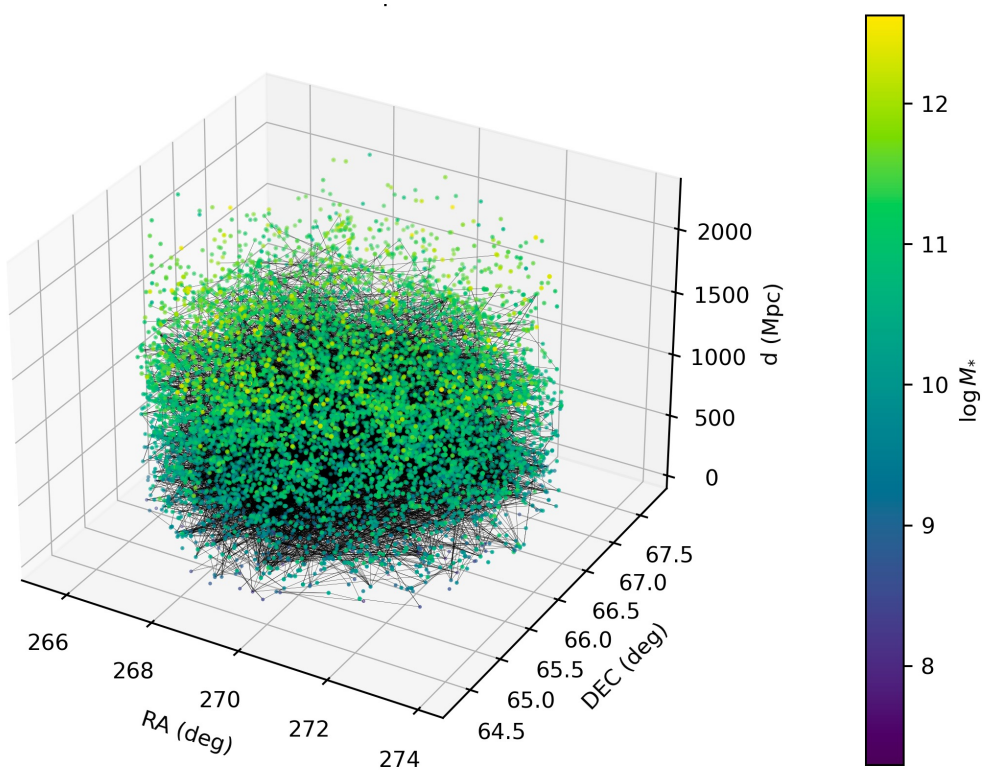
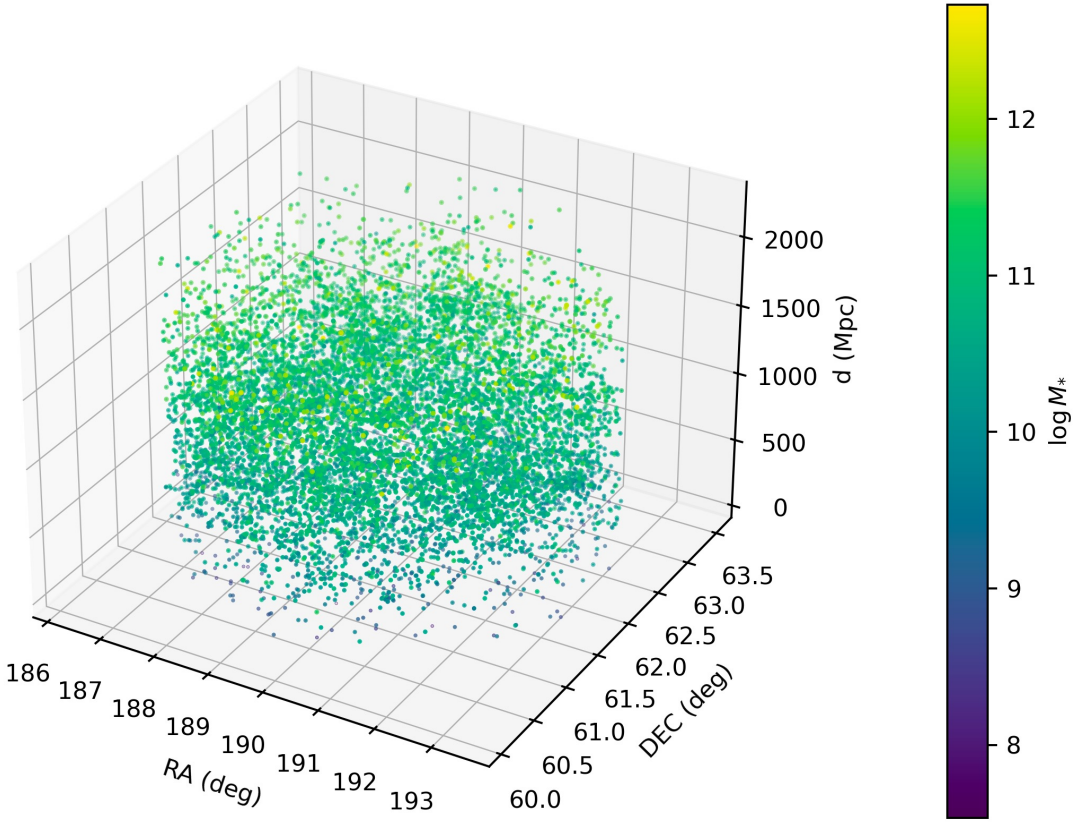
Galaxy stellar mass Estimation with Graph Neural Networks



Graph Neural Network (GNN)

Graph Neural Networks excel in handling irregular and sparse data, making them well-suited for tasks involving the spatial distribution of galaxies [2]

The GNN not only harnesses the intrinsic features of each galaxy, such as **fluxes in the R, G, Z, W1, and W2 filters**, alongside its **redshift Z**, but also incorporates the spatial relationship among galaxies through graph construction.



DESI (Dark Energy Spectroscopic Instrument)

DESI will measure the effect of dark energy on the expansion of the universe, constructing a 3D map spanning the nearby universe to 11 billion light years.

Spectra of stellar and extragalactic targets from Survey Validation constitute the first major data sample from the DESI survey [1]. The public release includes **428,758** objects targeted as part of good-quality spectral information from the **Bright Galaxy Survey.**

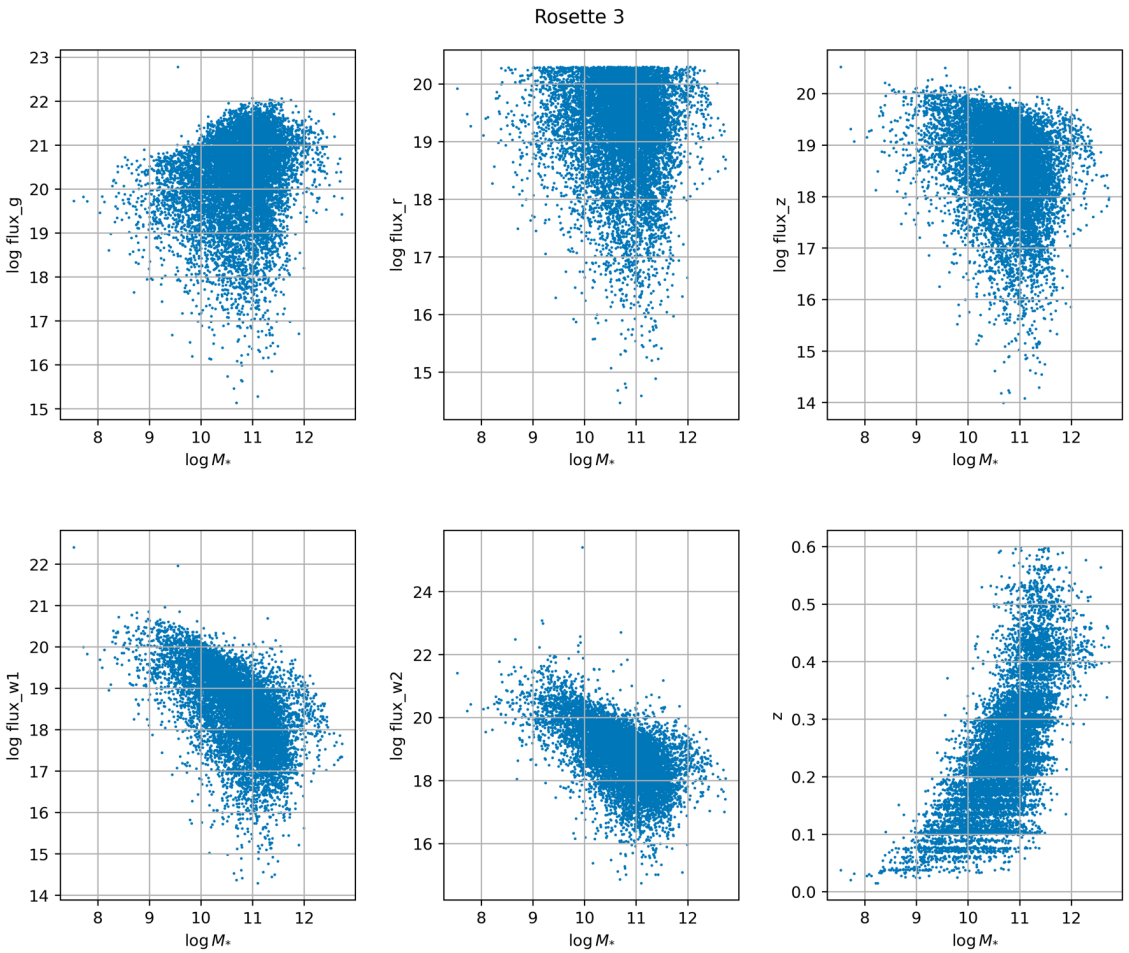
DESI EDR (Early Data Release)

