Classification of cosmic structures for galaxies with machine learning

Shigeki Inoue
(Hokkaido Univ.)

Xiaotian Si

Takashi Okamoto

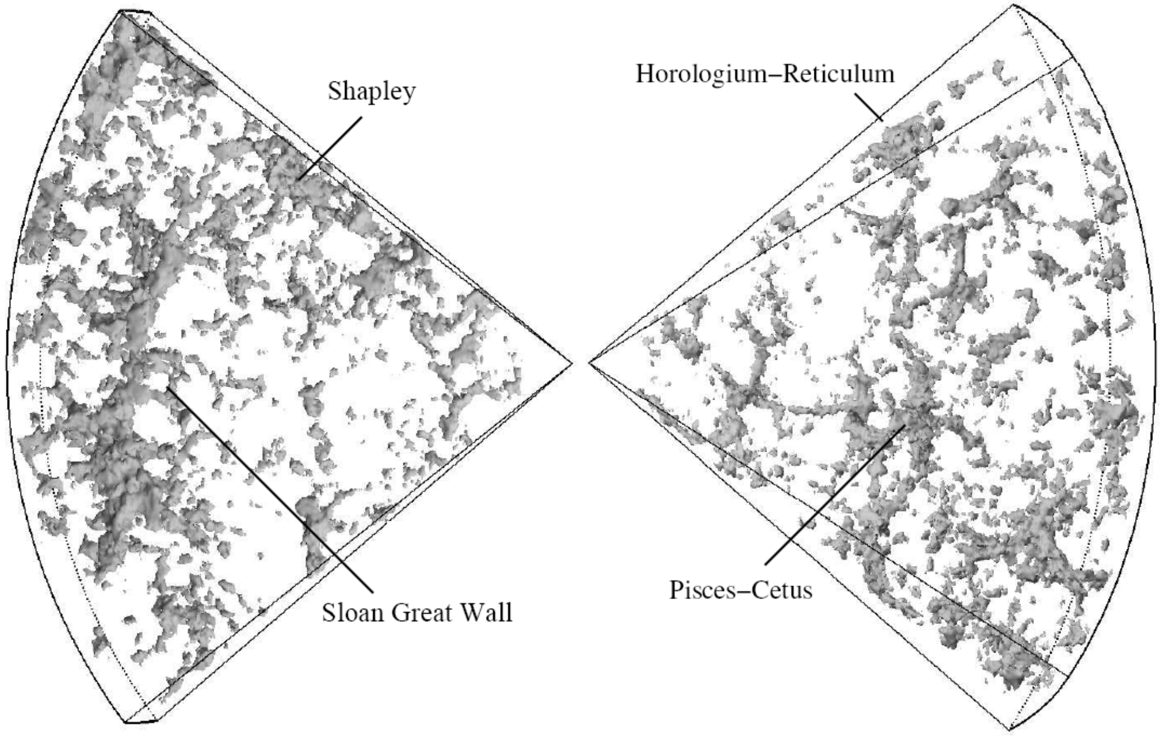
Moka Nishigaki

Large-scale structure / Cosmic structure / Cosmic web

- Spongy structures in large scales >~10 Mpc

- Galaxy formation/evolution
  - Environmental dependence of galaxies
  - cluster region v.s. field environment
    - Gas supply to filament galaxies

- Cosmology
  - Test for the current cosmology
    - Lengths of filaments
    - Sizes of voids



How can we define voids and filaments?

How to define the cosmic structures

- In theory (cosmological simulations),
  - Using DM velocity fields, compute a gradient tensor

(see Hoffman et al. 2012)

dark matter

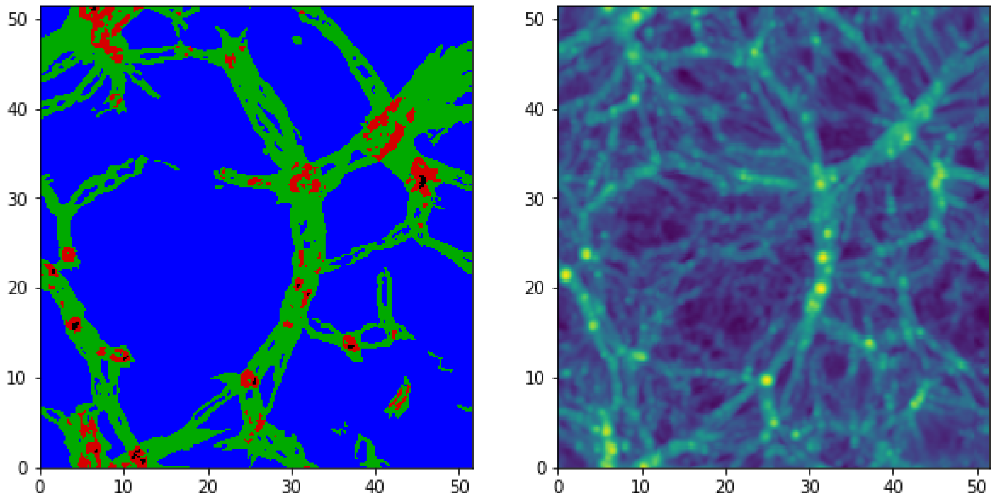
$$\Sigma_{ij} = -\frac{1}{2H_0}\left(\frac{\partial v_i}{\partial r_j} + \frac{\partial v_j}{\partial r_i}\right)$$

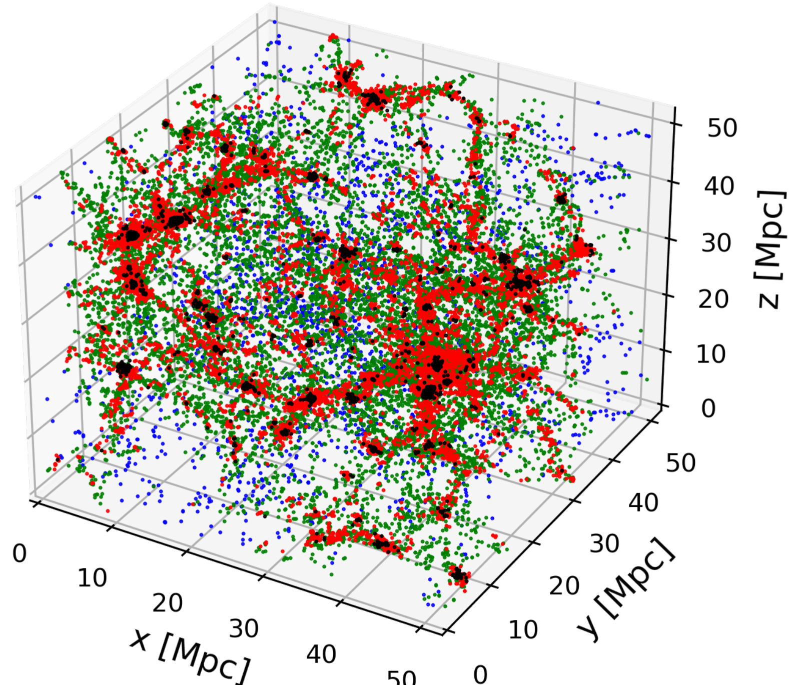
Gradients of DM density and potential are also often used, but basically the same in the linear regime.

