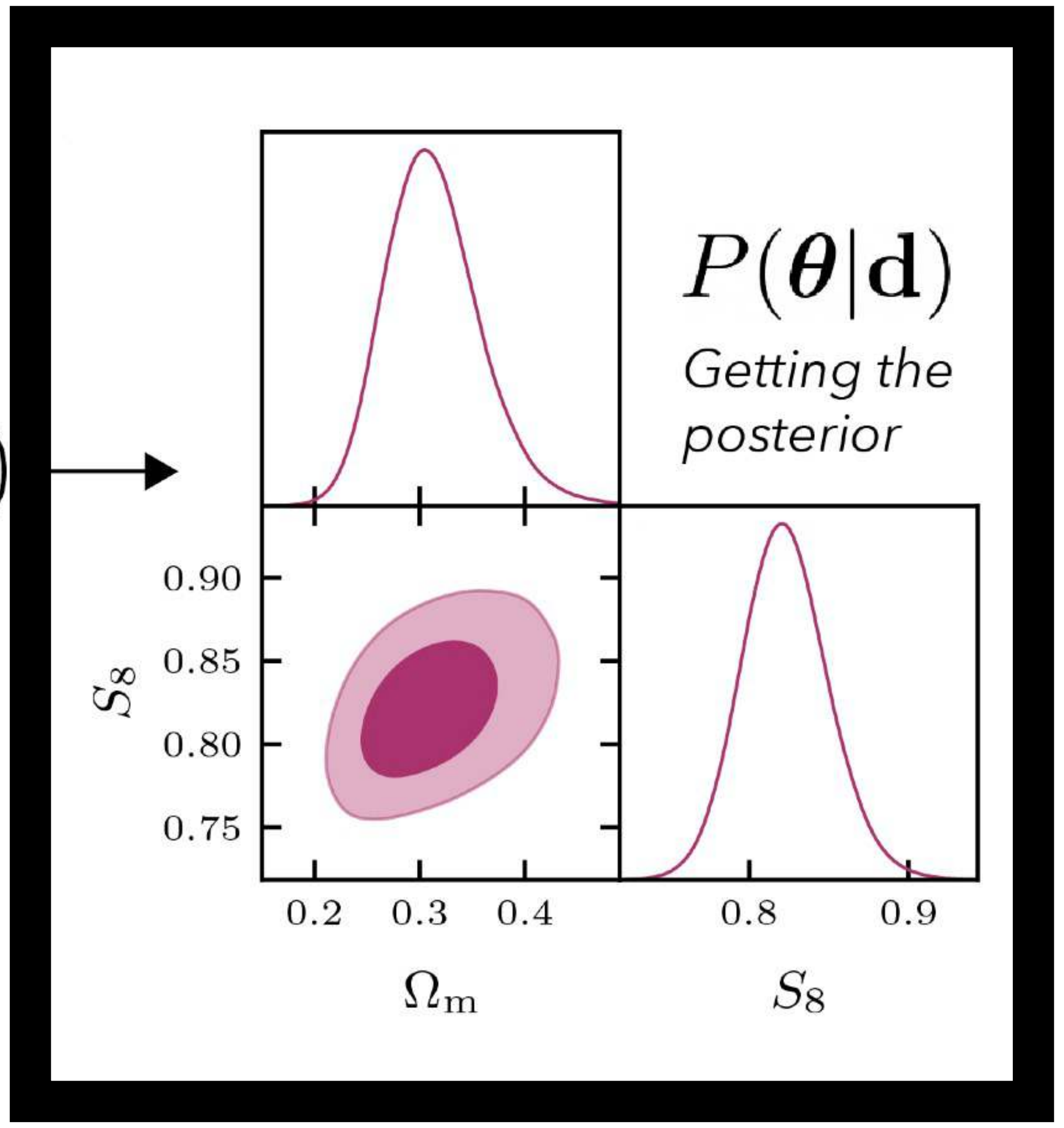
Compression with CNN

$\{\mathbf{t}, \boldsymbol{\theta}\}$ → NDE → $P(\mathbf{t}|\boldsymbol{\theta})$ →

Low-dimensionality summary statistic

Neural density estimation with normalizing flows

Learning the likelihood



Advantage is that this is very fast, the MCMC runs in ~1 min, compared to ~days in regular analyses with an analytical likelihood

Evaluating the learned likelihood (Blinded analysis)

