INAF
ISTITUTO NAZIONALE DI ASTROFISICA
OSSERVATORIO ASTROFISICO DI CATANIA

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

UNDERSTANDING GRAPH NEURAL NETWORKS (GNN)

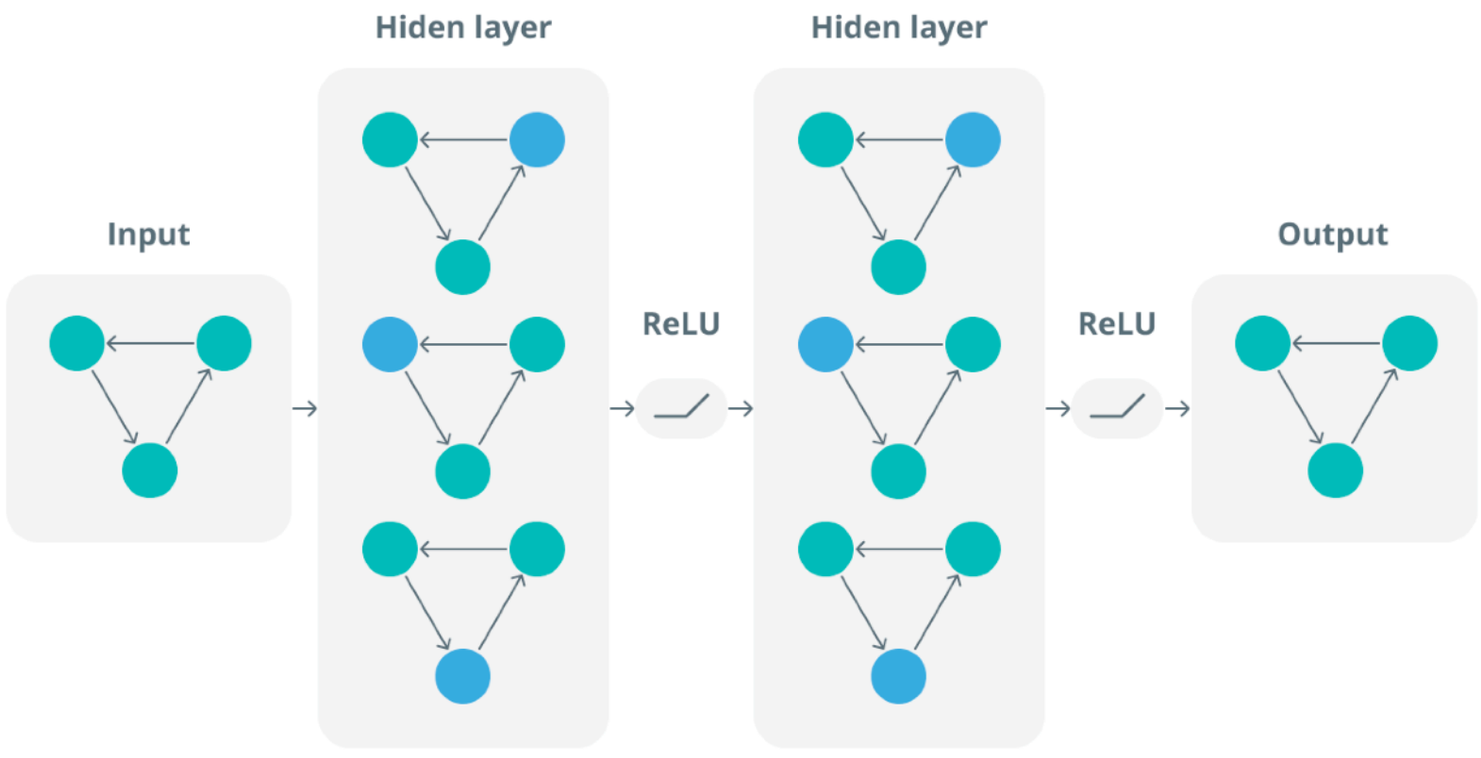
Farida Farsian

AI in Astronomy Workshop, Catania, 23/05/2025

Introduction to GNN

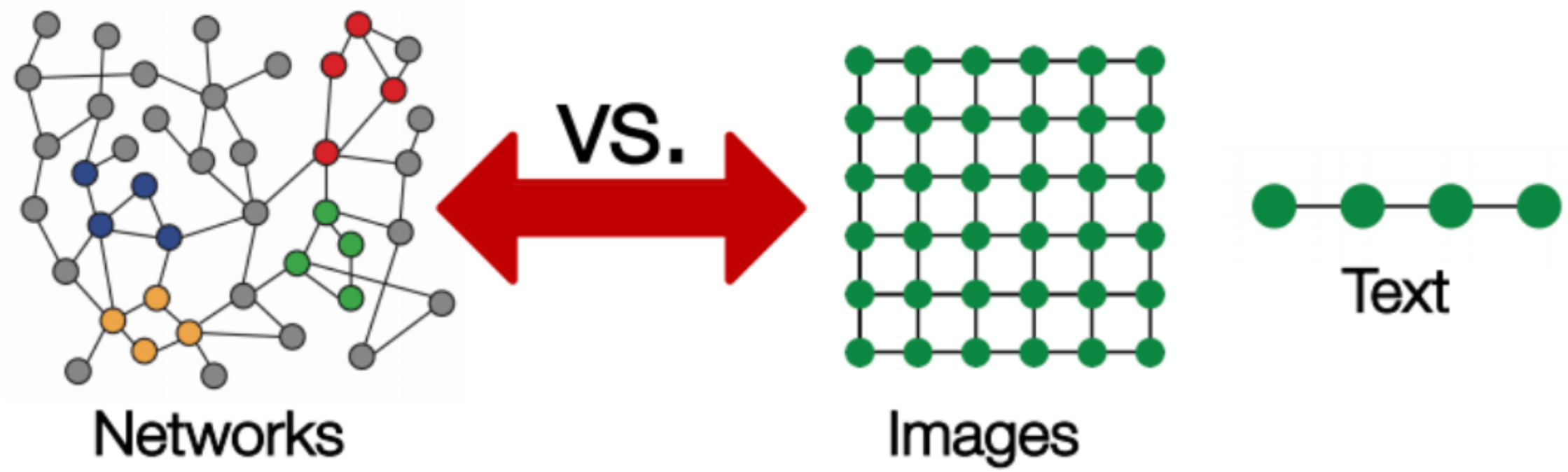
WHAT ARE GRAPH NEURAL NETWORKS

- GNNs are a class of neural networks designed to work with **graph-structured data**.
- A **graph** consists of **nodes** (entities) and **edges** (relationships).
- GNNs learn representations for nodes, edges, or whole graphs.
- They generalize traditional deep learning beyond regular data formats like images or sequences.
- GNNs capture both **features** and **relationships**, making them ideal for complex systems.



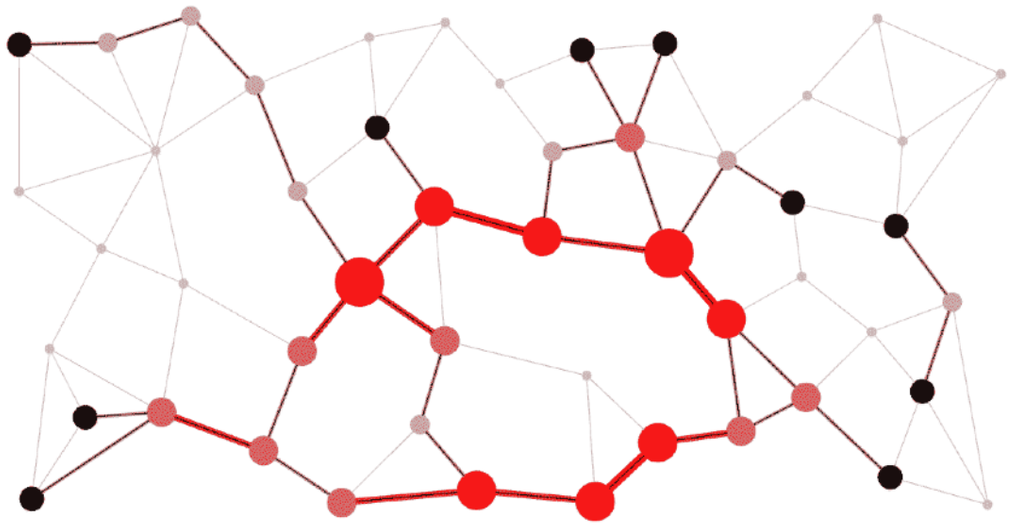
WHY GRAPH NEURAL NETWORKS

- Many real-world data types are best modeled as graphs: molecules, social networks, citations, cosmic structures.
- Traditional neural networks struggle with **non-Euclidean** structures and arbitrary connectivity.
- GNNs enable learning directly on graphs, maintaining the topology and interactions.
- They reduce the need for manual feature engineering in relational domains.
- Enable better generalization in problems with limited data by leveraging graph structure.