  - Compute the 3 eigenvalues of the tensor
  - How many eigenvalues are larger than $\lambda_{th} = 0.44$
    - **3: knot**
    - **2: filament**
    - **1: sheet**
    - **0: void**

Dimension of contraction

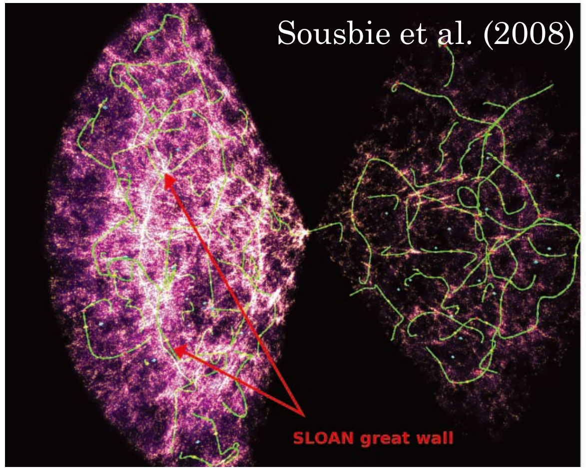
2D thin slice





How to define the cosmic structures

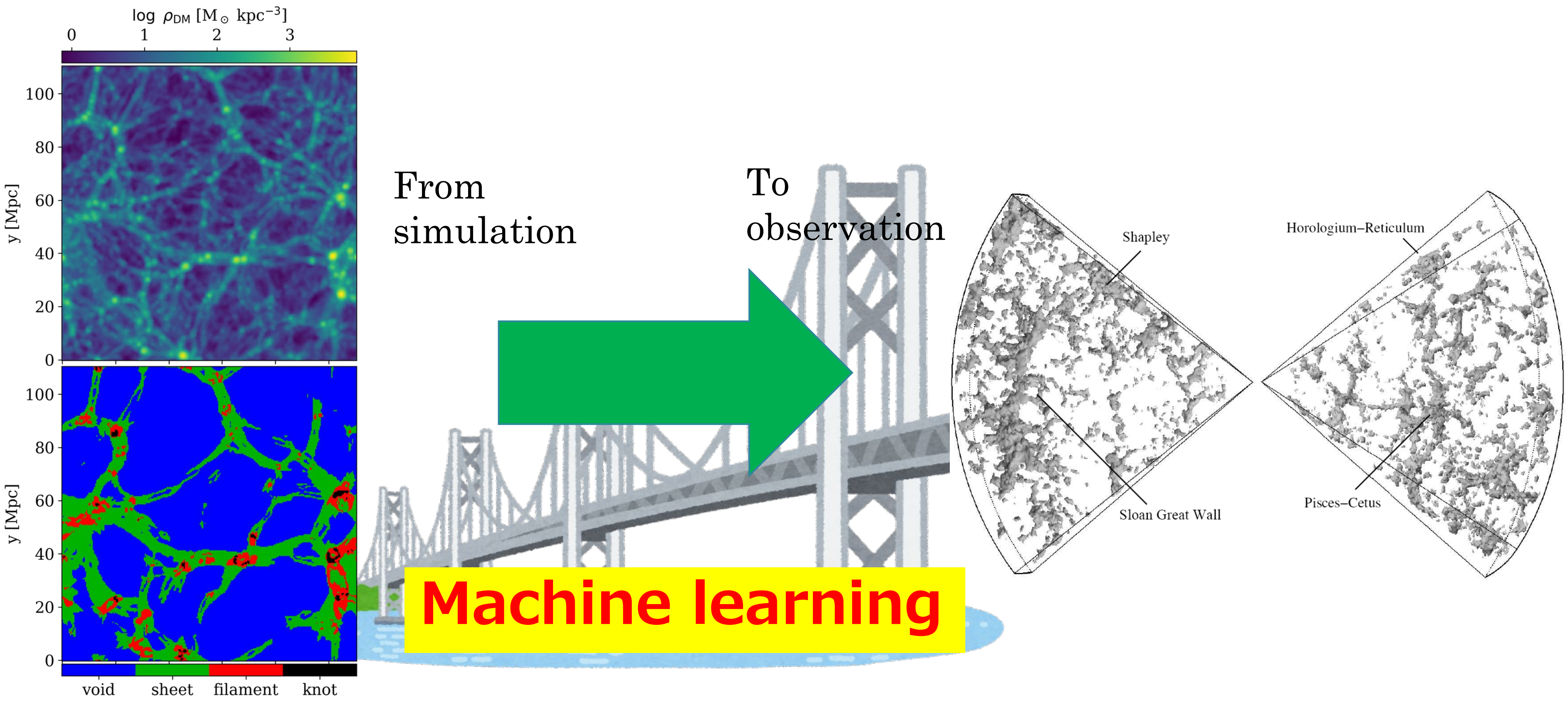
- In observations, DM is not usable.
  - Using galaxy distribution instead, with various methods,
    - **Knot** (cluster)
      - Overdensity of galaxies with high velocity dispersion
    - **Filament**
      - Connecting saddle point of galaxy density (Sousbie 2008)
      - Concatenated cylinders with a constant width (Tempel et al. 2014)
    - **Void**
      - Watershed algorithm (e.g. Sutter et al. 2012)
      - Maximum sphere devoid of galaxy (Hoyle & Vogeley 2002)



In observations,
- The detecting methods are not consistent between the structures
- They assume that galaxy distribution traces DM density fields

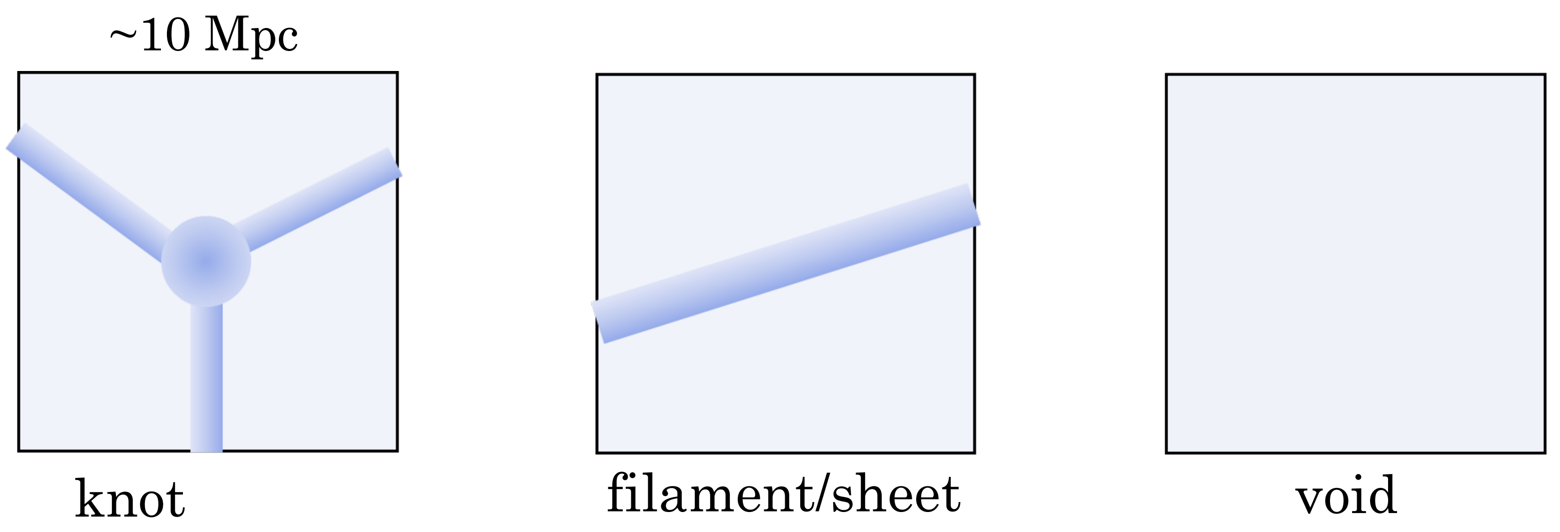
From simulations to observations

- The cosmic structures are formed by gravity.
  - Therefore, DM-based analysis in theory is thought to be plausible.
  - It is ideal to classify observed galaxies with the DM-based analysis.

