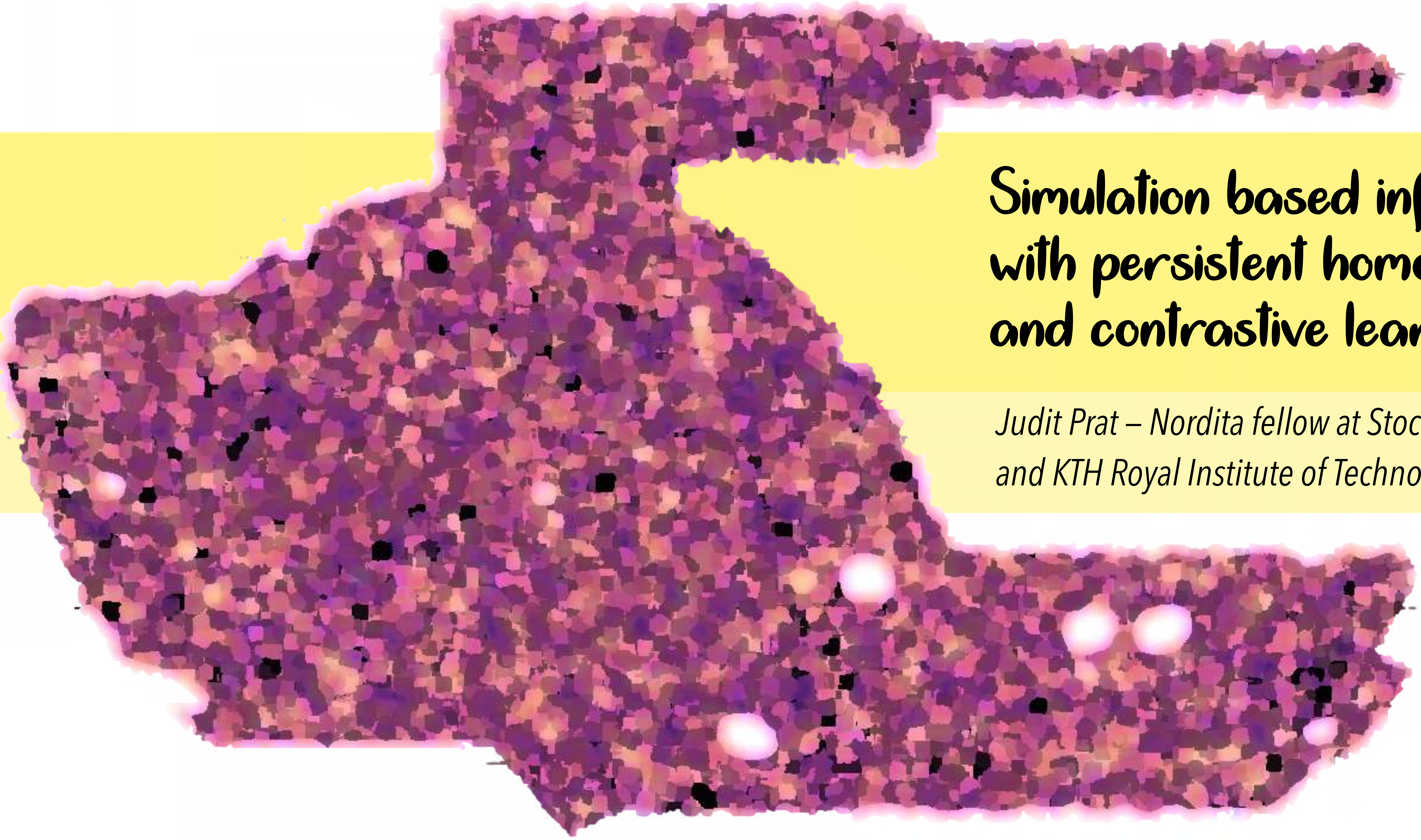


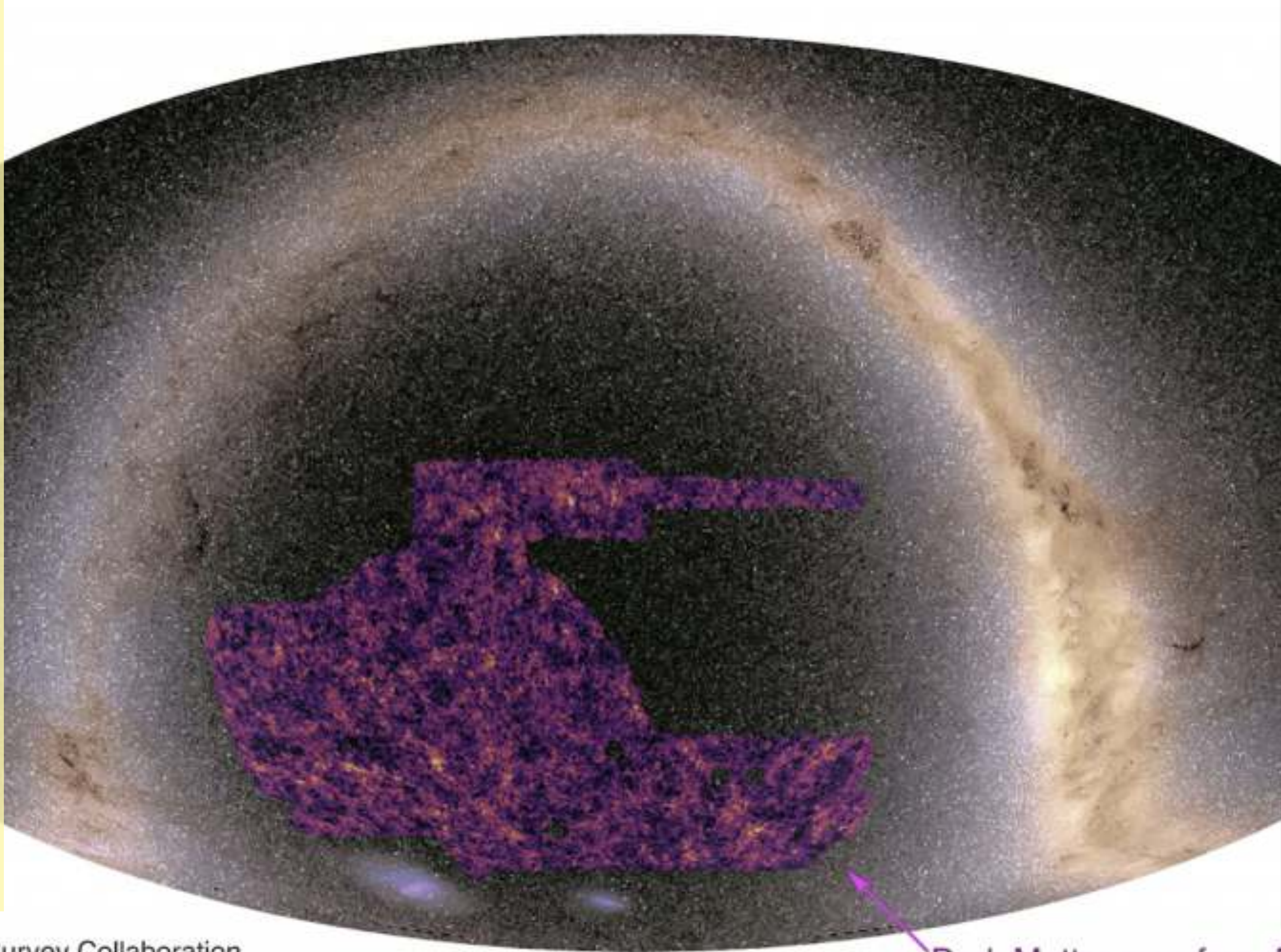
# Simulation based inference with persistent homology and contrastive learning

*Judit Prat – Nordita fellow at Stockholm University and KTH Royal Institute of Technology. May 2024.*

NORDITA
The Nordic Institute for Theoretical Physics

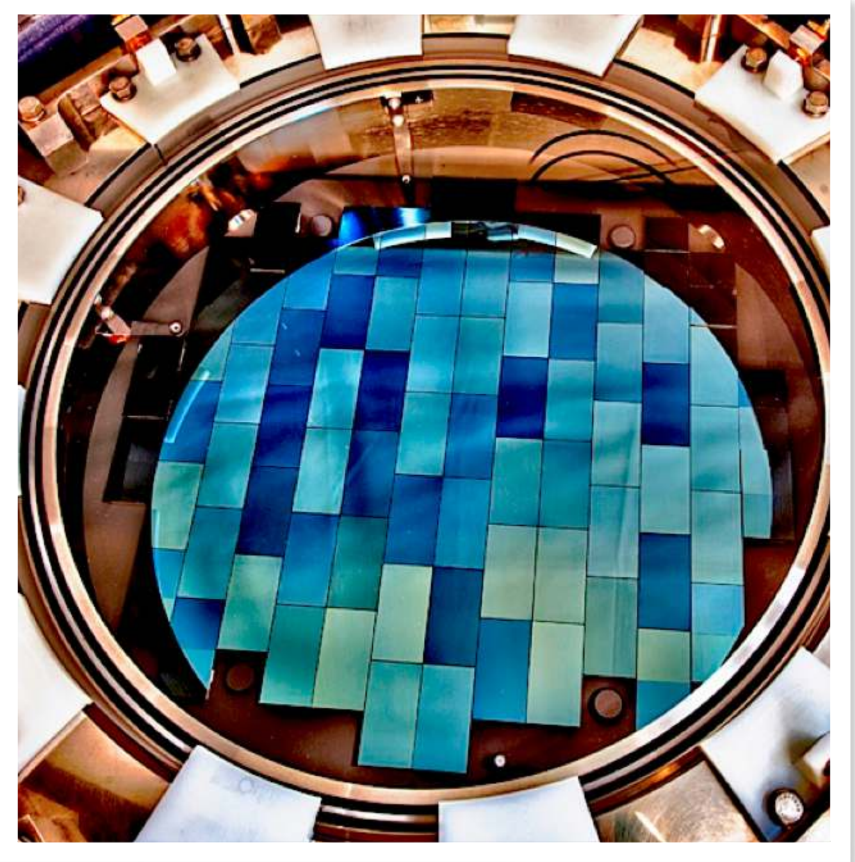
# **The Dark Energy Survey:** *a photometric lensing survey*

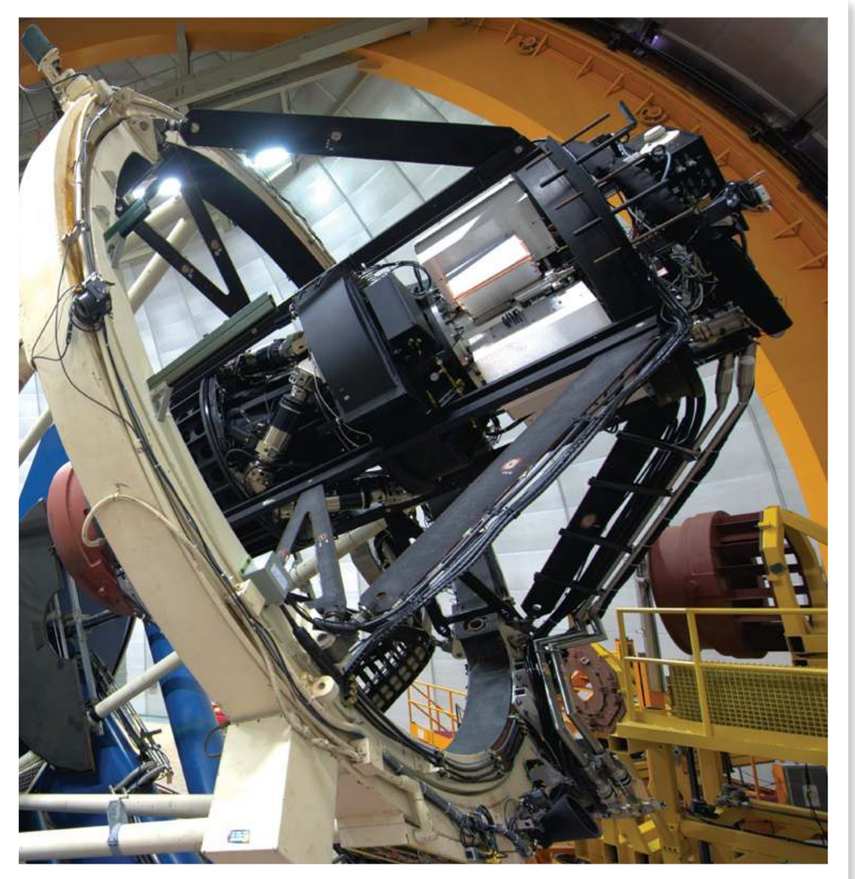
- 570 Megapixel camera for the Blanco 4m telescope in Chile.
- Full survey 2013-2019 (Y3 2013-16).
- Wide field: 5000 sq. deg. in 5 bands. ~23 magnitude.
- DES Y3: Positions and shapes of > 100M galaxies.



N. Jeffrey; Dark Energy Survey Collaboration

Dark Matter map from DES observations



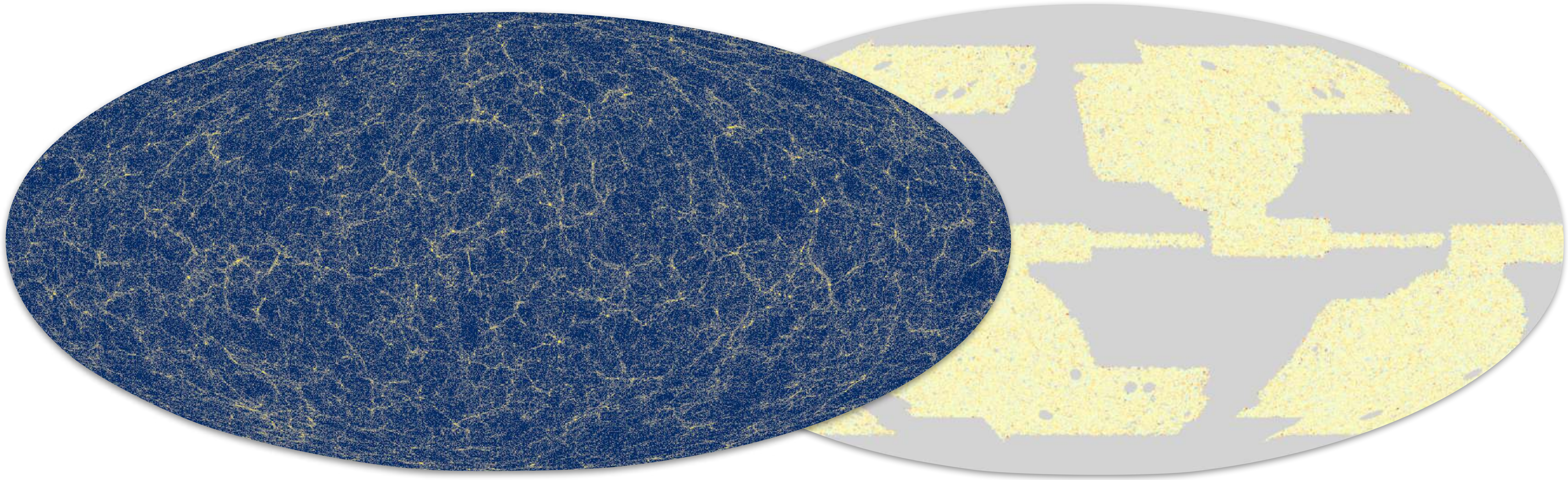


# ~800 full sky dark matter simulations

## Gower St simulations

Simulation effort led by Niall Jeffrey and Marco Gatti

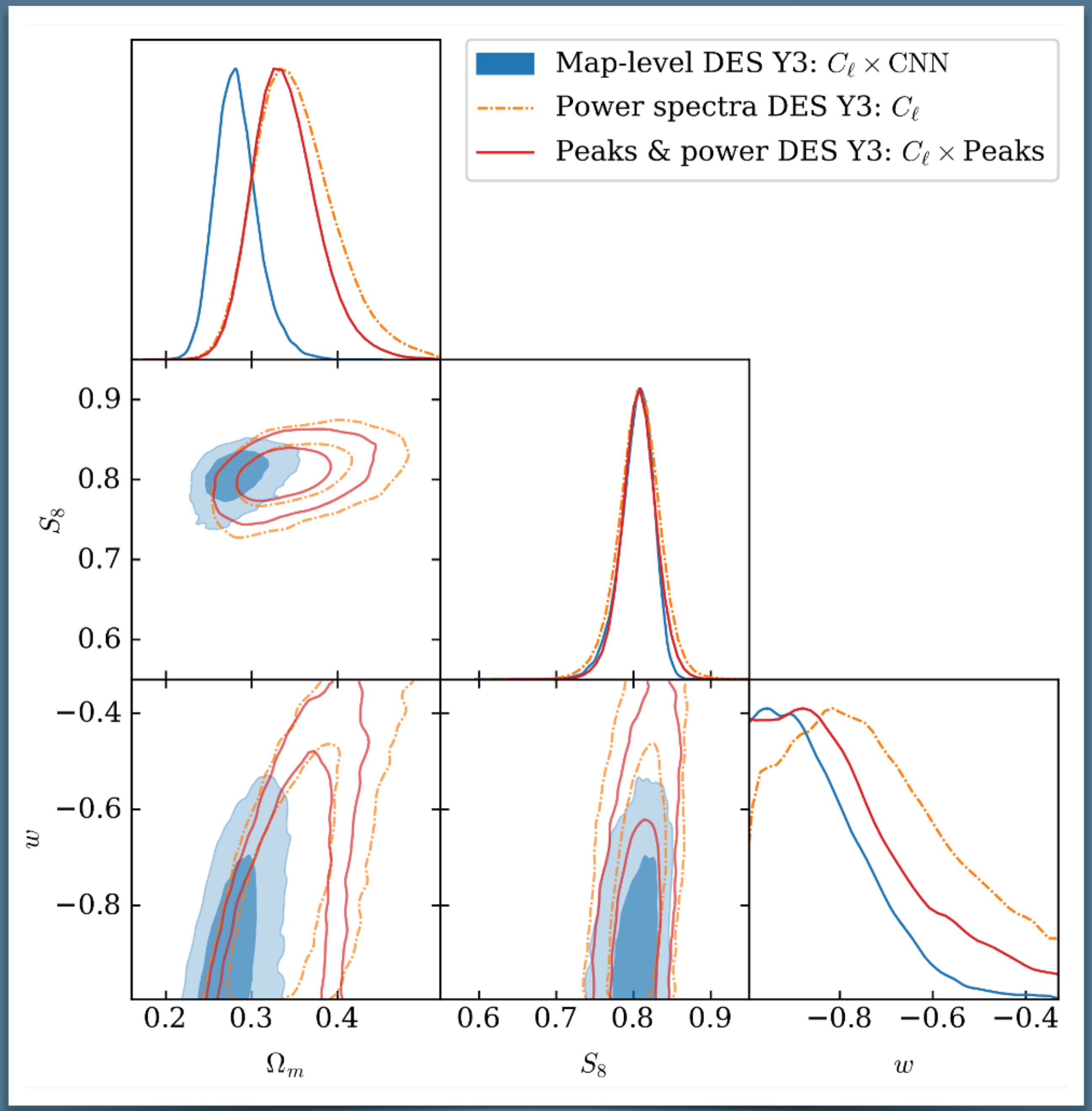
- Matching DES Y3 properties.
- **Varying nuisance parameters:** redshift full shape distributions, IA (NLA), multiplicative shear biases.



| | $N$ | Input varied |
|---|---|---|
| Full-sky $\gamma-$maps | 791 | $\Omega_i$ |
| Independent DES $e-$maps | $791 \times 4 = 3164$ | $n_i$ |
| Quasi-independent DES $e-$maps | $3164 \times 10 = 31640$ | $n_i$, $e_{\rm SN}$ |