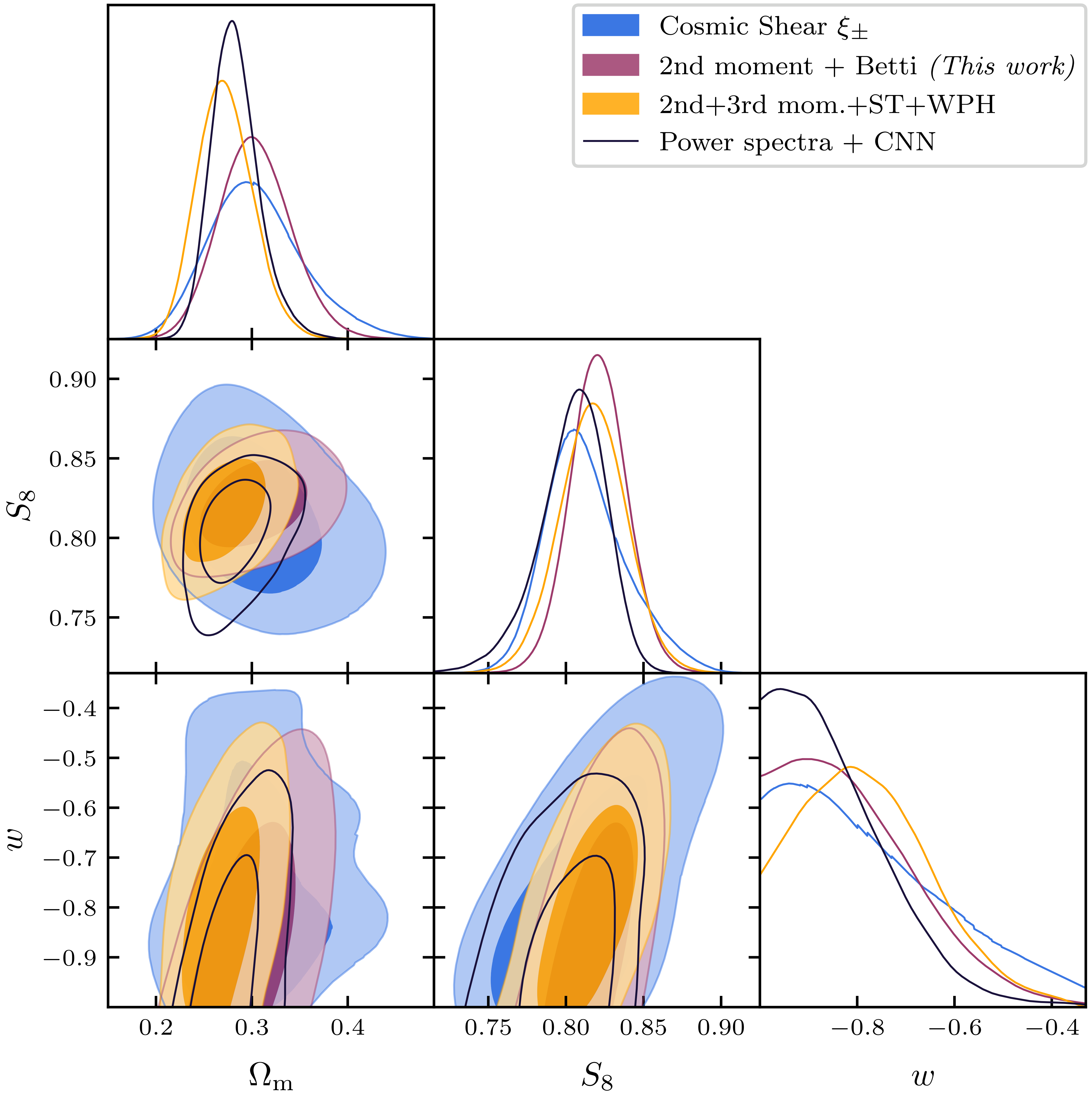
TALK BASED ON:
arXiv:2506.13439

DARK ENERGY SURVEY YEAR 3 RESULTS: wCDM COSMOLOGY FROM SIMULATION-BASED INFERENCE WITH PERSISTENT HOMOLOGY ON THE SPHERE

ABSTRACT

We present cosmological constraints from Dark Energy Survey Year 3 (DES Y3) weak lensing data using persistent homology, a topological data analysis technique that tracks how features like clusters and voids evolve across density thresholds. For the first time, we apply spherical persistent homology to galaxy survey data through the algorithm TopoS2, which is optimized for curved-sky analyses and HEALPix compatibility. Employing a simulation-based inference framework with the Gower Street simulation suite—specifically designed to mimic DES Y3 data properties—we extract topological summary statistics from convergence maps across multiple smoothing scales and redshift bins. After neural network compression of these statistics, we estimate the likelihood function and validate our analysis against baryonic feedback effects, finding minimal biases (under 0.3σ) in the $\Omega_m - S_8$ plane. Assuming the wCDM model, our combined Betti numbers and second moments analysis yields $S_8 = 0.821 \pm 0.018$ and $\Omega_m = 0.304 \pm 0.037$—constraints 70% tighter than those from cosmic shear two-point statistics in the same parameter plane. Our results demonstrate that topological methods provide a powerful and robust framework for extracting cosmological information, with our spherical methodology readily applicable to upcoming Stage IV wide-field galaxy surveys.

Bringing analyses end to end on precursor data to identify challenges and mature methods for upcoming next-generation surveys...



OUTLOOK AND PERSPECTIVES

- Simulation-based inference using persistent homology as a summary statistic is a powerful, feasible and scalable way of extracting robust cosmological information from weak lensing data. The pipeline is ready to be applied to upcoming surveys.

- It's important to target tools to data structure, and sometimes that's not easy. We had to borrow tools from other contexts, and develop new CNN architectures to tackle our particular case.

- It's important to stay close to data to understand what are the main roadblocks.