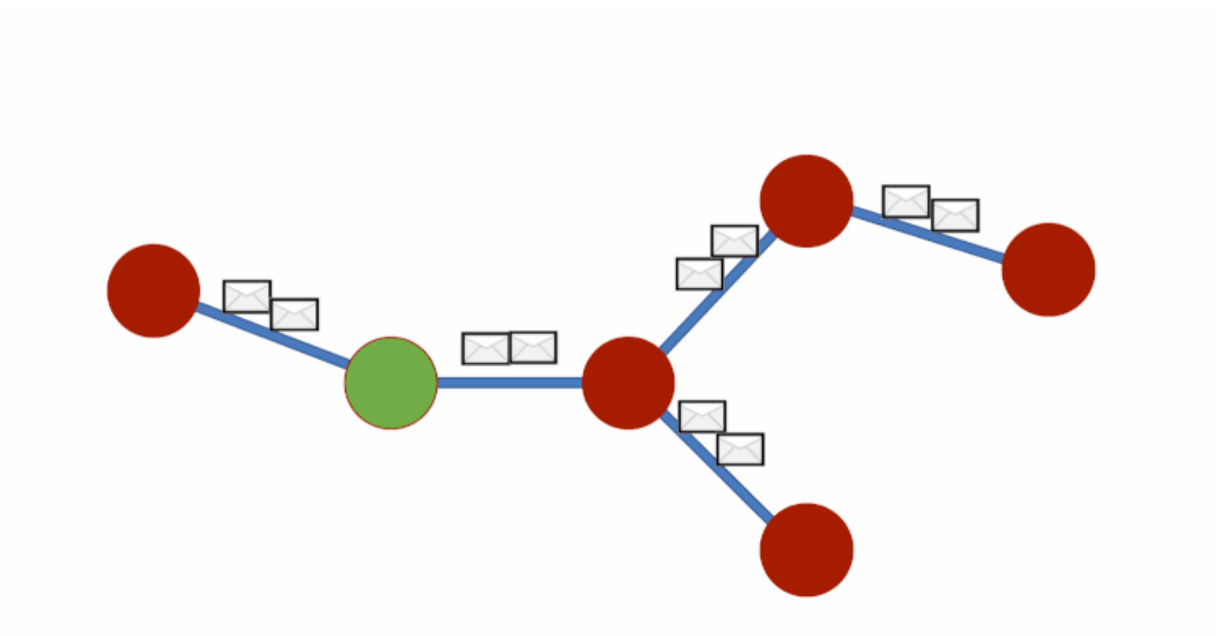
HOW DO GNNS WORK?

- GNNs iteratively **propagate information** between nodes in a graph.
- Each node updates its state based on its **neighbours' features**.
- These updates typically combine **aggregation + transformation**.
- After several layers, a node has access to **multi-hop** neighbourhood information.
- The final embeddings can be used for tasks like **classification**, **regression**, or **clustering**.



MESSAGE PASSING SCHEMA

1. For each node in the graph, all the neighbouring node messages are gathered;
2. Then, all messages are aggregated via an aggregate function (like sum).
3. Lastly, all pooled messages are passed through an update function, usually a learned neural network

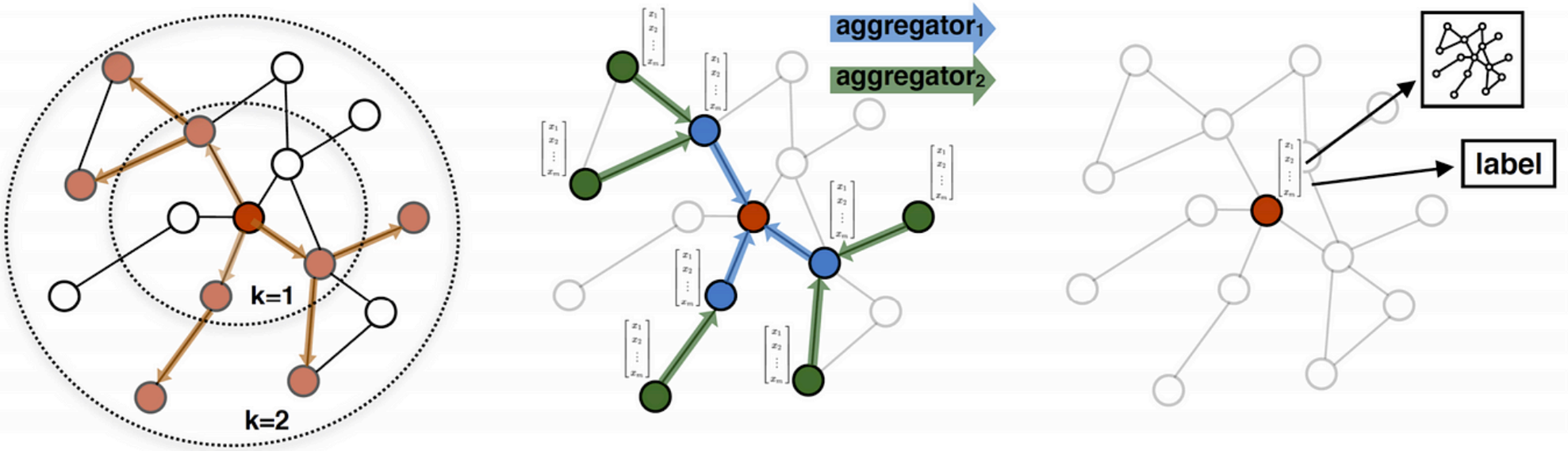


MATHEMATICAL OVERVIEW

- Let $h_v^{(k)}$ be the feature vector of node v at layer k:

$$h_v^{(k)} = \text{UPDATE}^{(k)}\left(h_v^{(k-1)}, \text{AGGREGATE}^{(k)}\left(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}\right)\right)$$

- Aggregation ensures permutation invariance over neighbours.
- Update functions often include linear layers and non-linearities (ReLU, tanh).
- The graph structure is encoded via an **adjacency matrix or edge list**.
- Final output depends on the task: node-level, edge-level, or graph-level prediction.



1. Sample neighborhood

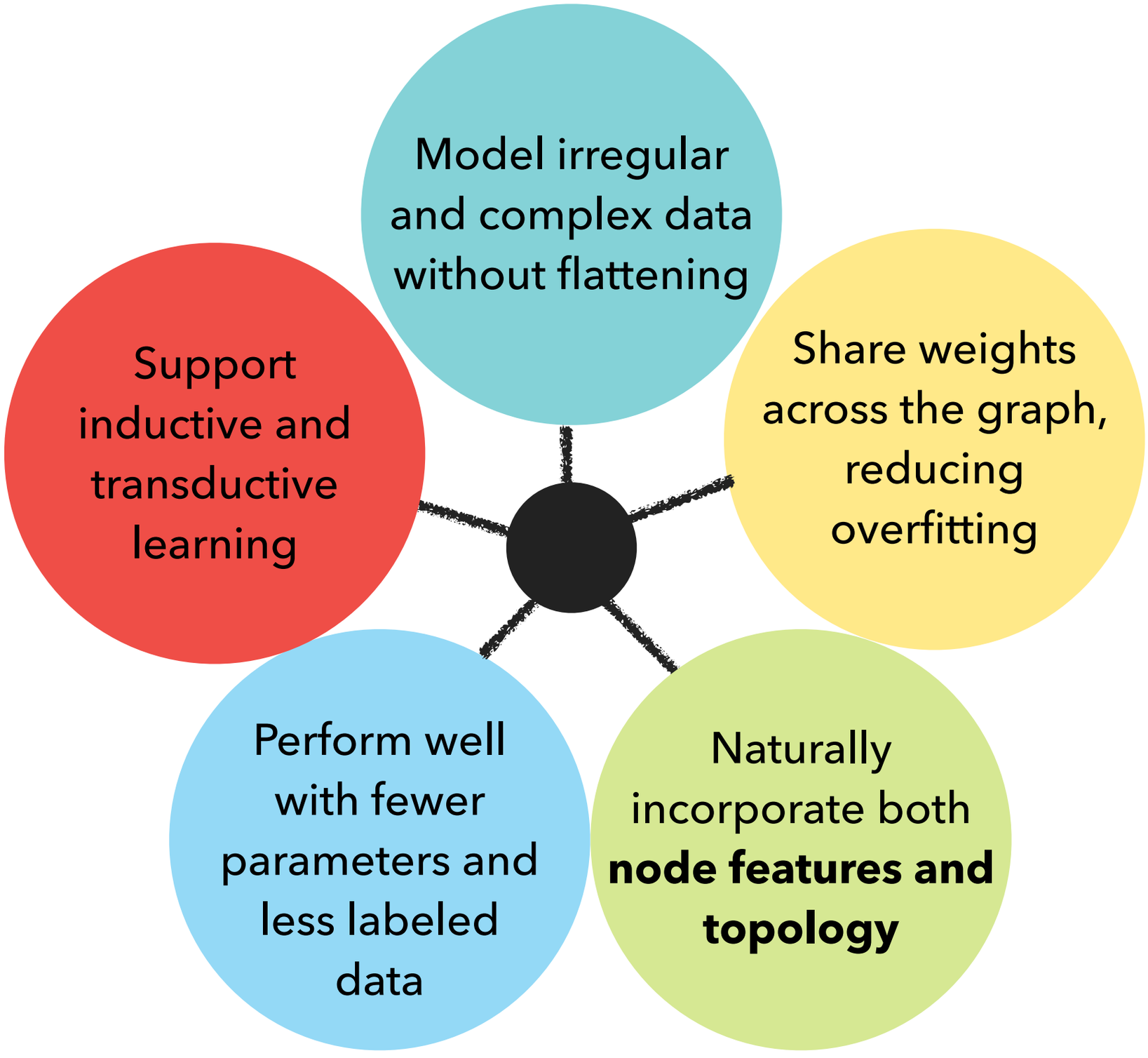
2. Aggregate feature information from neighbors

3. Predict graph context and label using aggregated information

DIFFERENT TYPES OF GNNS

- **GCN (Graph Convolutional Network)**: performs convolution on graphs by approximating spectral filtering, enabling nodes to aggregate information from their neighbours
- **GraphSAGE**: Samples neighbours and learns aggregation functions for scalability.
- **GAT (Graph Attention Network)**: Assigns different weights to neighbours via attention.
- **GIN (Graph Isomorphism Network)**: Maximizes discriminative power over graph structures.
- **MPNN (Message Passing Neural Network)**: General framework for molecules and physics with customizable message/update functions.

