Universidad Autónoma de Madrid

From Projection to Reconstruction: Predicting Peculiar Velocities and Internal Distances in Galaxy Clusters via Graph Neural Networks

Sheng-Chieh Lin[1] & Weiguang Cui[1]

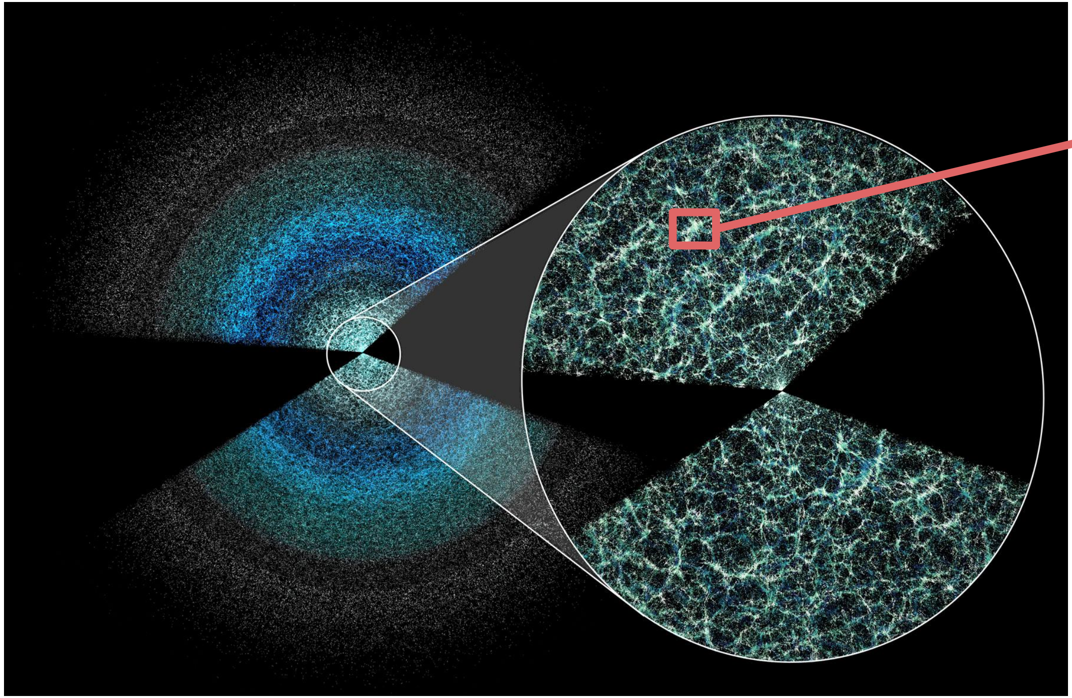
the300 collaboration

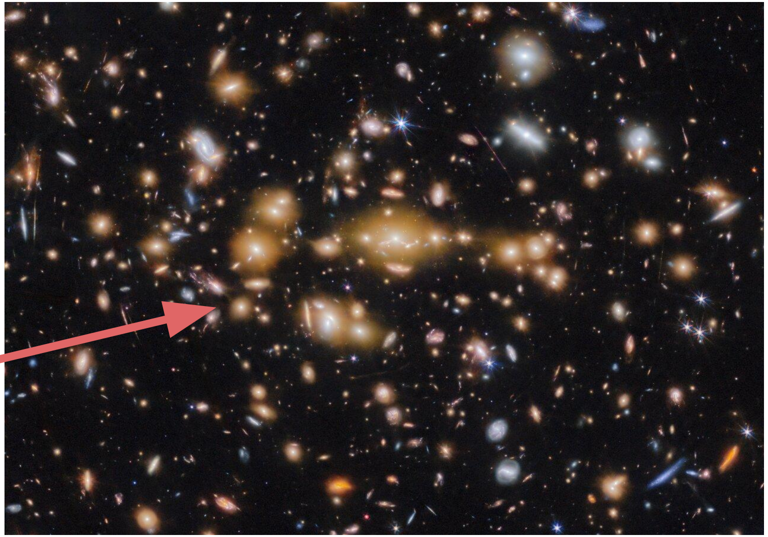
[1]Universidad Autónoma de Madrid Ciudad Universitaria de Cantoblanco. 28049. Madrid

ML4ASTRO3 | Malta | 09/02/2026

Galaxies are not randomly distributed

Galaxies live in overdense regions like galaxy clusters are greatly shaped by the environments