What is needed to make this **the headline result?**

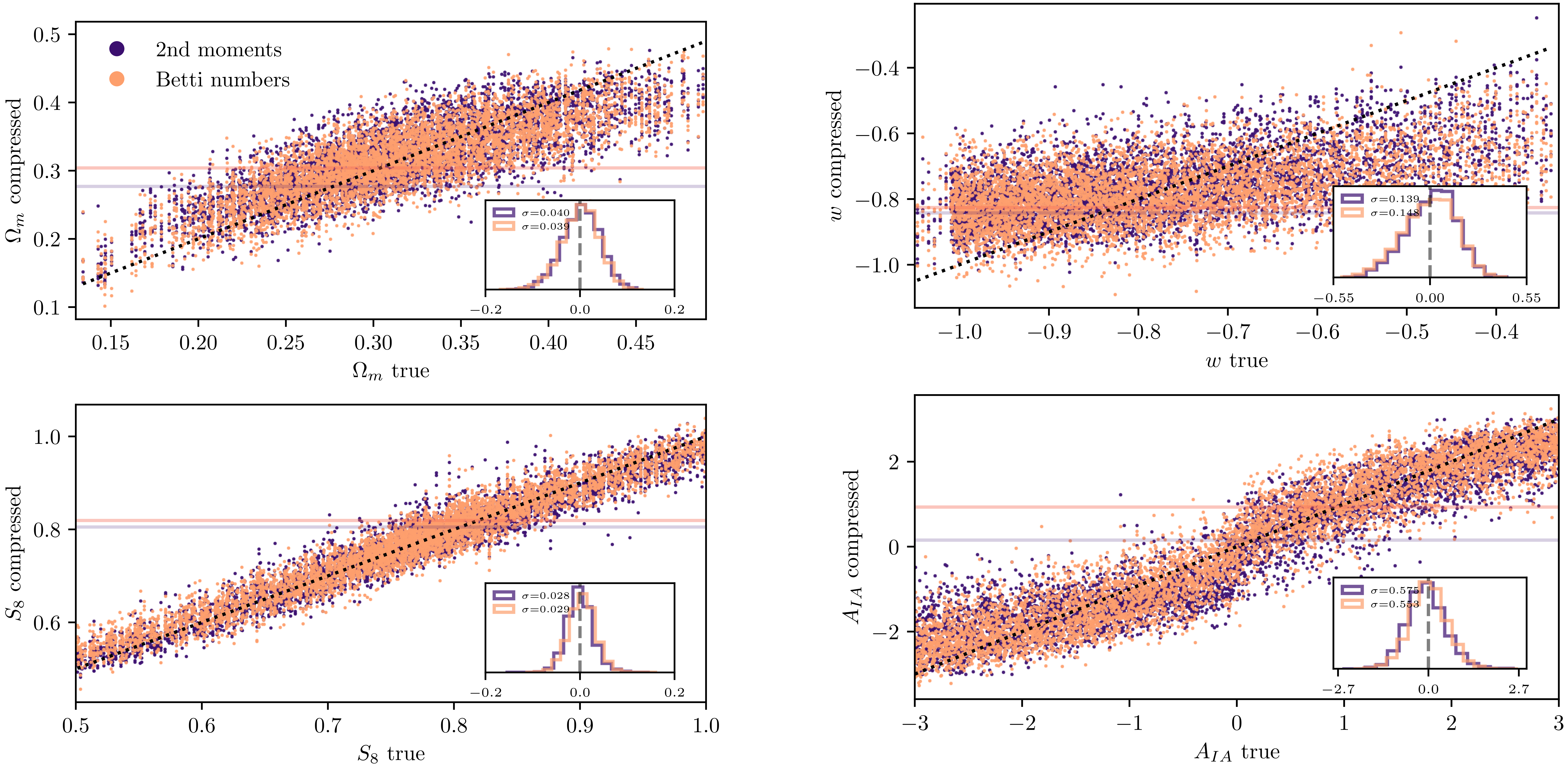
- More testing against systematics (validate against more simulations with baryons, simulations with better IA models), etc. (Developing better simulations)

- Develop simulations with other cosmological models


THANK YOU!