ADVANTAGES OF GNNS



GNN VS MLP

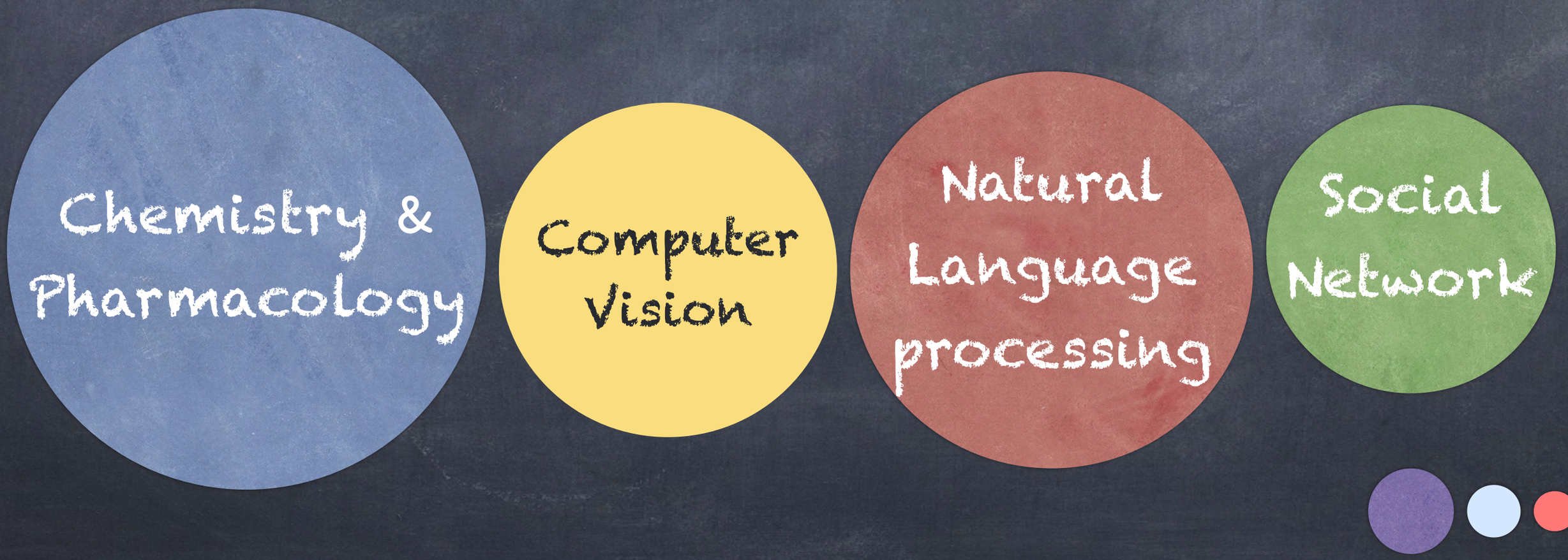
Aspect	MLP	GNN
Input Format	Vectors	Graphs (nodes + edges)
Relational Modeling	None	Encodes relationships via edges
Feature Interaction	Only local input features	Neighborhood features shared
Parameter Sharing	Across layers only	Shared across all nodes and layers
Use Case	Tabular data	Social networks, molecules, etc.

GNN VS CNN

Aspect	CNN	GNN
Input Format	Regular grids (images)	Arbitrary graphs
Neighborhood	Fixed-size, local	Variable-size, graph-defined
Connectivity	Spatial adjacency	Topological adjacency
Feature Aggregation	Convolutions over pixels	Aggregation over neighbors
Applications	Vision, speech	Molecules, social networks, physics

APPLICATION OF GNNS

- **Node classification**: Labeling users, articles, or web pages.
- **Link prediction**: Recommender systems, fraud detection.
- **Graph classification**: Molecule activity, protein function.
- Clustering and community detection in networks.
- **Scientific discovery**: Modeling dynamics or symbolic laws.



APPLICATION OF GNNS

Chemistry & Pharmacology

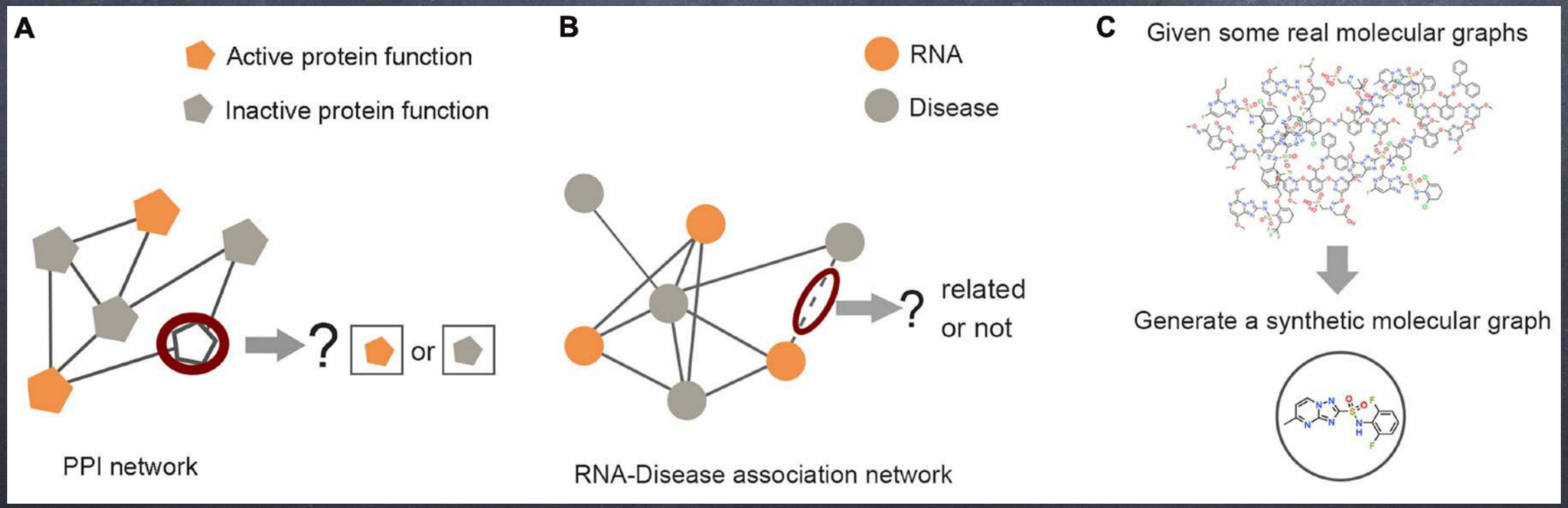
Molecules are naturally graphs: atoms (nodes), bonds (edges).

Predict properties like solubility, reactivity, toxicity.

Drug discovery: Screen chemical libraries efficiently.

Outperform traditional methods.

GNNs support zero-shot generalization to unseen compounds.



Zhang et al 2021

APPLICATION OF GNNS

Computer Vision

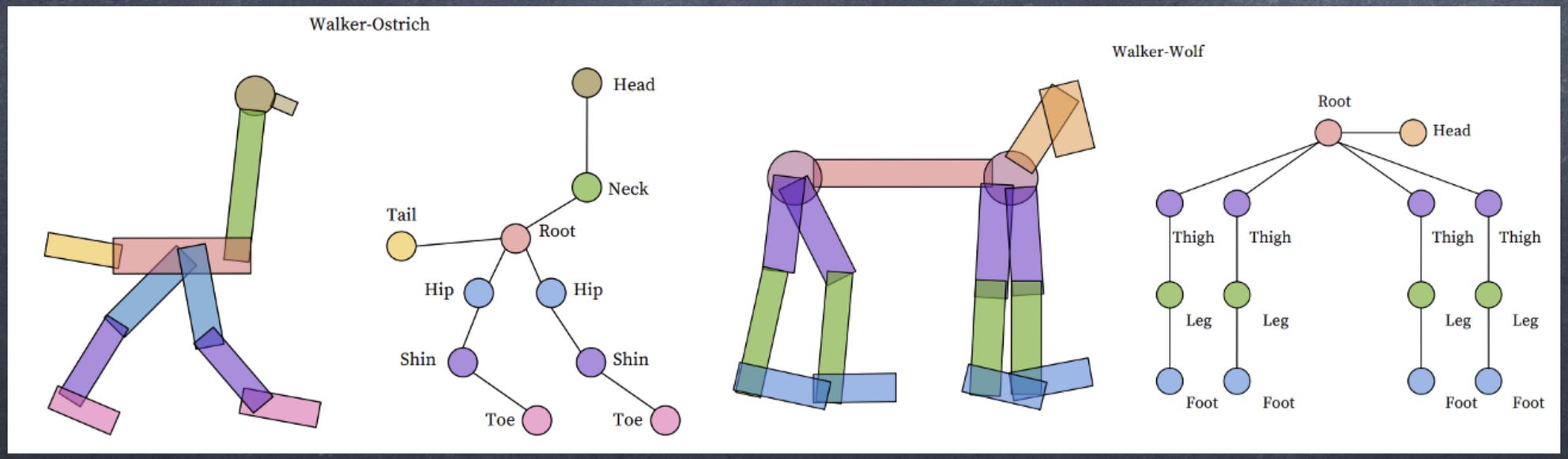
Scene graphs: Represent objects and their relationships in an image.