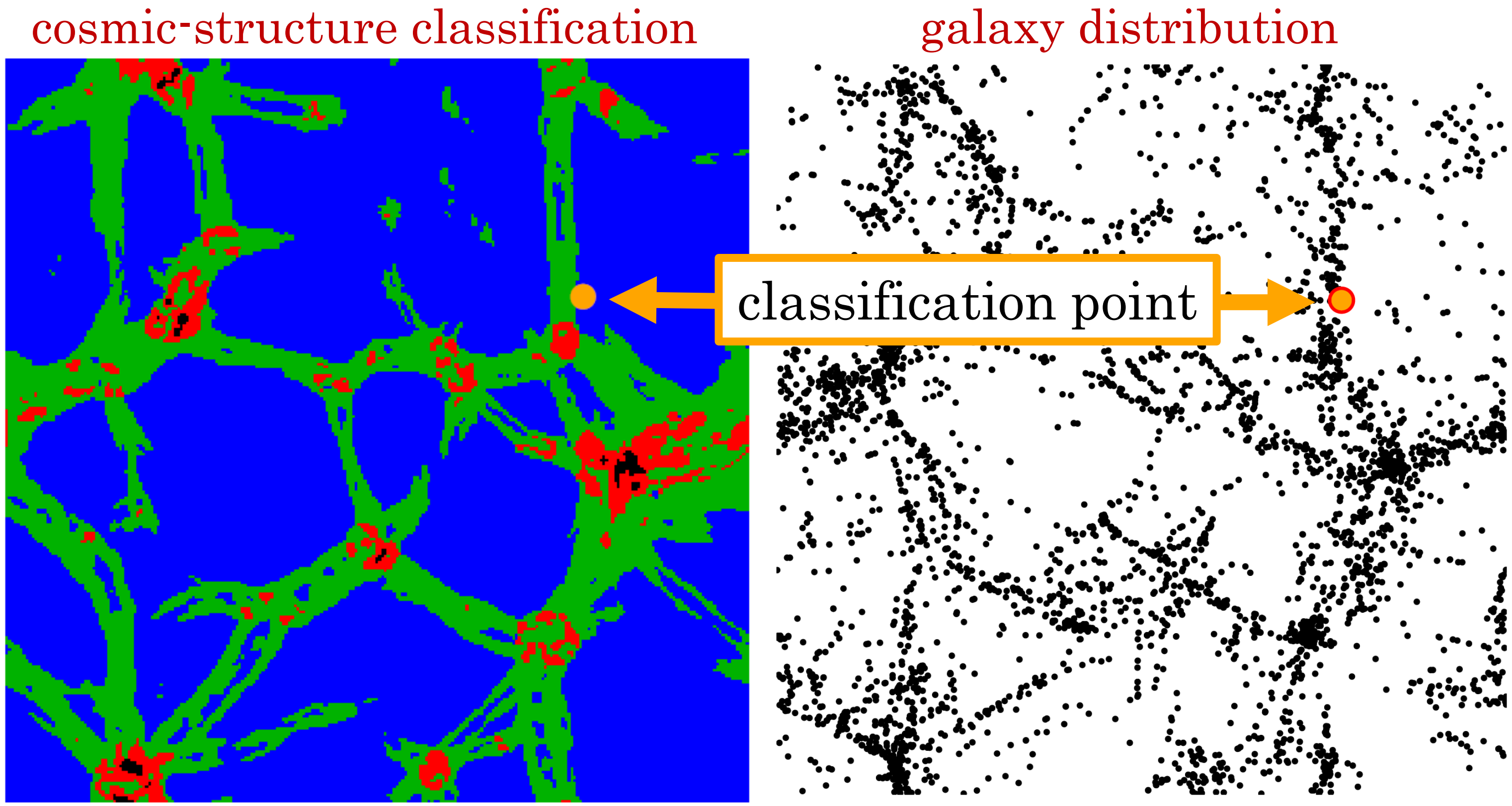


From simulations to observations

- Both DM and galaxies are accessible in cosmological simulations with baryons.
  - We use IllustrisTNG (TNG100-1) @ z=0
- Build machine-learning models trained with
  - **Classification (labelling) based on DM**
  - **Distribution of galaxies**
  - **3D-CNN**
    - The models trained with simulation can be applied to observations such as SDSS