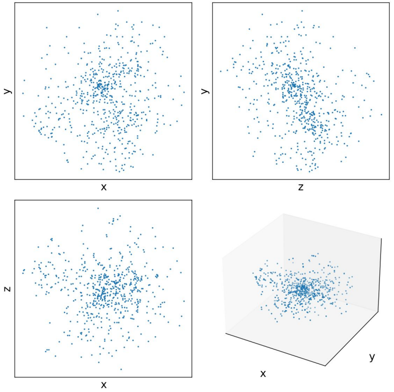


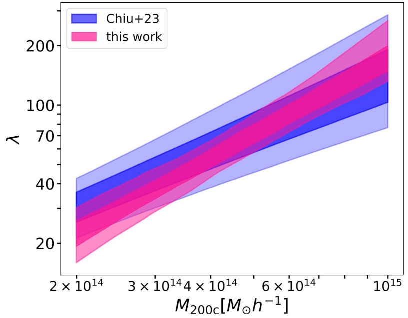
Galaxies trace the characteristics of these regions individually, offering an unique probe into the formation and evolution of cosmic structures.

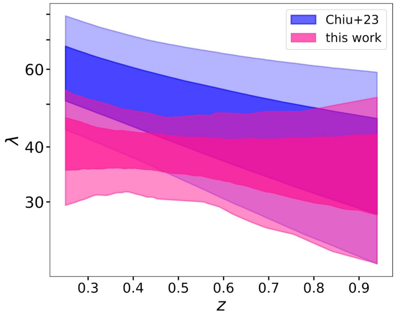
Left: DESI
Right: ESA/Webb, NASA & CSA

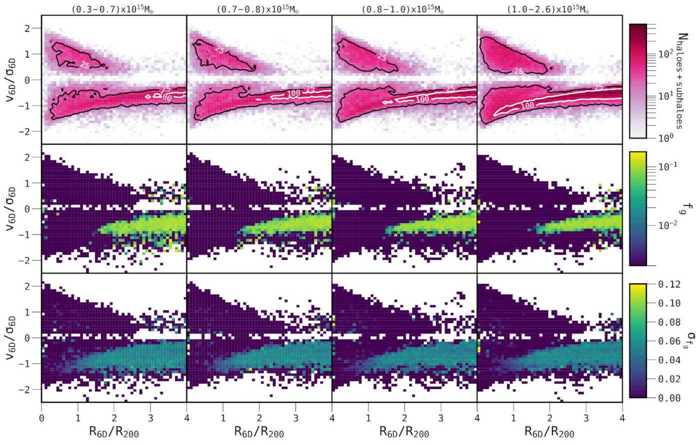
Information enabled by the phase space distribution of cluster galaxies

1. True shapes of galaxy clusters
    a. the triaxial, filamentary morphology
2. Cluster density profile
    a. Richness
    b. Anisotropy and mass reconstructions
    c. Caustic mass
3. Dynamics of galaxy clusters
    a. Infaller v.s. Backsplash
    b. Environmental quenching
    c. Substructure