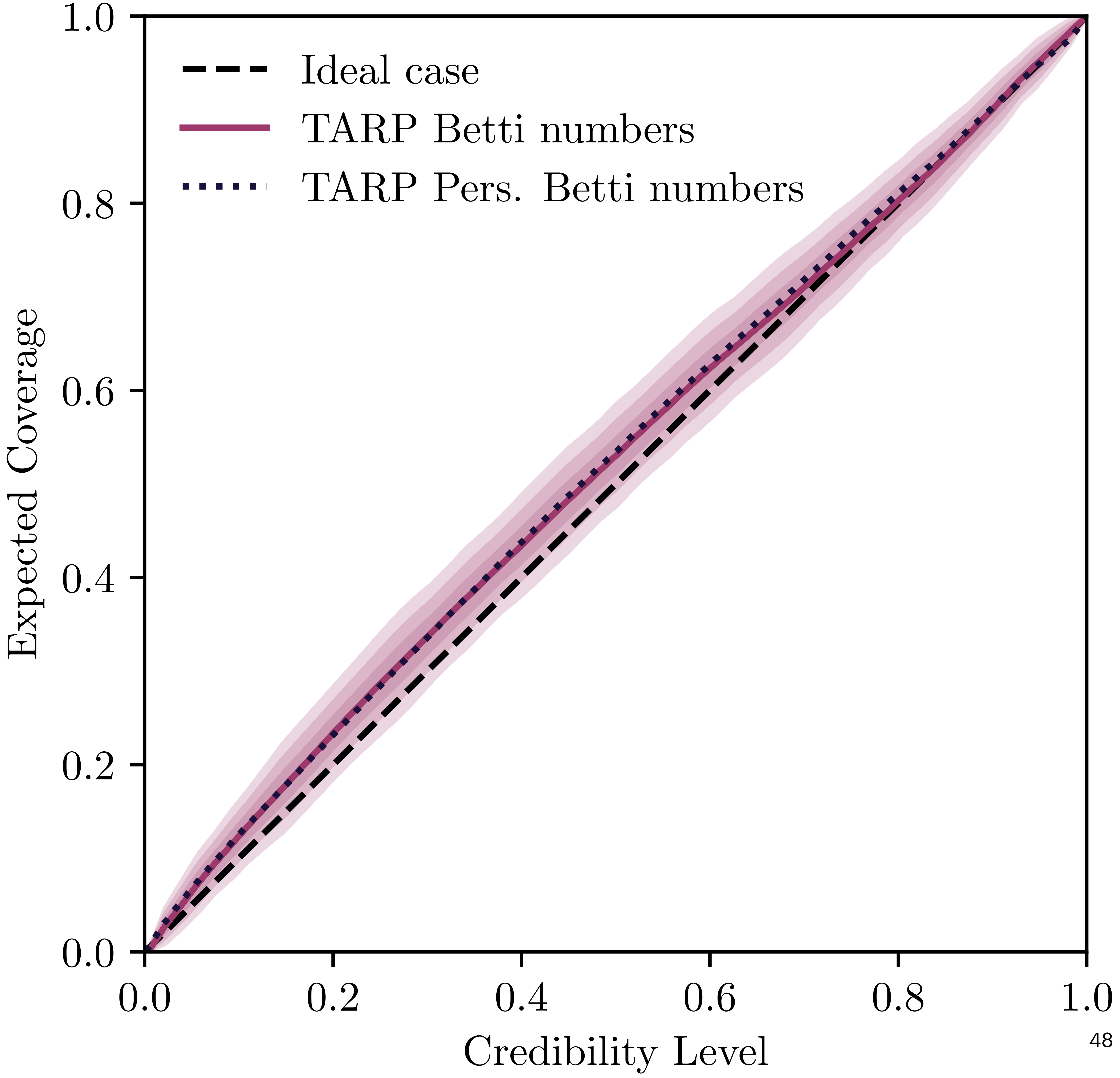
COMPRESSION

BETTI NUMBERS AND 2ND MOMENTS



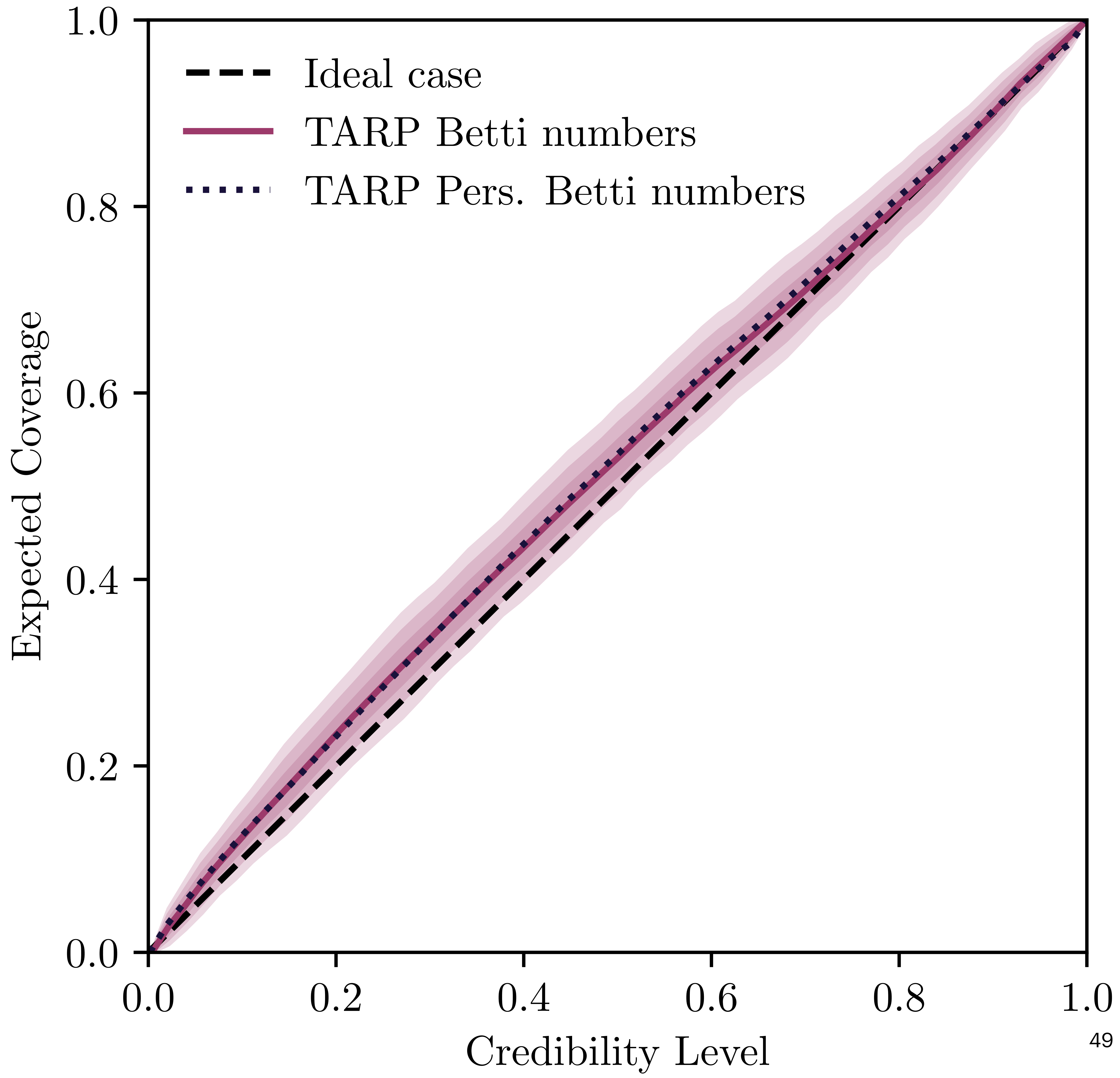
COVERAGE TEST

TARP