Skeleton-based action recognition using graph structures of human joints.

Instance segmentation with relational reasoning.

3D object recognition and point cloud processing.

Combining CNNs with GNNs for better **contextual understanding**.



A graph representation of animal skeletons (from Wang et al 2017)

APPLICATION OF GNNS

Social Network

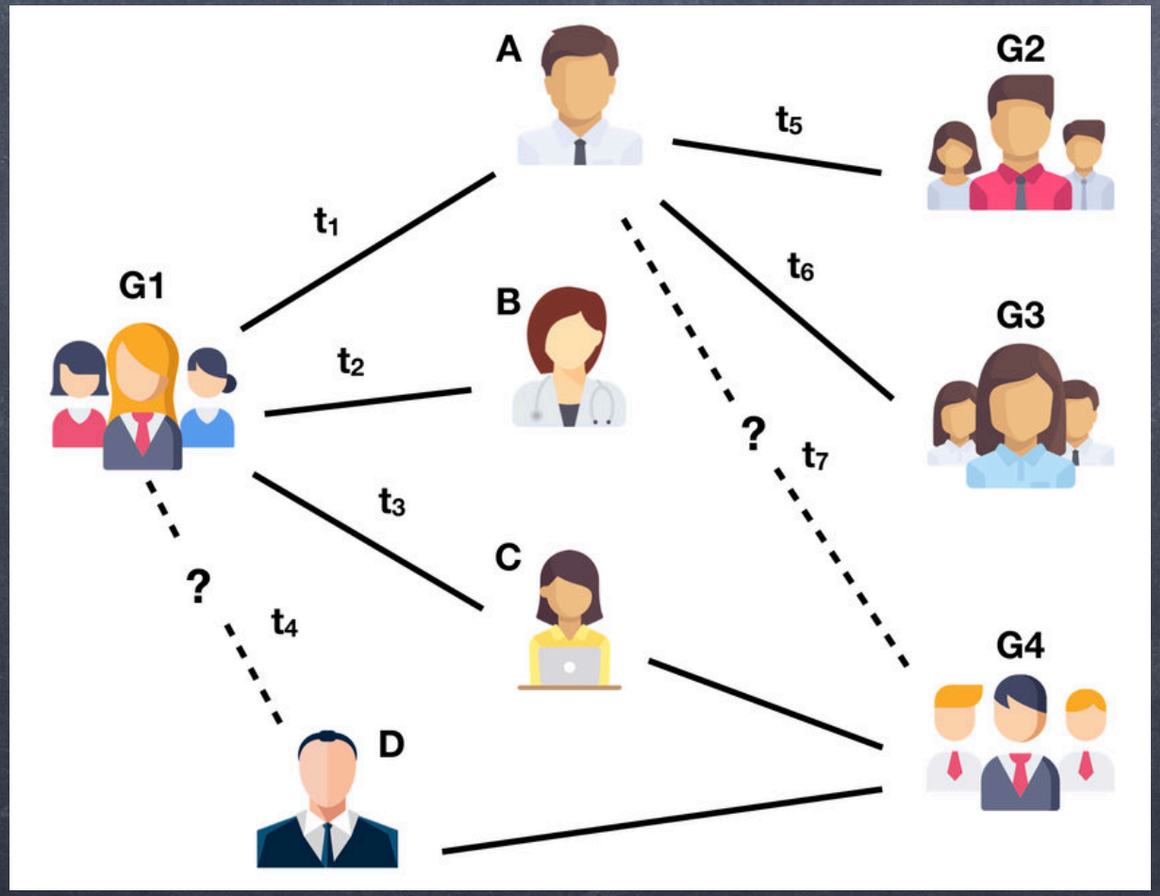
User classification (e.g., spam detection, interest prediction).

Community detection by learning from graph structure.

Friend suggestion via link prediction.

Information diffusion modeling (e.g., viral content spread).

Fake news detection through graph-based feature propagation.

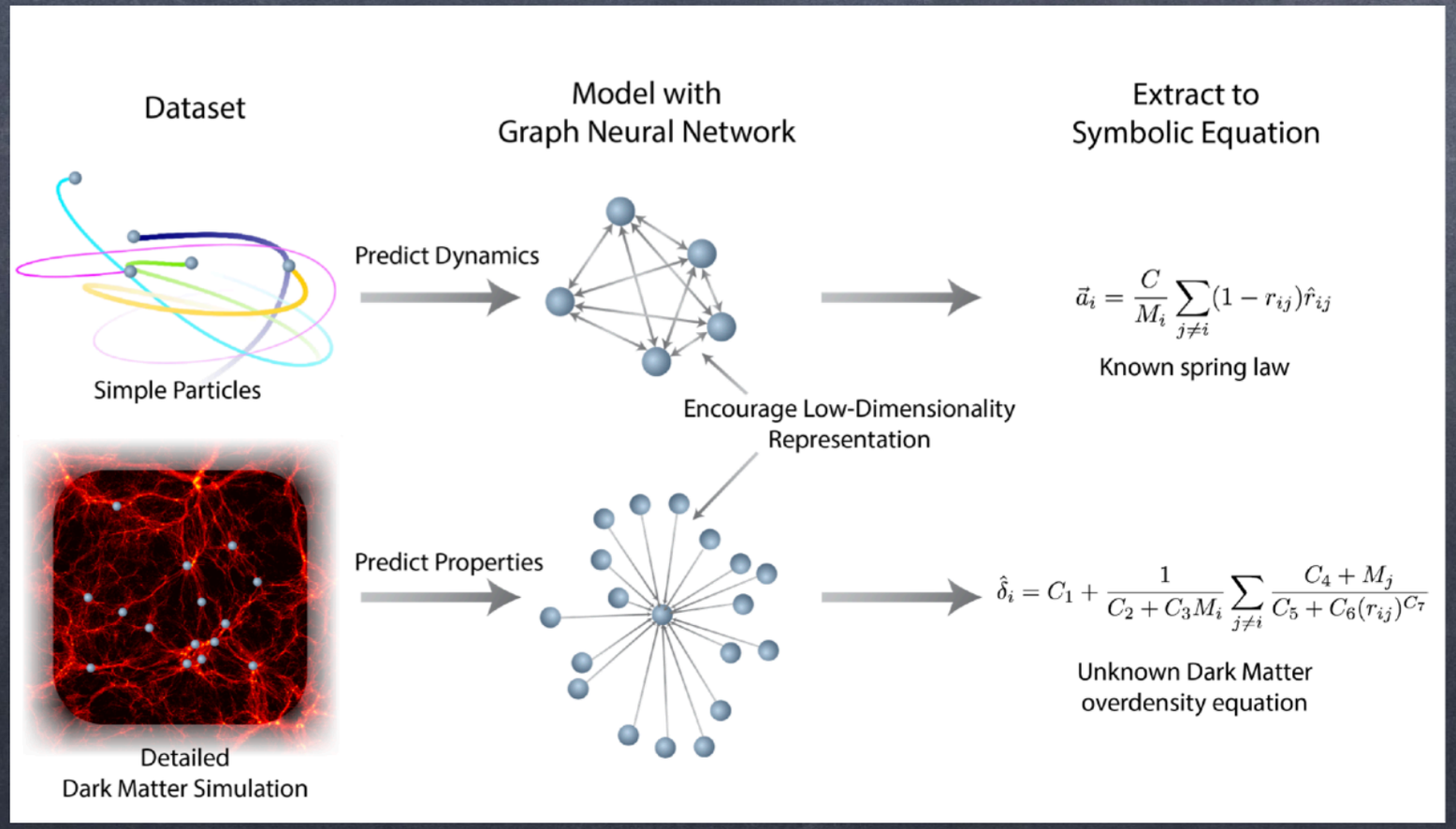


Group Link Prediction Sha et al, 2021

GNNS IN COSMOLOGY

Discovering Symbolic Models from Deep Learning

- GNN trained to model **particle interactions** in simulations.
- After training, symbolic regression distills learned dynamics.
- Recovered symbolic expressions match known physical laws.
- Approach blends deep learning with **interpretable science**.