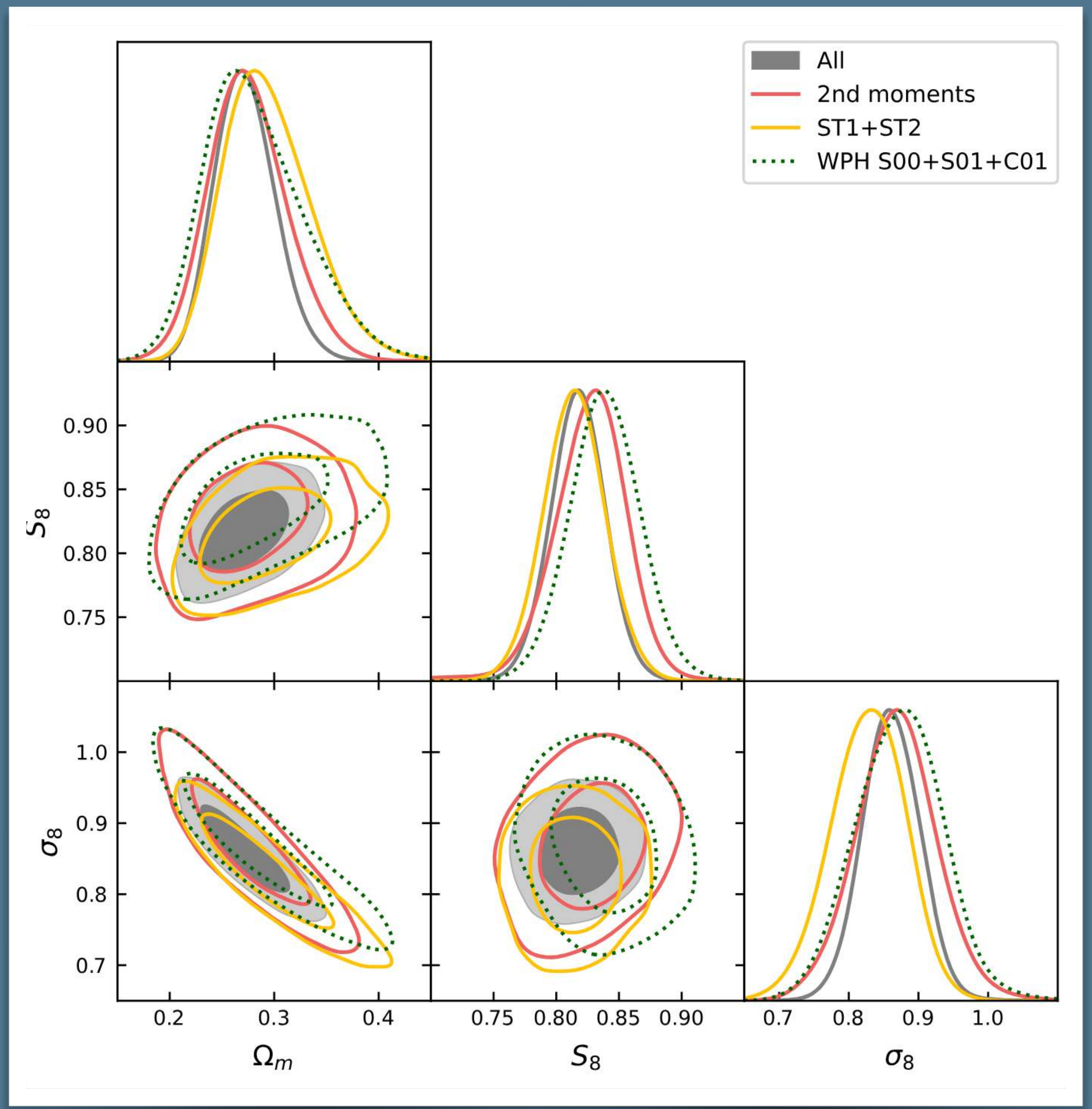
# DES Y3 beyond 2pt on weak lensing maps

(1) Peaks
(2) CNN + C(l)



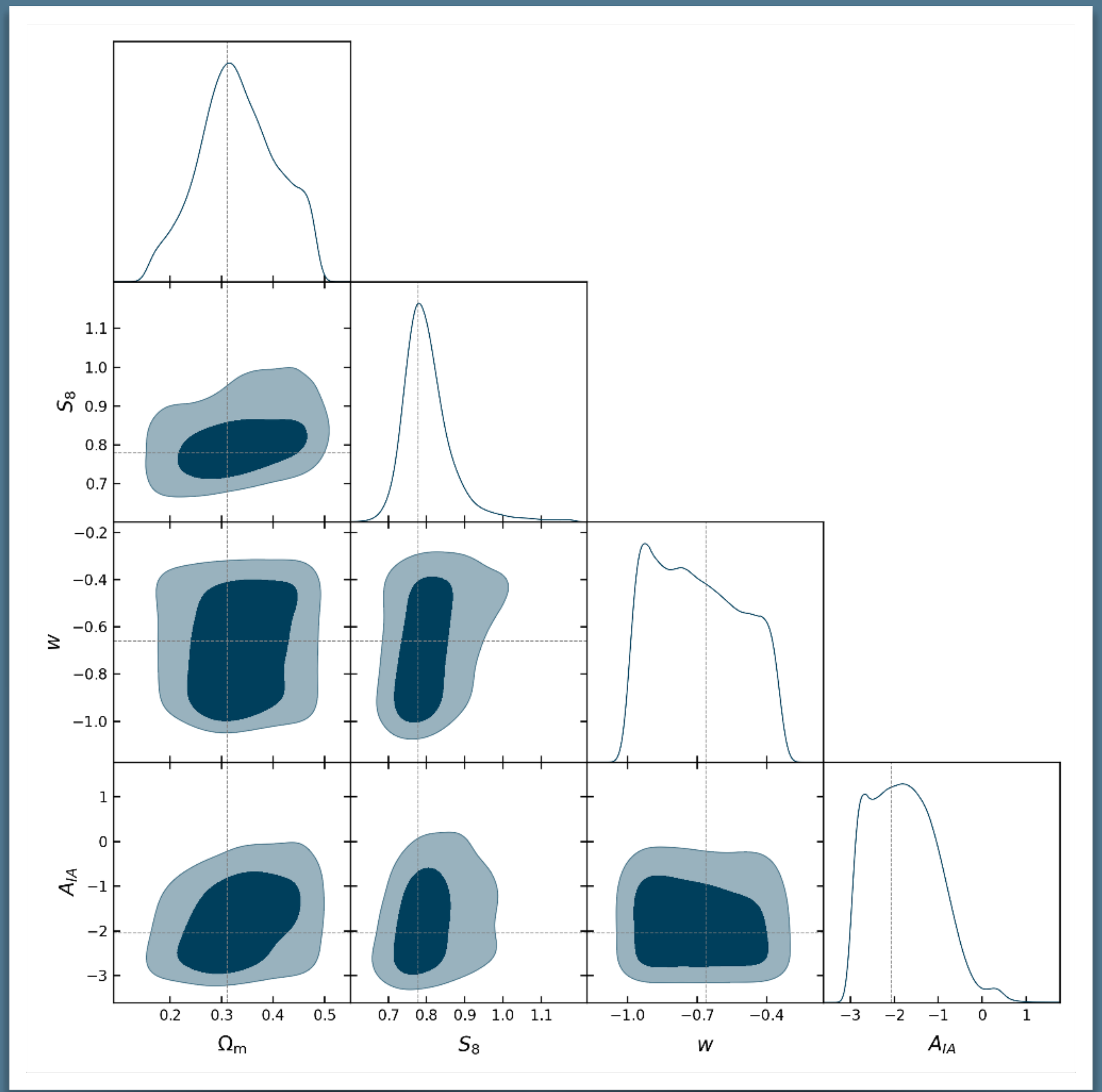

Sims & Data:
**Jeffrey et al. (2024):** 2403.02314

(3) 2nd & 3rd moments
(4) Scattering transform (ST)
(5) Wavelet phase harmonics




Sims: **Gatti et al. (2024):** 2310.17557
Data: **Gatti et al. (2024):** 2405.10881

(6) Topological statistics:
Persistent homology

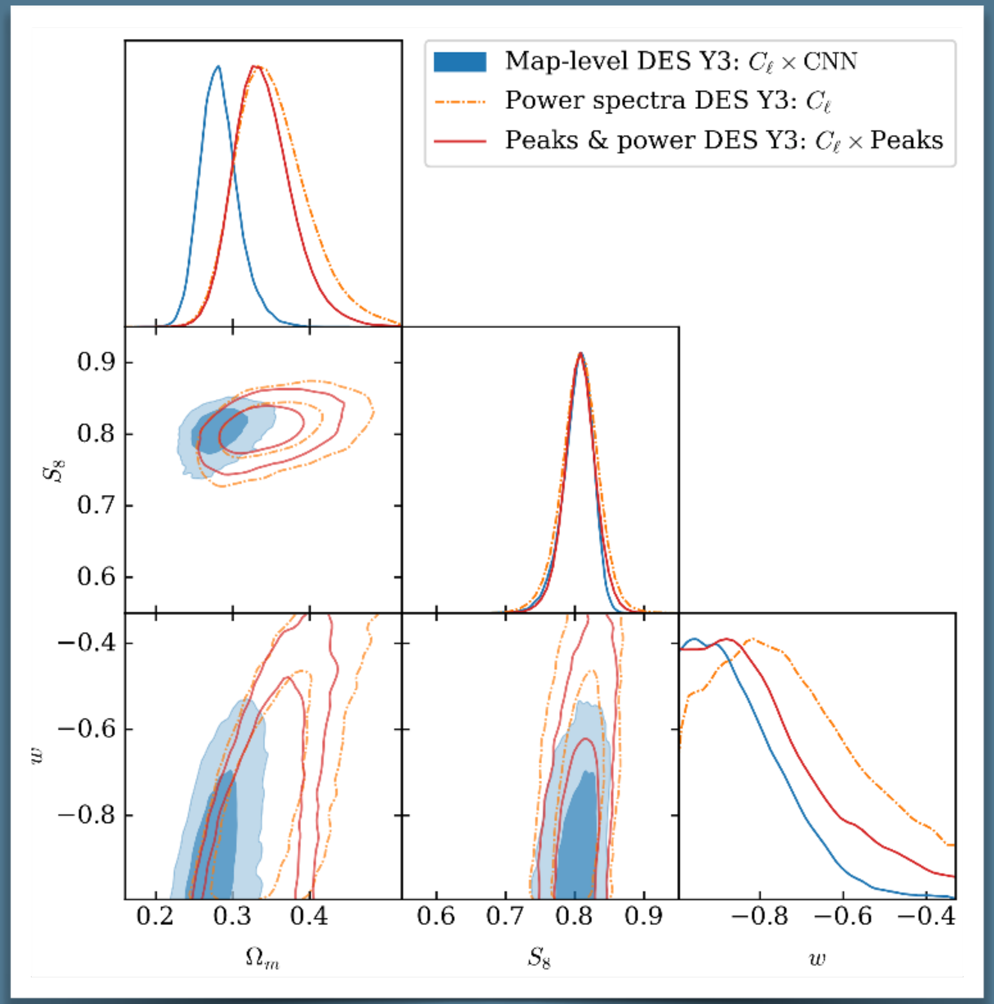



Prat et al. (in prep)

*All using the **same** data, the same simulation based inference **framework** & the **Gower Street** Simulation suite.*

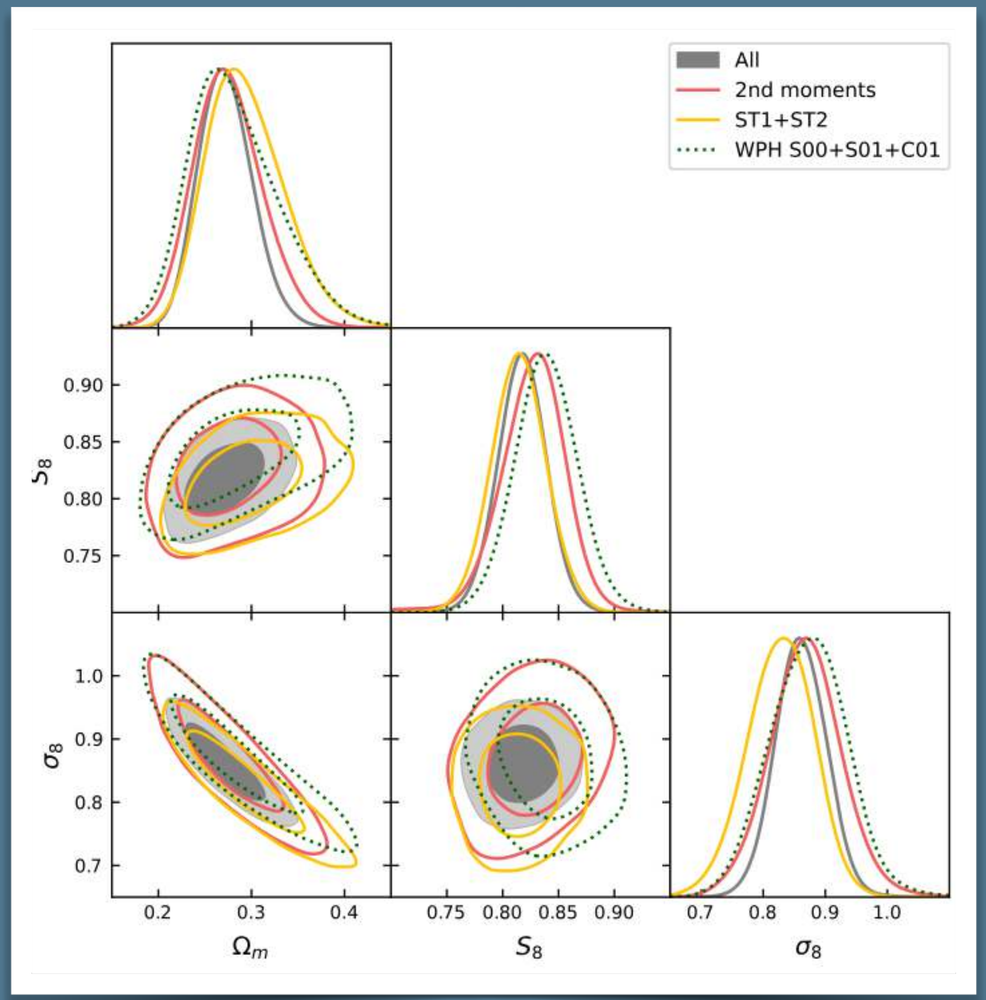
# DES Y3 beyond 2pt on weak lensing maps

(1) Peaks
(2) CNN + C(l)



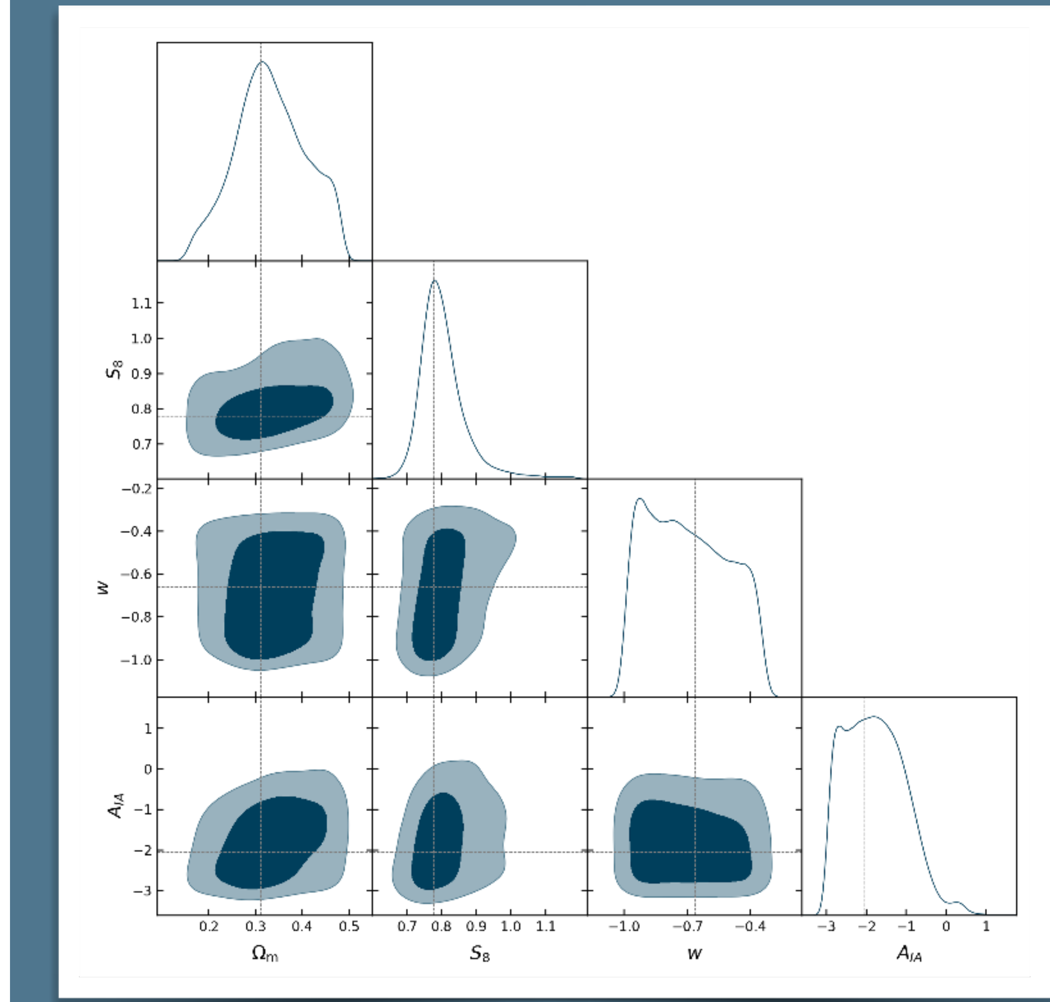

Sims & Data:
**Jeffrey et al. (2024):** 2403.02314

(3) 2nd & 3rd moments
(4) Scattering transform
(5) Wavelet phase harmonics



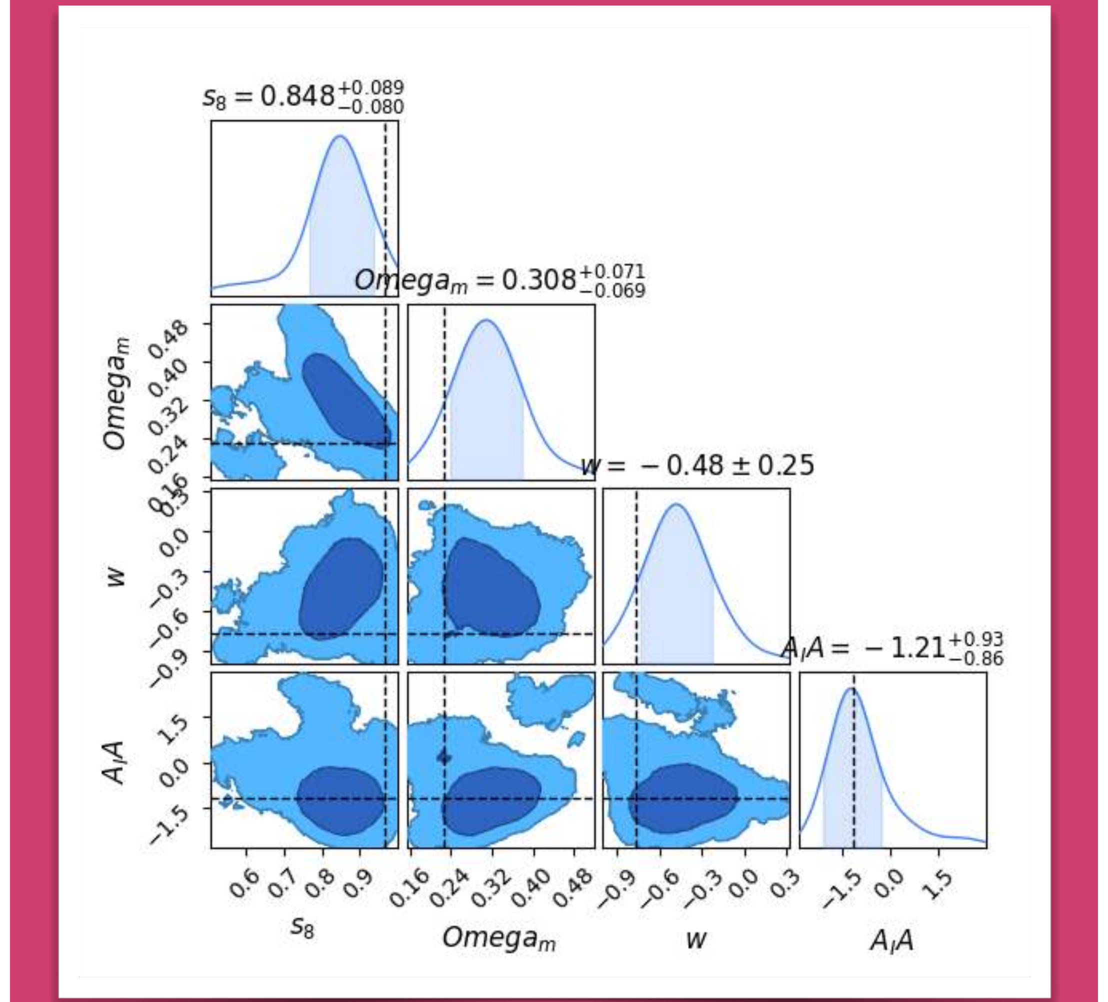

Sims: **Gatti et al. (2024):** 2310.17557
Data: **Gatti et al. (2024):** 2405.10881