It includes spectra from 1.8 million observed targets, taken from December 2020 to June 2021 [1]

Summary of EDR	
Number of useful spectra	1,769,157
Number of useful spectra of unique targets	1,712,004
Galaxies	1,125,635
Quasars	90,241
Stars	496,128
Spectral coverage	360–982.4 nm
Spectral resolution R	2000 (at 360 nm)–5500 (at 980 nm)
Wavelength system	vacuum
Photometric bands	*g*, *r*, *z*, *W*1, *W*2, *W*3, *W*4
Approximate area	~1390 square degrees

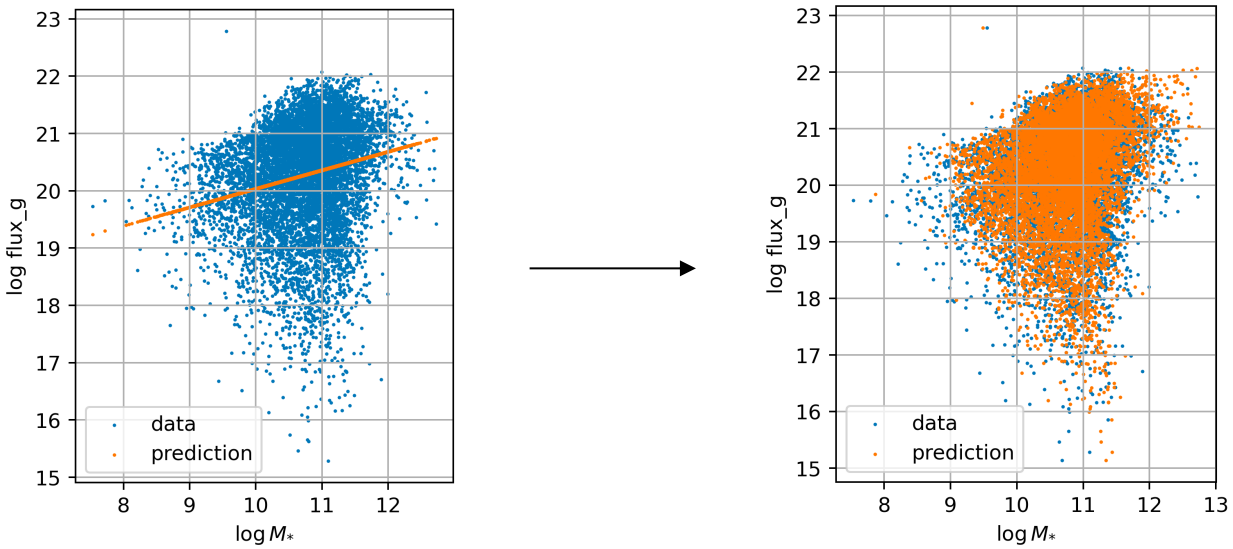
Linear model

Multivariable linear model with fluxes G, R, Z, W1, W2 and redshift Z



sklearn.linear_model.LinearRegression

Adjust a linear model with coefficients $w = (w1, \ldots, wp)$ to minimize the residual sum of squares between the observed targets and the ones predicted by the linear approximation [3]



Loss function: Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

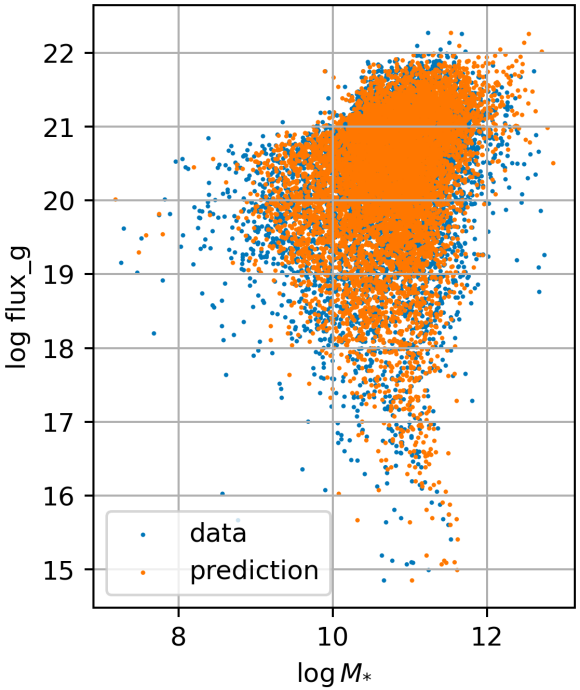
Mean squared error between the model predictions and the actual values

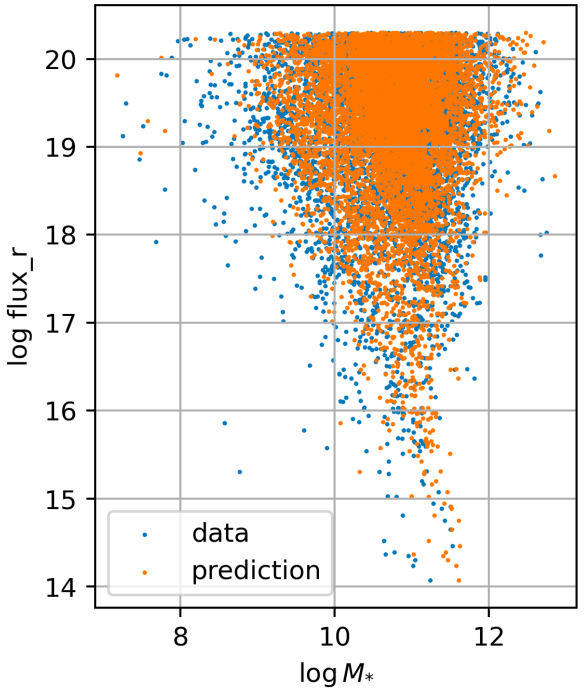
Coefficient of determination R^2

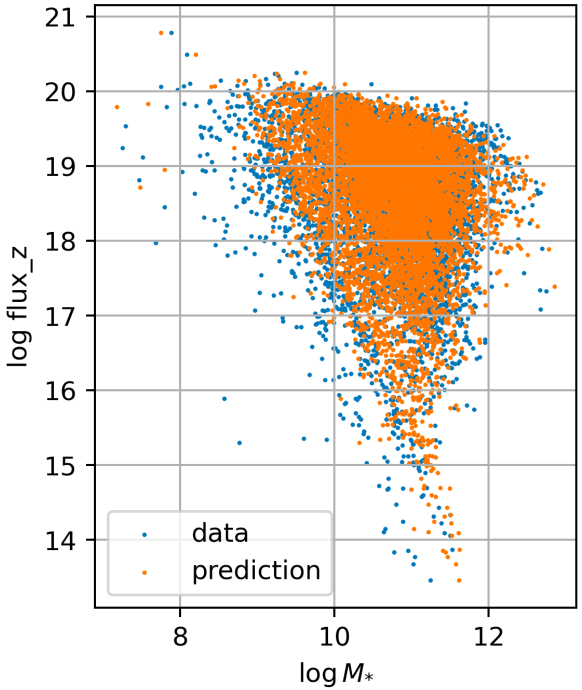
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