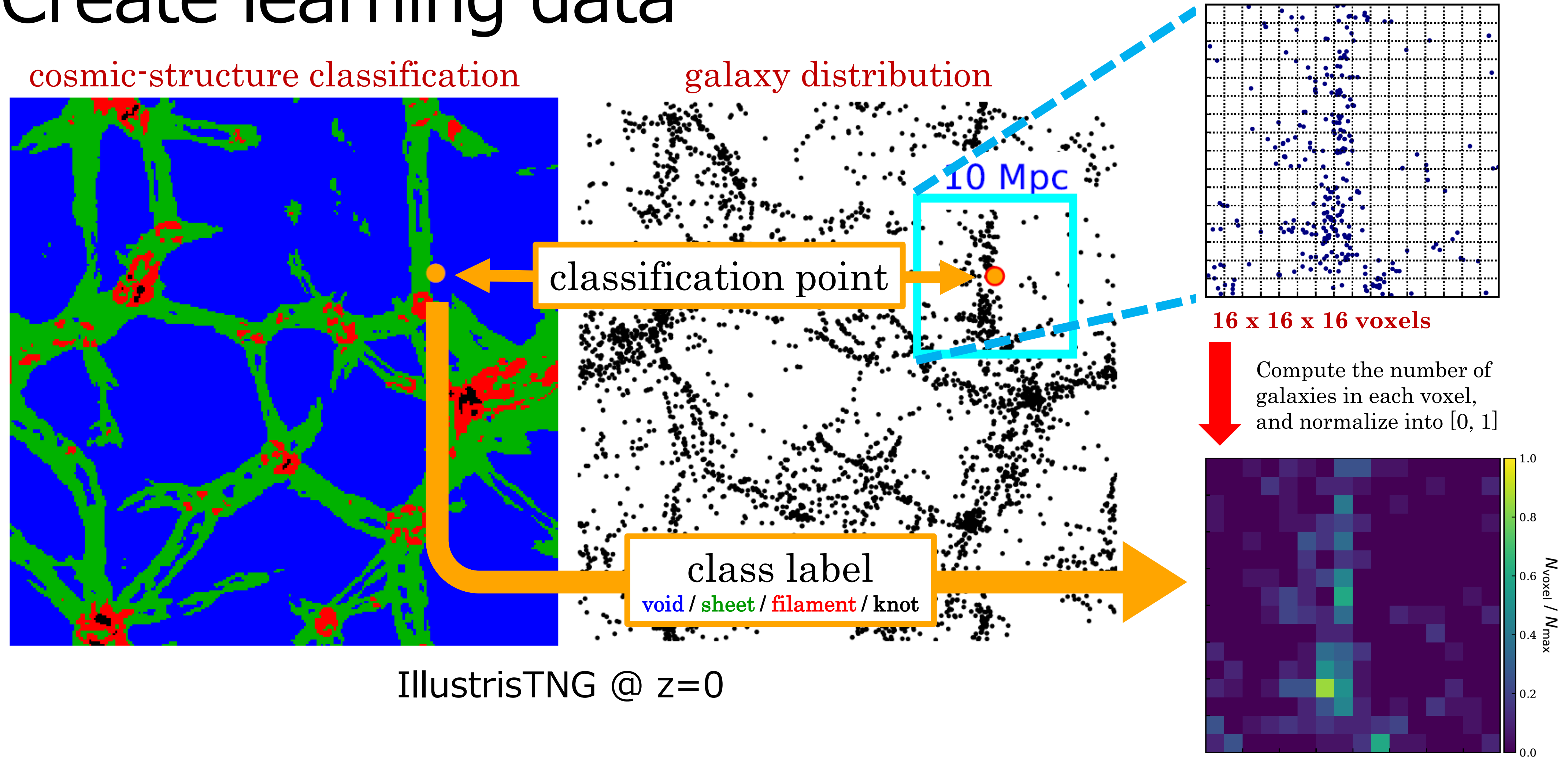


Create learning data



IllustrisTNG @ z=0

Create learning data



- We create 10000 cubic data for each class
    - 6400, 1600 and 2000 are used as training, validation and test data
  - The data have only a single channel of number distribution of galaxies

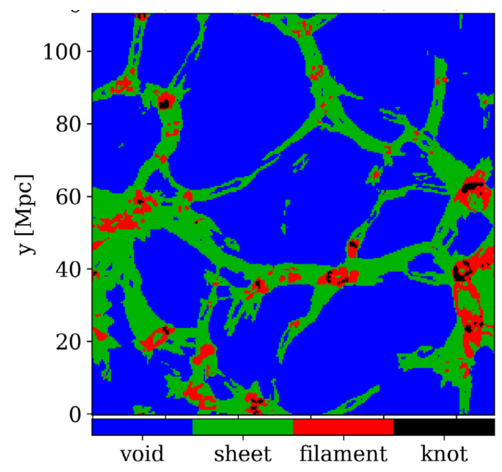
Our 3D-CNN classifier

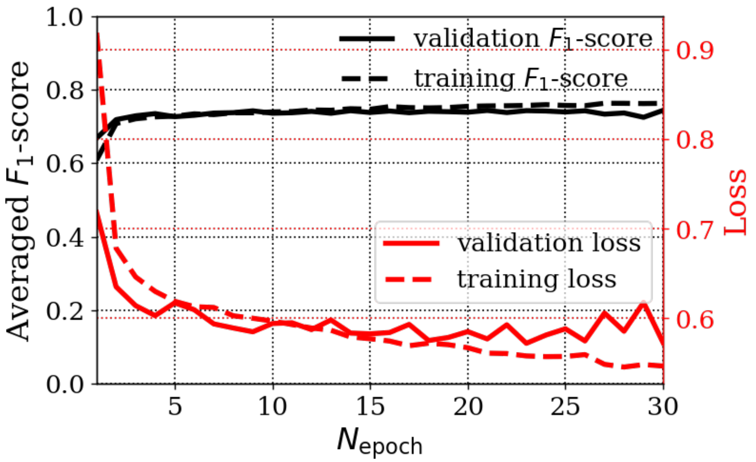
- Simple networks with only two convolution layers works enough

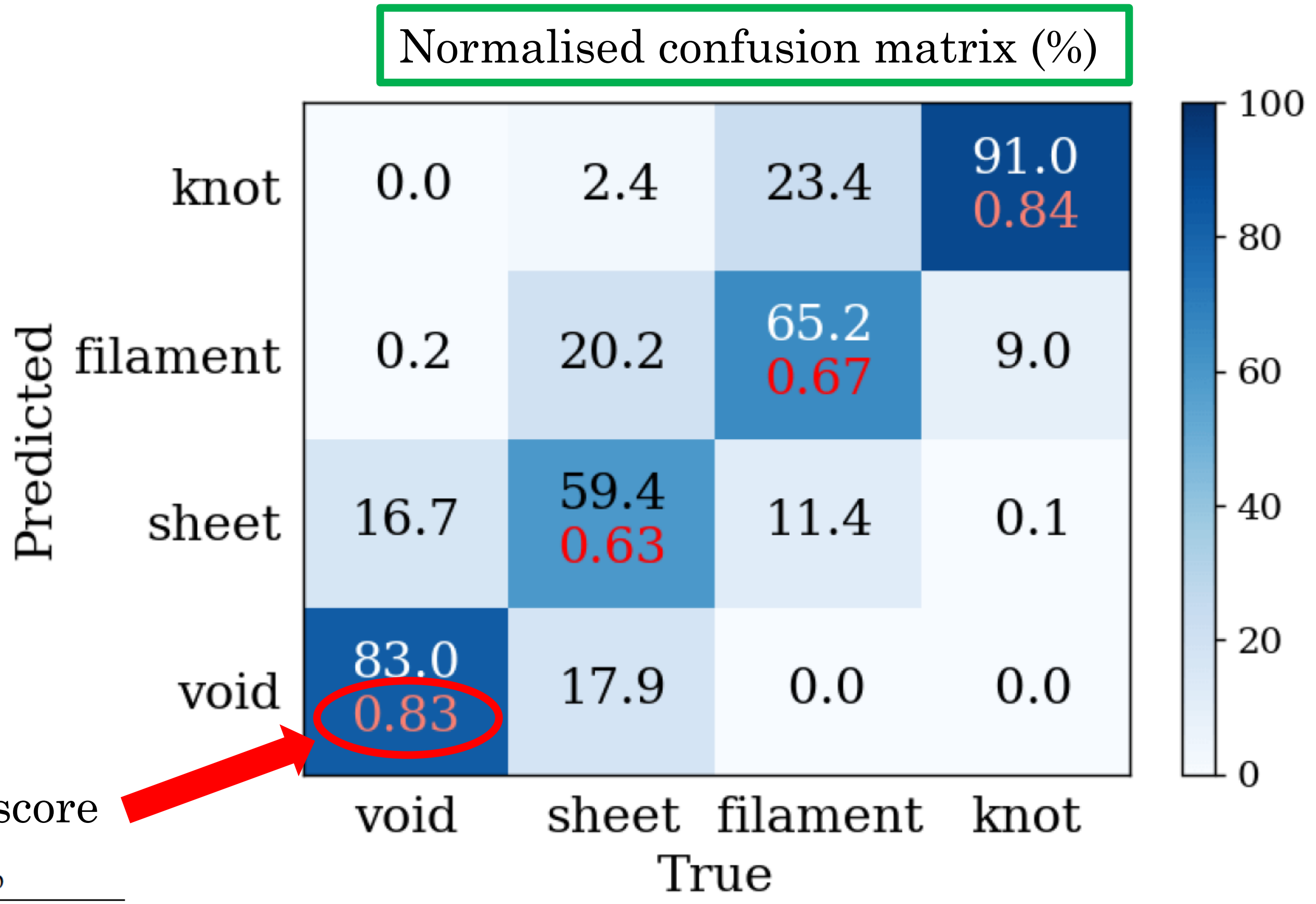
layer type	kernel size	N_{filter} / N_{out}	remarks
Convolution 1	$3 \times 3 \times 3$	$N_{\text{filter}} = 32$	ReLU / He uniform
Maxpool 1	$2 \times 2 \times 2$	-	-
Dropout 1	-	-	$R_{\text{drop}} = 0.25$
Convolution 2	$3 \times 3 \times 3$	$N_{\text{filter}} = 64$	ReLU / He uniform
Maxpool 2	$2 \times 2 \times 2$	-	-
Dropout 2	-	-	$R_{\text{drop}} = 0.25$
Flatten	-	-	-
Fully connected 1	-	$N_{\text{out}} = 512$	ReLU / He uniform
Fully connected 2	-	$N_{\text{out}} = 512$	ReLU / He uniform
Output	-	$N_{\text{out}} = N_{\text{class}}$	softmax