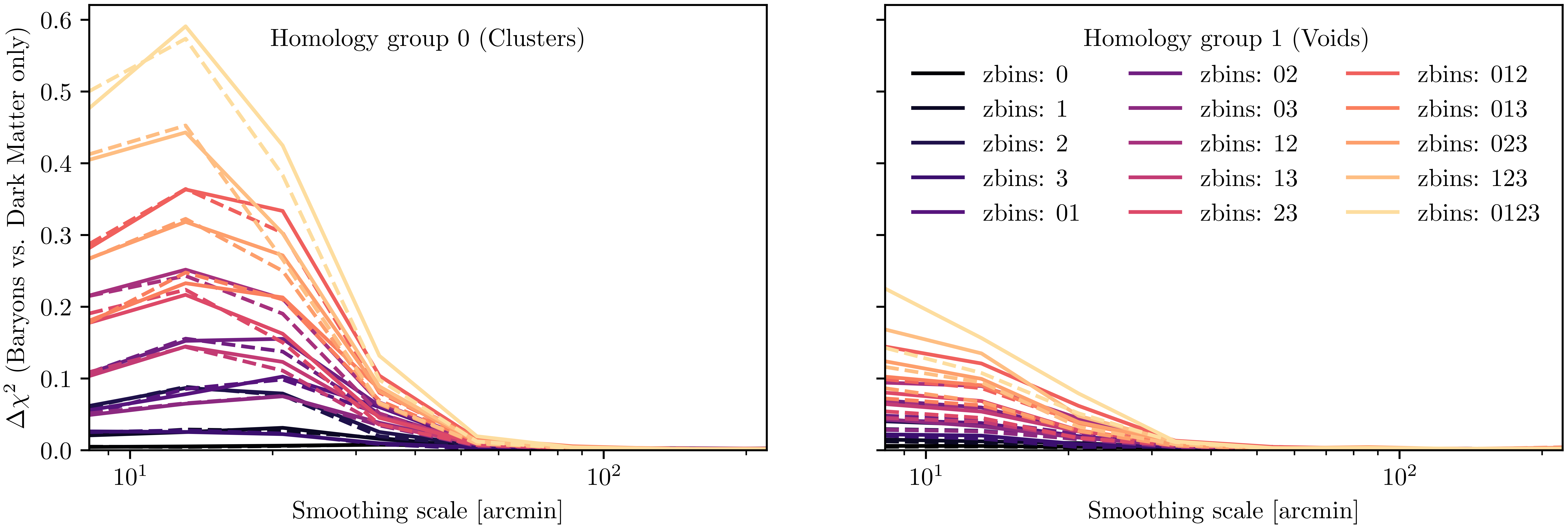


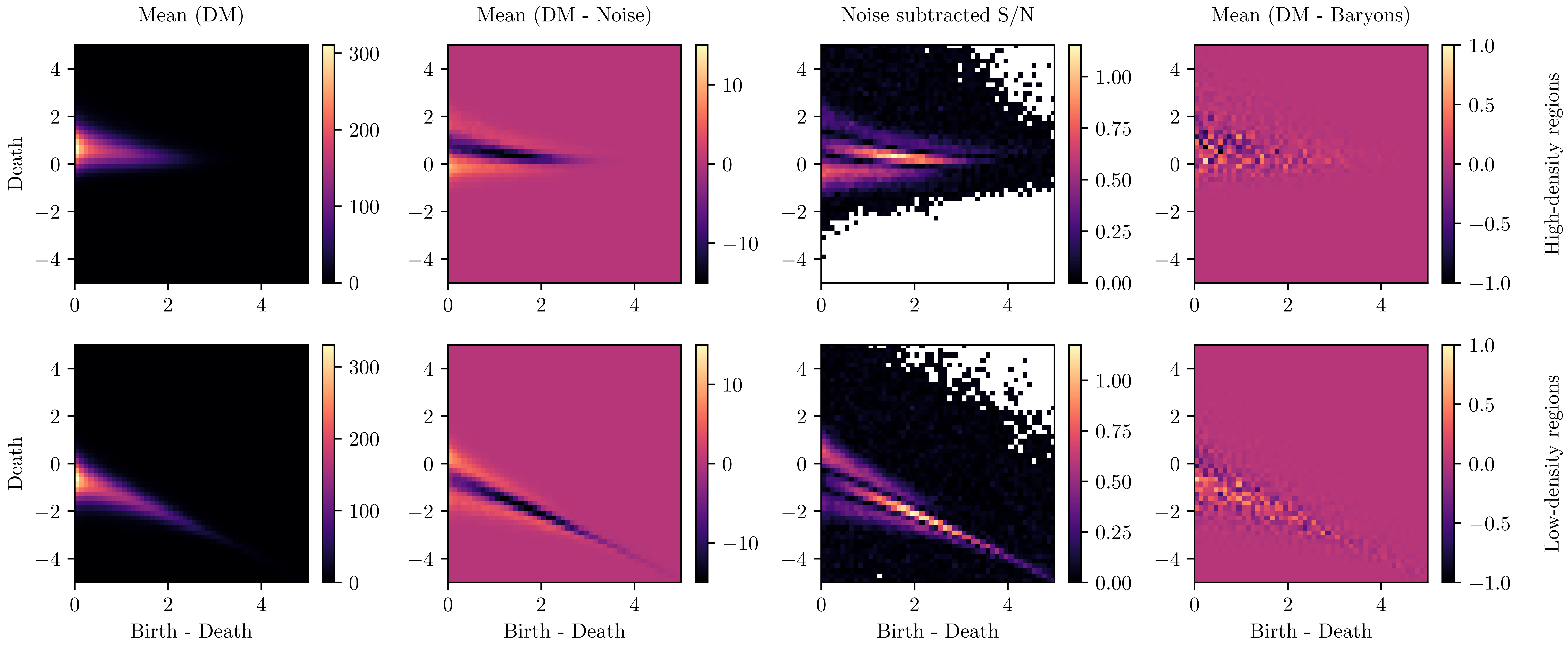
COVERAGE TEST

TARP



BARYONIC EFFECTS IMPACT ON BETTI NUMBERS