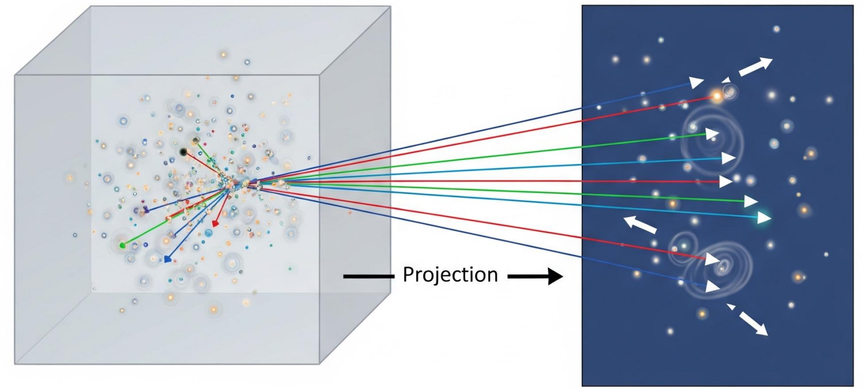


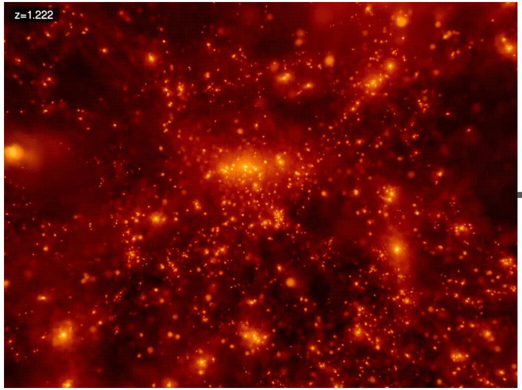


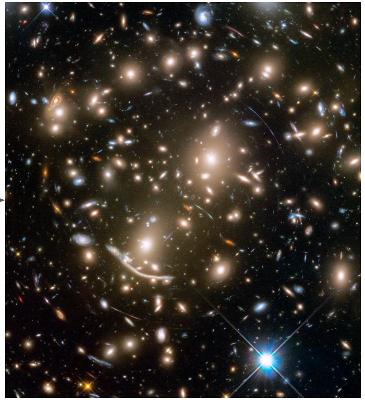


Bottom left: Arthur et al. 2019
Right: Singh et al. 2025

But we only see projections…







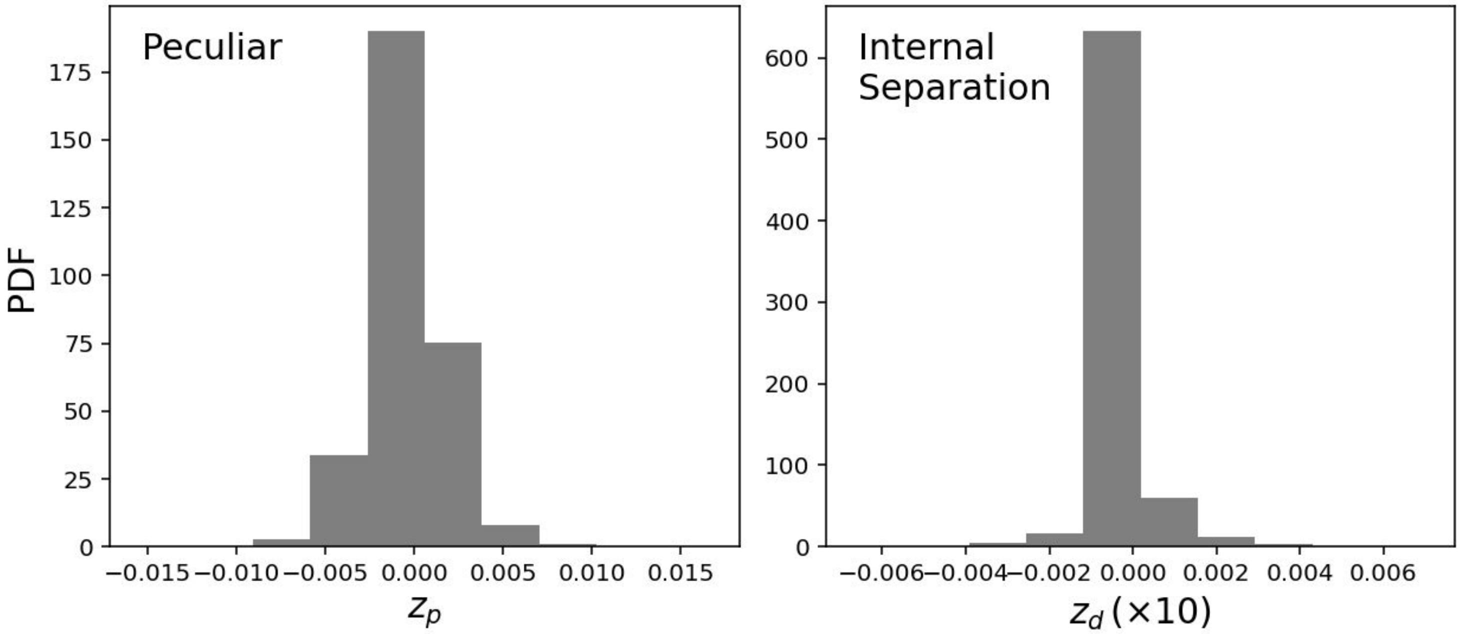
Left: MultiDark resimulation
Right: JWST image (credit: *NASA, ESA, CSA, STScI)*

The tanglement of peculiar velocity and internal distance

The magnitudes of both components are on different scales where the peculiar velocity significantly dominate over the internal separation