(6) Topological statistics: Persistent homology



Prat et al. (in prep)

(7) Contrastive learning CNN

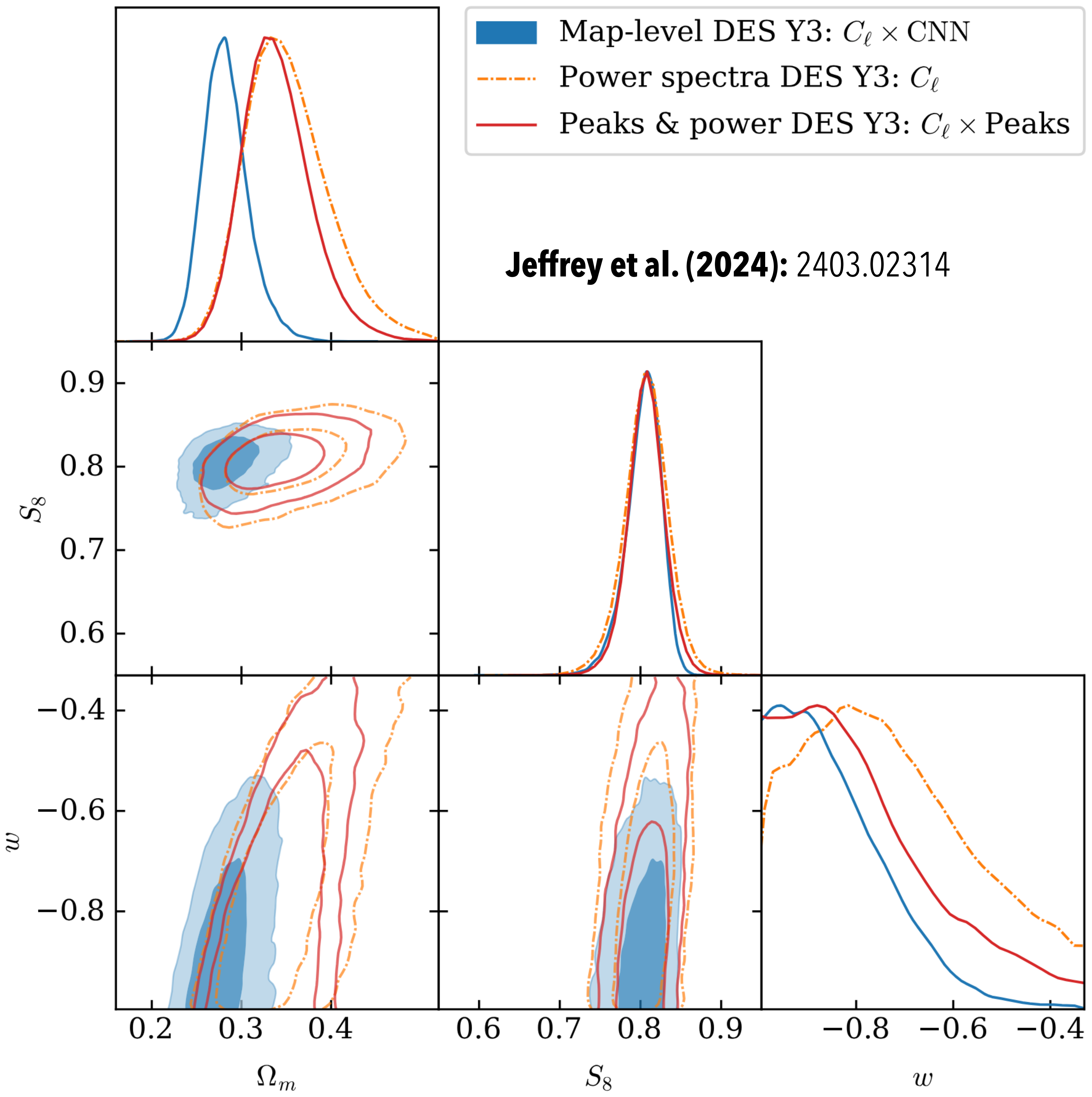



Recent project

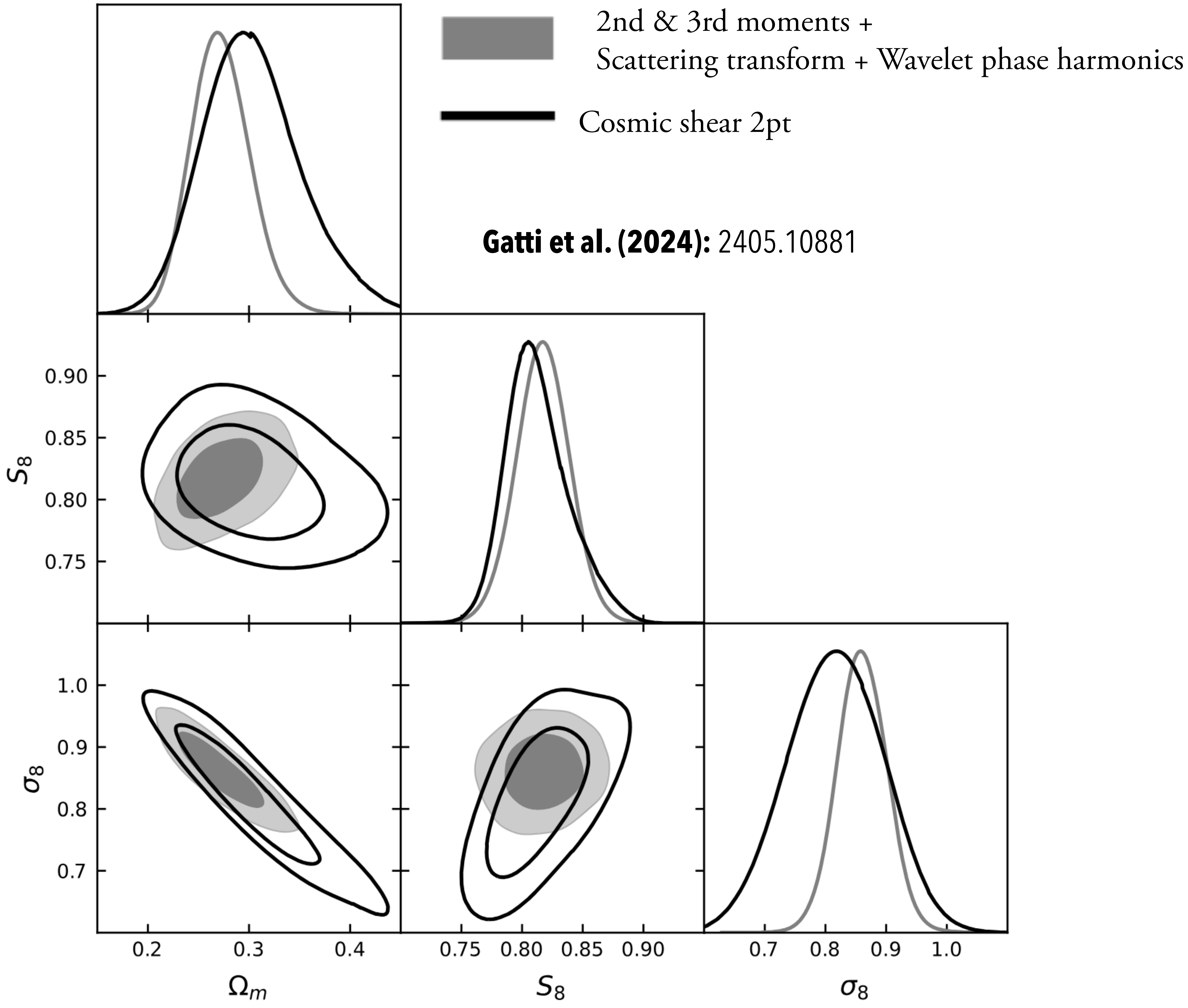
*All using the **same** simulation based inference **framework** & the **Gower Street** Simulation suite.*

# Where is the gain?

*Improvement is on the 2D plane, $S_8$ is not a good metric to compare methods anymore*




**Jeffrey et al. (2024):** 2403.02314




**Gatti et al. (2024):** 2405.10881

# The persistent homology part

*Judit Prat, Chihway Chang (UChicago), Pratyush Pranav (Bennet University) , Marco Gatti (UPenn), Niall Jeffrey (UCL), Manon Ramel (CNRS), Cyrille Doux (CNRS), Sven Heydenreich (Santa Cruz), ++*

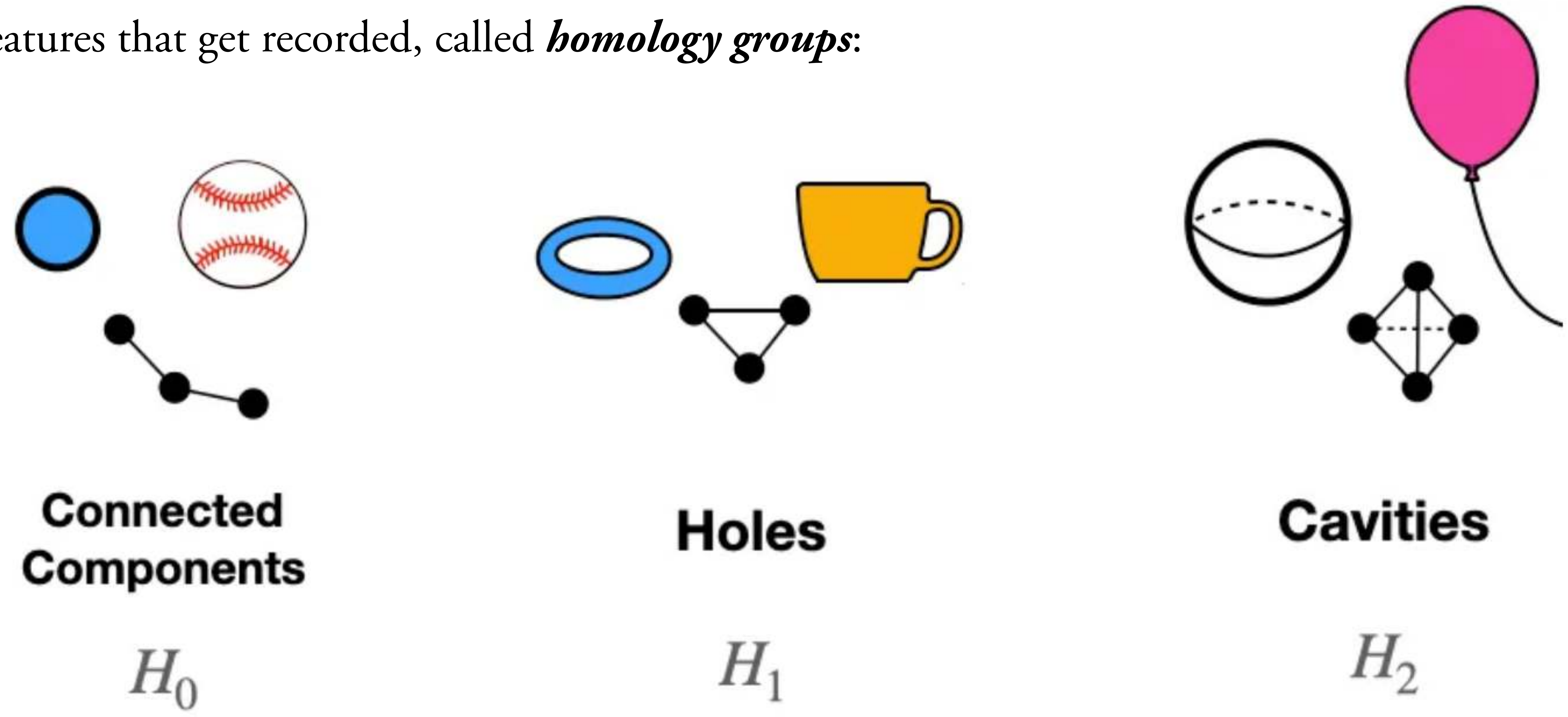
NORDITA
The Nordic Institute for Theoretical Physics

# Persistent homology

## What is it?

Persistent homology is a method for computing topological features of a space. More persistent features are detected over a wide range of spatial scales and are deemed more likely to represent true features of the underlying space rather than artifacts of sampling, noise, or particular choice of parameters.

There are different features that get recorded, called ***homology groups***:



# Persistent homology applied to WL mass maps

*How do we construct a persistent diagram?*

To build the persistent diagram, we:

- Convert to a dimensionless map, which is the mean-subtracted and variance-scaled field derived from the original convergence field.

$$\nu = (f - \mu_f)/\sigma_f$$

- Then we build a filtration, which means you cut the map at many different thresholds.

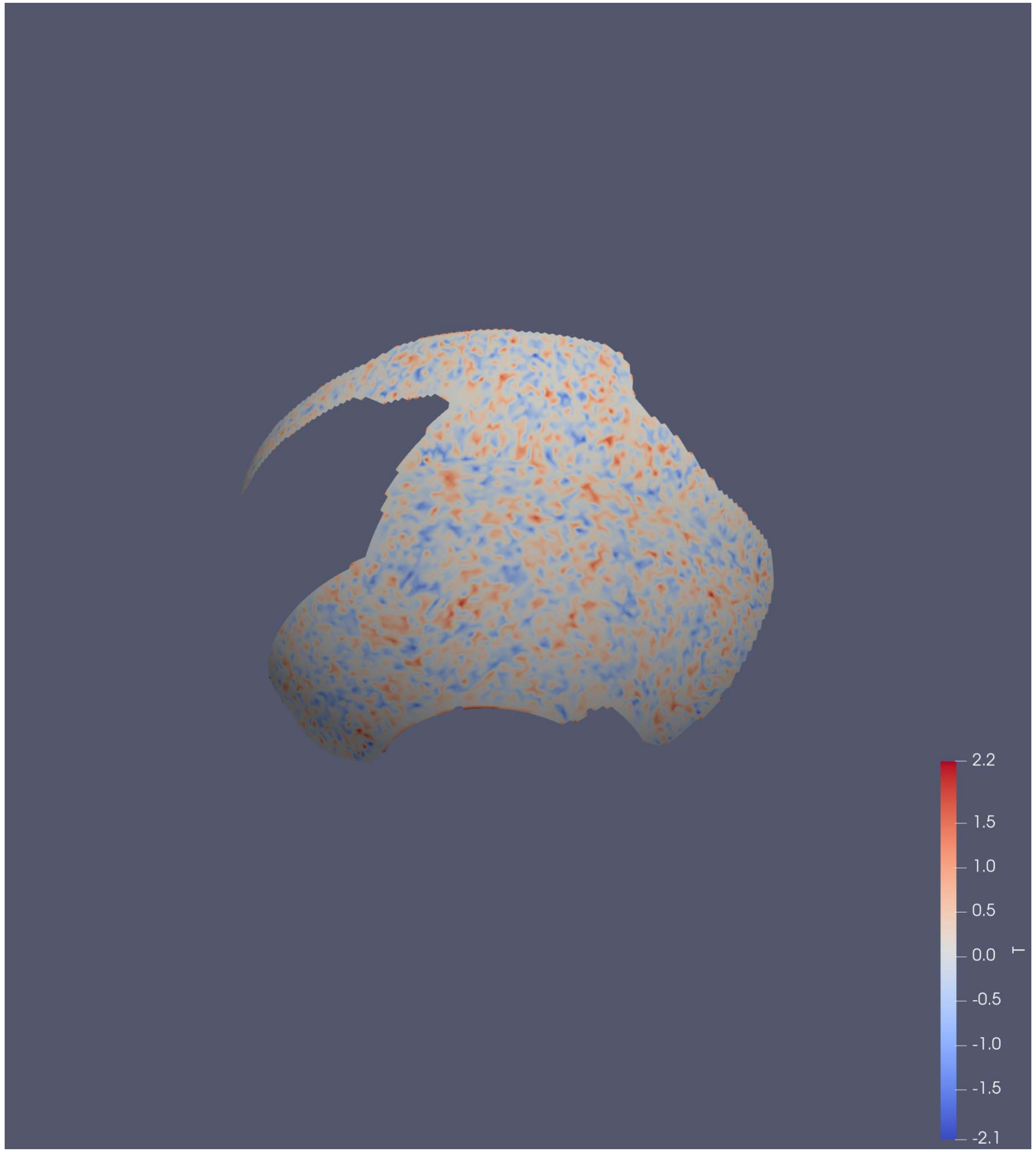
- Then at each threshold you record if a feature (either hole or connected component) has appeared (its birth), and then you record at which threshold it disappears.