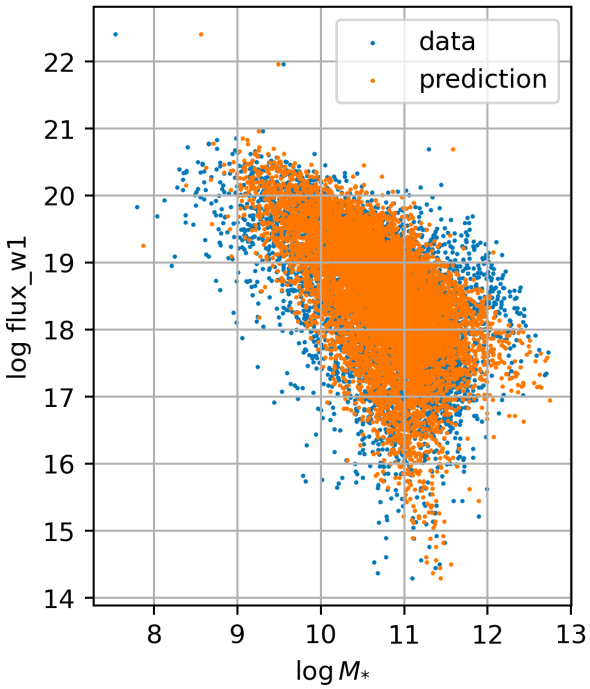
Proportion of variance in the dependent variable that is predictable from the model independent variables

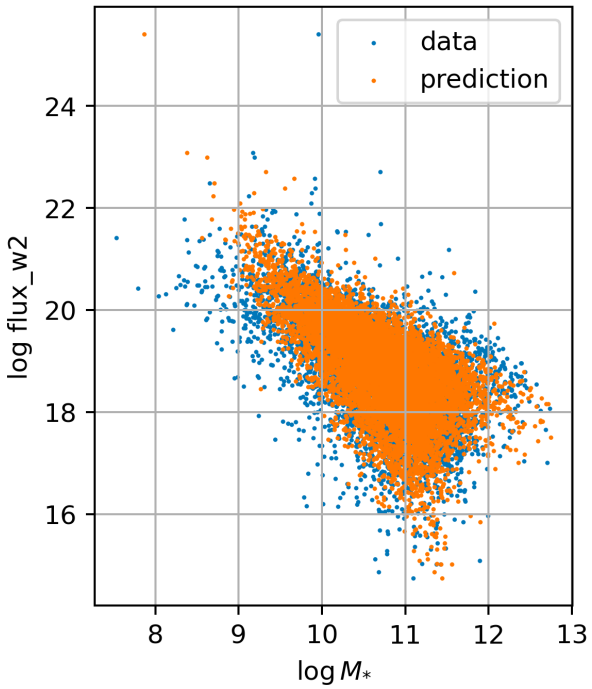
$$R^2 = 1.0000$$
$$mse = 0.0515$$

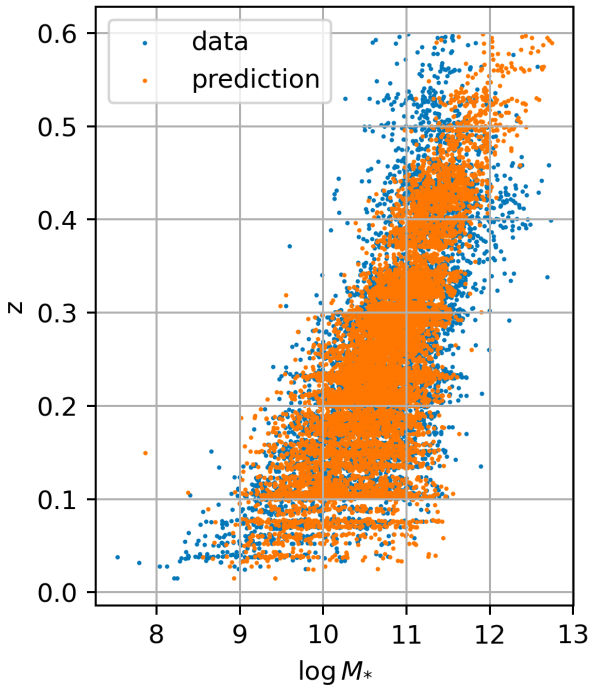







data
prediction
log flux_w1
log M_*


data
prediction
log flux_w2
log M_*


data
prediction
z
log M_*

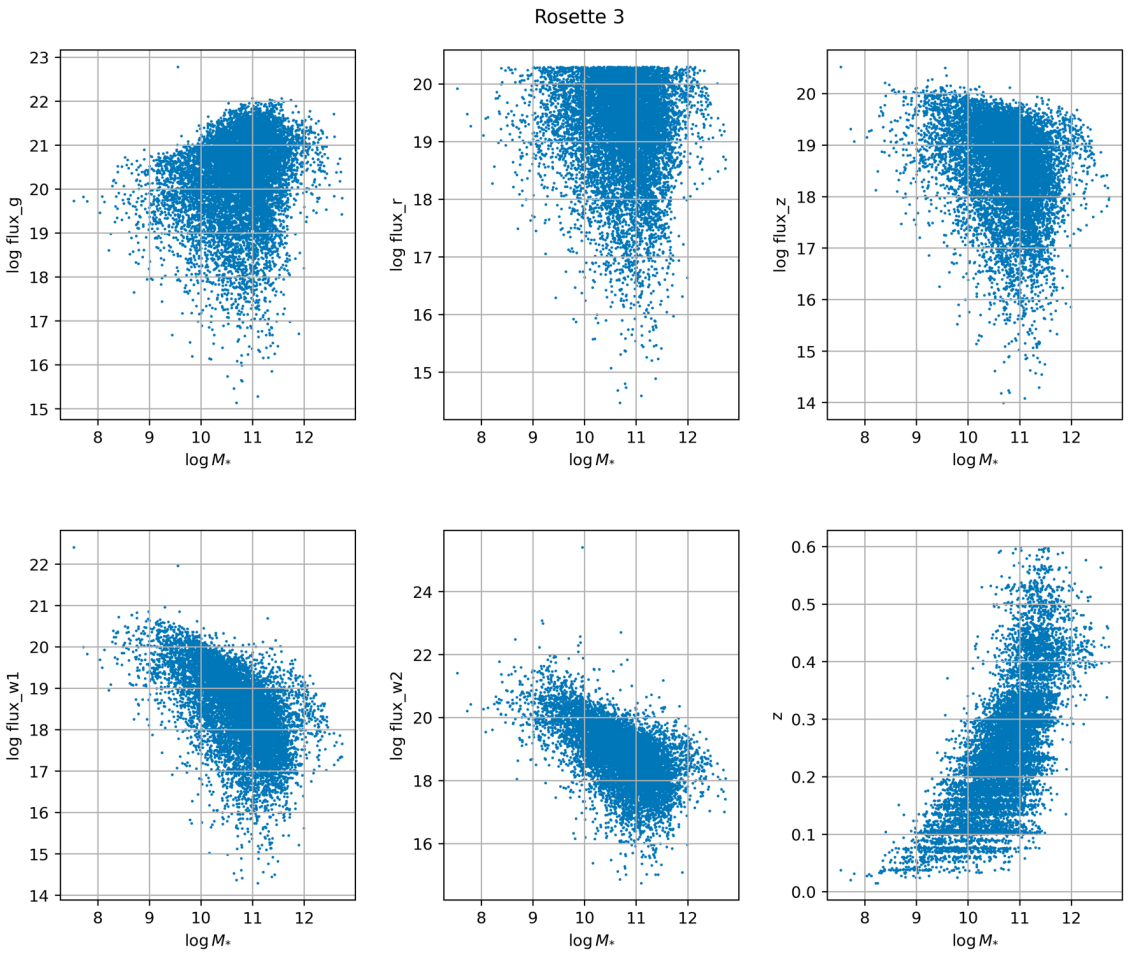
Linear model

Some testing results

$\boldsymbol{Rosette\ number}$	$\boldsymbol{R^2}$	$\boldsymbol{mse}$
$3\ (train)$	1.000	0.0515
$6\ (test)$	0.8612	0.0596
$7\ (test)$	0.8504	0.0637