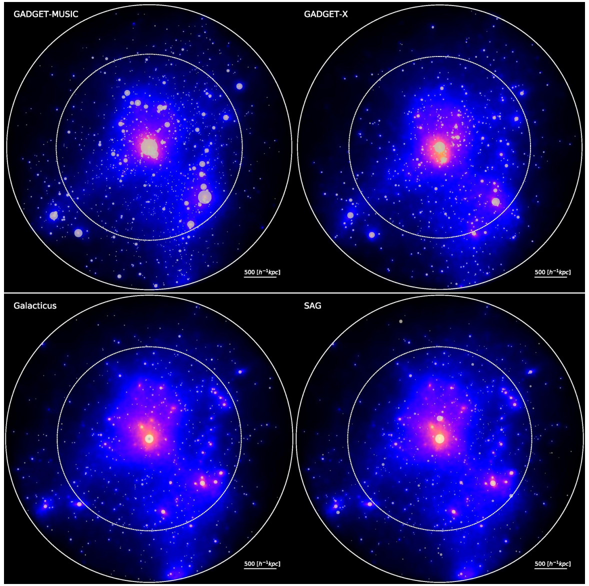
Assuming no gravitational redshift:

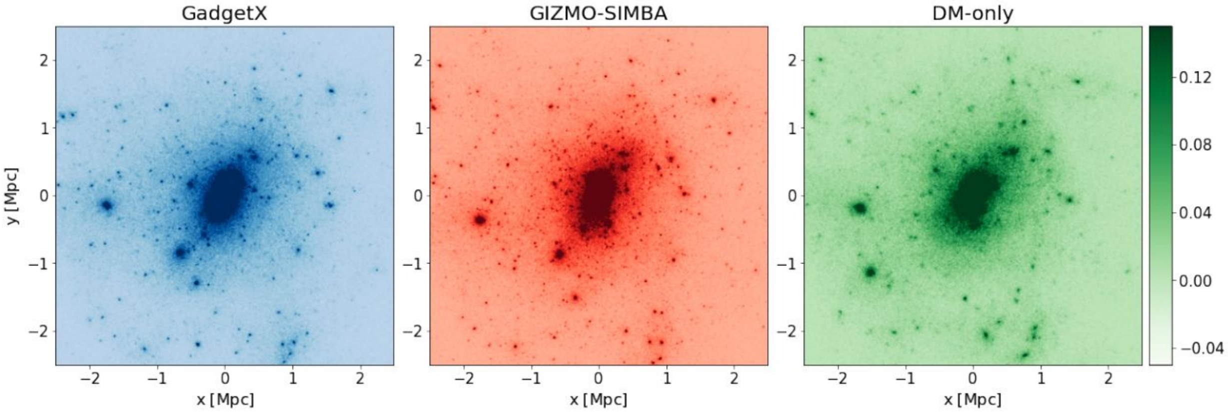
$$(1 + z_{gal}) = (1 + z_{cls})(1 + z_d)(1 + z_p)$$

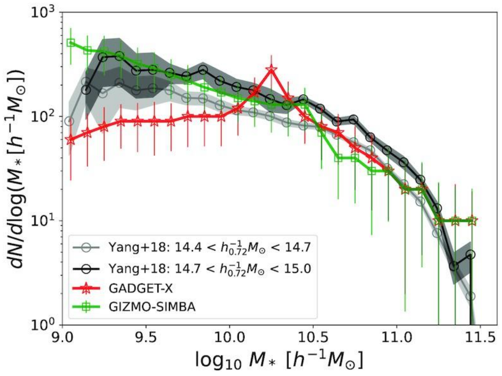


THE THREE HUNDRED (the300)

1. Re-simulations with box size of 15 Mpc extracted from the MultiDark-Planck N-body simulation incorporating different hydro-solvers and semi-analytic models
2. Centered on most massive clusters in the N-body simulation with mass above 10^14.8 Msun



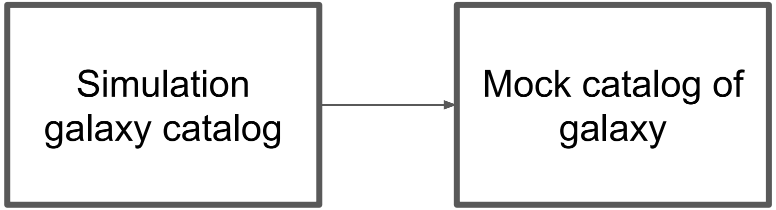




Left: Giocoli et al. 2025
Top right: Cui et al. 2018
Top left: Cui et al. 2022

The catalog-level data

- **Data basics:** simulated galaxies from the300 suites with the GIZMO hydro code
- **Input features:** galaxy properties with projected sky coordinates and redshift (mock); photometry can be used as features as well
- **Output labels:** redshift from peculiar velocity Z_p and the cosmological distance of galaxies Z_h