Classification for spatial points

- We randomly select and classify a spatial point in the simulation
  - Consider haloes having stars to be “galaxies”







F1-score

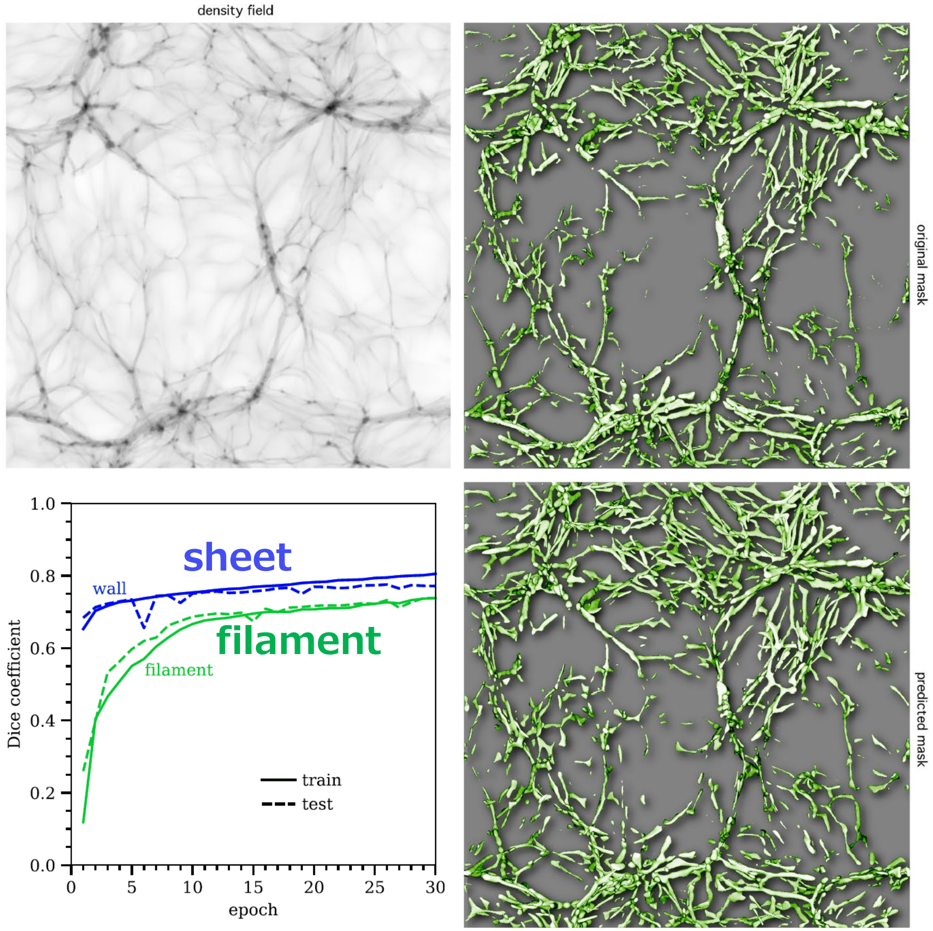
$$F_1 \equiv \frac{N_{\mathrm{tp}}}{N_{\mathrm{tp}} + \frac{1}{2}\left(N_{\mathrm{fp}} + N_{\mathrm{fn}}\right)},$$

Macro-averaged F1-score: 0.74

DM vs galaxies

- A similar previous study: **Aragon-Calvo (2019)**
  - U-Net
  - **Using DM density fields** for learning and labelling, rather than galaxies
  - **Binary classification**: filament and sheet

  - F1-score ~ 0.7-0.8



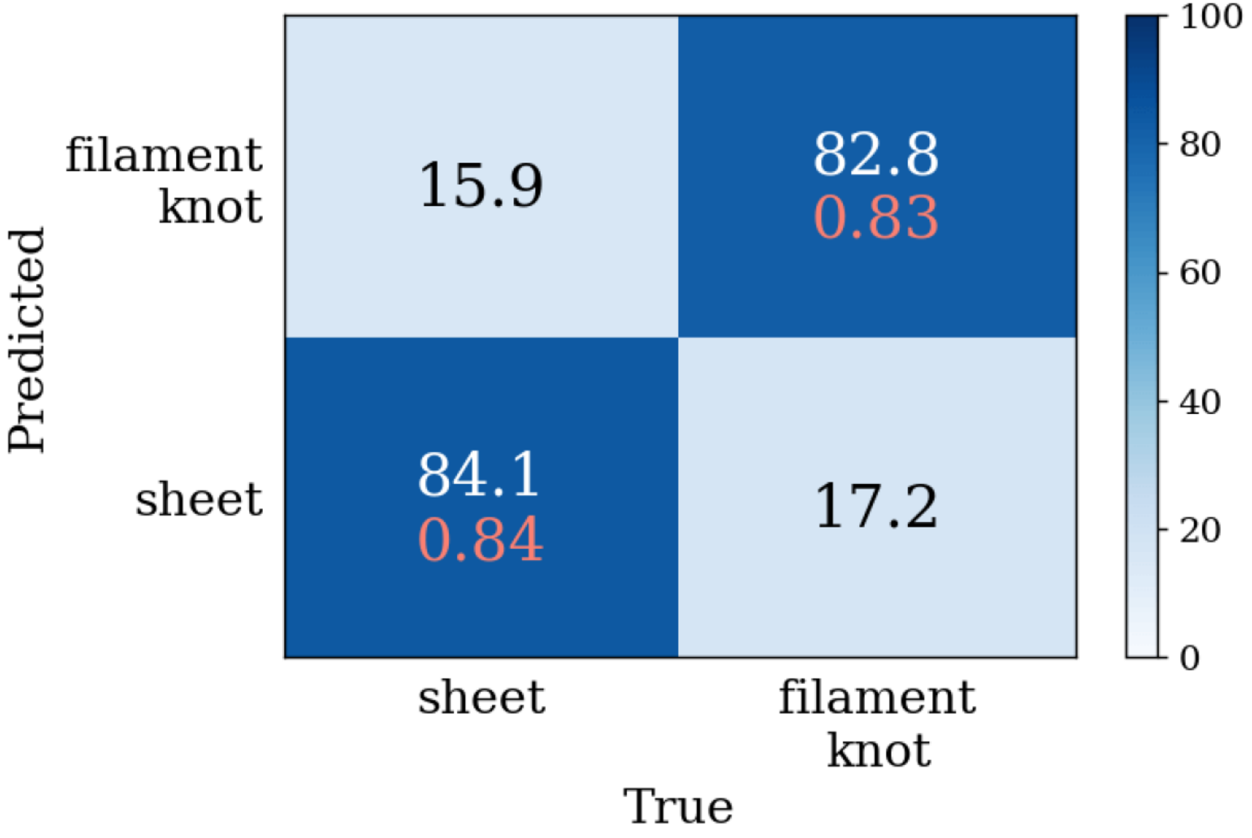
DM vs galaxies

- Our model learns galaxy distribution, rather than DM density fields.