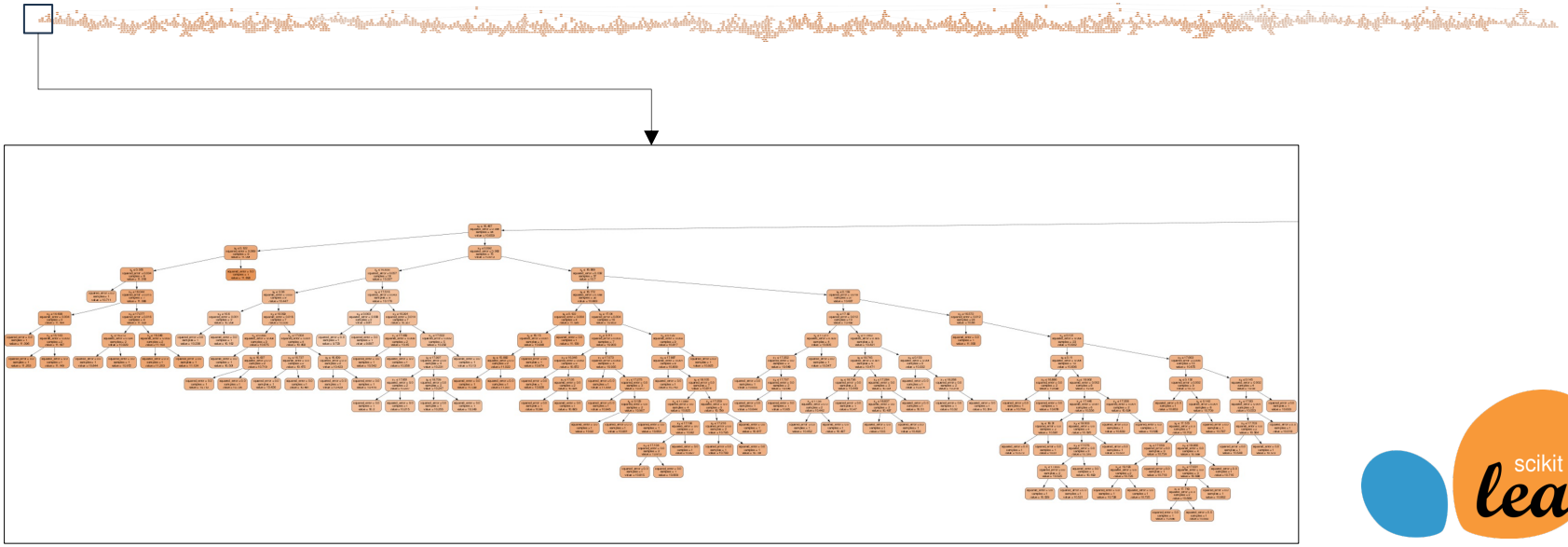
Random Forest

RF model with fluxes G, R, Z, W1, W2 and redshift Z



sklearn.ensemble.RandomForestRegressor

Meta-estimator that fits decision trees on different subsets, using averaging to enhance prediction and control overfitting [3]

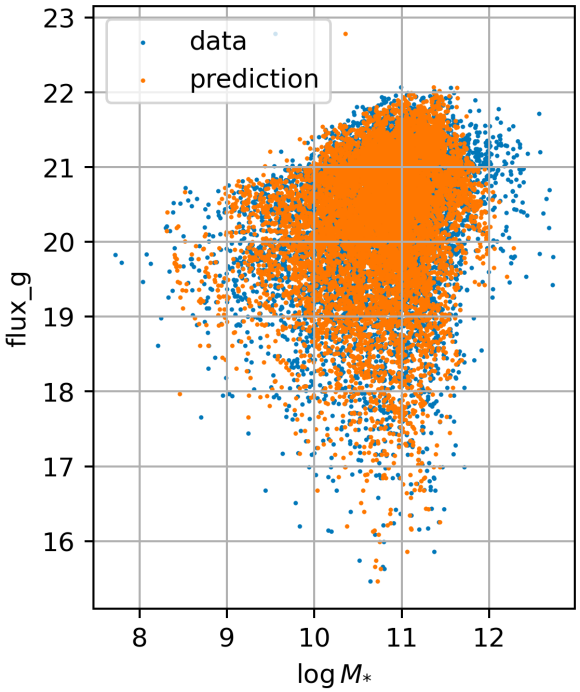


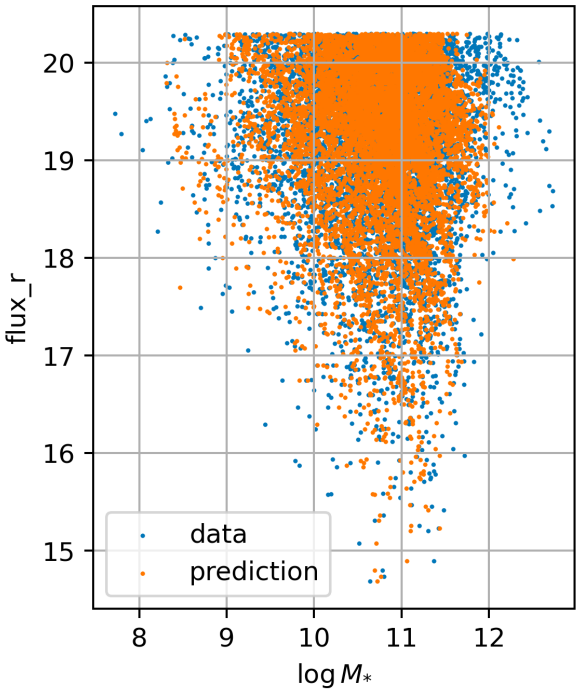
Loss function: Mean Squared Error (MSE)

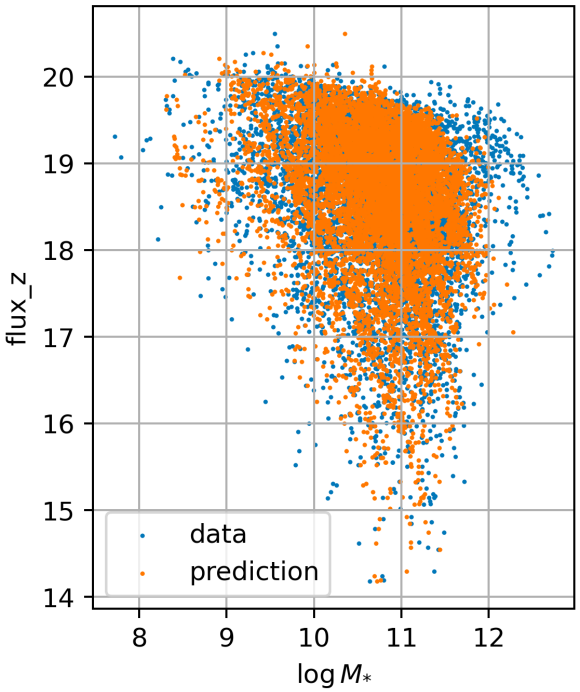
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

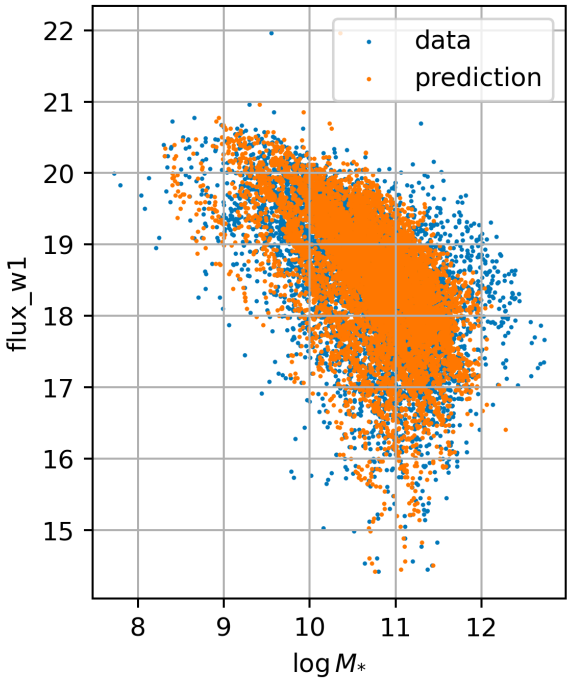
Mean squared error between the model predictions and the actual values