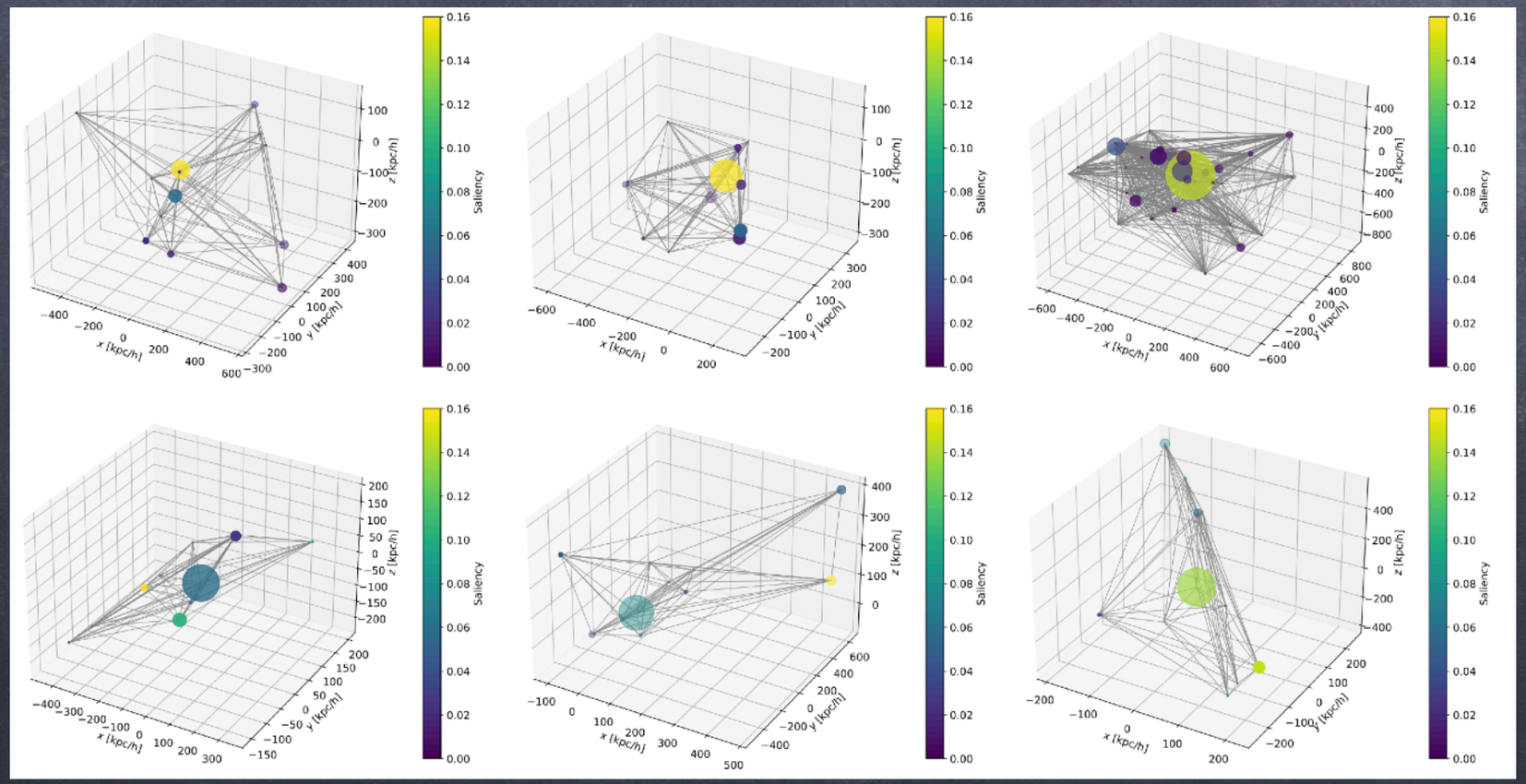


Cranmer et al. 2020

GNNS IN COSMOLOGY

Inferring Halo Masses with GNNs

- Task: Predict mass of dark matter halos using halo graphs.
- Graphs built using nearest-neighbor relationships in 3D space.
- GNNs outperform MLPs by capturing **local cosmic environment**.
- Input includes positions, velocities, and other halo features.
- Shows that GNNs can extract **cosmological information** from structure.



Villanueva-Domingo et al. 2022

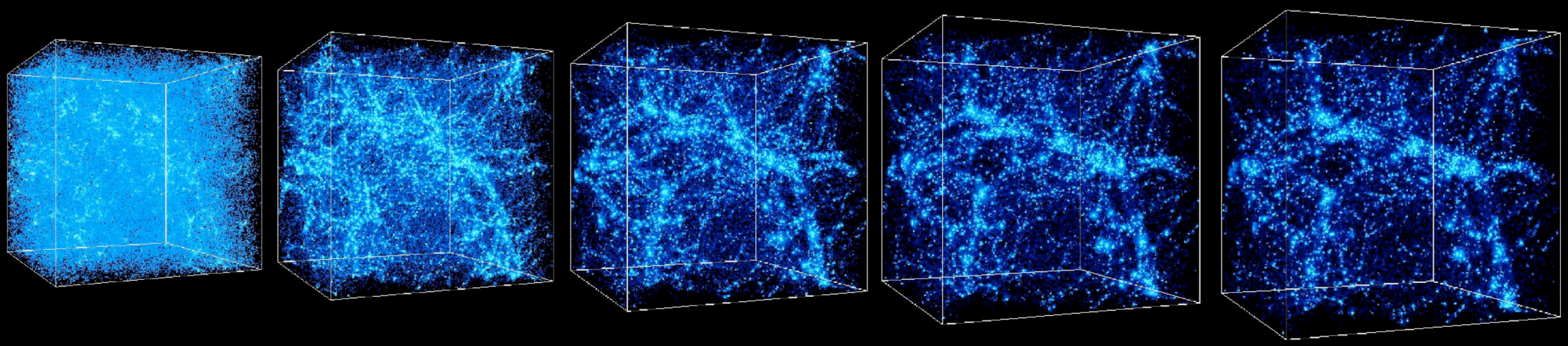
SUMMARY AND TAKE AWAYS

- **GNNs** extend deep learning to graph-structured data, enabling powerful modeling of systems where relationships and topology matter.
- They operate by **message passing**, where each node iteratively aggregates information from its neighbors, allowing the model to learn from both **node features** and the **structure of the graph.**
- Compared to traditional neural networks like MLPs and CNNs, GNNs handle **irregular, non-Euclidean data**, offering more expressive power for structured problems.
- GNNs promote **parameter sharing**, require **fewer features to be manually engineered**, and generalize better when data is sparse or noisy — especially when leveraging **relational inductive biases**.
- Despite their power, challenges remain in terms of **scalability** to large graphs, **interpretability** of learned representations, and **training stability**.

An example

SOME WORDS ABOUT LSS

- Large Scale Structure (LSS) is a network of filaments of dark matter, to form the skeleton of the universe, known as cosmic web
- These structures are often described by a matter density field, or by its statistical properties through the matter power spectrum.