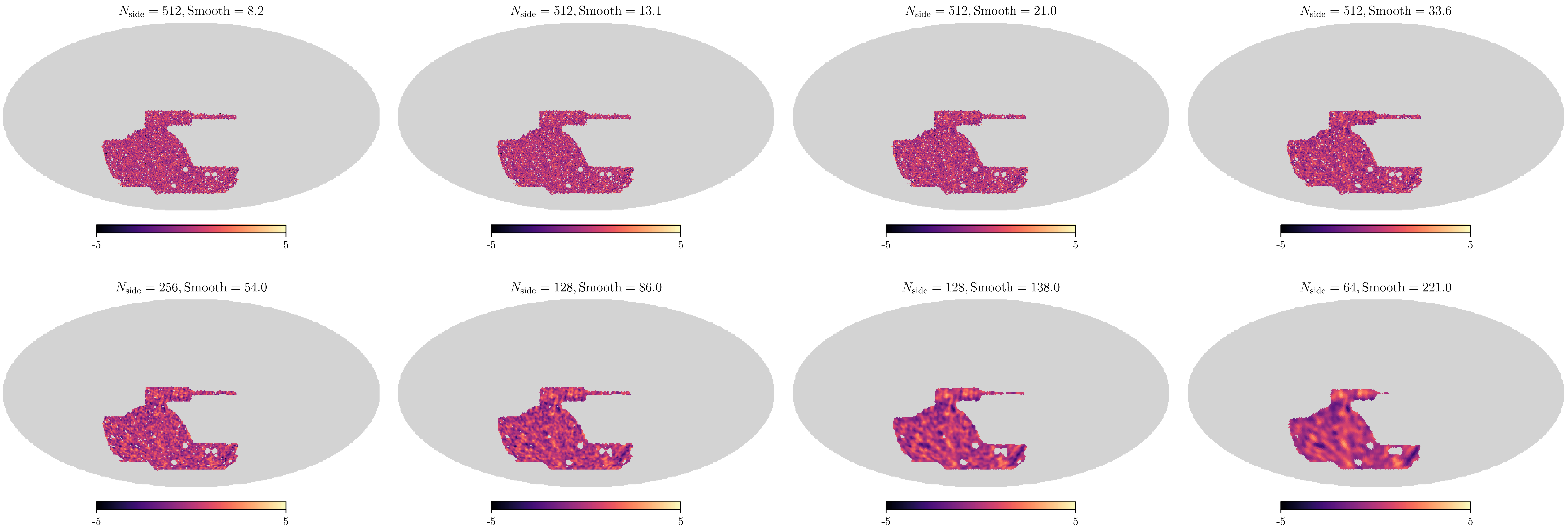




Mean (DM)
Mean (DM - Noise)
Noise subtracted S/N
Mean (DM - Baryons)
High-density regions
Low-density regions
Death
Birth - Death