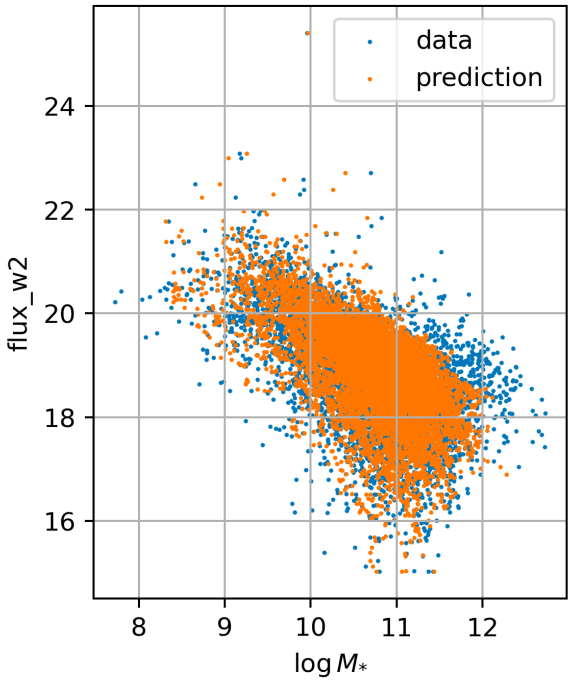
$$mse = 0.0357$$

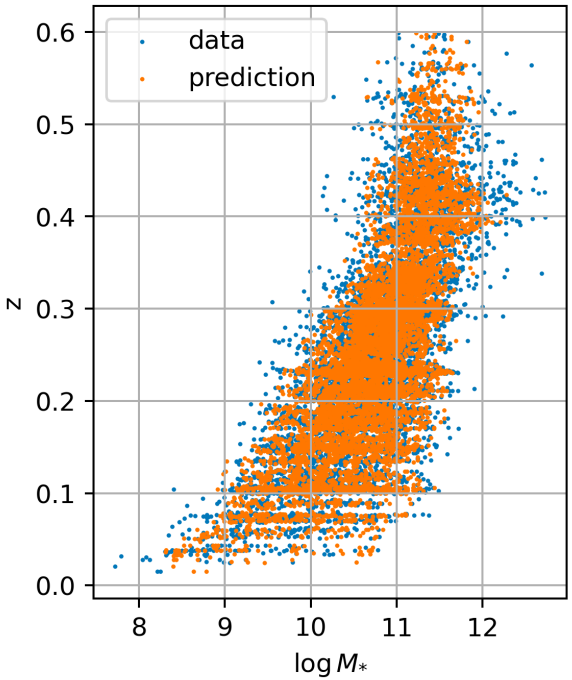

data
prediction
flux_g
log M_*


data
prediction
flux_r
log M_*


data
prediction
flux_z
log M_*

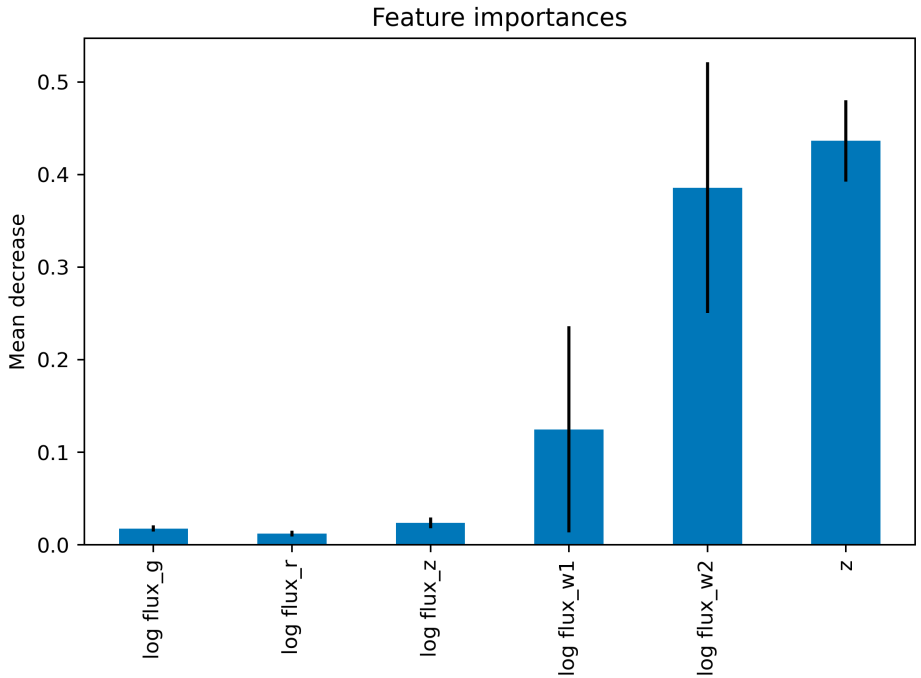

data
prediction
flux_w1
log M_*


data
prediction
flux_w2
log M_*

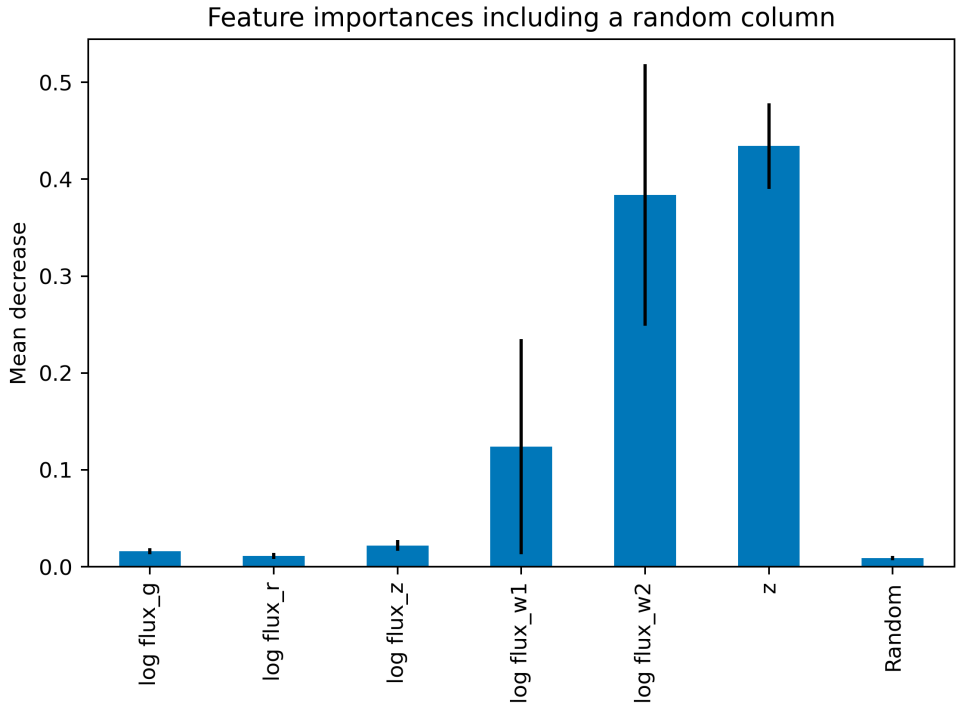

data
prediction
z
log M_*

Random Subset Feature Selection (RSFS)

Introduces additional randomness in the model-building process to enhance generalization and reduce overfitting



- Reduction of correlation between trees
- Overfitting prevention
- Increased robustness