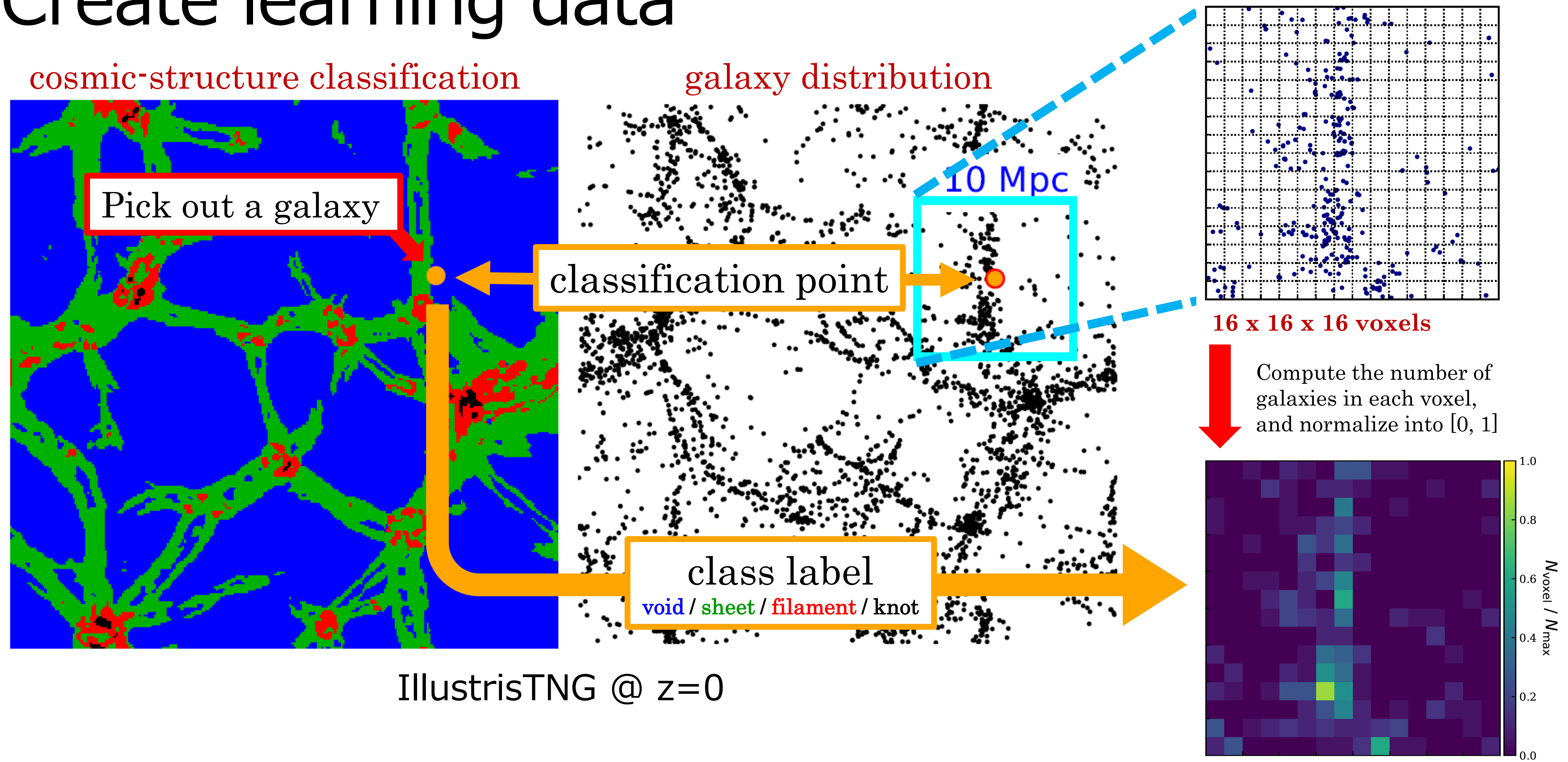
- Our model is as accurate as Aragon-Calvo (2019)

- **Galaxy distribution can be a substitution for DM density**

- **Galaxies are observable.**
- **The classification can be preformed with galaxy observations!!**



Create learning data



- We create 10000 cubic data for each class
    - 6400, 1600 and 2000 are used as training, validation and test data
  - The data have only a single channel of number distribution of galaxies

Observational restriction

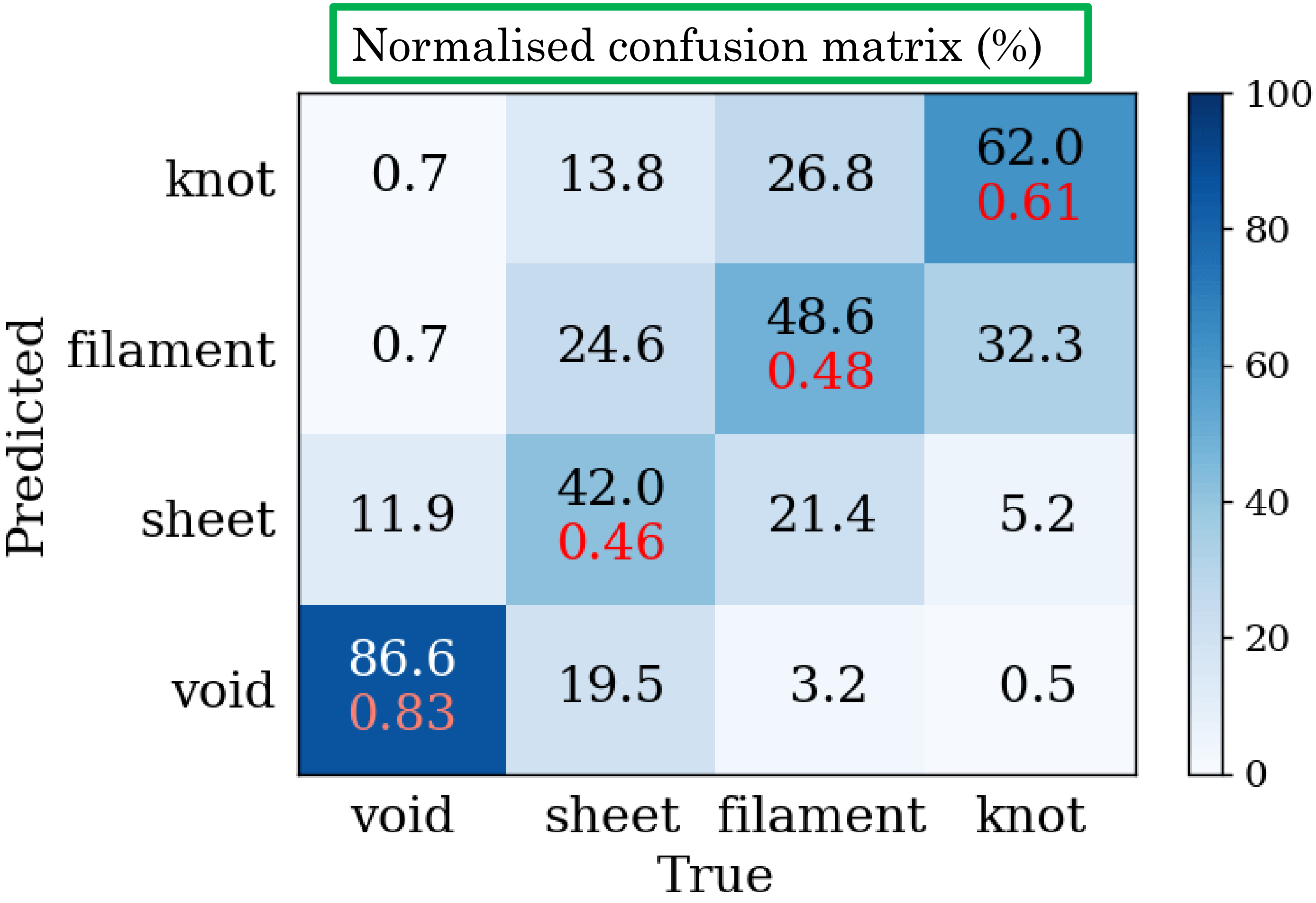
- **Limiting magnitude**
  - r-band magnitude for SDSS spectroscopy
    - $m_r = 17.75$ mag → $M_r = -17.25$ mag