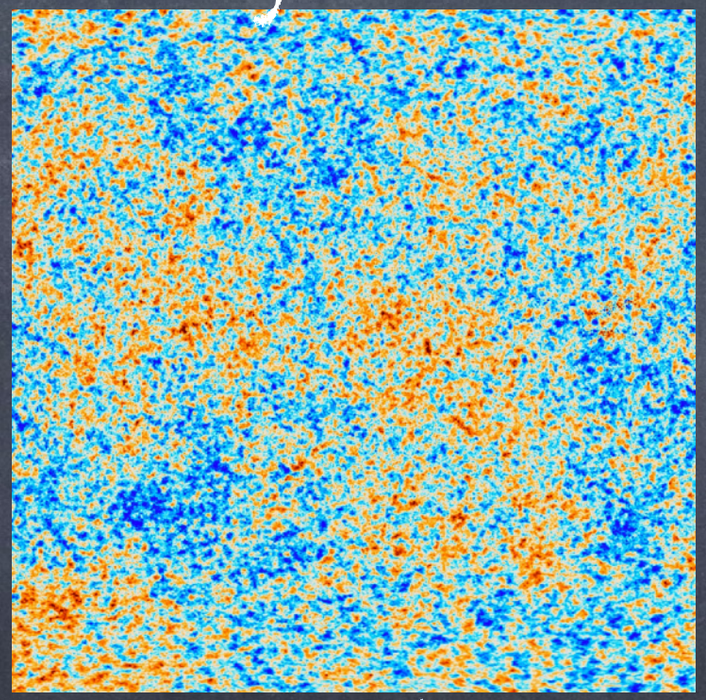


by Andrey Kravtsov

PROBES OF LARGE SCALE STRUCTURE

TWO POINT CORRELATION FUNCTION (2PCT) AND POWER SPECTRUM

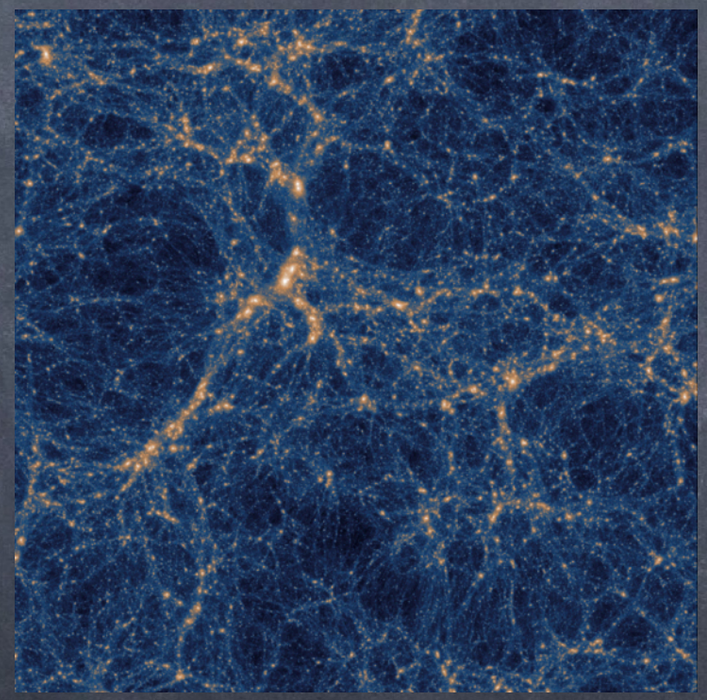
Early Universe



Planck 2018

Gravitational Collapse

Late time



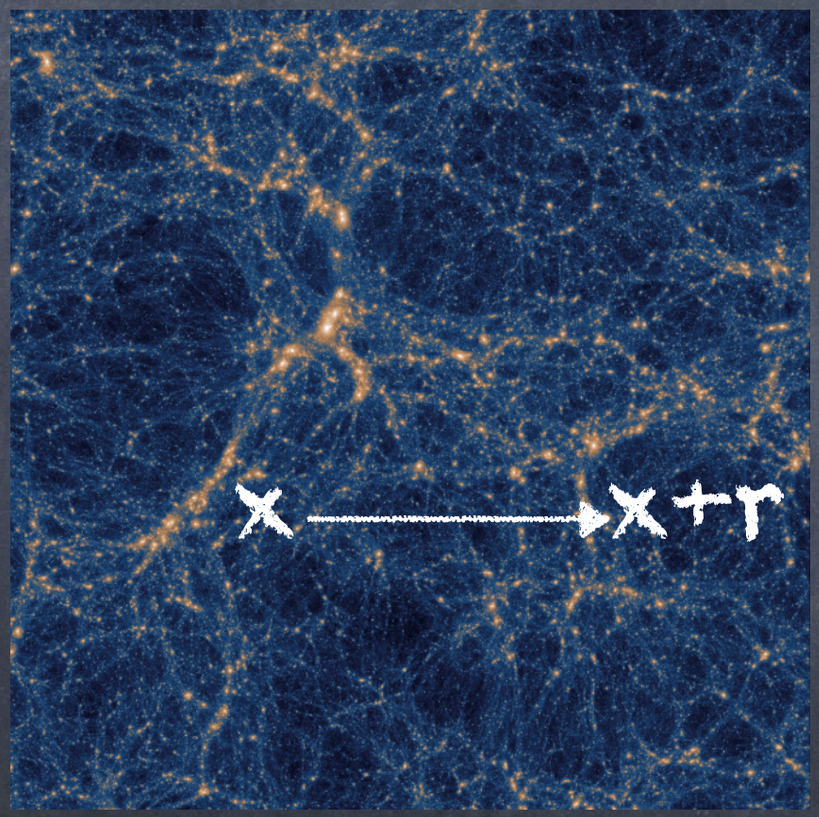
TNG300

$$\delta(\mathbf{k}) \sim \mathcal{N}(0, P(\mathbf{k}))$$

$$\delta(\mathbf{k}) \nsim \mathcal{N}(0, P(\mathbf{k}))$$

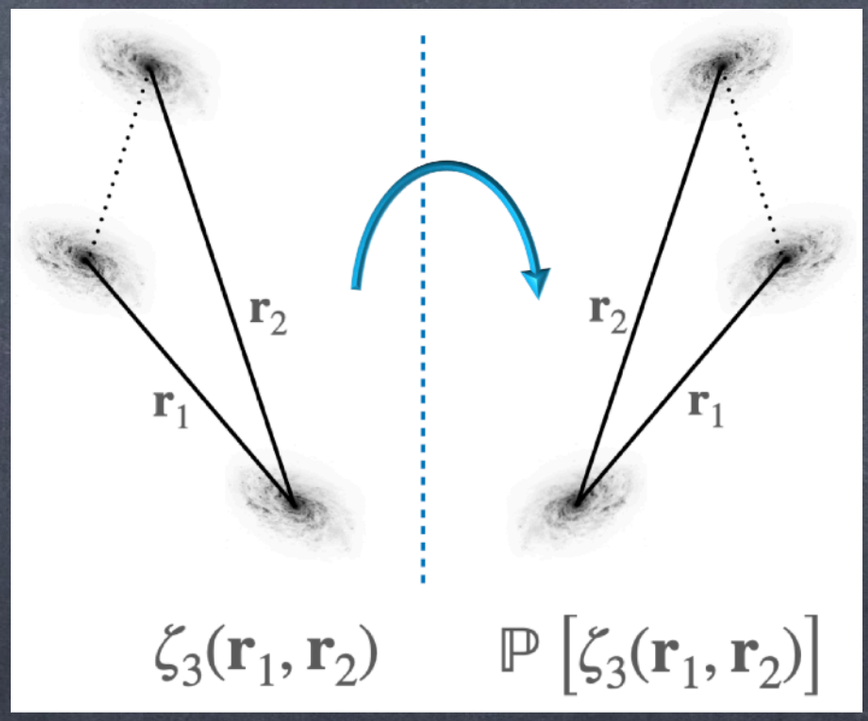
POWER SPECTRUM AND HIGHER ORDER STATISTICS

- Not all the information contained in the power spectrum (bispectrum statistics is needed)



TNG300

- 2PCT and 3PCT are not sensitive to parity, so moving to higher order statistics...



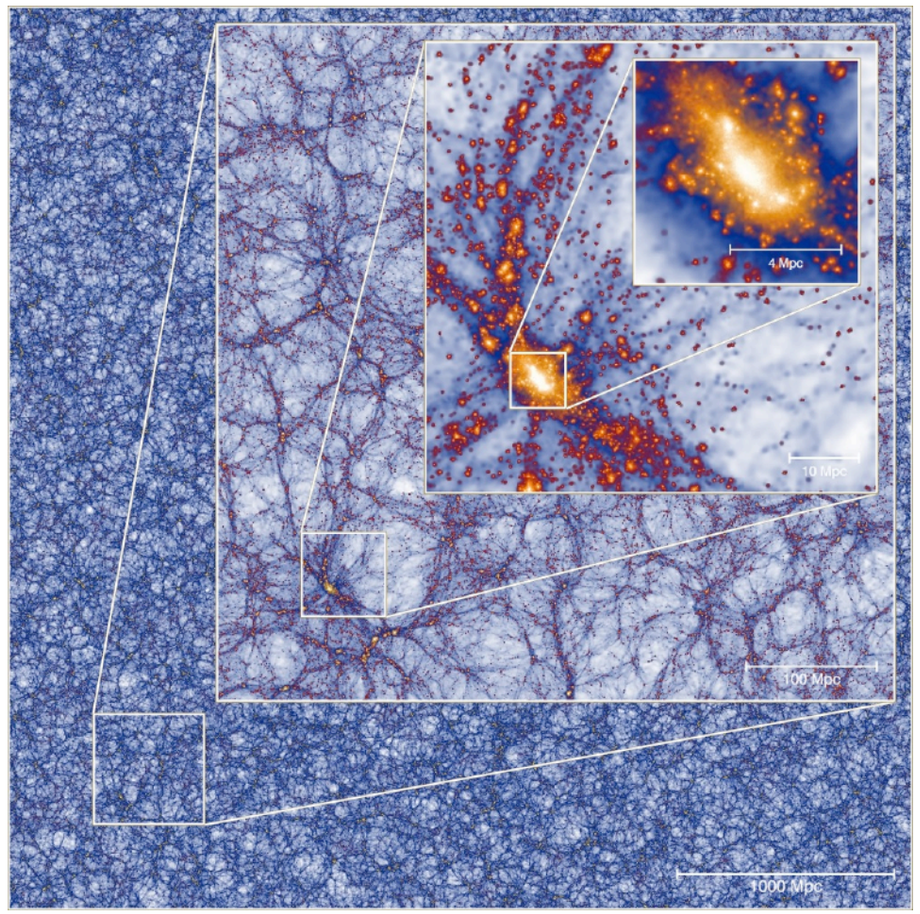
From Oliver Philox

MOTIVATION

- Standard cosmological analyses based on abundances, two-point and higher-order statistics, for extracting the information encoded in the Large Scale Structure (LSS), have been widely used up

- They can only exploit a sub-set of the whole information content available.

- The need of extracting maximum information from Dark Matter (DM) halo spatial distribution without using compressed statistics.



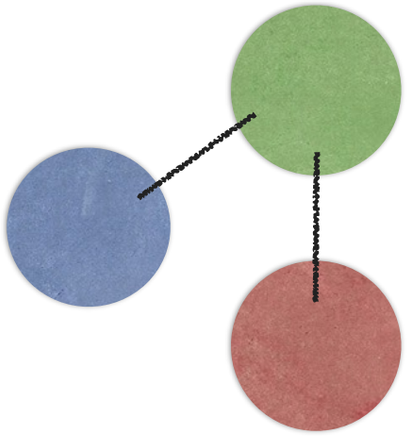
MIllenium XXL, A.Smith et al 2017

Question:
How to extract the whole information available without compressing it?