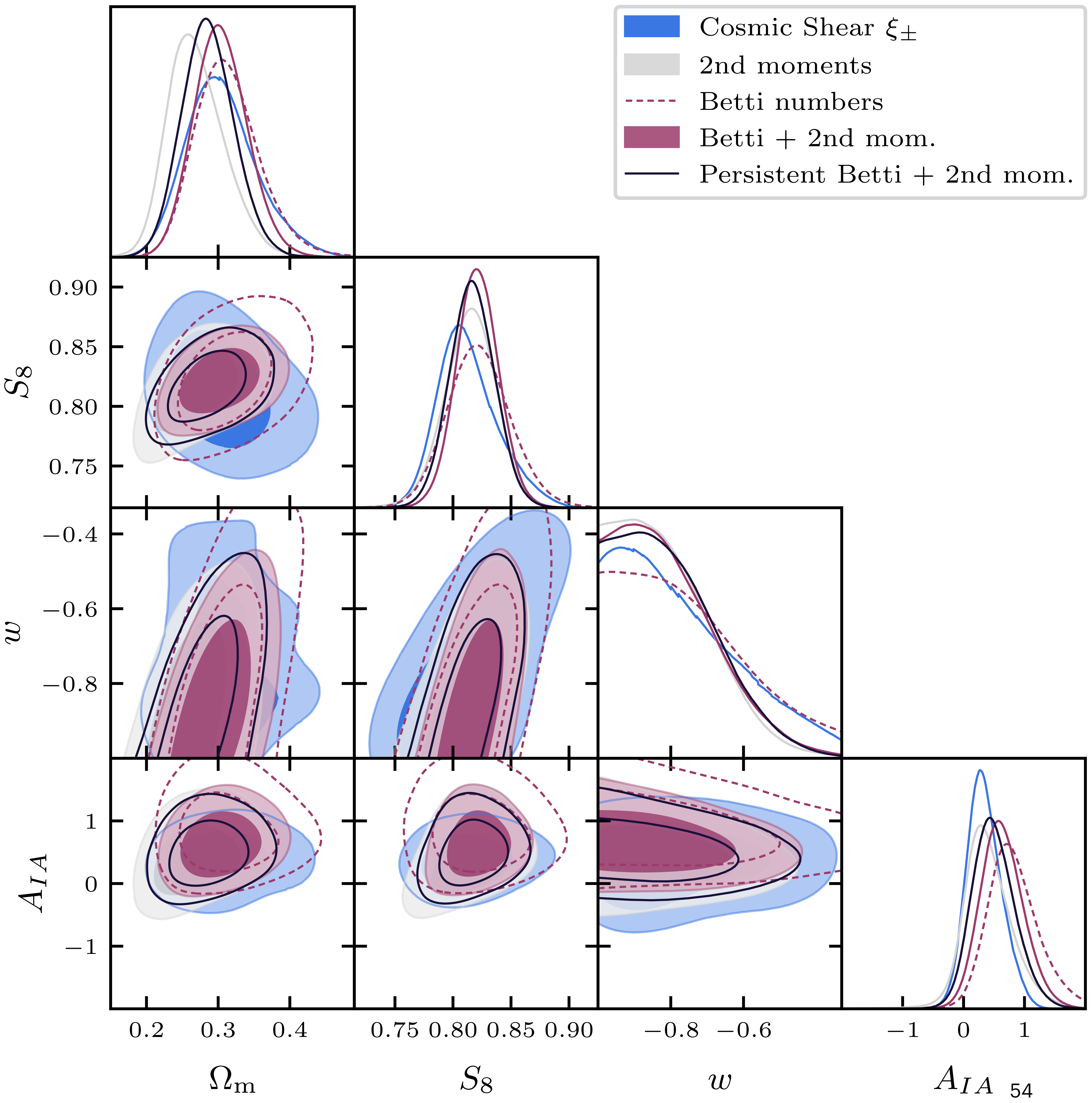

0.10
0.08
0.06
0.04
0.02
0.00
$\Delta\chi^2$
z-bin: 1
z-bin: 2
z-bin: 3
z-bin: 4
z-bin: 34
β_0
β_1
Persistent β_0
Persistent β_1
10^1
10^2
Smoothing scale [arcmin]

SMOOTHING SCALES



FULL PARAMETER SPACE

POSTERIORS



SEPARABLE CONV 2D

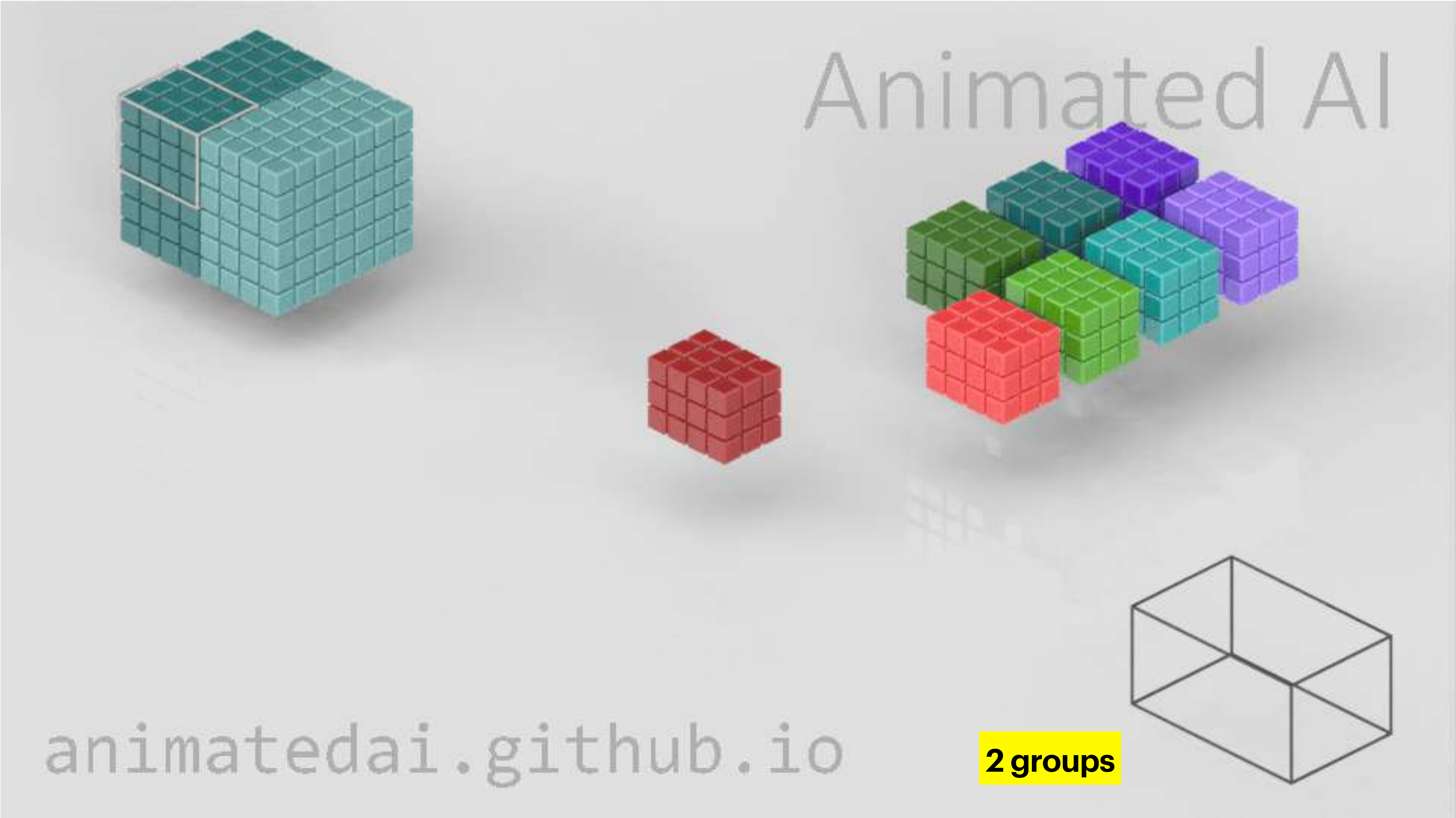
WHAT IS IT?