      Assuming d=100 Mpc
    - exclude galaxies fainter than the limit from the simulation data

- **Distance measurement error**
  - Distance (line-of-sight position) is measured from spectroscopic redshift
    - affected by proper motion of a galaxy

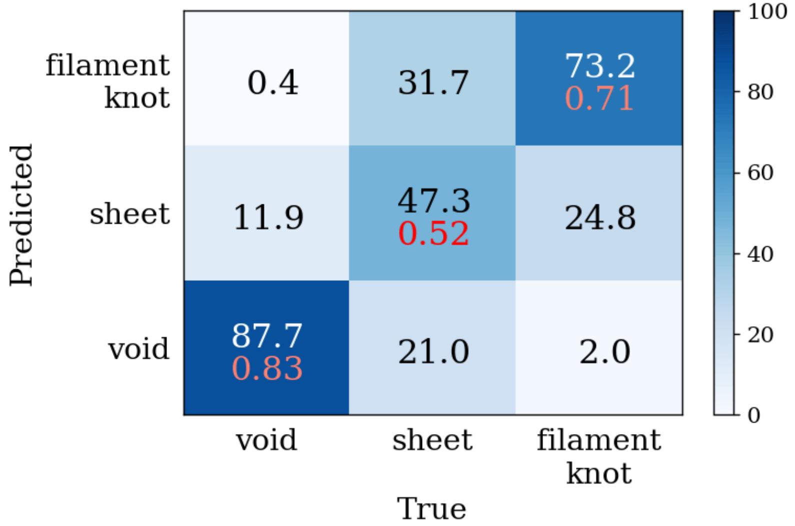
$$x_{\mathrm{obs}} = x_{\mathrm{true}} + \frac{v_{\mathrm{los}}}{H_0},$$