- 6D phase space
- Simulated photometry of various observations from Caesar algorithm

- Projecting phase space distributions into sky
- Sky coordinates (ra, dec) + z
- Redshift includes both cosmological distance and peculiar velocity

*Cosmological distance is easier to predict compared to the internal distance

Experimented algorithms: XGBoost, LightGBM, TabPFN (prediction only)

The catalog-level data - results

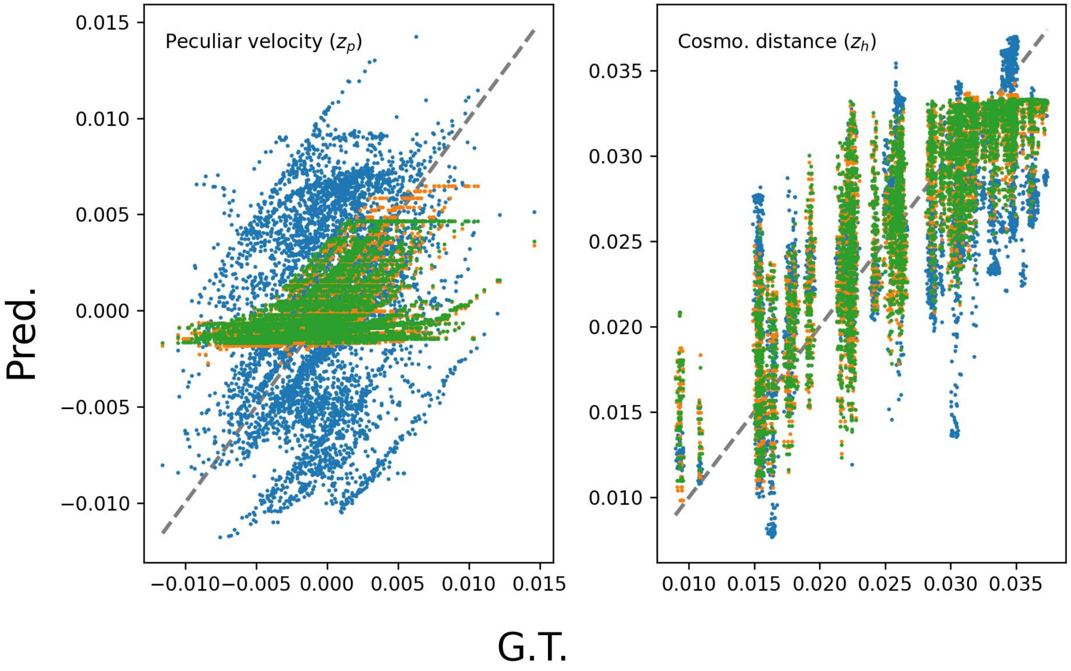
Experimented algorithms: XGBoost, LightGBM, TabPFN (prediction only)

- **Data basics:** simulated galaxies from the300 suites with the GIZMO hydro code
- **Input features:** galaxy properties v sky coordinates and redshift (mock can be used as features as well
- **Output labels:** redshift from peculi and the cosmological distance of g

*Cosmological distance is easier to predict com internal distance

Boosting: RMSE ~ 0.3% and ~ 0.3% for z_p and z_h, respectively
TabPFN: ~0.5% for both labels