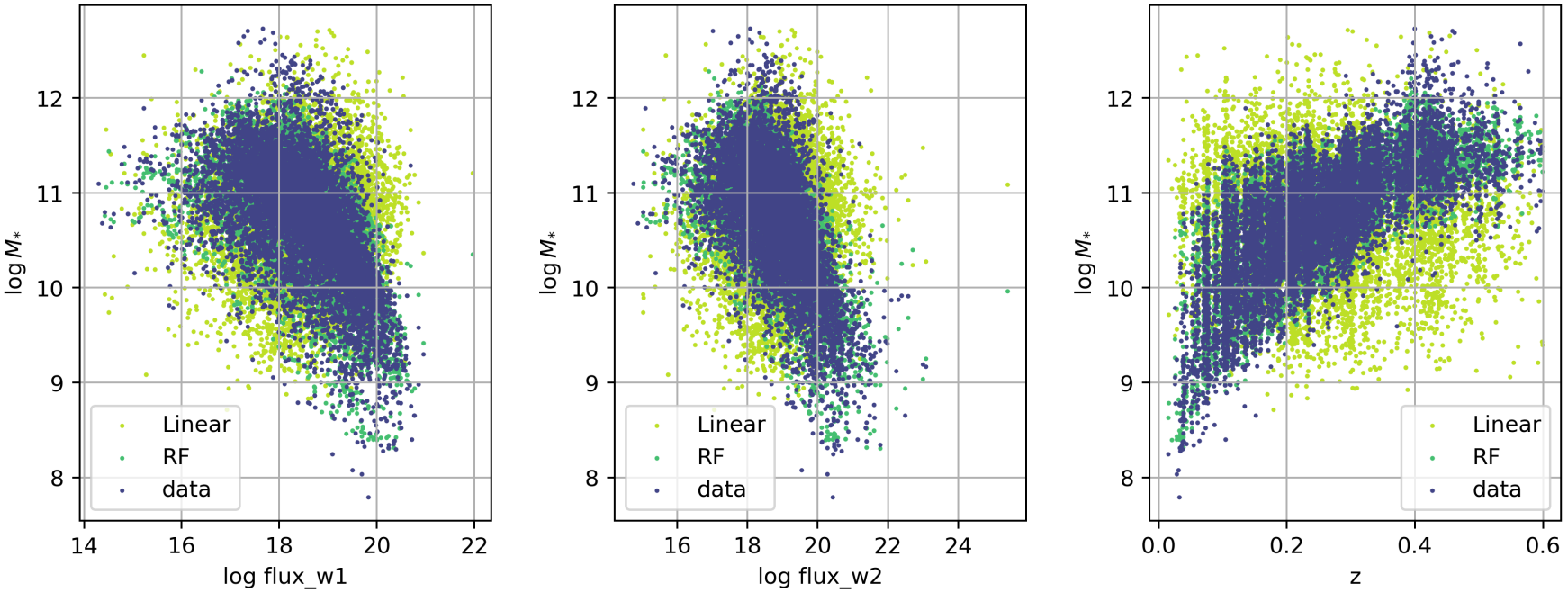


Random Forest model

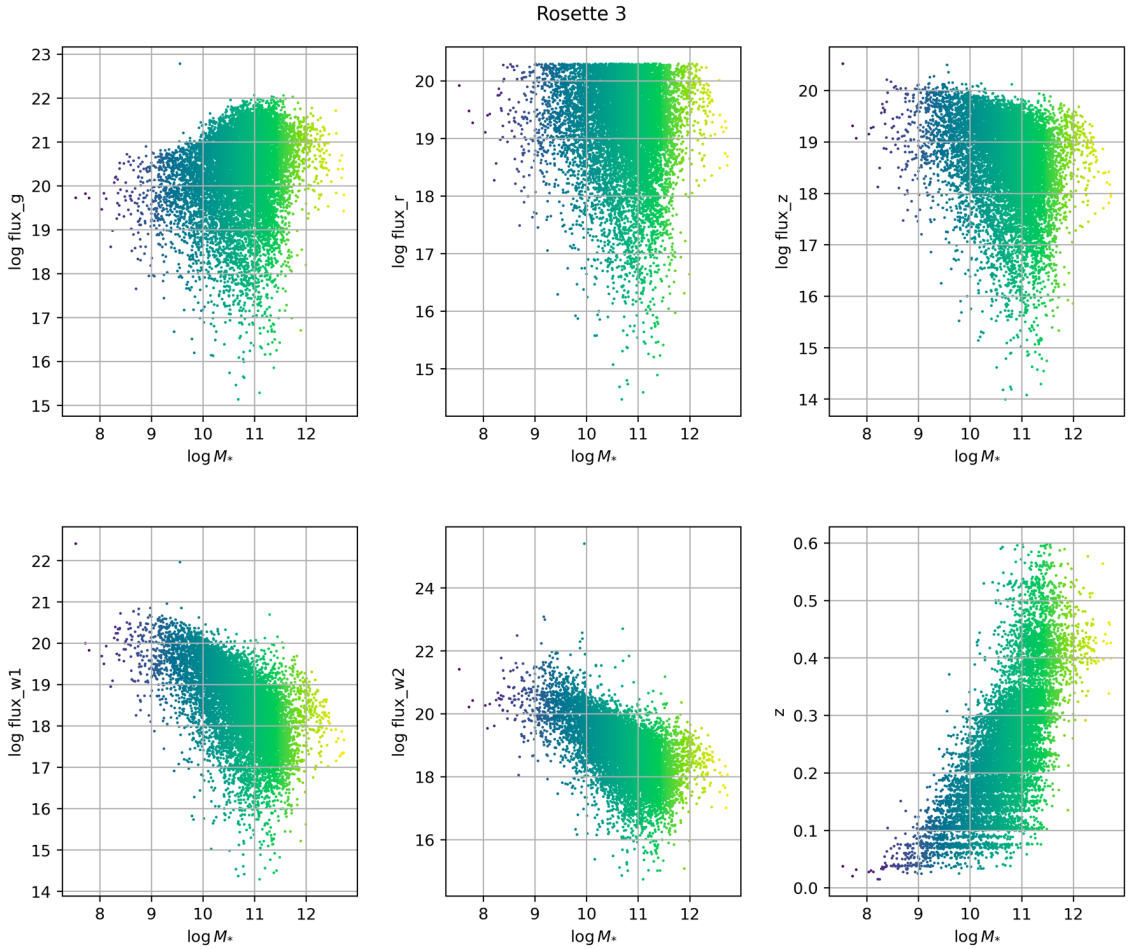
Some testing results

$\boldsymbol{Rosette\ number}$	$\boldsymbol{mse}$
$3\ (train)$	0.0357
$6\ (test)$	0.0420
$7\ (test)$	0.0447

RF and Linear



Graph Neural Network



About GNNs

$$h_i^{t+1} = f\left(h_i^t W + \sum_{j \in N(i)} \frac{1}{C_{i,j}} h_j^t U\right)$$

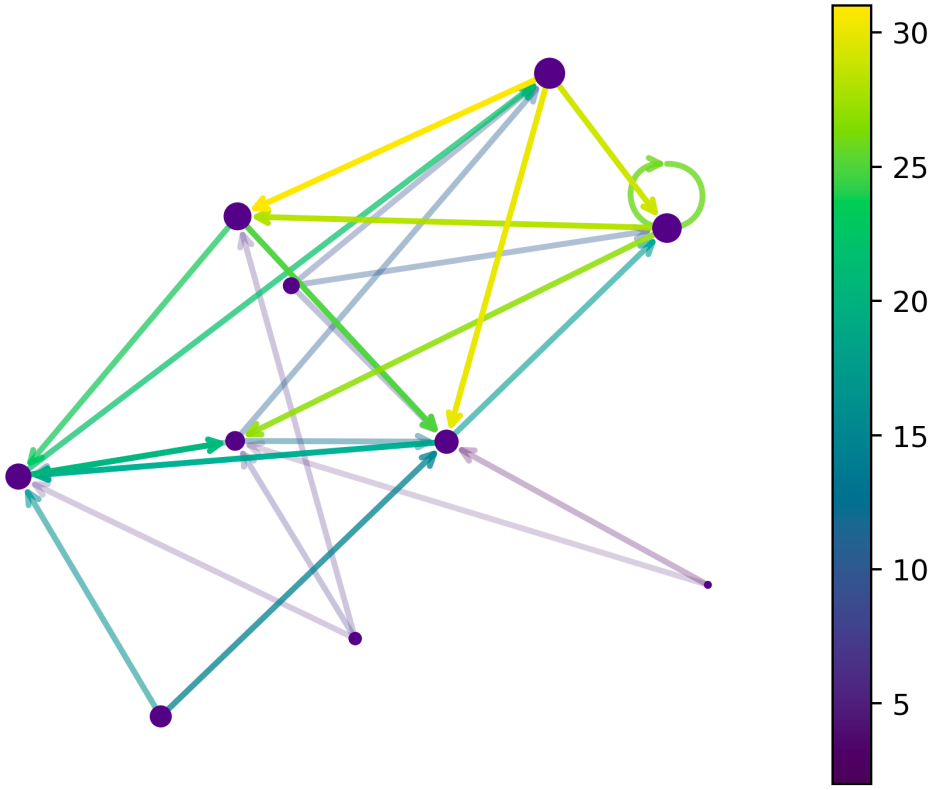
Old node representation times a weight matrix: $h_i^t W$