# Building a filtration:

*Visualization*

You start  building a filtration with the mask and the dimensionless (scaled) map:

$$\nu = (f - \mu_f)/\sigma_f$$

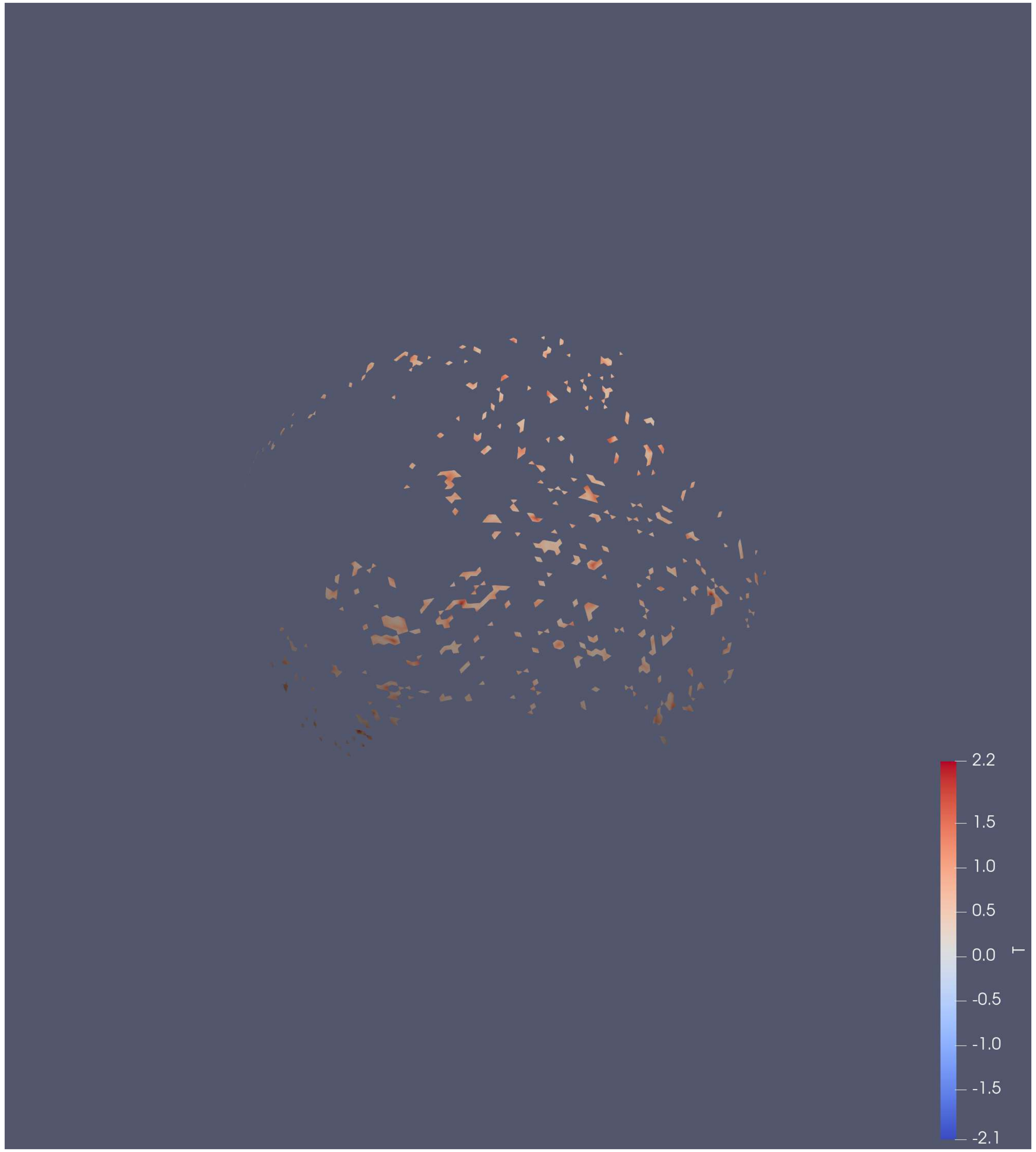


# Building a filtration:
## *Visualization*

Then you apply a threshold, and keep everything above it (this is called a *superlevel filtration*, and this the convention that we adopt).

Connected components will appear first (*H0 homology group*) and they represent the highest *peaks* of the density field.

When a features appears, you record the threshold where it appears and call it *birth*.

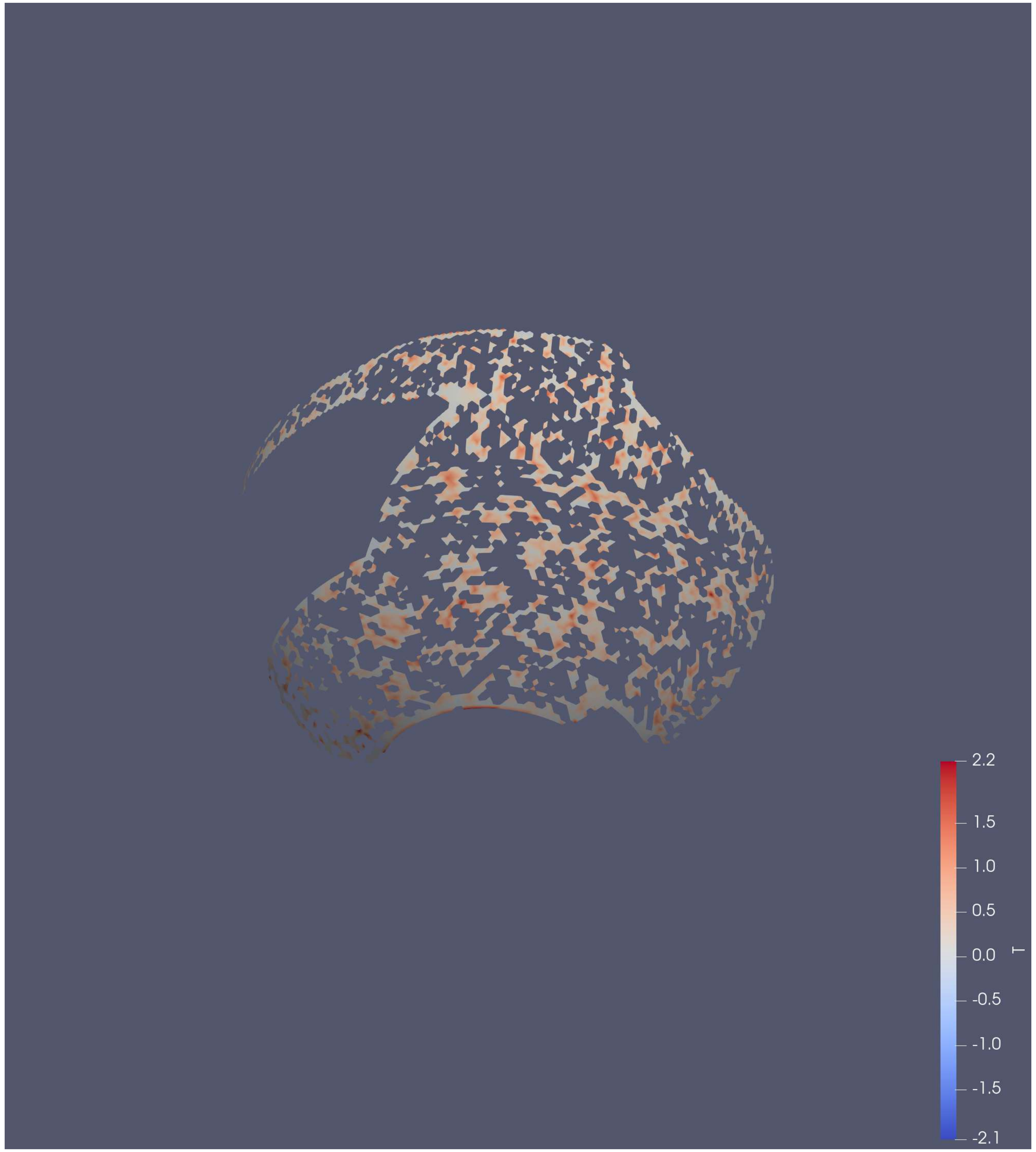


# Building a filtration:

*Visualization*

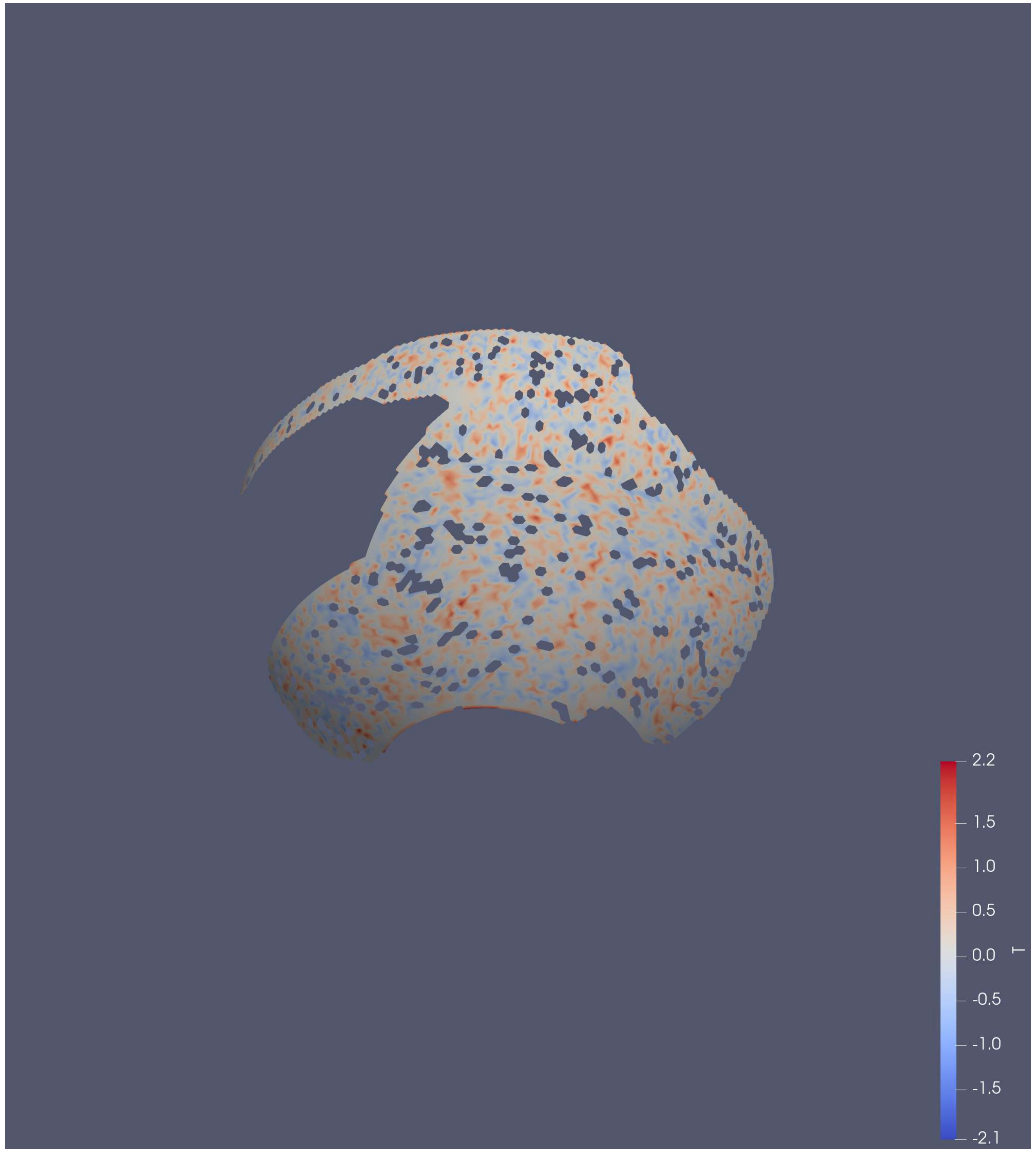
You continue decreasing the threshold, and recording when connecting components appear and disappear, and when holes appear. When a feature disappears, you call it death.

***What happens with merging?*** When two components merge, the one that was born at an earlier threshold is alive, the younger one dies.



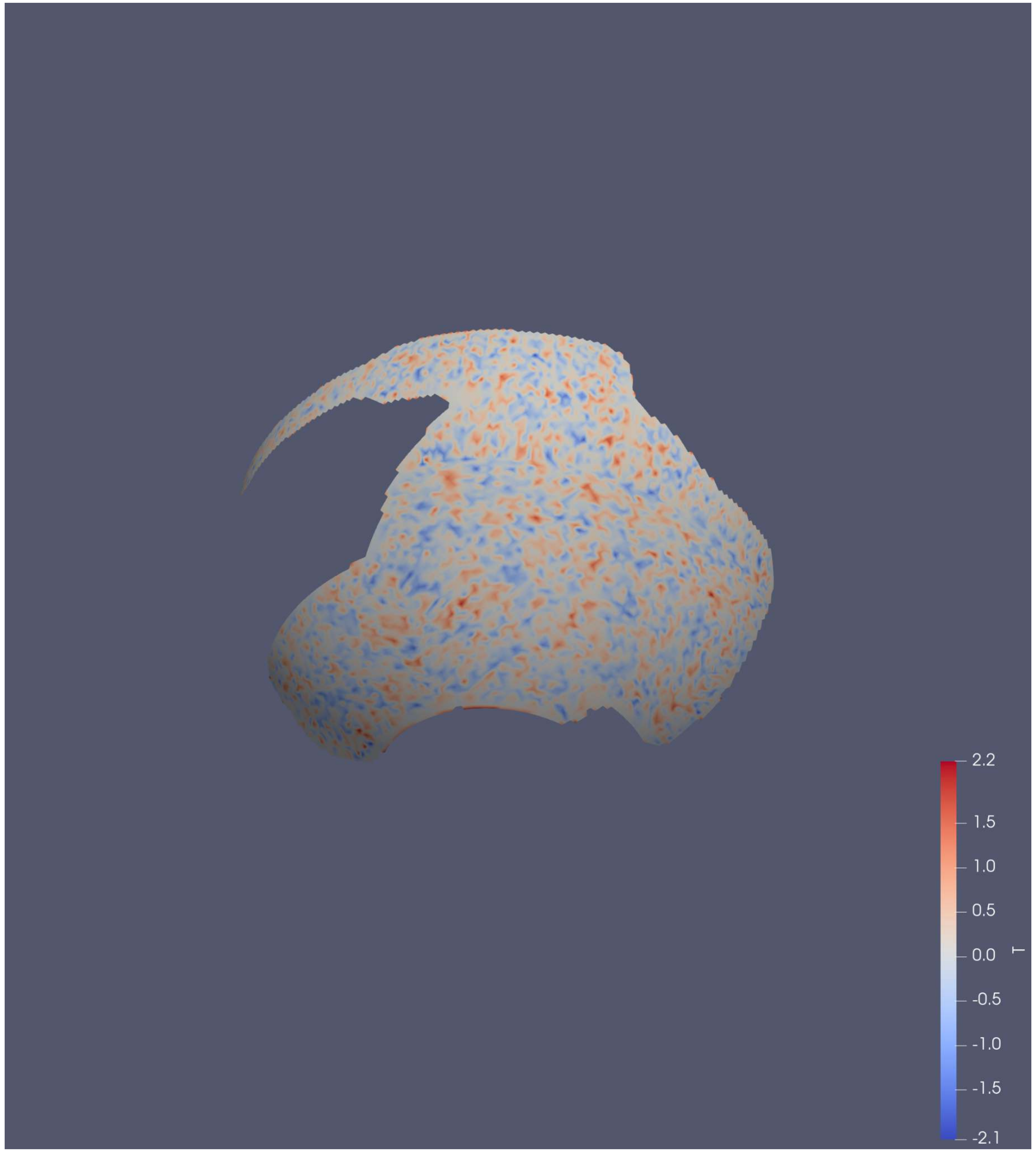
# Building a filtration:
## *Visualization*

*Holes* (*H1 group*) correspond to the *voids*.



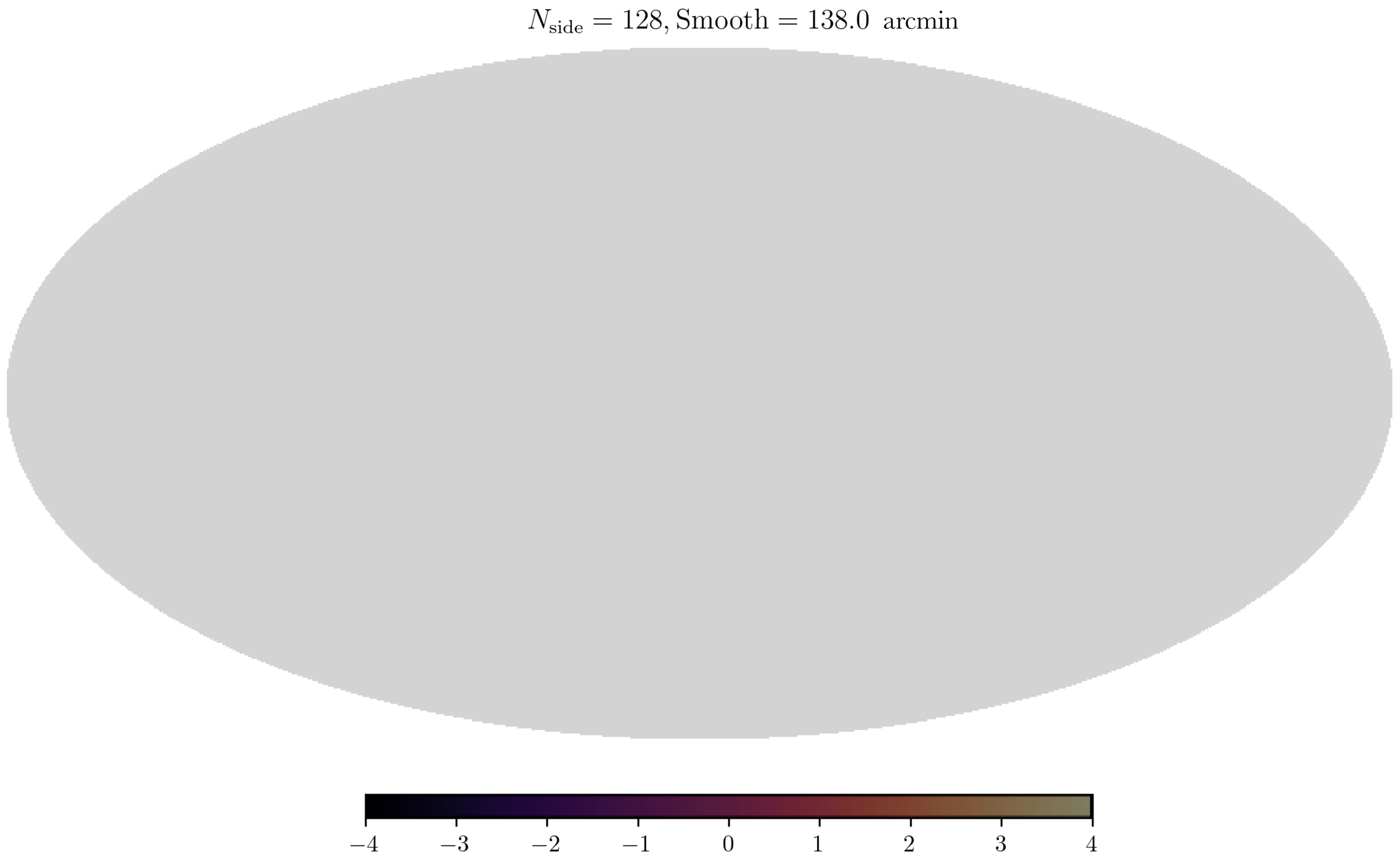
# Building a filtration:
*Visualization*

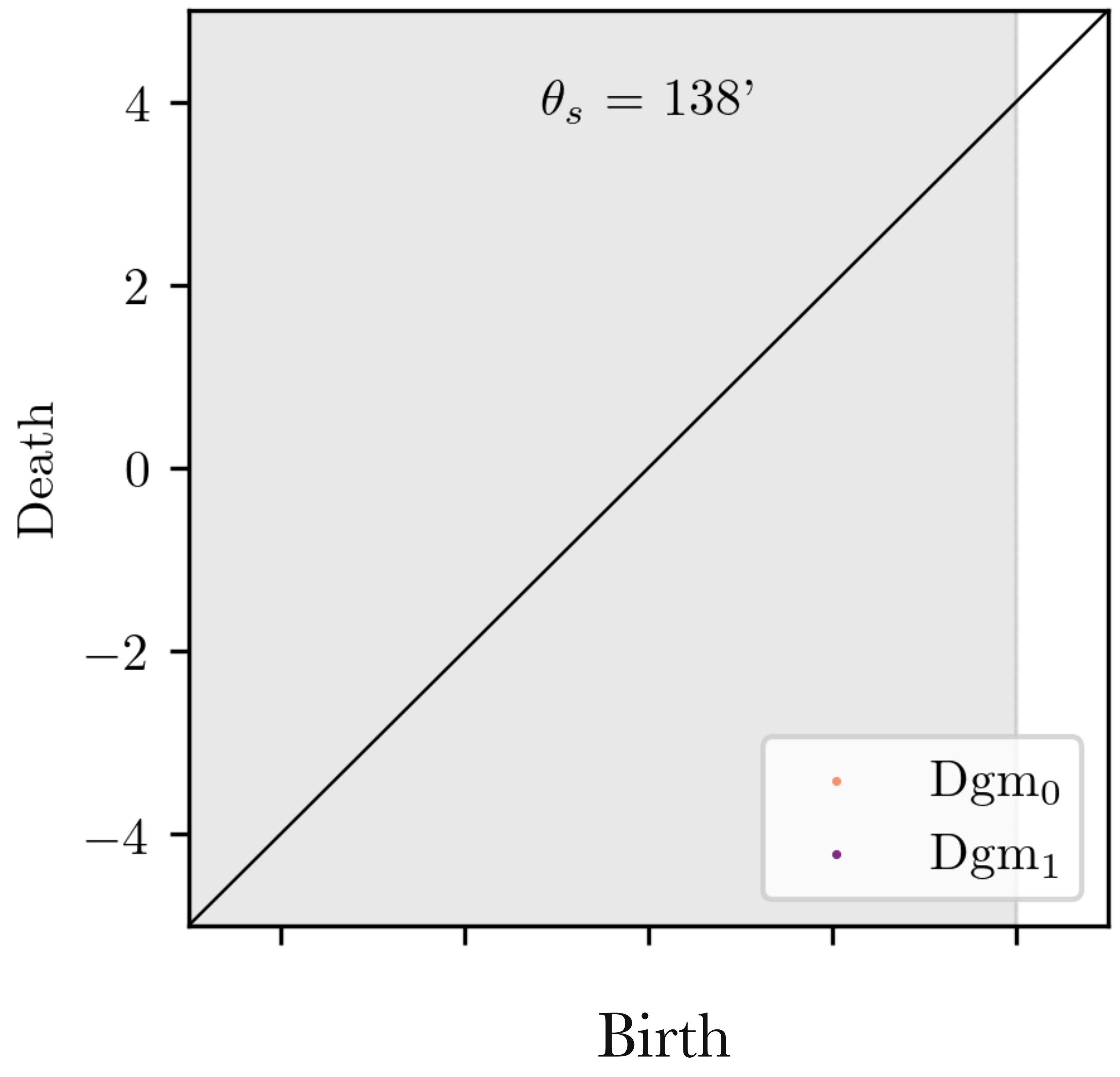
You go until you recover the full map.