Possible Answer:
using the row data in the halo/galaxy catalogue

Challenge:
facing a sparse data, grasping larger scale statistics (the relations between halos/galaxies)

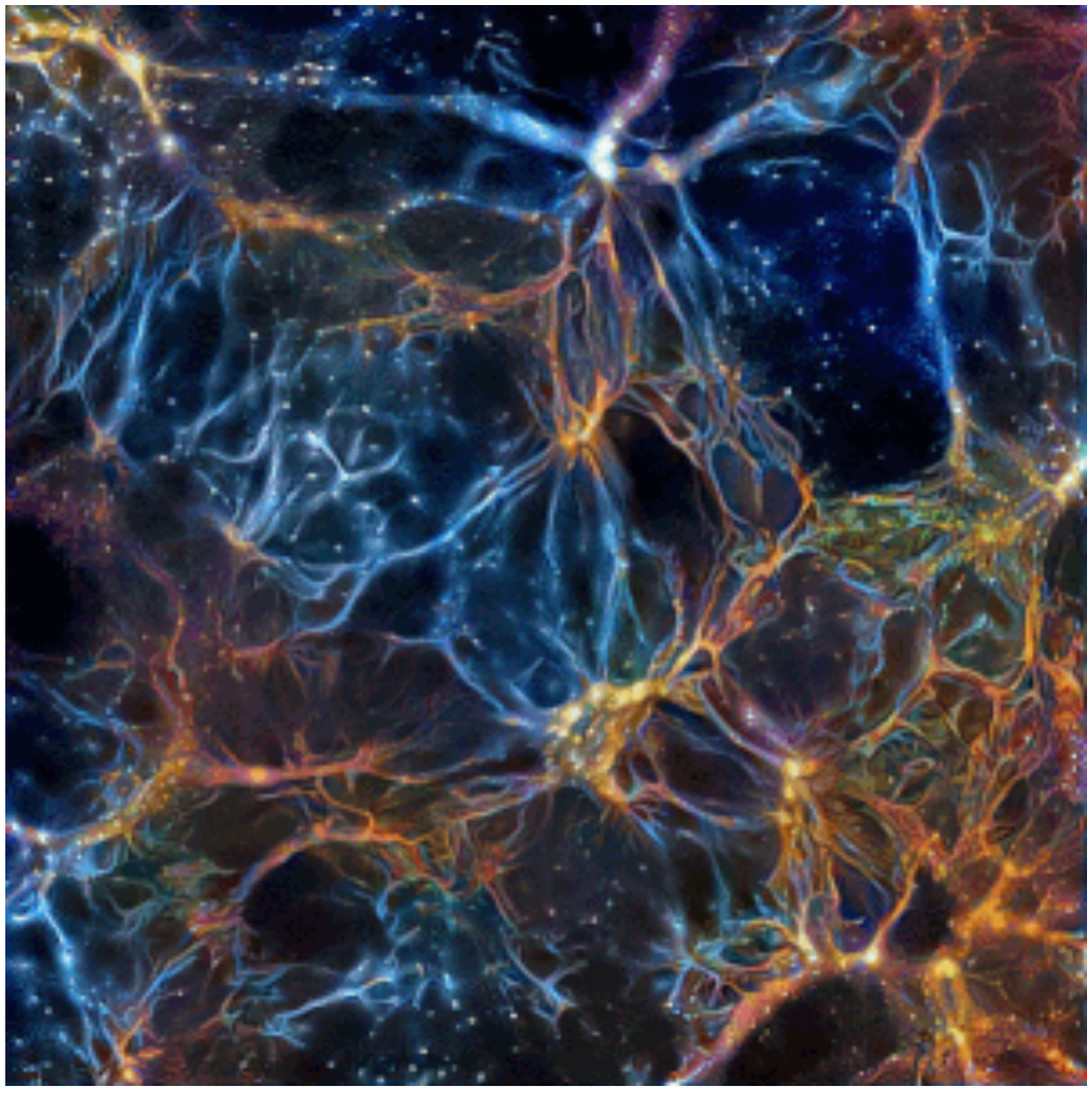
PROPOSED SOLUTION: GRAPHS



- Considering **row information** of DM field, such as mass and coordinates of halos.
- Representation of **cosmic web data in the form of graphs** contains the clustering information automatically.
- Using **Graph Neural Network** to capture the graph structure of data.
- Such method can be used to extract cosmological parameters in **likelihood-free** manner.

SIMULATION

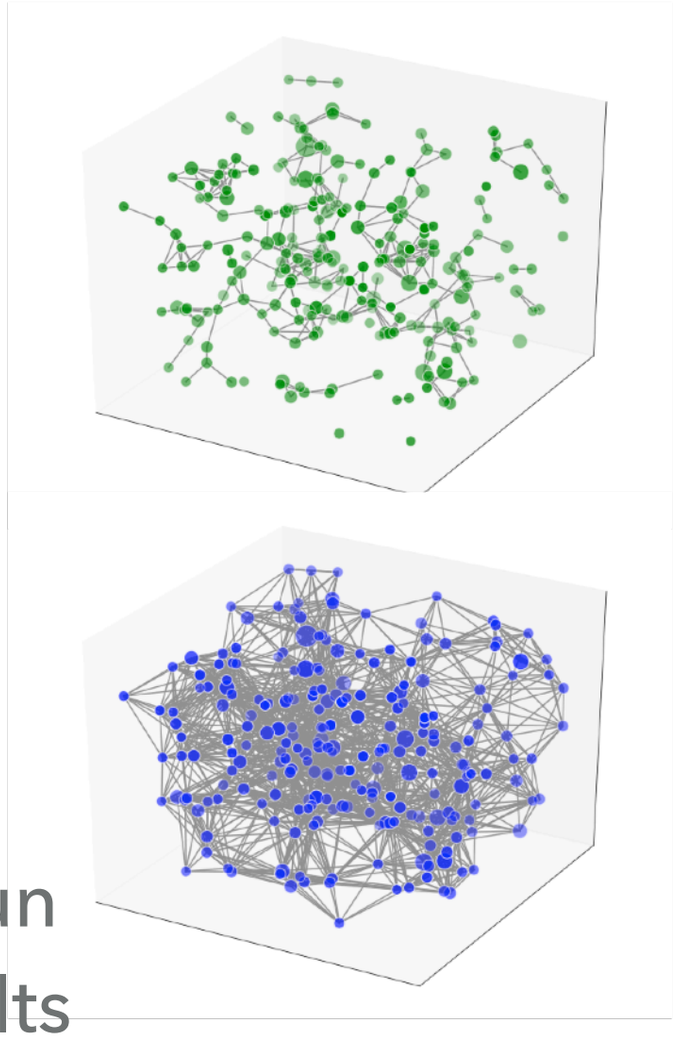
- Quijote simulation, full N-body, Boxes of 1 Gpc/h, more than 8.5 trillions of particles
- 500 realizations at redshift 0
- The DE parameter, **w0**, changes in the range of **[-1.05, -1, -0.95]**.



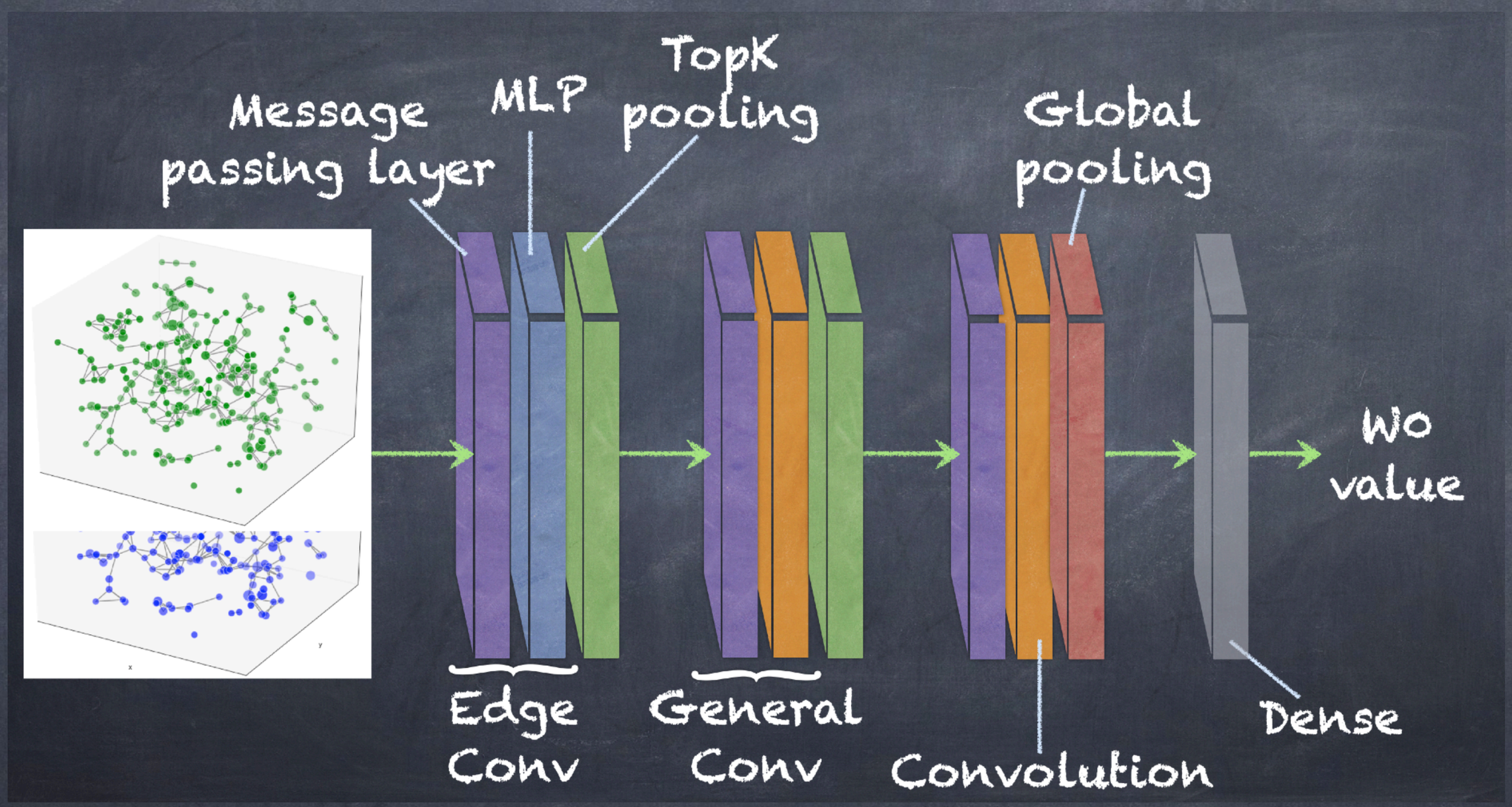
Quijote, arXiv:1909.05273 FVN et al 2020

DATA PREPRATION

- Only **mass and coordinates** of halos are give to the network as features.
- We have applied mass cut of 7×10^{14} for each catalogue.
- Two nodes i and j are connected by an **edge** if they are closer than a certain distance r.
- In our analysis r is a hyper-parameter and we run the analysis to find the optimum value the results presented here are based on **r = 100** Mpc/h.