An efficient approximation that reduces computational cost and model parameters by separating the filtering and combination steps:

- **Standard convolutions** perform filtering and feature combination simultaneously.

- **Depthwise-separable** convolutions first use a depthwise convolution that acts separately on channels and then a pointwise convolution to combine the channels.

This approach allows the network to initially learn features specific to each redshift bin configuration and topological structure (clusters or voids) separately.



SEPARABLE CONV 2D

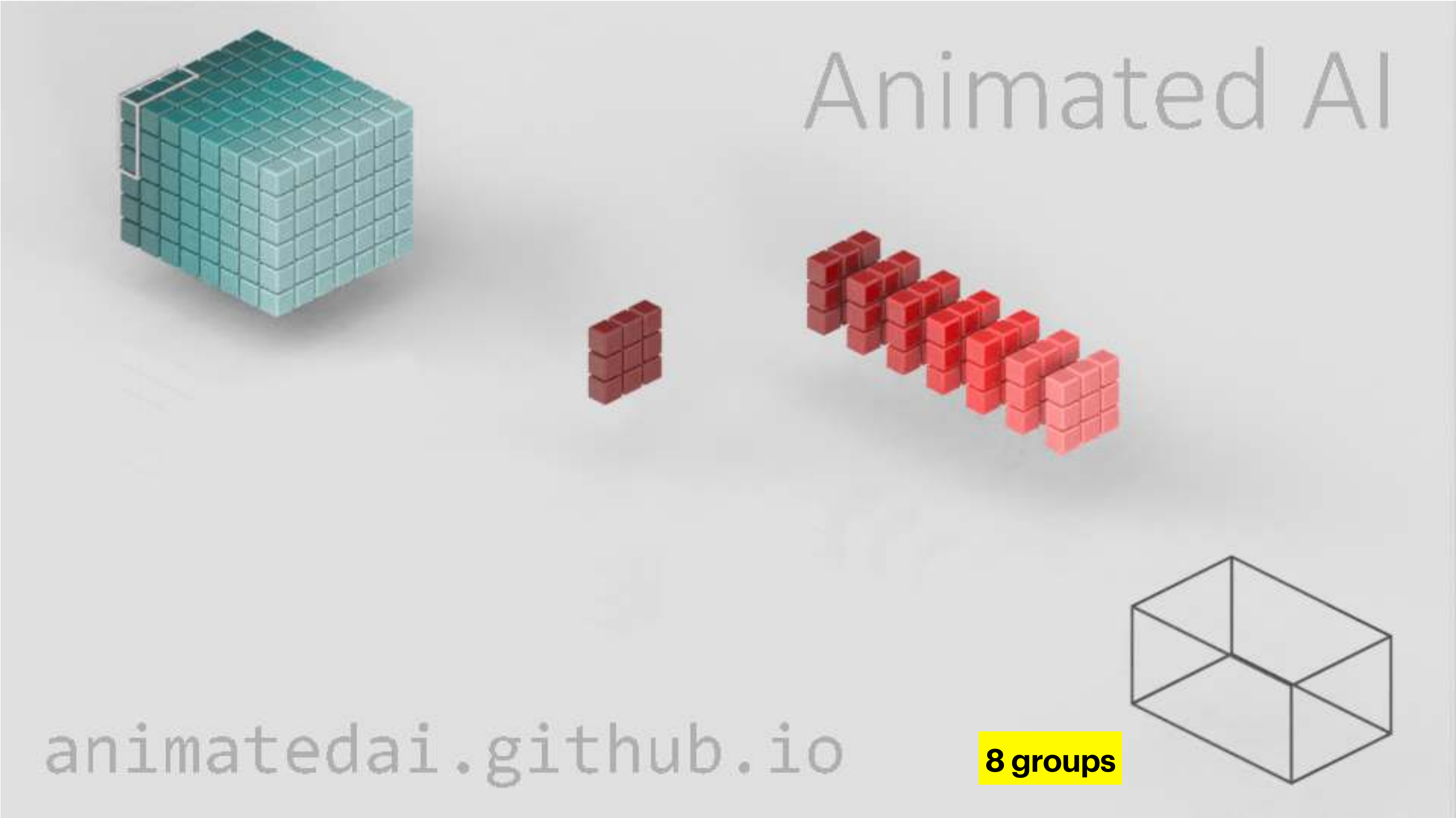
WHAT IS IT?

An efficient approximation that reduces computational cost and model parameters by separating the filtering and combination steps:

- **Standard convolutions** perform filtering and feature combination simultaneously.

- **Depthwise-separable** convolutions first use a depthwise convolution that acts separately on channels and then a pointwise convolution to combine the channels.

This approach allows the network to initially learn features specific to each redshift bin configuration and topological structure (clusters or voids) separately.



SEPARABLE CONV 2D

WHAT IS IT?

An efficient approximation that reduces computational cost and model parameters by separating the filtering and combination steps:

- **Standard convolutions** perform filtering and feature combination simultaneously.

- **Depthwise-separable** convolutions first use a depthwise convolution that acts separately on channels and then a pointwise convolution to combine the channels.

This approach allows the network to initially learn features specific to each redshift bin configuration and topological structure (clusters or voids) separately.