# Building a filtration

## Persistent diagram and Betti numbers

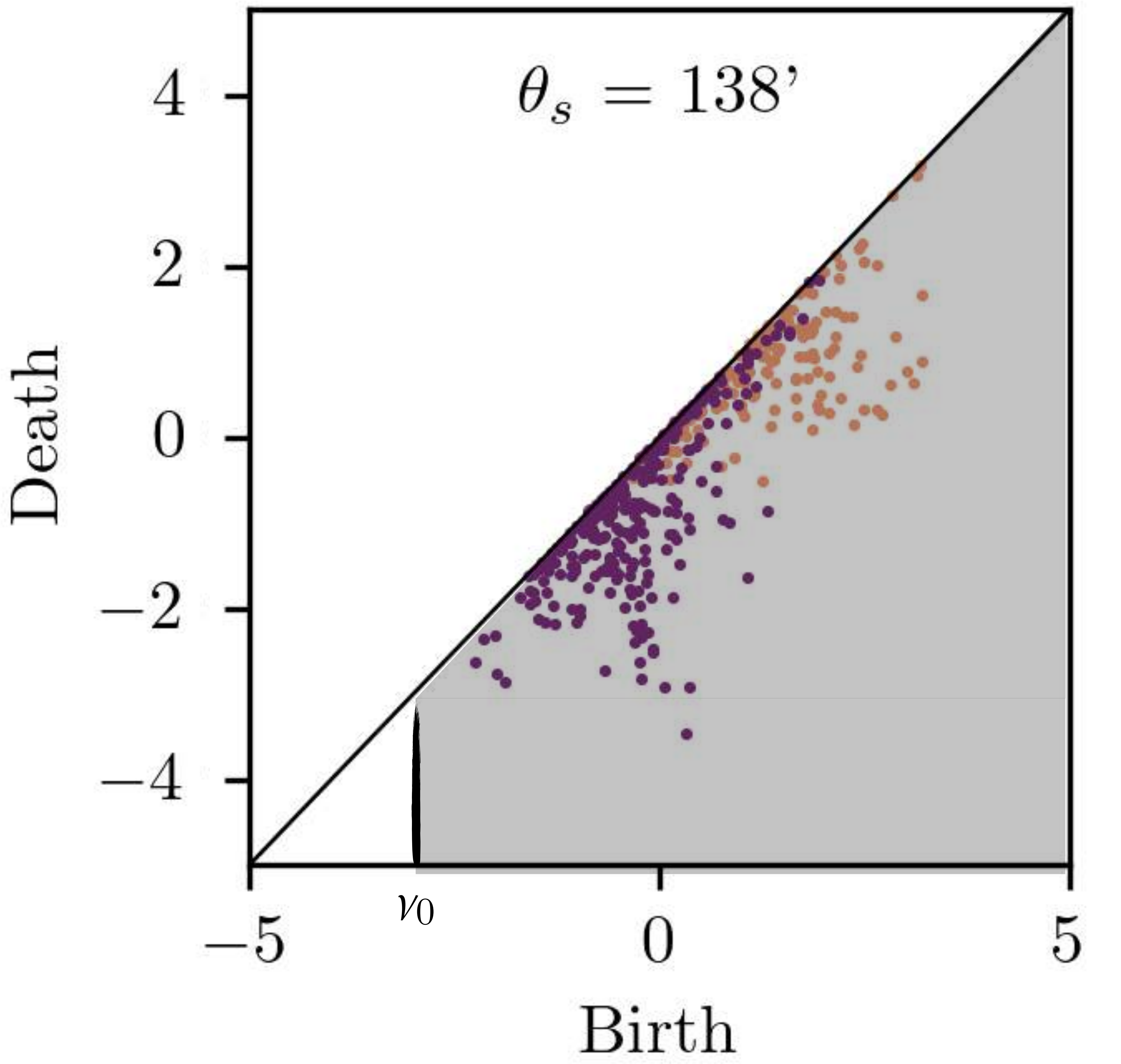


# From a persistent diagram to Betti number

Betti numbers count the number of connected components and holes that are present above a given threshold $\nu_0$. These are called the 'Betti numbers' $\beta_0$ (for connected components) and $\beta_1$ (for holes).

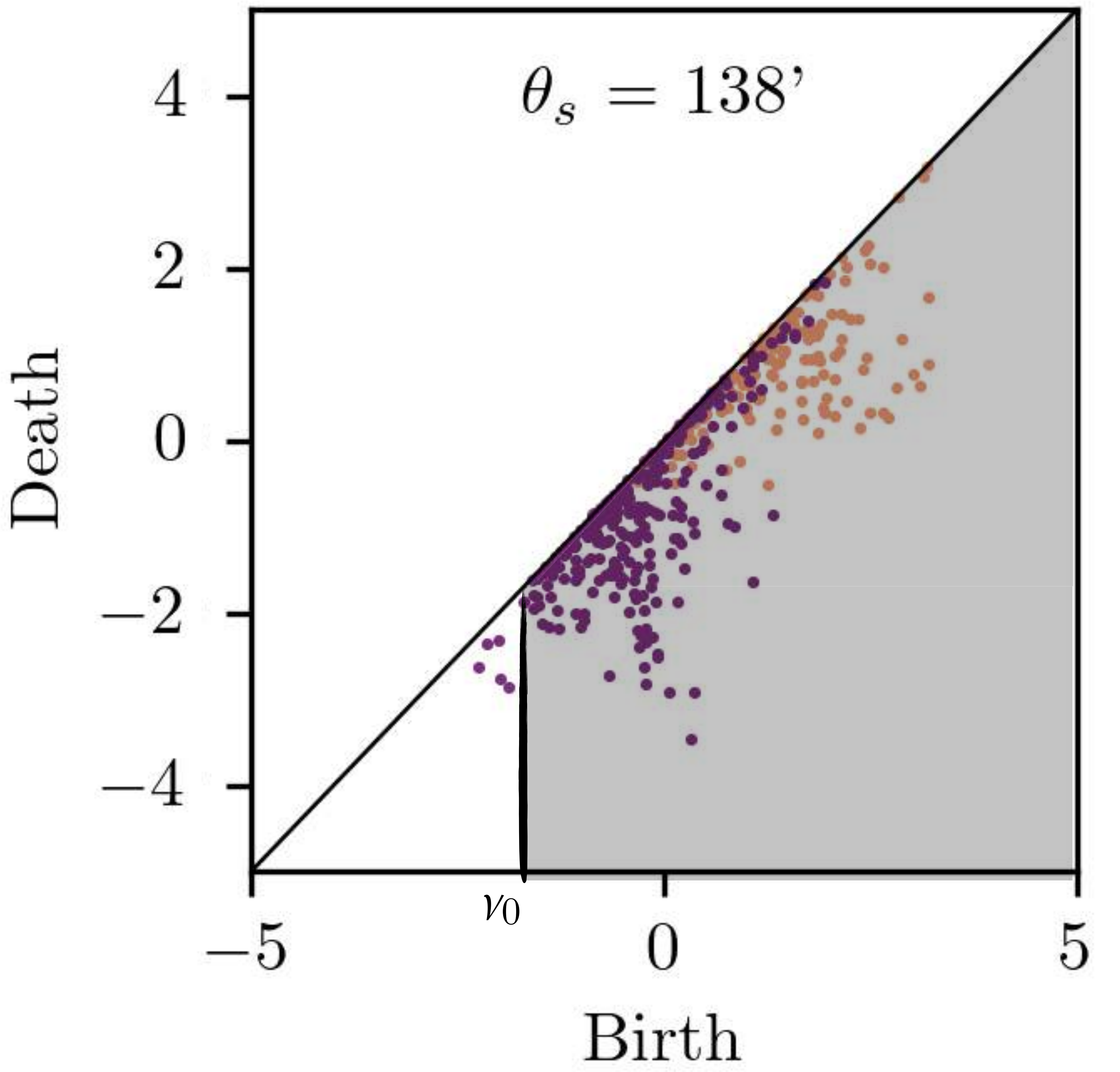
They are a compression of the persistent diagram.



# From a persistent diagram to Betti number

Betti numbers count the number of connected components and holes that are present above a given threshold $\nu_0$. These are called the 'Betti numbers' $\beta_0$ (for connected components) and $\beta_1$ (for holes).

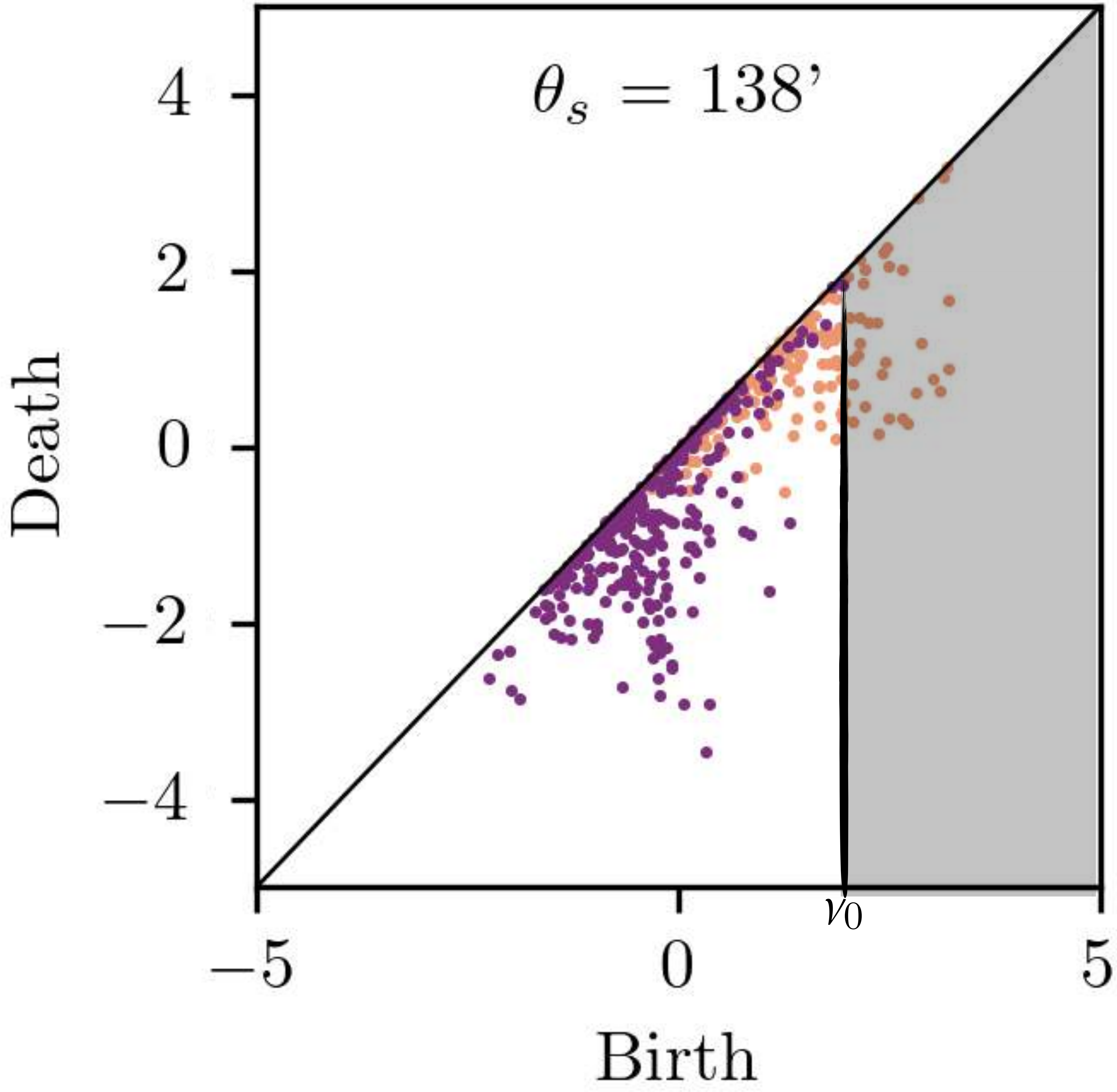
They are a compression of the persistent diagram.



# From a persistent diagram to Betti number

Betti numbers count the number of connected components and holes that are present above a given threshold $\nu_0$. These are called the 'Betti numbers' $\beta_0$ (for connected components) and $\beta_1$ (for holes).

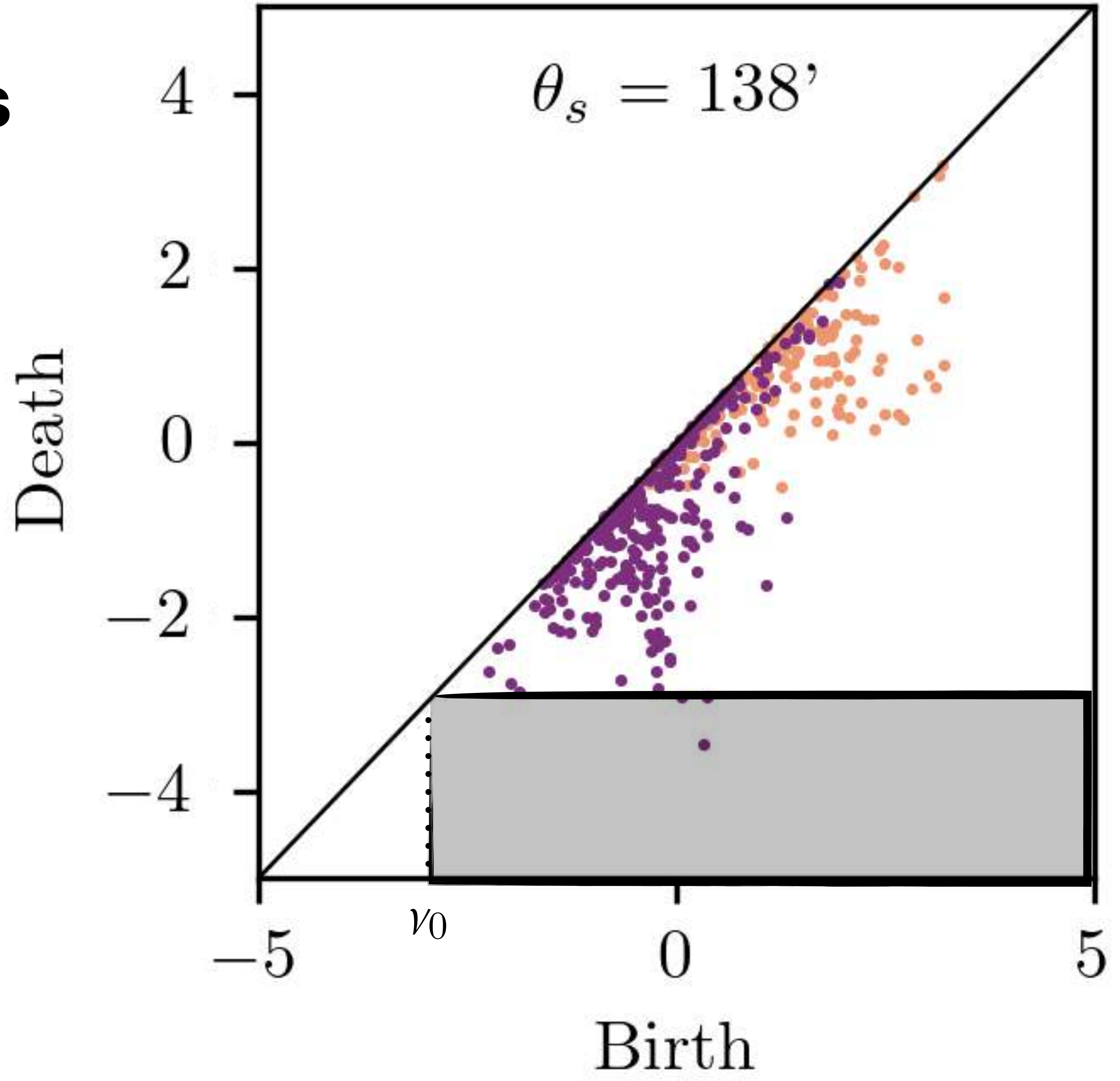
They are a compression of the persistent diagram.



# From a persistent diagram to Persistent Betti numbers

Persistent Betti numbers count the number of connected components and holes that are in a given region of the persistent diagram.

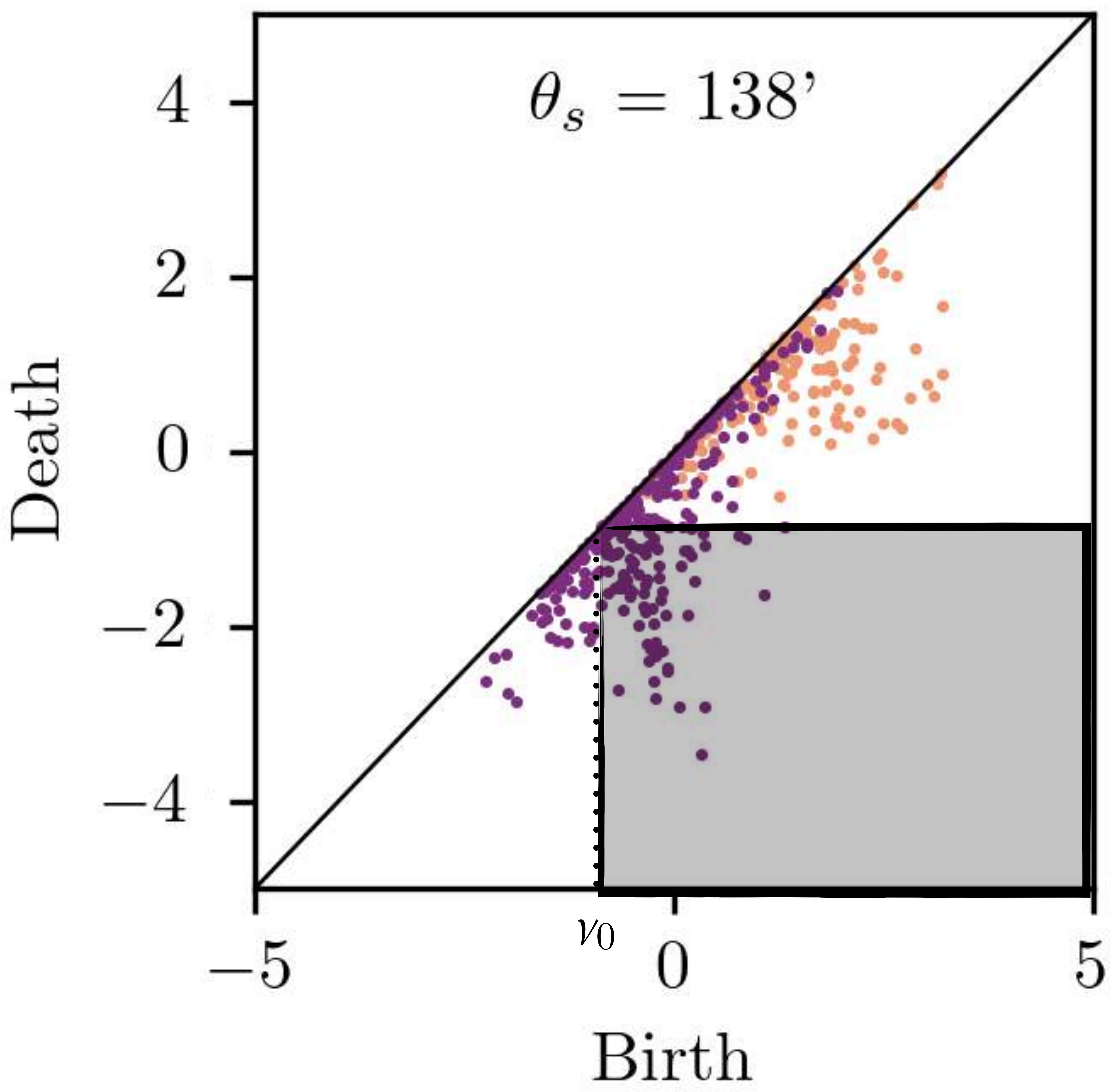
They are also a compression of the persistent diagram, capturing the most persistent features.