Information from neighbors times a weight matrix: $h_j^t U$

The sum is a **permutation-invariant** aggregation function

Graphs

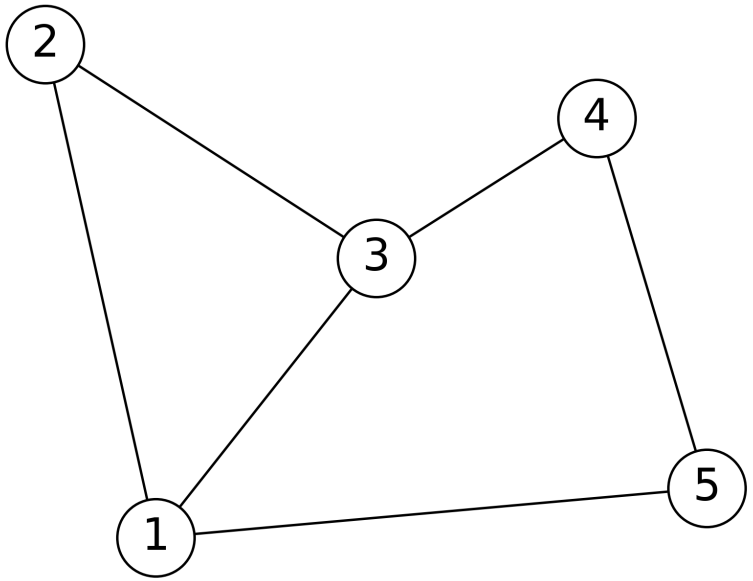
A simple graph G consists of a non-empty finite set V(G) of elements called vertices (or nodes), and a finite set E(G) of distinct unordered pairs of distinct elements from V(G) called edges (or arcs) [4]



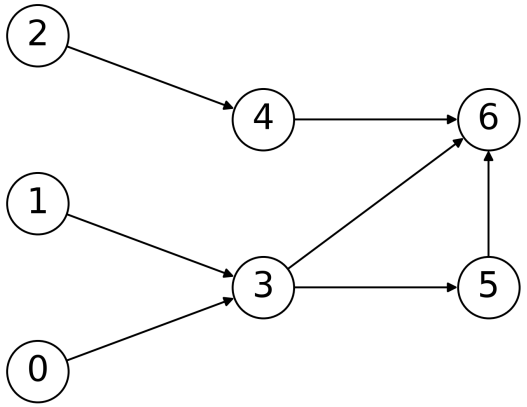
Some definitions

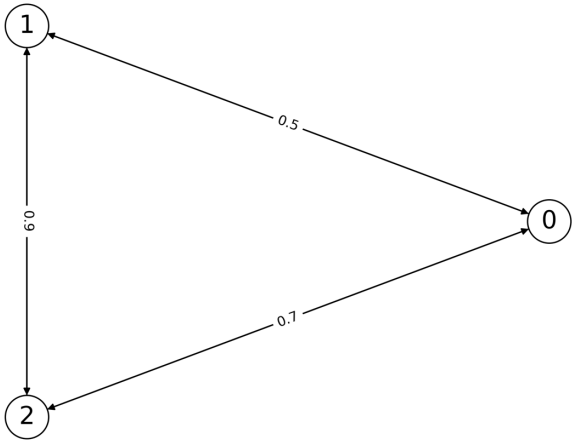
Adjacency: two vertices *v* and *w* of a graph *G* are adjacent if there is an edge *vw* joining-them [4]

Degree: the degree of a vertex *v* of *G* is the number of edges incident with *v* [4]



Directed graph (digraph): a graph in which edges have a direction, indicating a one-way relationship between nodes



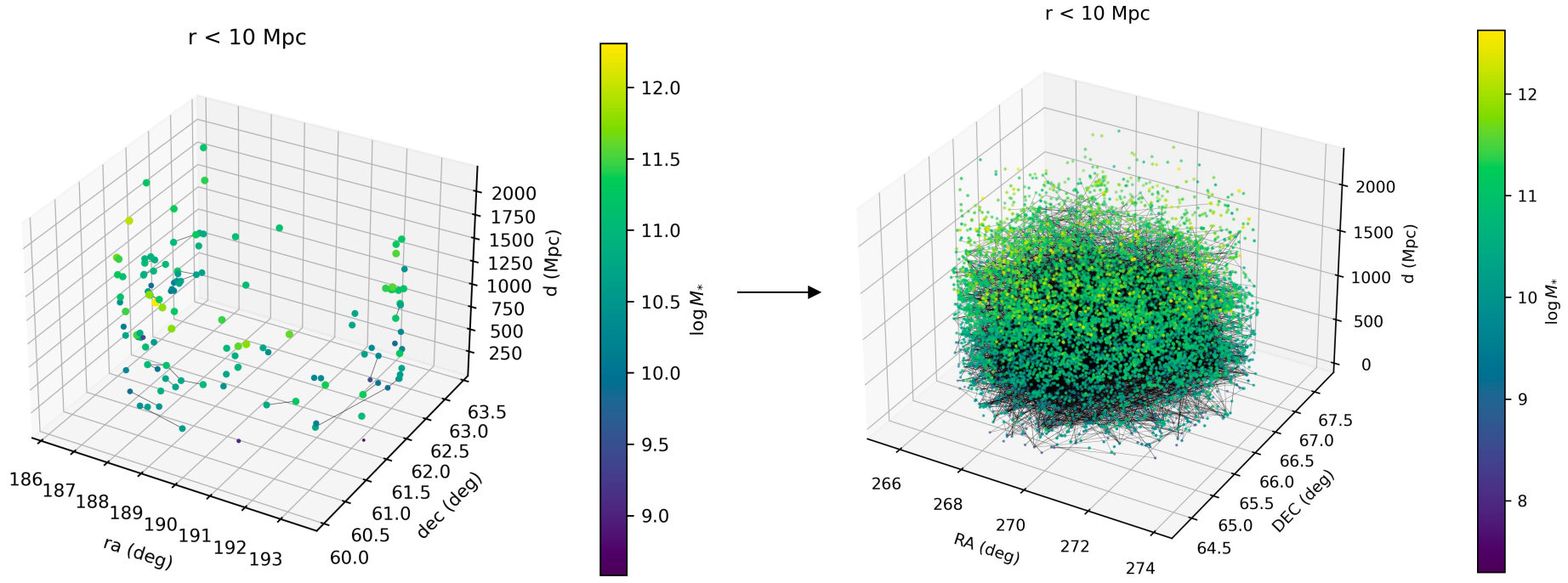


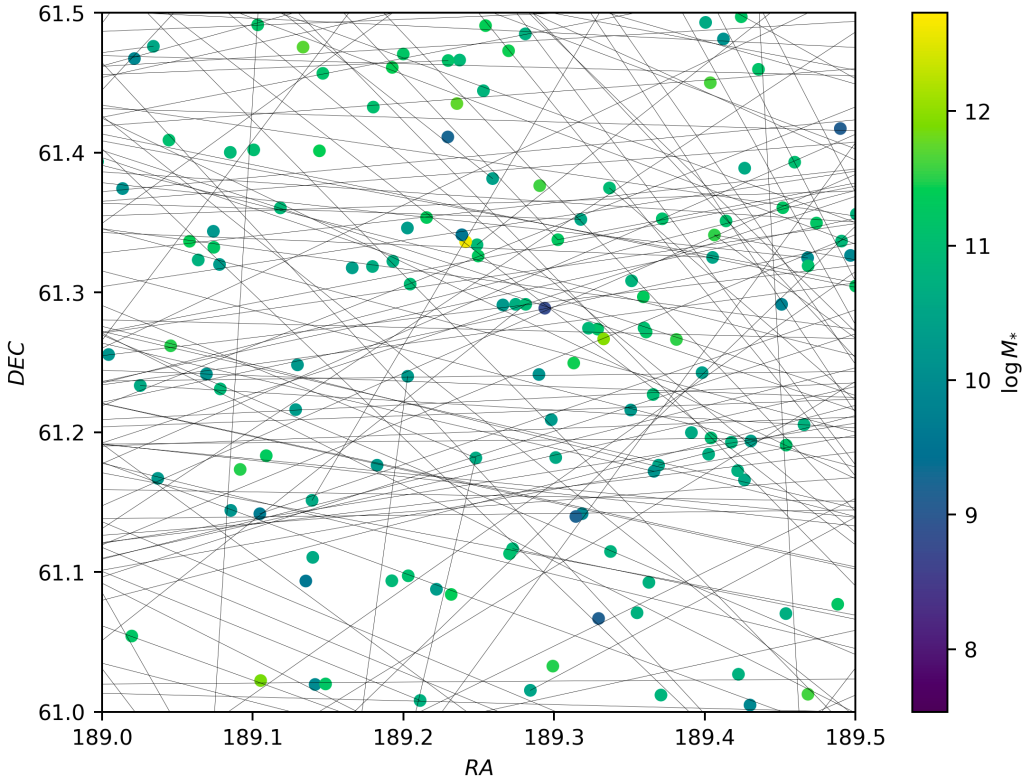
Weighted graph: when edges are assigned some weights which represent cost, distance, etc.