NETWORK ARCHITECTURE



Yue Wang et al. 2018 Jiaxuan You et al. 2020

RESULTS: CLASSIFICATION

- **99%** of accuracy for **Binary classification,** to distinguish between w0 = -1.05 & w0 = -0.95.
- **97%** of accuracy for **Multi-class classification,** to distinguish between three values of w0.

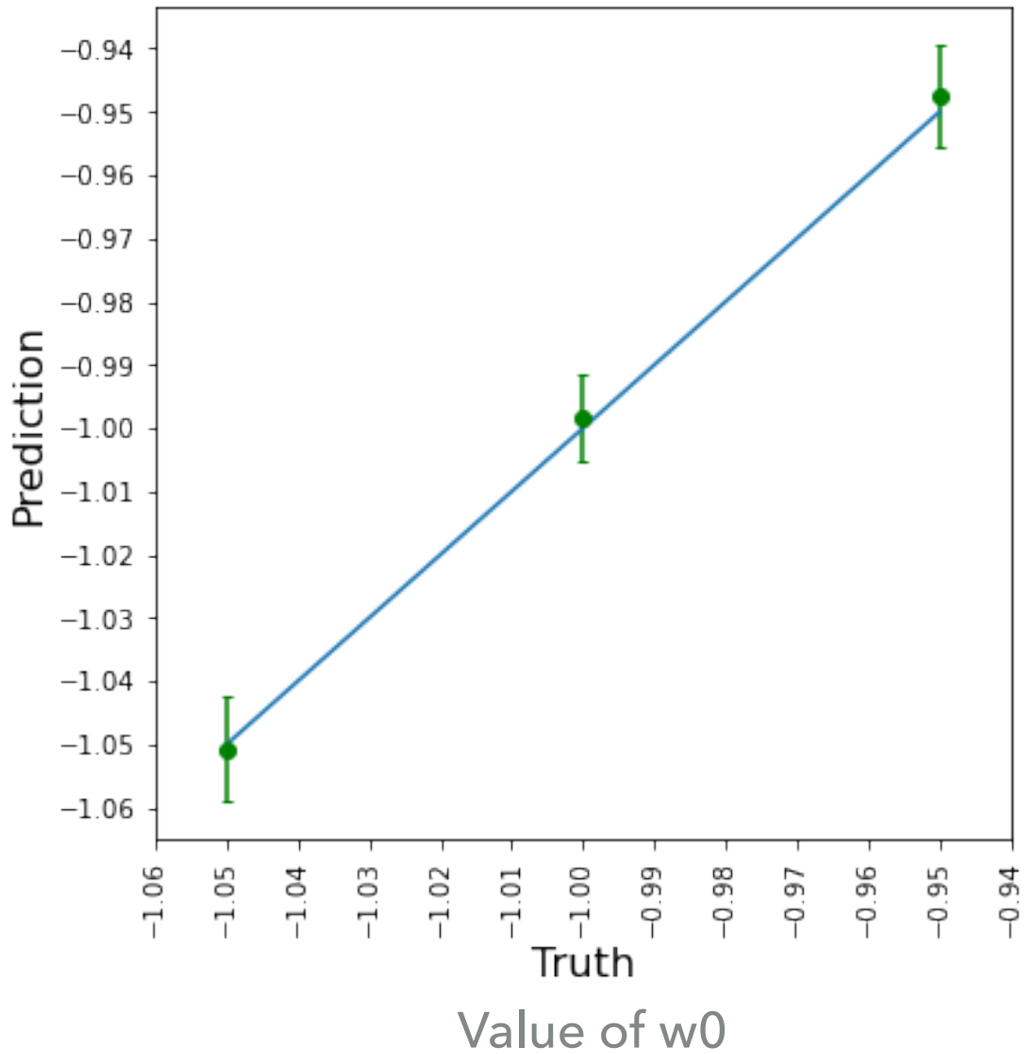
	Redshift	# halo (per realization)	r	Train Acc (360 realizations)	Valid acc (40 realizations)	Test acc (100 realizations)
Binary classification	Z= 0	~1000	[0, 100]	100%	100%	99%
Multi classification	Z= 0	~1000	[0, 100]	99%	99%	97%

F.Farsian et al. 2022

RESULTS: REGRESSION

The GNN is able **to predict** the value of **w0** correctly with only **2% error**.

The error-bars here represent the variation of GNN prediction for 100 test set realizations.



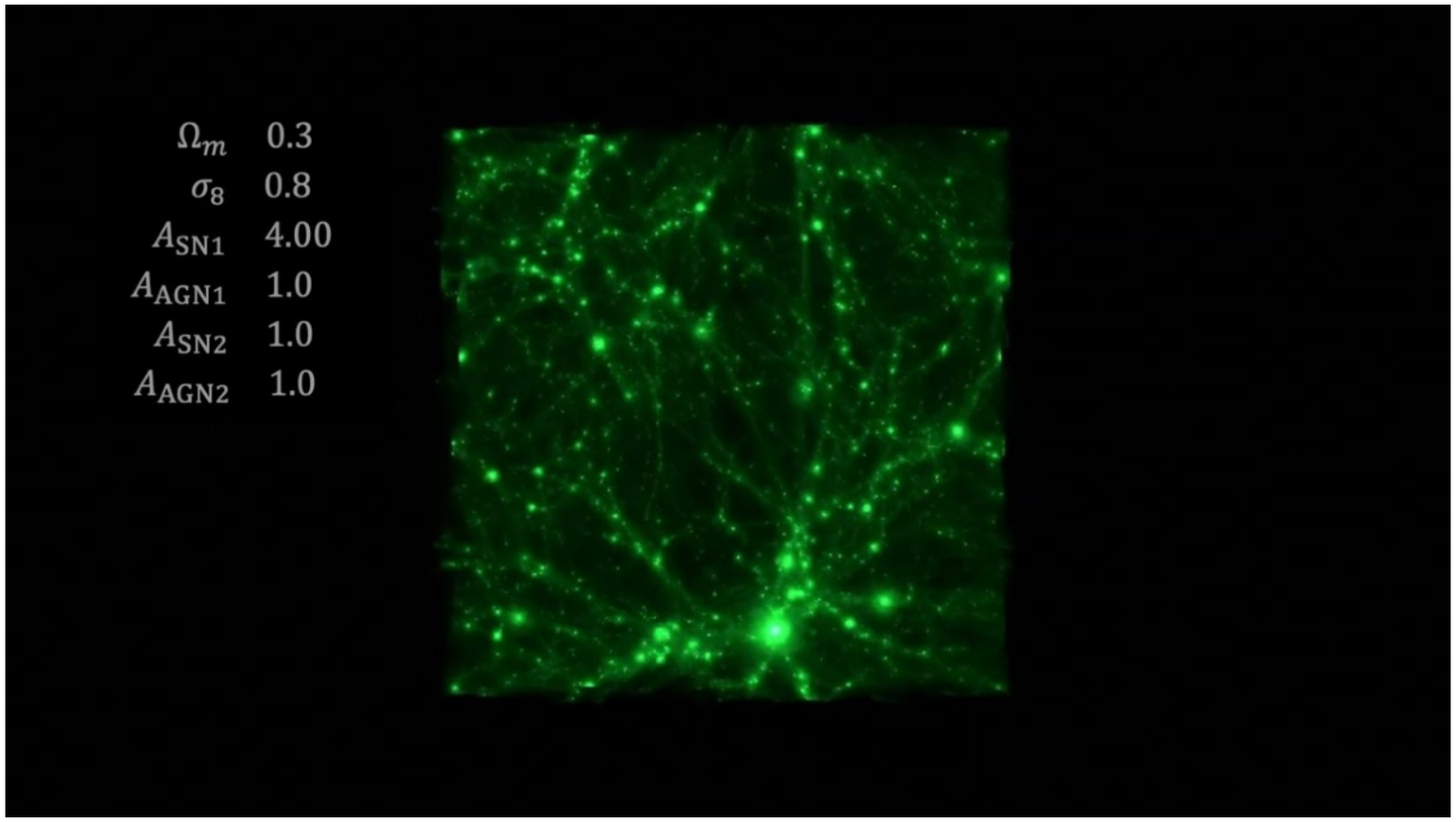
Value of w0

F.Farsian et al. 2022

The problem

SIMULATION

- CAMEL simulation, 4,000 numerical simulations, both N-body and (magneto-)hydrodynamic
- More than 100 billion dark matter particles and fluid elements in a combined volume of ~ (400 Mpc/h)^3
- Span thousands of different cosmological and astrophysical models, represent a large dataset to train machine learning algorithms.



The CAMEL project, arXiv:2010.00619 FVN et al 2020

DATA PREPRATION

- Only **mass and coordinates** of halos are give to the network as features.
- We could applied different mass cut for each catalogue.
- Two nodes i and j are connected by an **edge** if they are closer than a certain distance r (Mpc/h).
- In our analysis r is a hyper-parameter and we run the analysis to find the optimum value.