# From a persistent diagram to Persistent Betti number

Persistent Betti numbers count the number of connected components and holes that are in a given region of the persistent diagram.

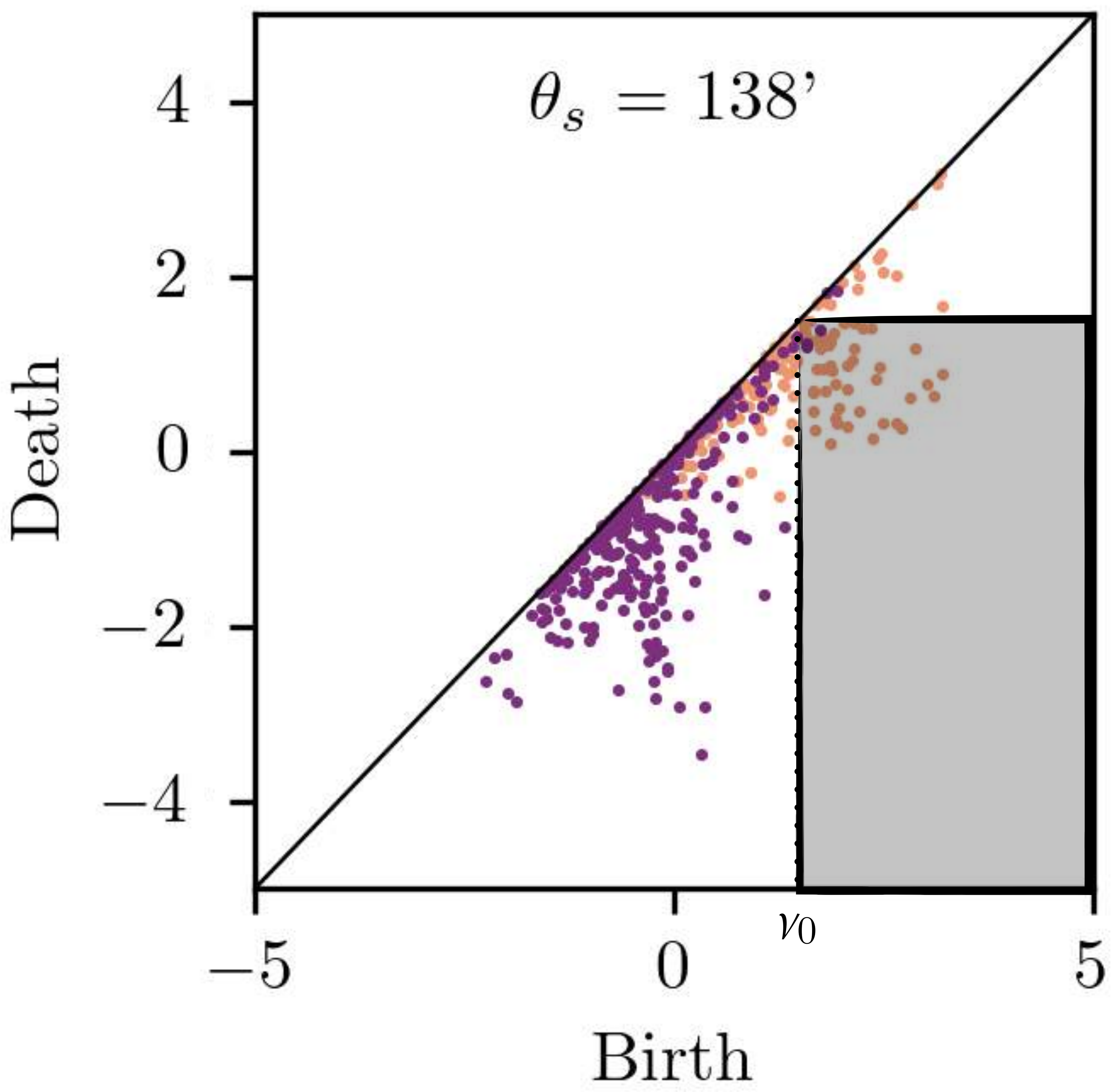
They are also a compression of the persistent diagram, capturing the most persistent features.



# From a persistent diagram to Persistent Betti number

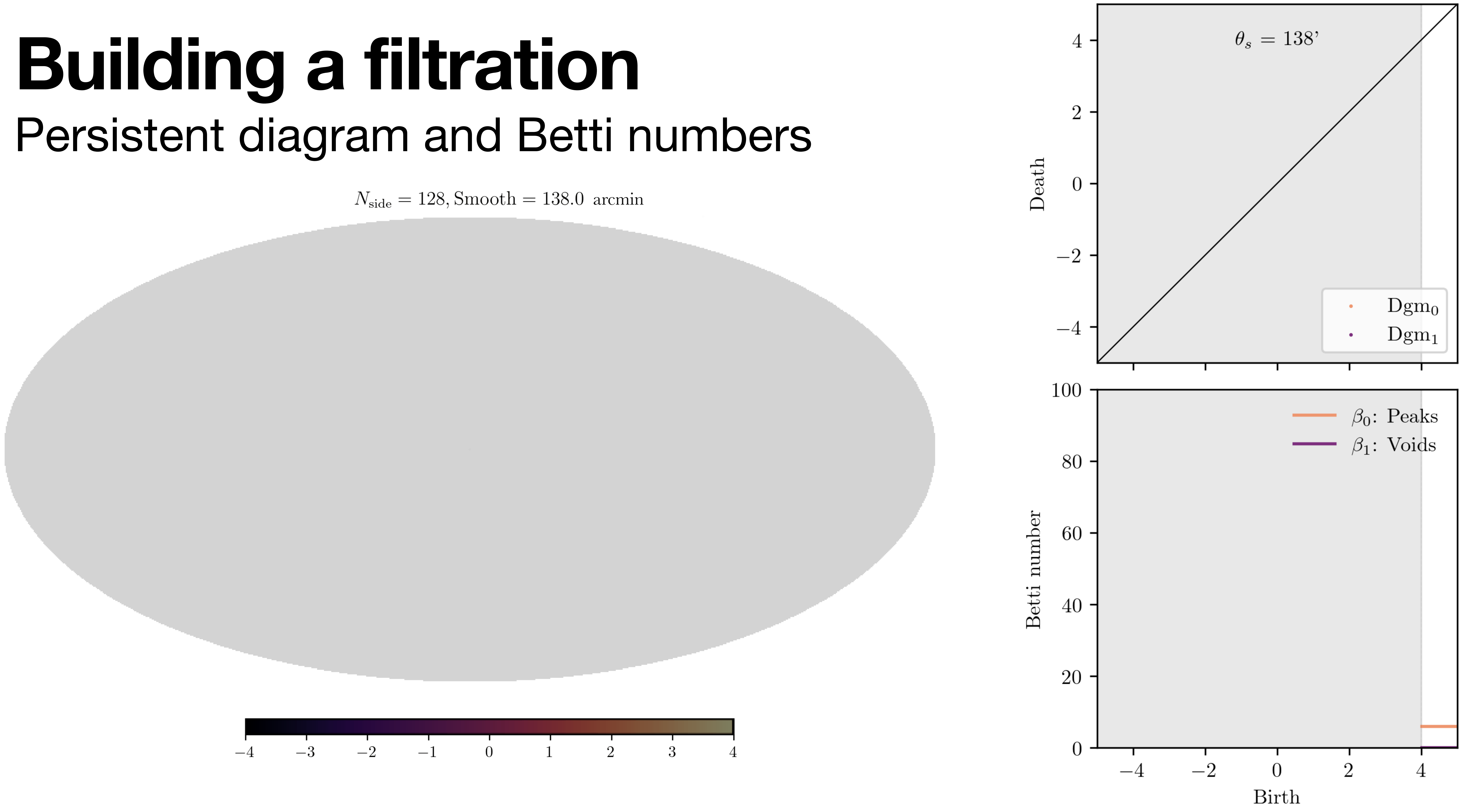
Persistent Betti numbers count the number of connected components and holes that are in a given region of the persistent diagram.

They are also a compression of the persistent diagram, capturing the most persistent features.



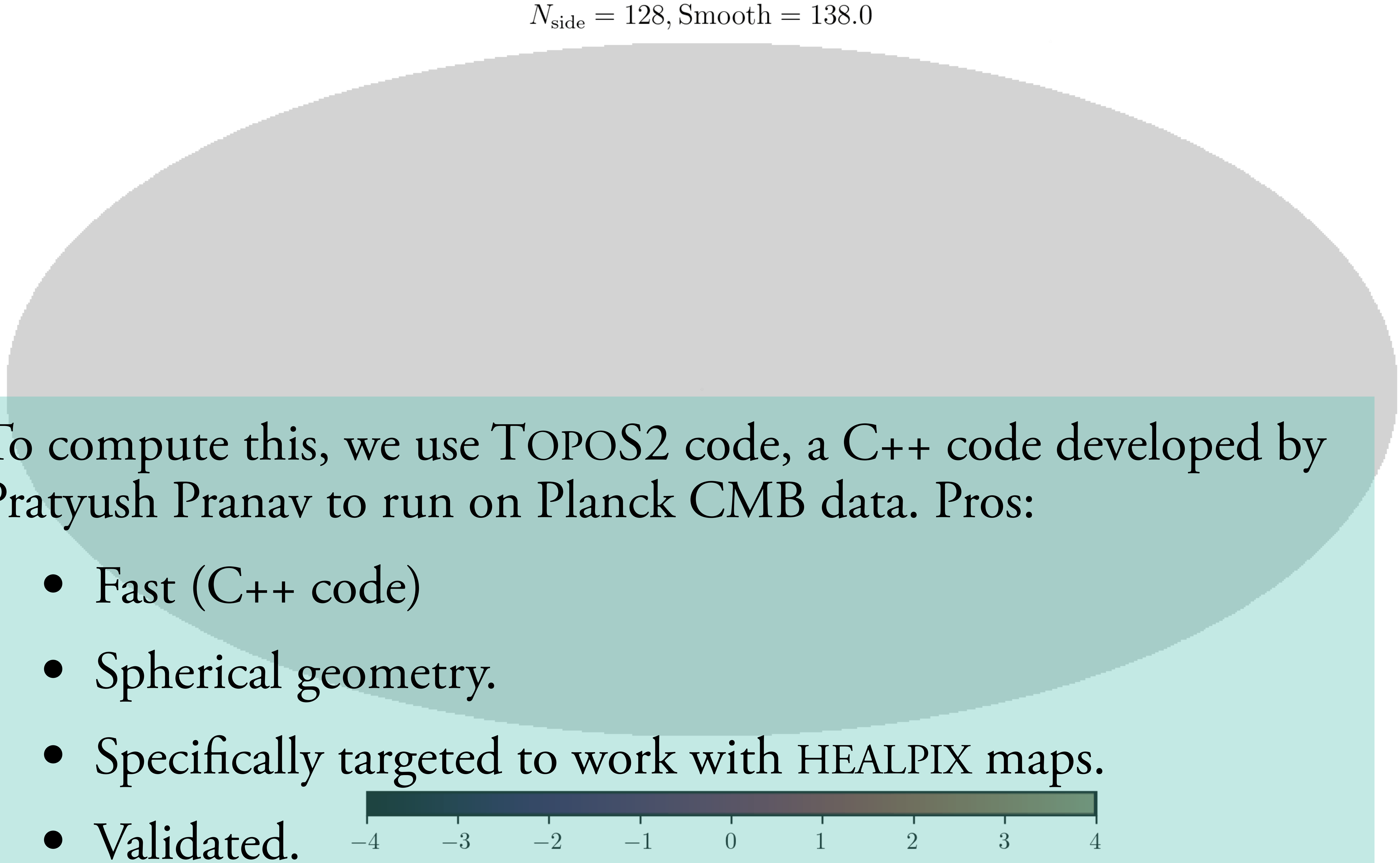
# Building a filtration

## Persistent diagram and Betti numbers



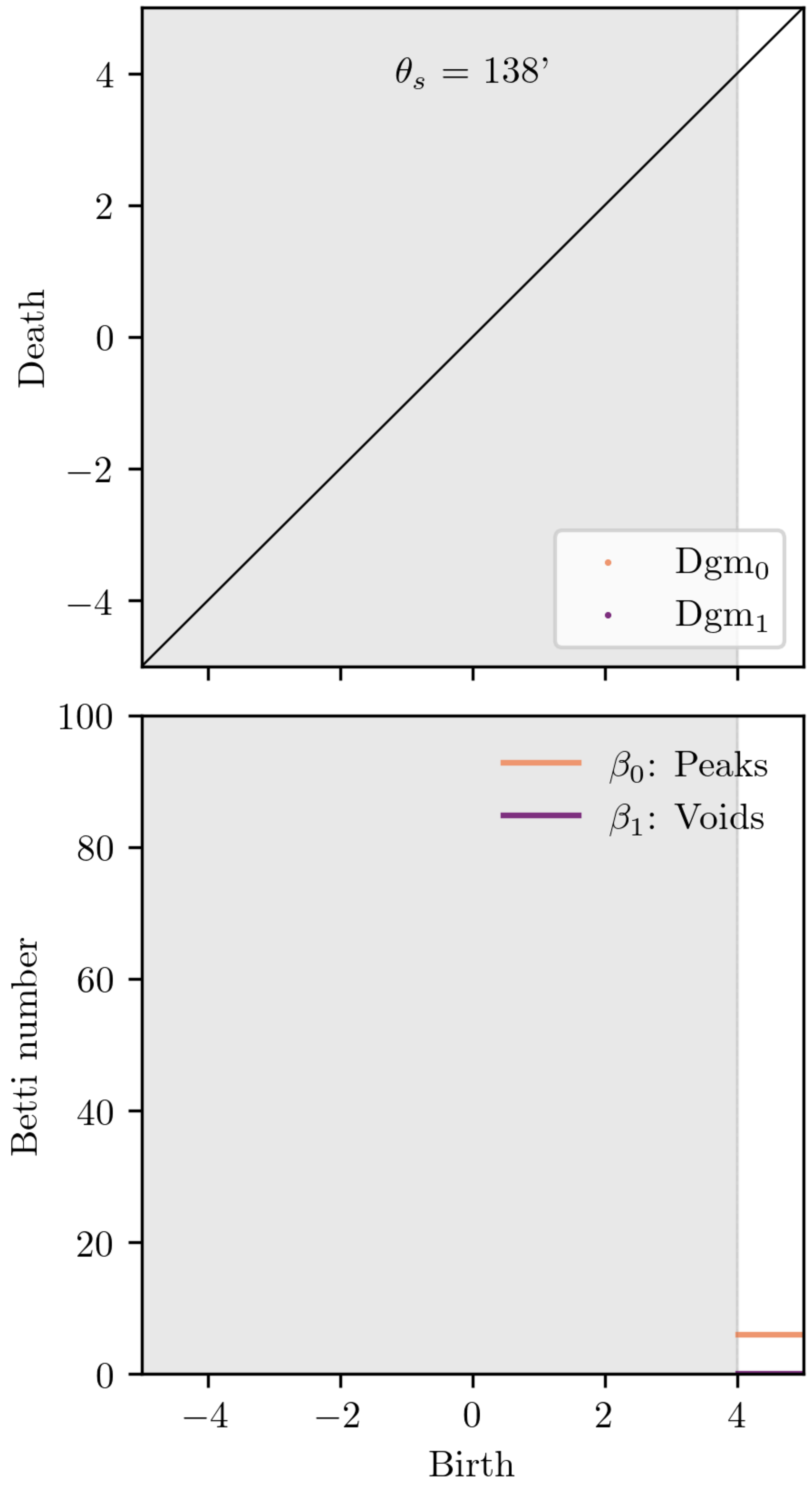
# Building a filtration

## Persistent diagram and Betti numbers

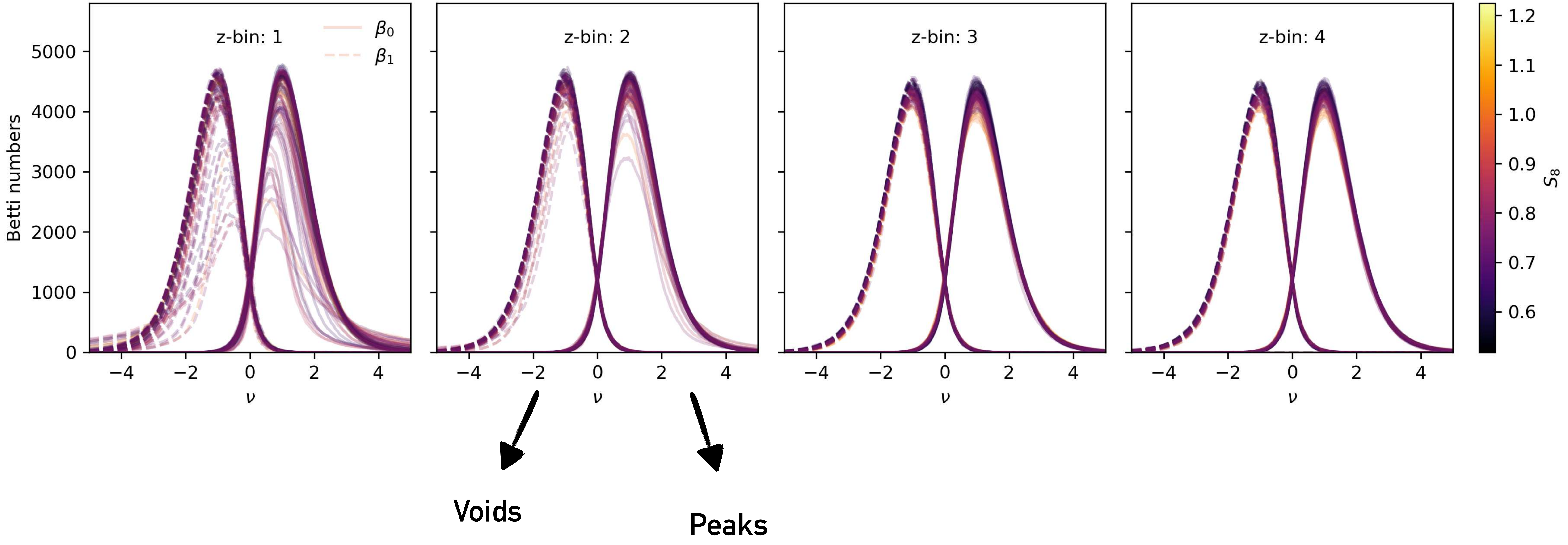



To compute this, we use TOPOS2 code, a C++ code developed by Pratyush Pranav to run on Planck CMB data. Pros:

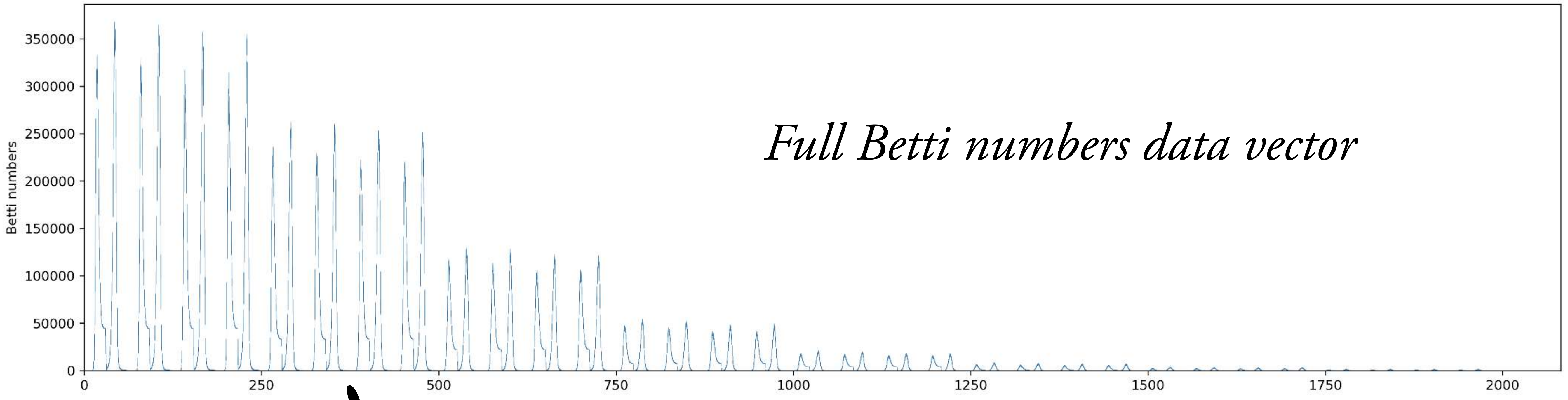
- Fast (C++ code)
- Spherical geometry.
- Specifically targeted to work with HEALPIX maps.
- Validated.