Orange: XGBoost; Green: LightGBM; Blue: TabPFN

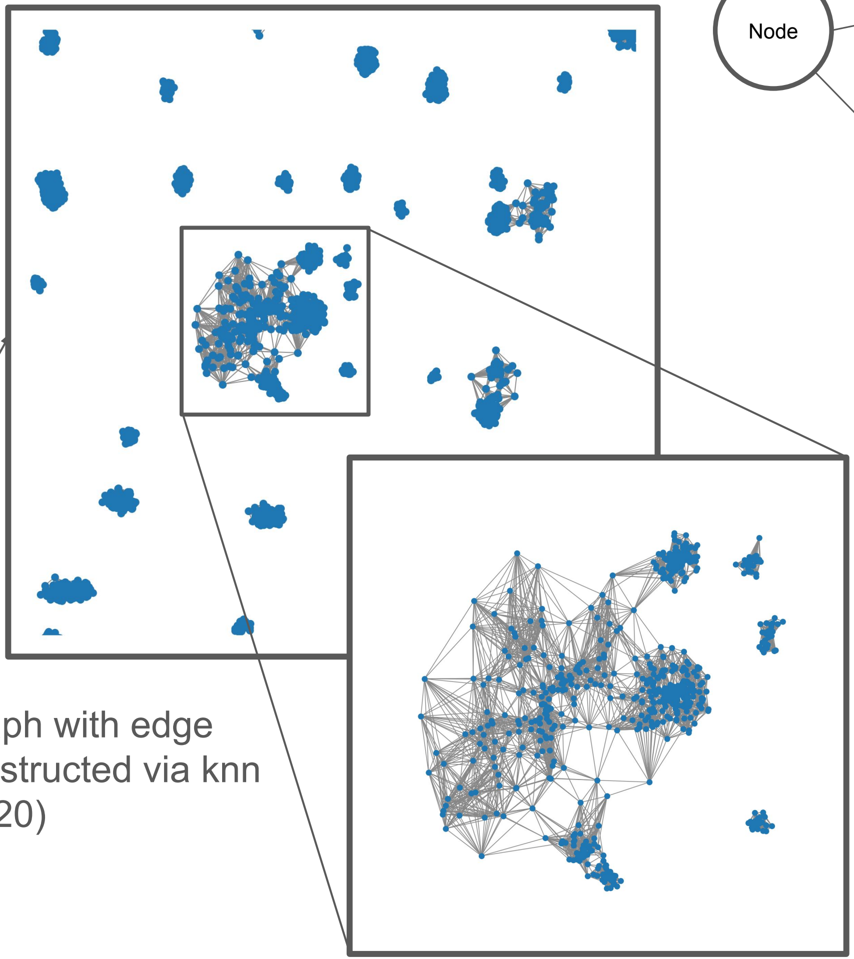


From catalog to graph…

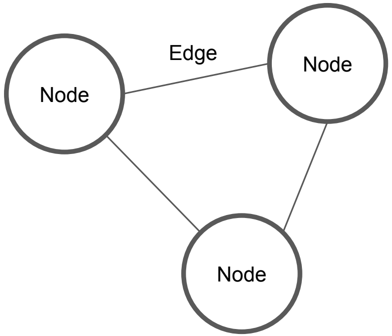
Galaxy catalog

↓

Mock catalog of galaxy



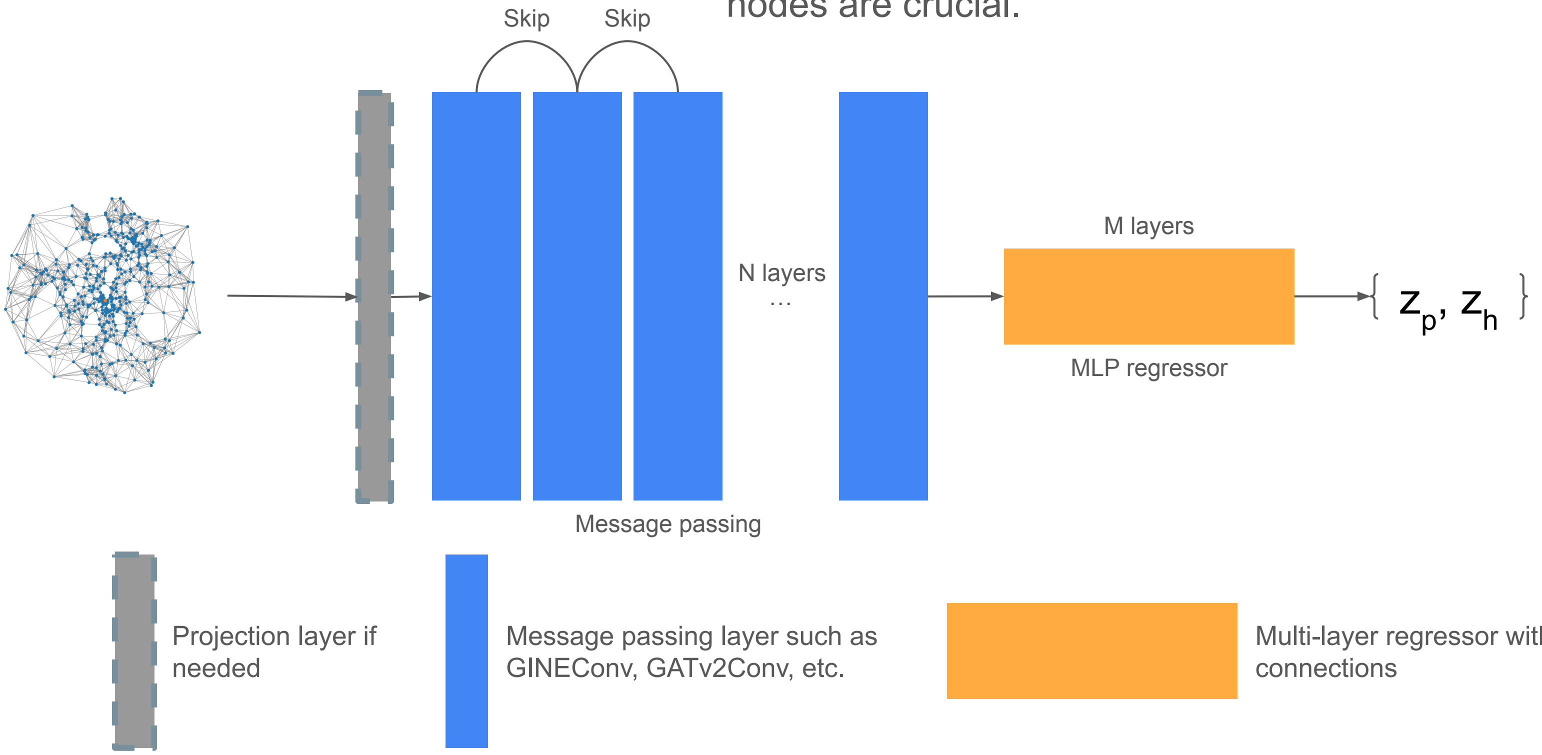
Graph with edge constructed via knn (k=20)



- **Nodes** represent individual galaxies containing information about the galaxy properties
- **Edges** link the galaxies by computing the Gaussian Radial Basis Function distance between galaxy's sky coordinates

Graph Neural Networks

A type of network architectures specifically designed for learning relations between data points. GNNs are widely adapted in the tasks of working with relational data where connections between nodes are crucial.