Classification for mock SDSS



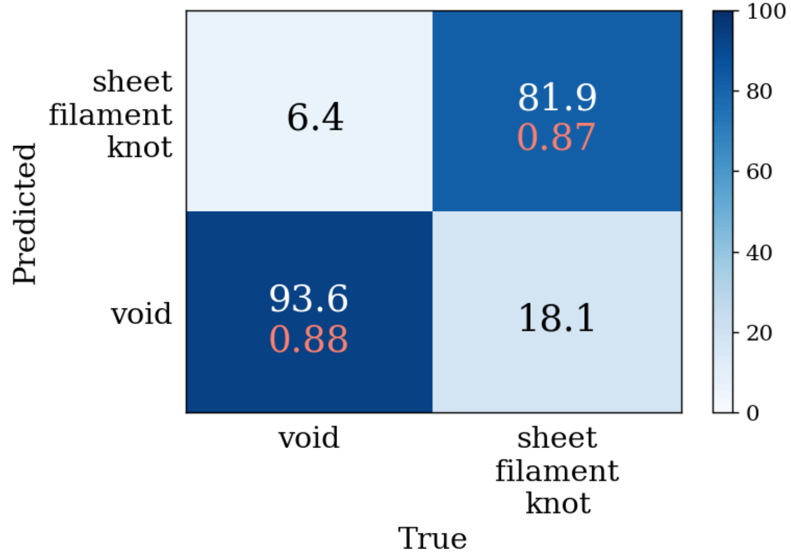
Macro-averaged F1-score: 0.60

Knot is merged with filament



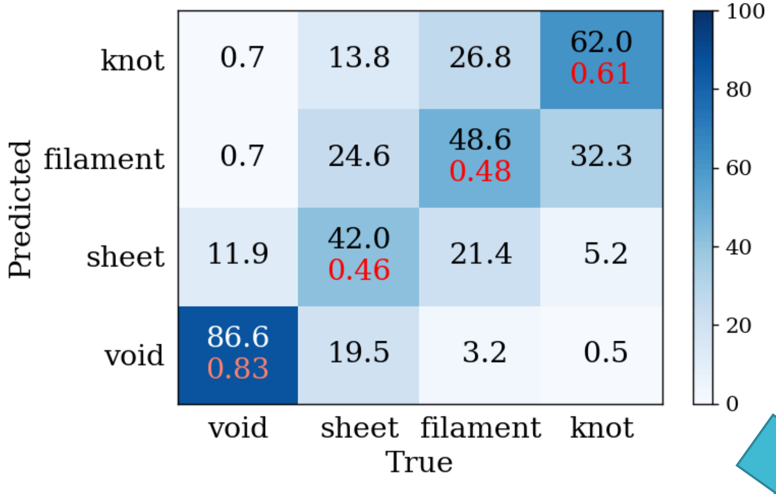
Macro-averaged F1-score: 0.68

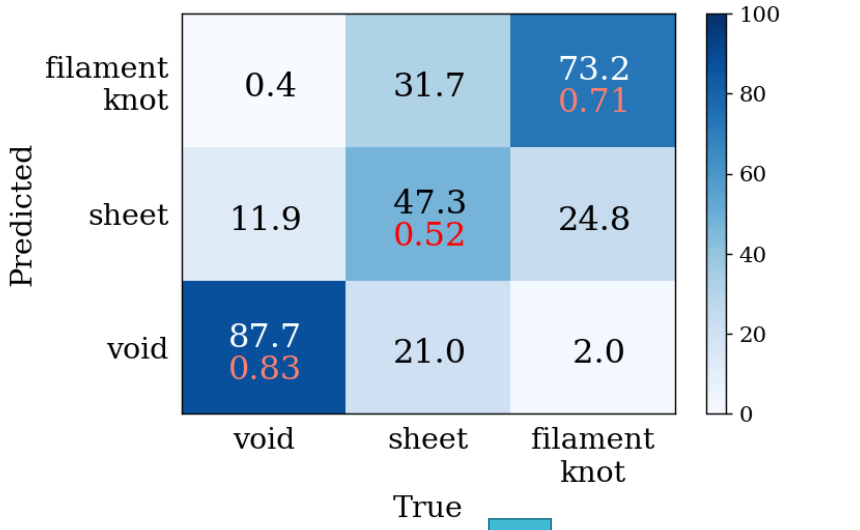
Knot and filament are merged with sheet

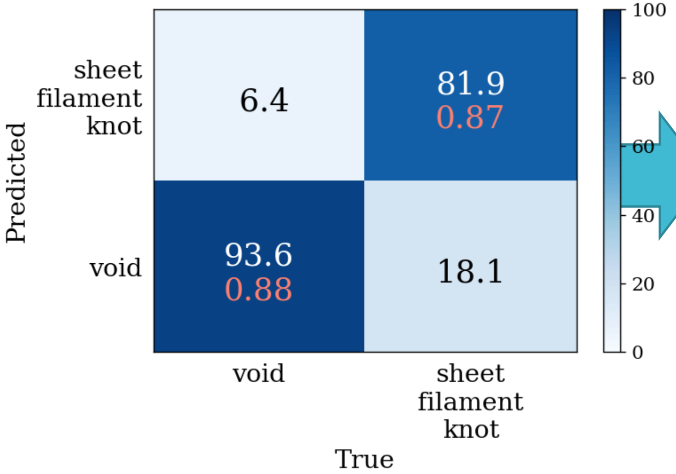


Macro-averaged F1-score: 0.88

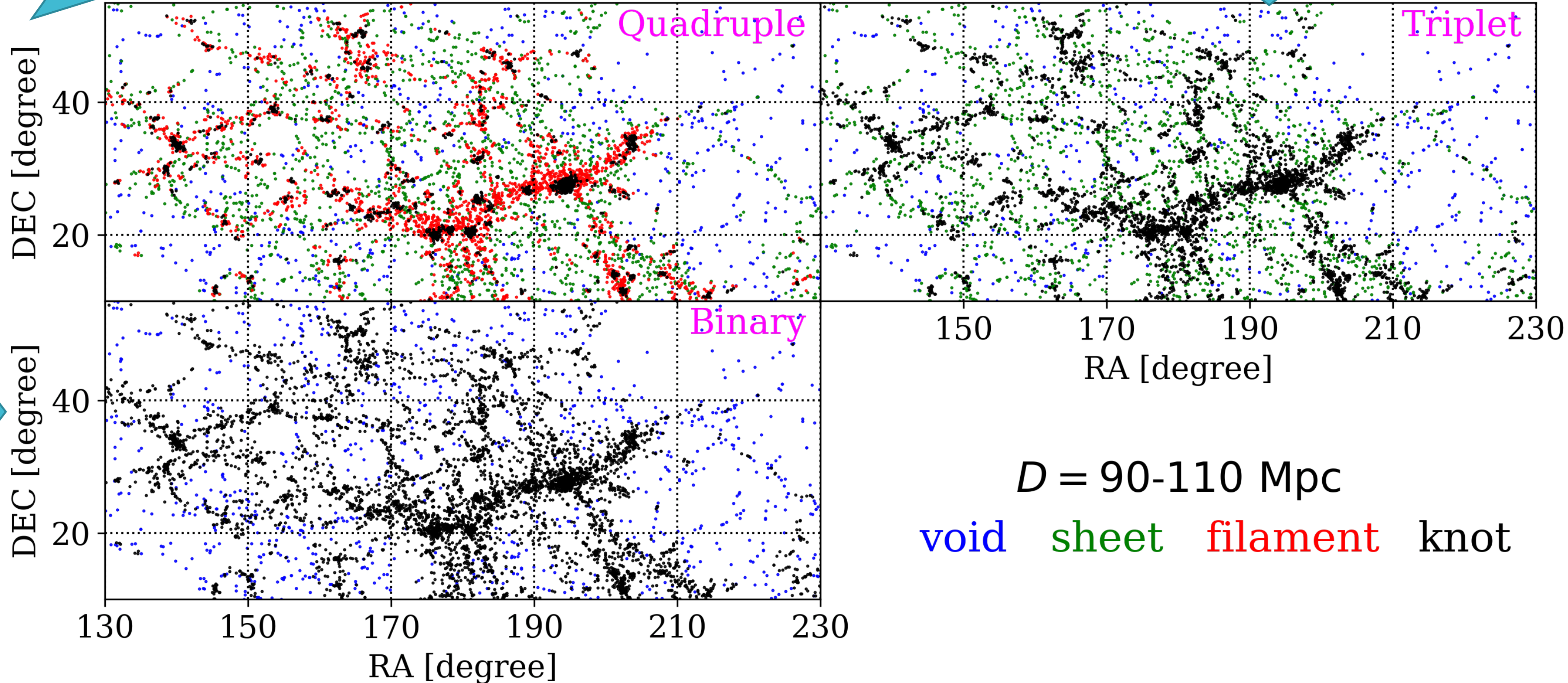
Applying to SDSS data







SDSS DR12



$D = 90\text{-}110$ Mpc

void sheet filament knot

Summary

- We explore the ability of 3D-CNN based on galaxies for the cosmic-structure classification.
- The class labels are obtained from DM distribution.

- Our models using galaxy distribution are as accurate as that using DM density fields.
  - **Galaxy number density can be a substitution for DM density fields.**

- **For classifying spatial grid points, our model can achieve the accuracy ~0.74.**
- **For classifying galaxies, without observational restriction, the accuracy is ~0.64.**
- **For classifying galaxies in "mock" SDSS, the accuracy is ~0.60.**

- It is the most difficult to distinguish sheet and filament.
- Our binary-classification model can classify void galaxies with an accuracy ~0.9.

- Proper motion does not matter, but the limiting magnitude lowers the accuracy.

Discussion: to improve the performance

- Limiting magnitude can be mitigated in future observations
  - If we ignore the limiting magnitude, the performance is improved.

Without distance error
Without limiting magnitude

Predicted \ True	void	sheet	filament	knot
knot	0.0	9.1	31.6	67.5 0.64
filament	0.1	23.5	45.5 0.46	31.3
sheet	7.7	51.6 0.56	22.6	1.2
void	92.3 0.89	15.8	0.3	0.0

Macro-averaged F1-score: 0.64

With distance error
Without limiting magnitude

Predicted \ True	void	sheet	filament	knot
knot	0.4	13.1	27.0	65.2 0.63
filament	0.2	28.0	54.8 0.52	31.5
sheet	8.5	43.1 0.50	17.4	3.3
void	91.0 0.88	15.8	0.7	0.1

Macro-averaged F1-score: 0.64

With distance error
With limiting magnitude

Predicted \ True	void	sheet	filament	knot
knot	0.7	13.8	26.8	62.0 0.61
filament	0.7	24.6	48.6 0.48	32.3
sheet	11.9	42.0 0.46	21.4	5.2
void	86.6 0.83	19.5	3.2	0.5

Macro-averaged F1-score: 0.60

- The distance errors by proper motions are unavoidable in observations.
  - However, the errors do not make the ML model inaccurate.