# **Betti Numbers:** *Cosmology dependence*

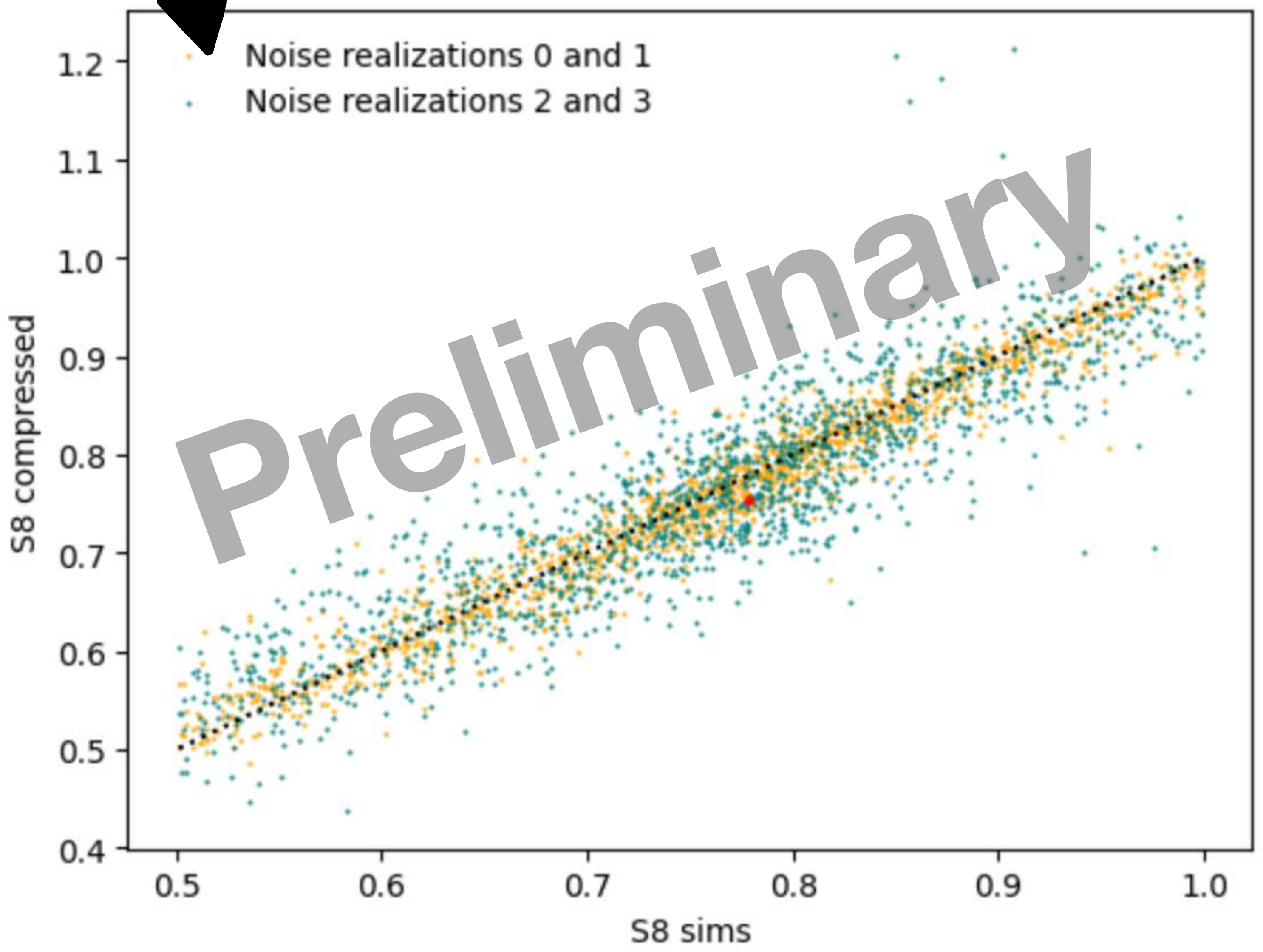


# **Betti Number:** *All the way to parameter constraints*

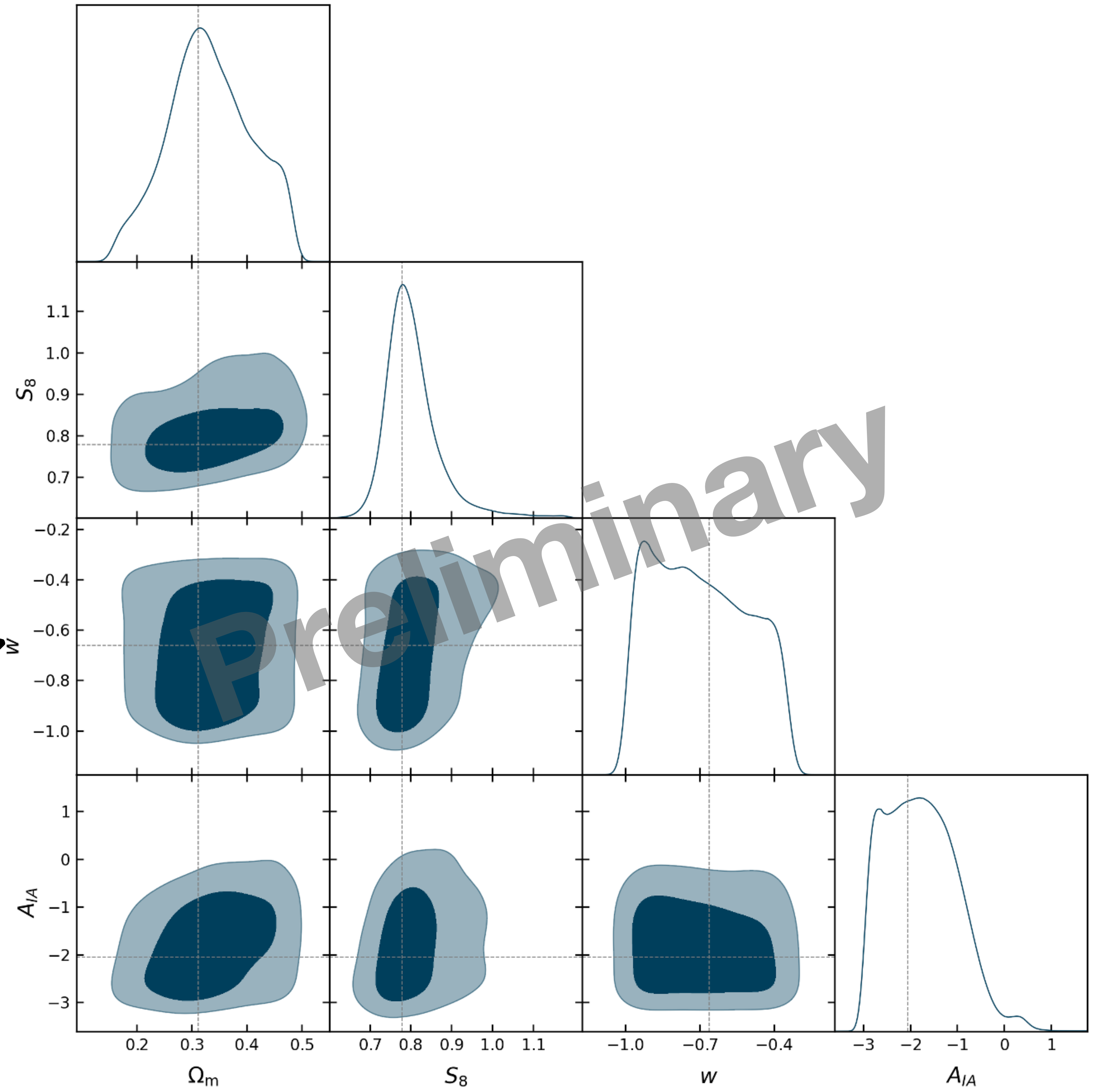



*Full Betti numbers data vector*

*Compression step using a neural network*




*Density estimation to learn the likelihood*



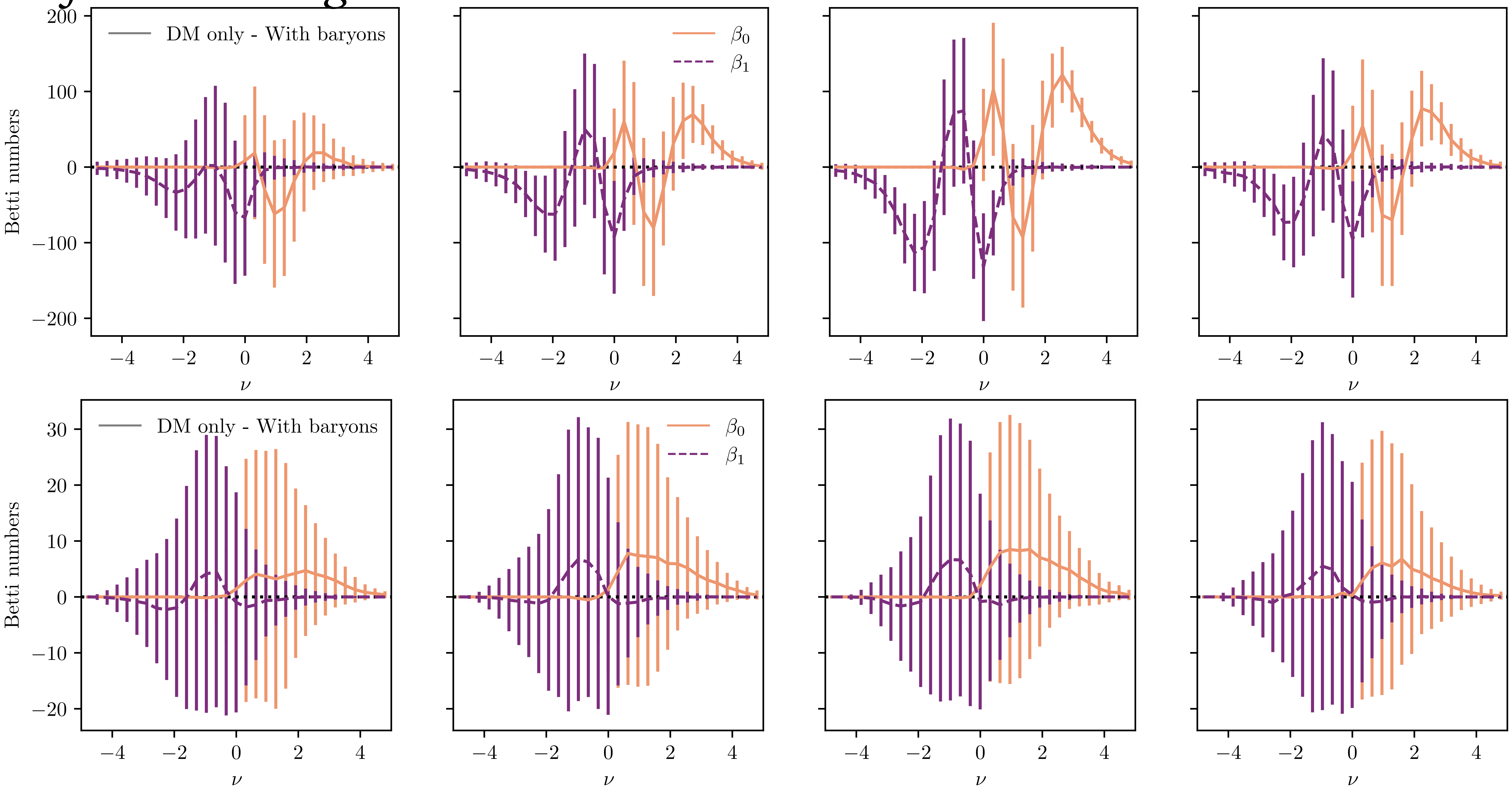

Caveat: Only 20% of the simulations have been used so far.

# **Betti Numbers:** *Baryonic feedback impact*

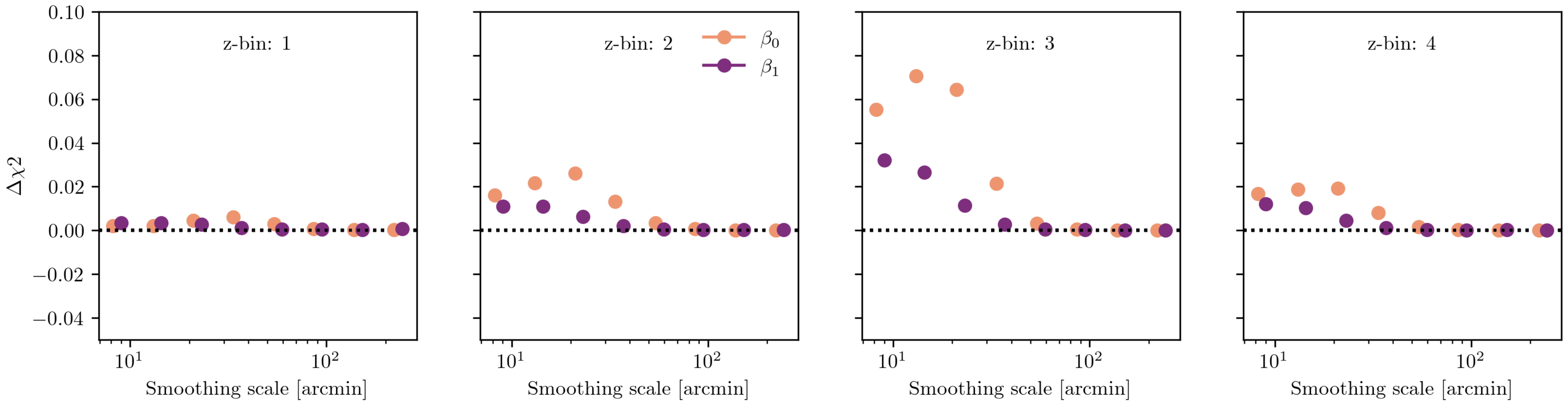
*Mean of 800 cosmo grid simulations*

8 arcmin

33 arcmin



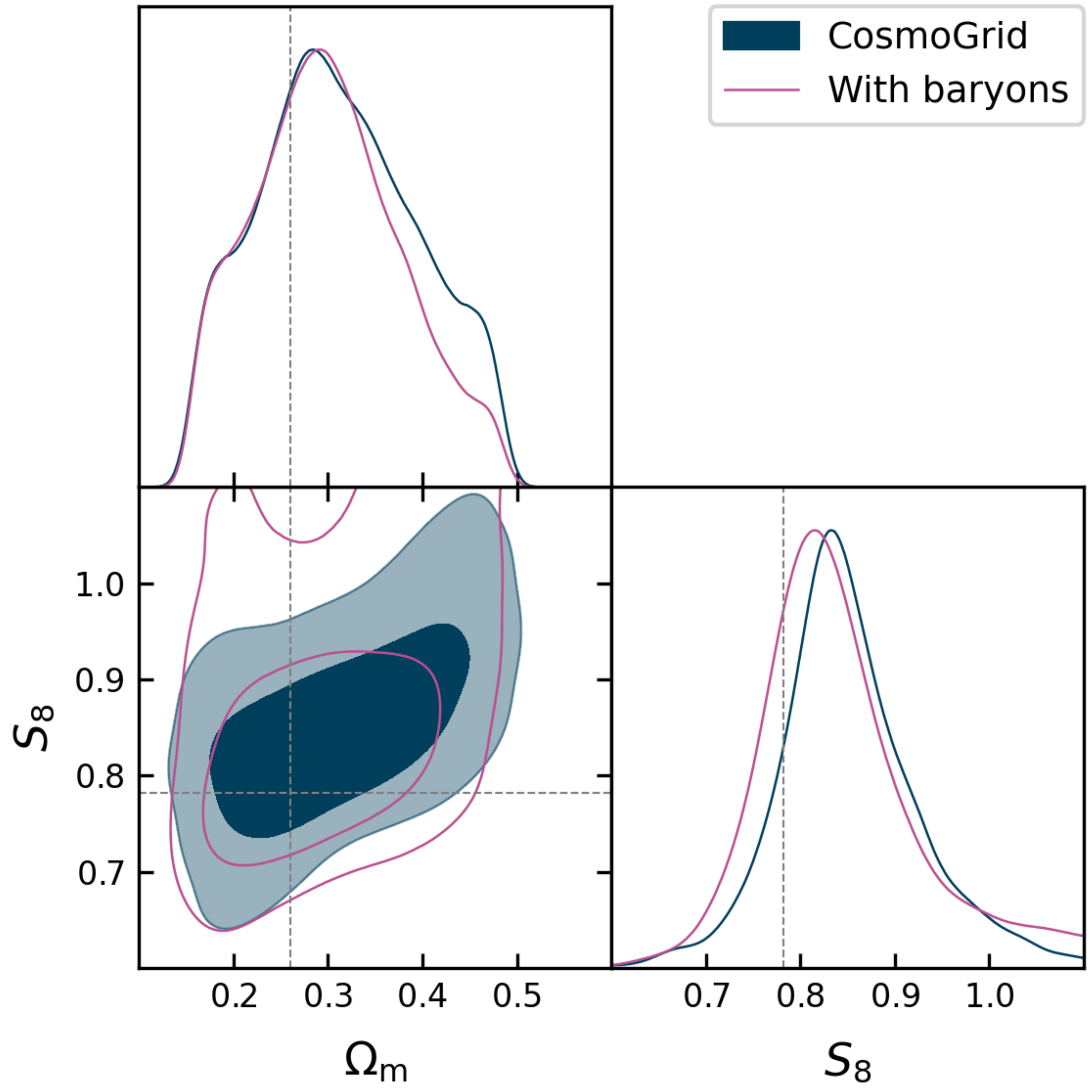
# **Betti Numbers:** *Baryonic feedback impact*



# **Betti Numbers Validation:** *Baryonic feedback impact*

*End-to-end test*

Caveat: Only 20% of the simulations have been used so far.