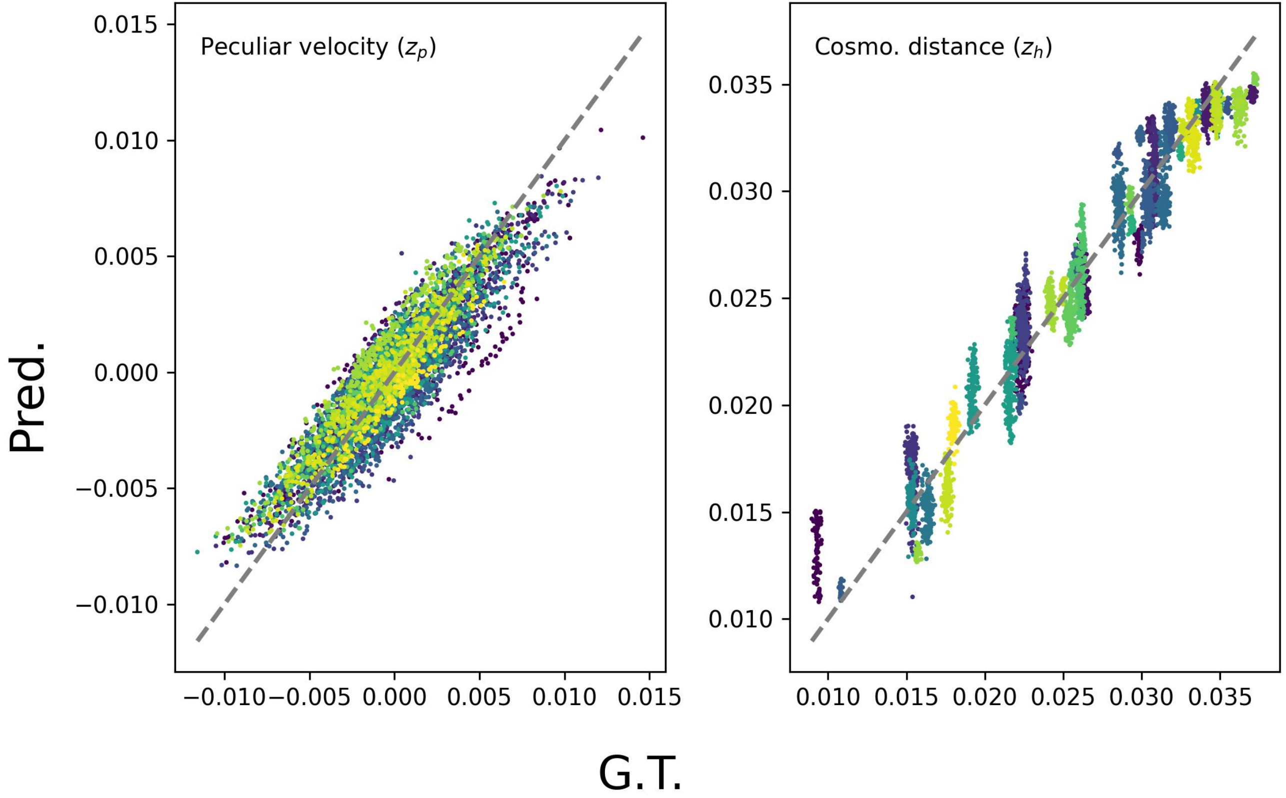


Models constructed using PyG

Results

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Experimented layers: GINEConv and GATv2Conv



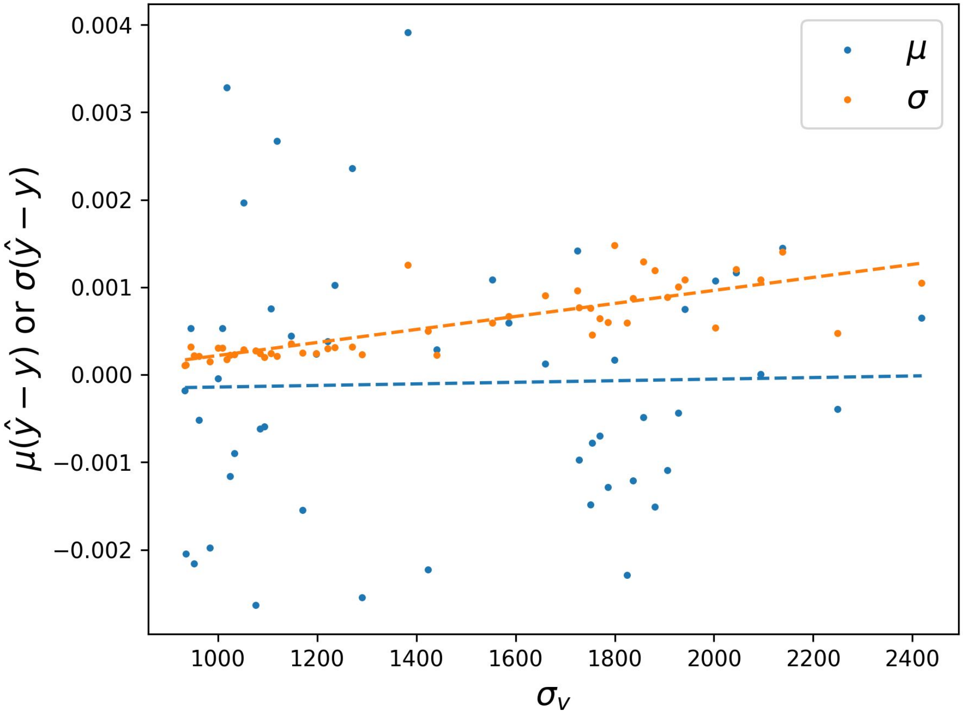
Peculiar velocity:

- z_p RMSE = 0.14%
- v_p RMSE ~ 431 km/s

Cosmological distance:

- z_h RMSE = 0.15%
- d_h RMSE ~ 6.6 Mpc

Predictions on cosmological distance



The scatter in prediction residual per halo positively correlates with the velocity dispersion stating that the per galaxy prediction still suffers from the FoG effect

This will require additional distance measures as features

Summary

1. GNNs provide a plausible path to break degeneracy between peculiar velocity and cosmological distance of galaxies
2. There are still lots of parameter space yet to be explored, physical-informed loss, dynamical knn, etc.
3. Realistic mock catalogs will be necessary for real-world applications