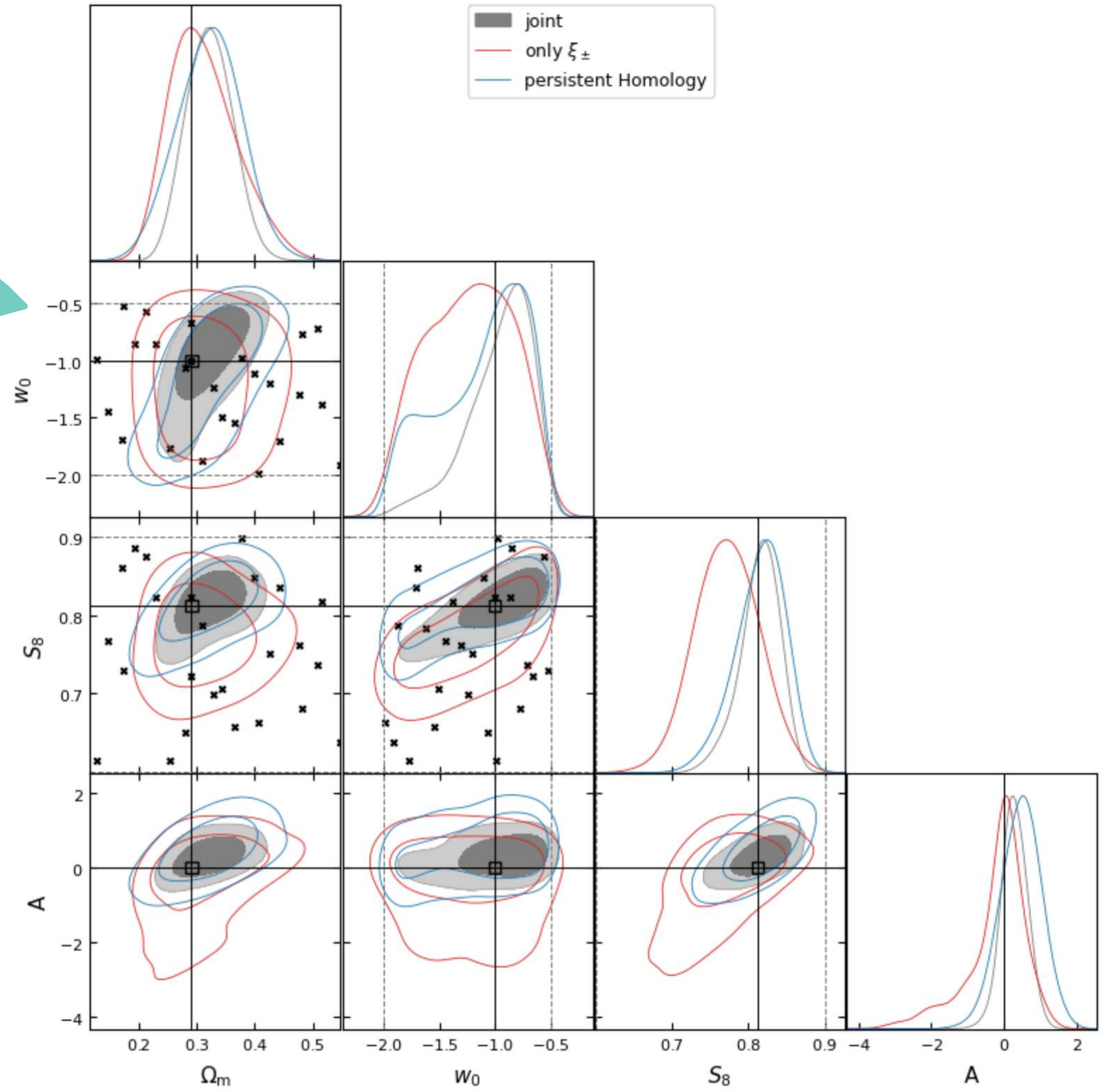


# This work in context

## Related persistent homology studies

- **Heydenreich et al. (2021, 2022)**: 2007.13724, 2204.11831.
  - Persistent homology applied to DES Y1 data. *—> We use DES Y3 data.*
  - Flat geometry (patches were used). *—> We use spherical geometry, full footprint can be analyzed at once.*
  - Cubical complexes *—> Complex (triangulation) is based on HEALPIX grid.*
- **Yip et al. (2024)**: 2403.13985.
  - Cosmology with Persistent Homology: a Fisher Forecast.
  - 3D LSS simulations, QUIJOTE SUITE.
- **Euclid Collaboration (2023):** 2301.12890.
  - Forecasts for ten different higher-order weak lensing statistics.
- **Pranav (2021):** 2111.15427
  - Using healpix complex applied to CMB data to find anomalies.

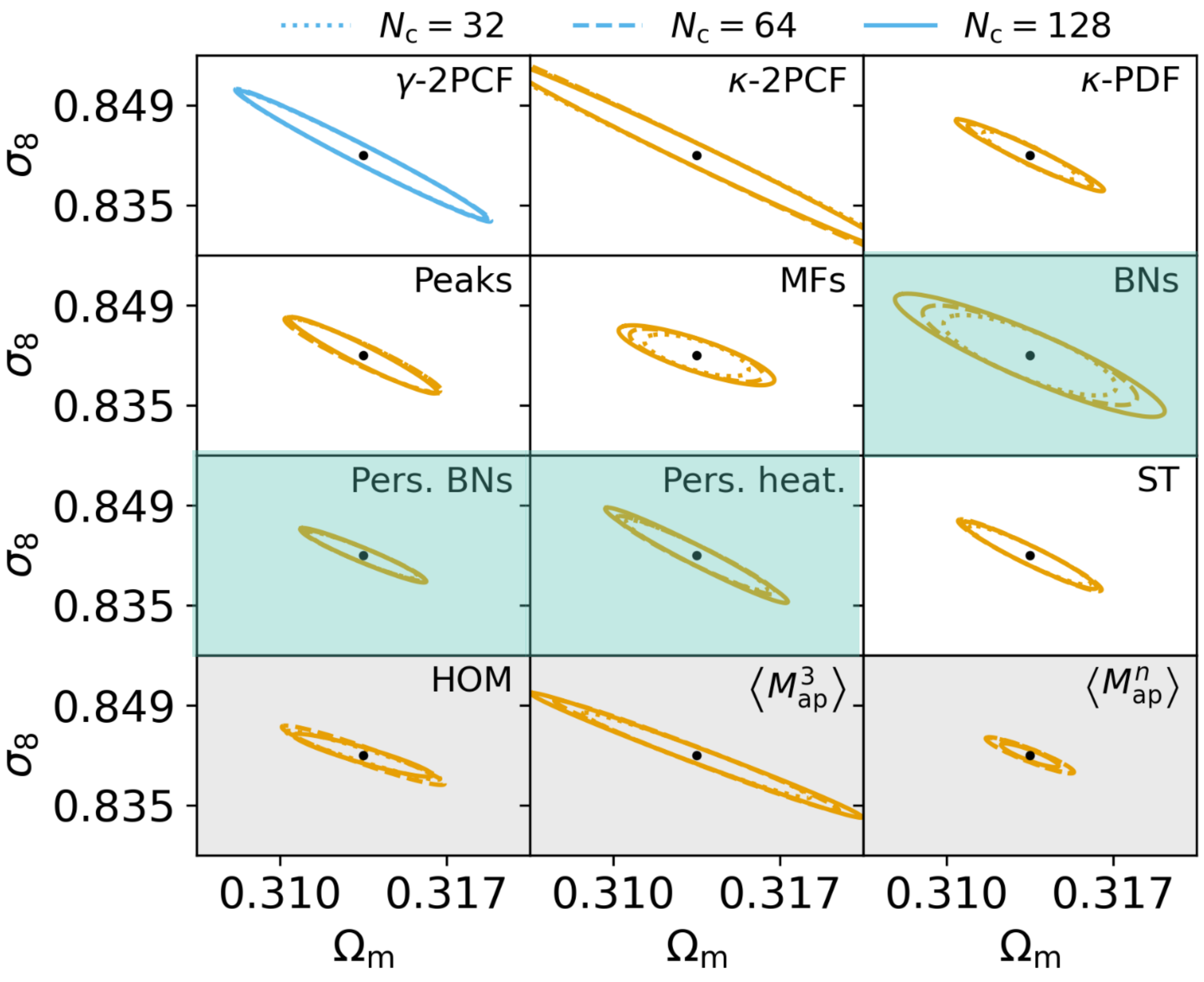
**Heydenreich et al. (2022)**: 2204.11831



# This work in context

## Related persistent homology studies

- **Heydenreich et al. (2021, 2022)**: 2007.13724, 2204.11831.
  - Persistent homology applied to DES Y1 data. —> *We use DES Y3 data.*
  - Flat geometry (patches were used). —> *We use spherical geometry, full footprint can be analyzed at once.*
  - Cubical complexes —> *Complex (triangulation) is based on HEALPIX grid.*
- **Yip et al. (2024)**: 2403.13985.
  - Cosmology with Persistent Homology: a Fisher Forecast.
  - 3D LSS simulations, QUIJOTE SUITE.
- **Euclid Collaboration (2023):** 2301.12890.
  - Forecasts for ten different higher-order weak lensing statistics.
- **Pranav (2021):** 2111.15427
  - Using healpix complex applied to CMB data to find anomalies.



# The contrastive learning part

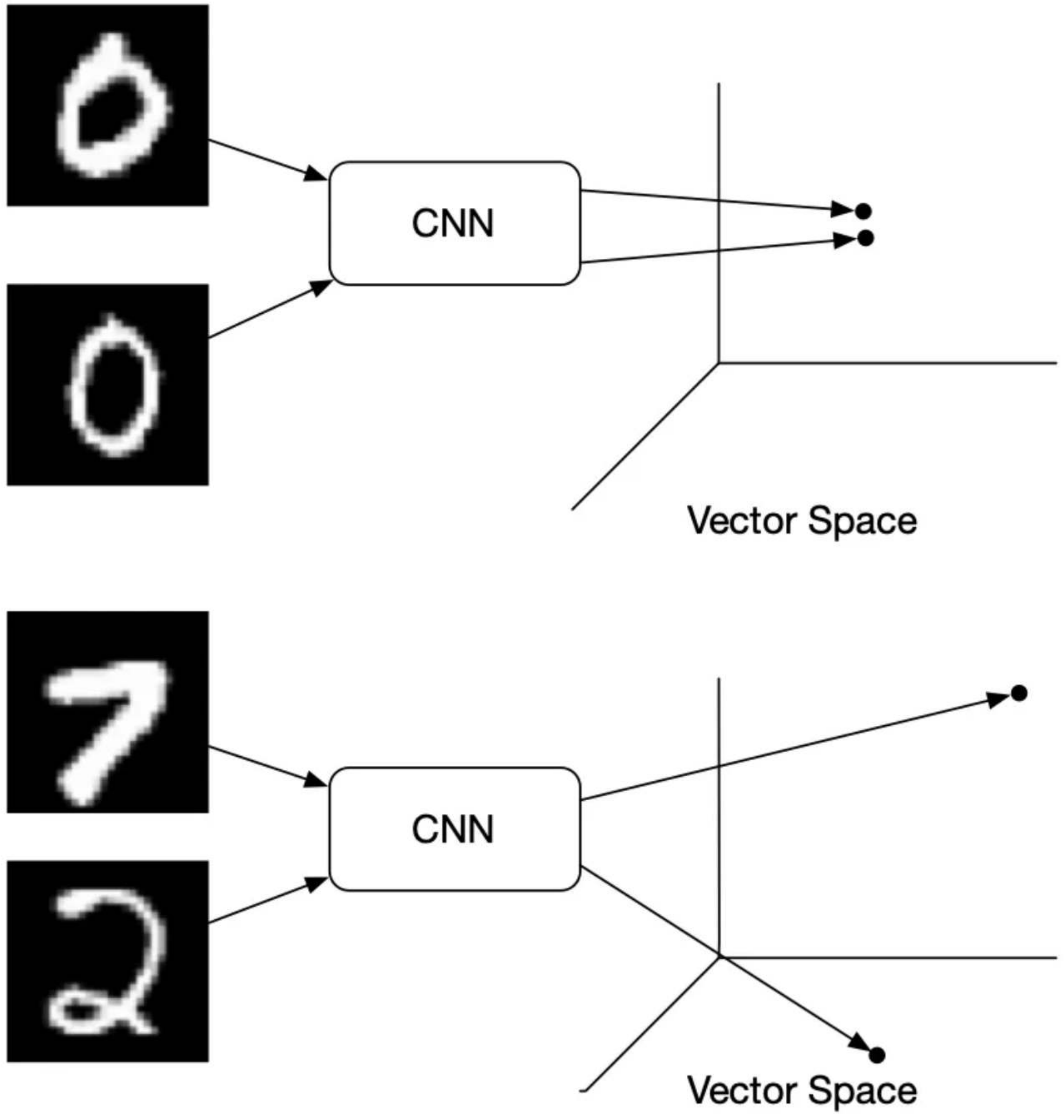
*Judit Prat, Ruoxi Jiang (UChicago), Peter Lu (UChicago), Joshua Williamson (UCL), Niall Jeffrey (UCL), Marco Gatti (UPenn), Chihway Chang (UChicago) ++*

NORDITA
The Nordic Institute for
Theoretical Physics

# Contrastive learning

## What is it?

The goal of contrastive learning is to learn an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart. Contrastive learning can be applied to both supervised and unsupervised settings.




Positive (top) and negative samples embedded into a vector space

# The EMBED & EMULATE framework

**Jiang & Willet et al. (2022)**: 2211.01554

## What is it?

EMBED & EMULATE is a recently proposed method that has shown that contrastive learning can be adapted into a powerful approach for parameter estimation and, more generally, simulation-based inference.

EMBED & EMULATE jointly trains two neural networks. The first, $f_{\varphi}$, maps an observation $X$ to a low-dimensional embedding.

The second, $g_{\varphi}$, emulates the map, and so maps a parameter vector $\Theta$ to the same low-dimensional embedding space.

Jiang & Willet et al. (2022): 2211.01554

# The Embed & Emulate framework

*Minimizing a symmetric multimodal InfoNCE loss function*

The InfoNCE identifies the positive sample amongst a set of unrelated samples.

Contrastive embeddings (Neural networks with parameters $\varphi$)

Input parameters of the simulation.

Pixels in map

Negative pairs: samples that are not equal to the main sample.

$$L_{\text{InfoNCE}}(\varphi) := \tag{6}$$

$$\mathop{\mathbb{E}}_{\substack{(X,\Theta)\sim p_D(X,\Theta)\\ \{\Theta_i^-\}_{i=1}^M \sim p_D(\Theta)}} \left[ -\log\left( \frac{e^{f_\varphi(X)\cdot g_\varphi(\Theta)/\tau}}{e^{f_\varphi(X)\cdot g_\varphi(\Theta)/\tau} + \sum_{i=1}^M e^{f_\varphi(X)\cdot g_\varphi(\Theta_i^-)/\tau}} \right) \right] \tag{7}$$

$$+ \mathop{\mathbb{E}}_{\substack{(X,\Theta)\sim p_D(X,\Theta)\\ \{X_i^-\}_{i=1}^M \sim p_D(X)}} \left[ -\log\left( \frac{e^{f_\varphi(X)\cdot g_\varphi(\Theta)/\tau}}{e^{f_\varphi(X)\cdot g_\varphi(\Theta)/\tau} + \sum_{i=1}^M e^{f_\varphi(X_i^-)\cdot g_\varphi(\Theta)/\tau}} \right) \right]$$

# The Embed & Emulate framework

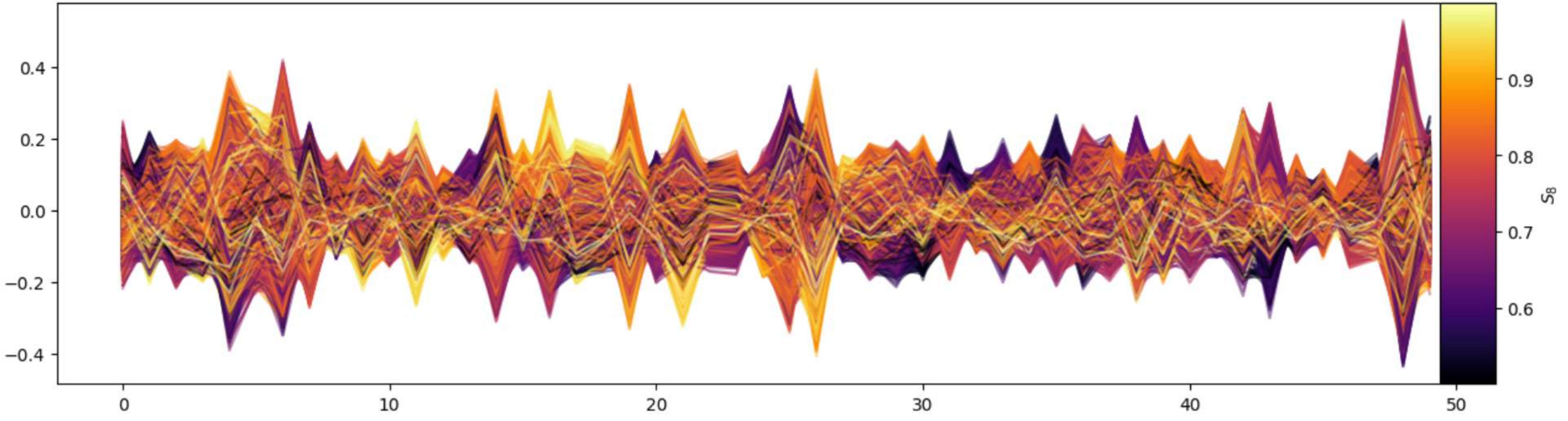
## *Contrastive embeddings as a summary statistic*

Then we define our summary statistic as

$$T(X) := f_{\varphi^*}(X)$$

*In the ideal case, where we have arbitrarily flexible embedding functions and we reach the global minimum during training, this learned summary statistic is guaranteed to be a sufficient statistic for $\Theta$.*

Then we can use the neural density estimation (NDE) framework to estimate the likelihood $p(T(X)|\,\Theta)$.



# Direct inference using the contrastive embeddings

The likelihood for the summary statistic $T(X)$ learned using EMBED & EMULATE can be written as:

$$p(T(X) \mid \Theta) \propto e^{T(X)\cdot g_{\varphi^*}(\Theta)/\tau}\; p_D(T(X))$$

Then, the posterior can be directly obtained as:

$$\begin{aligned} p(\Theta \mid X) = p(\Theta \mid T(X)) &\propto p(T(X) \mid \Theta)\, p(\Theta) \\ &\propto e^{T(X)\cdot g_{\varphi^*}(\Theta)/\tau}\; p(\Theta) \end{aligned}$$

# Conclusions and outlook

Simulation based inference is a promising method to extract higher order information from weak lensing mass maps.

Once the simulation infrastructure is in place, it has many benefits such as:

- Faster to run the likelihood with different priors.
- Easier to marginalize over high dimensionality parametrization of systematic uncertainties (such as redshift uncertainty).
- Make less assumptions about the likelihood form.

Moving forward, it will be useful to compare the different statistics on data using the same framework & characterize their sensitivity to systematic effects, preparing for applying these methods to Stage IV surveys.

# Thanks!

NORDITA
The Nordic Institute for
Theoretical Physics

Judit Prat
*COSMO21 Chania*
*May 2024*