Our Galaxy graph

Each node represents a galaxy interconnected with neighbors within a distance of $\mathbf{10}\ \boldsymbol{Mpc}$, establishing a cosmic web and encapsulating essential information.

The edge weights were determined based on the distance between each galaxy.


r < 10 Mpc
d (Mpc)
ra (deg)
dec (deg)
log M_*
r < 10 Mpc
d (Mpc)
RA (deg)
DEC (deg)
log M_*



Average degree: 1.265

Max degree: 14

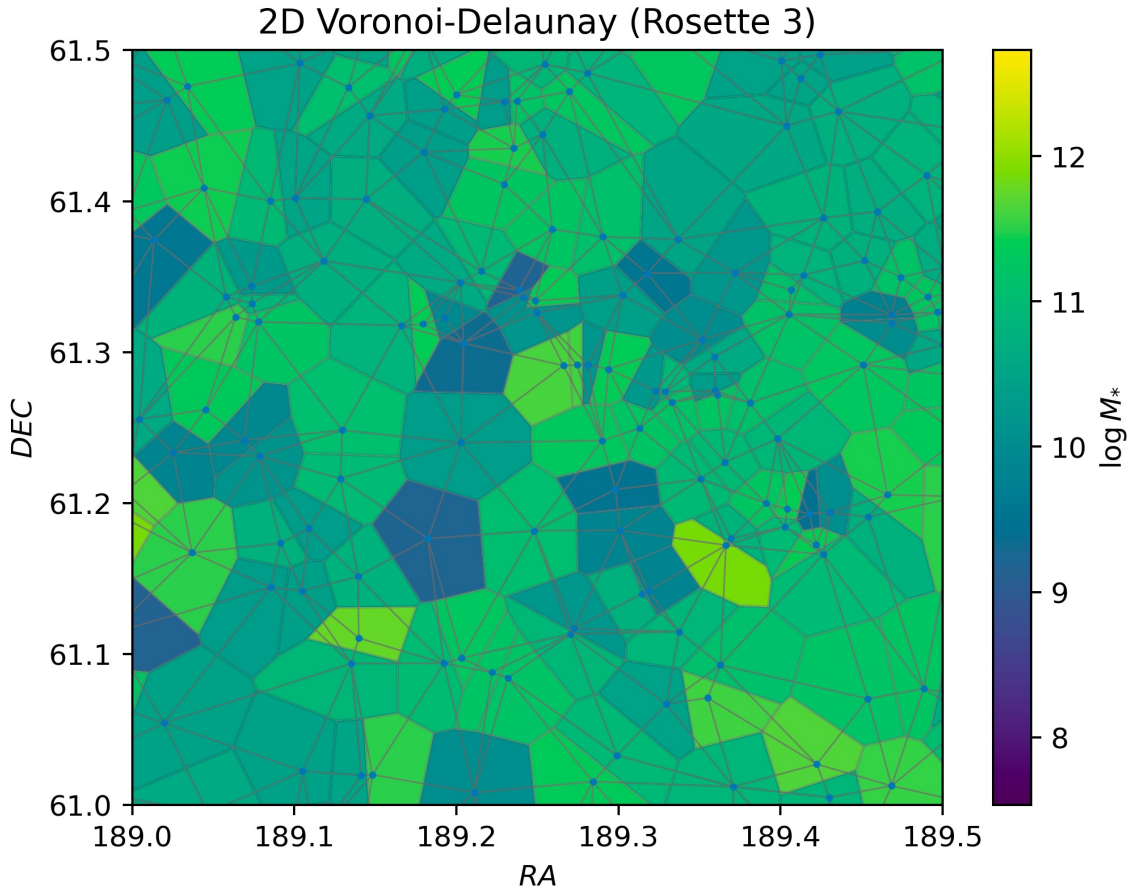
Min degree: 0

Density: 0.00011

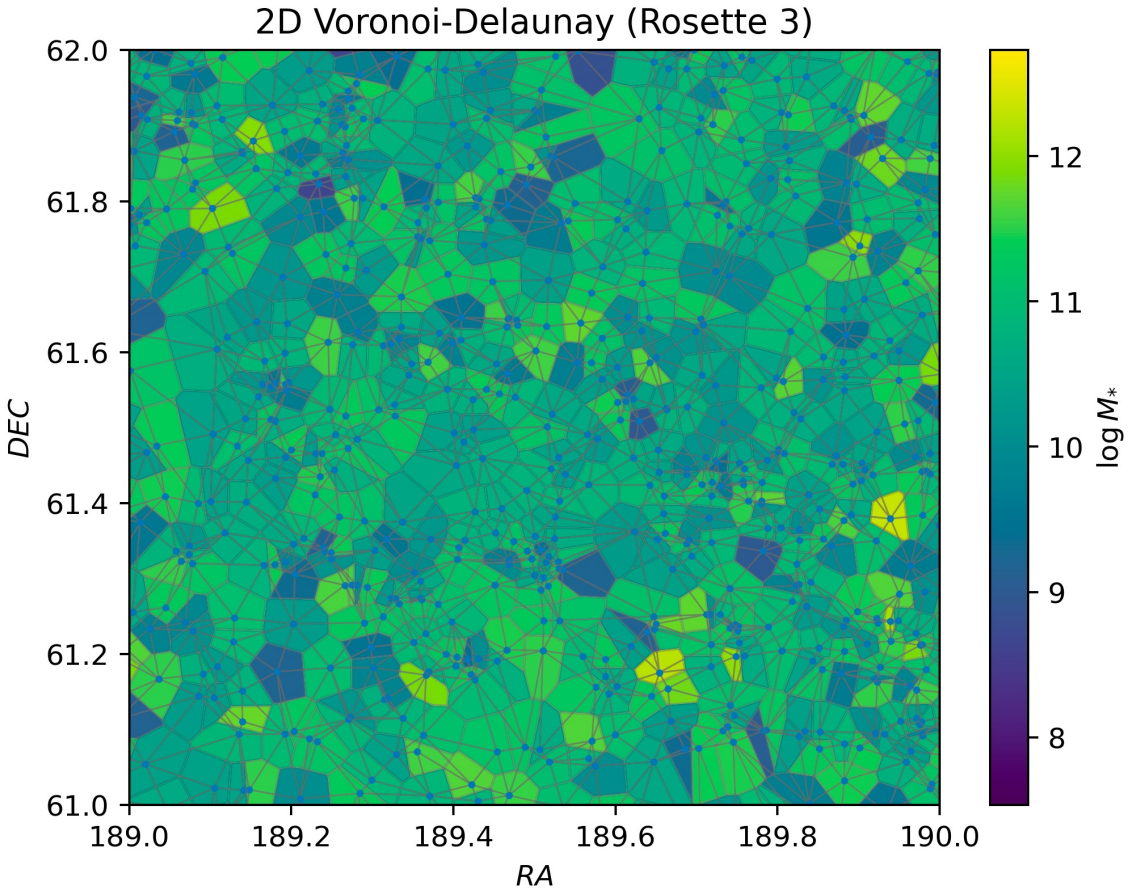
Has self-loops: False

Directed: False

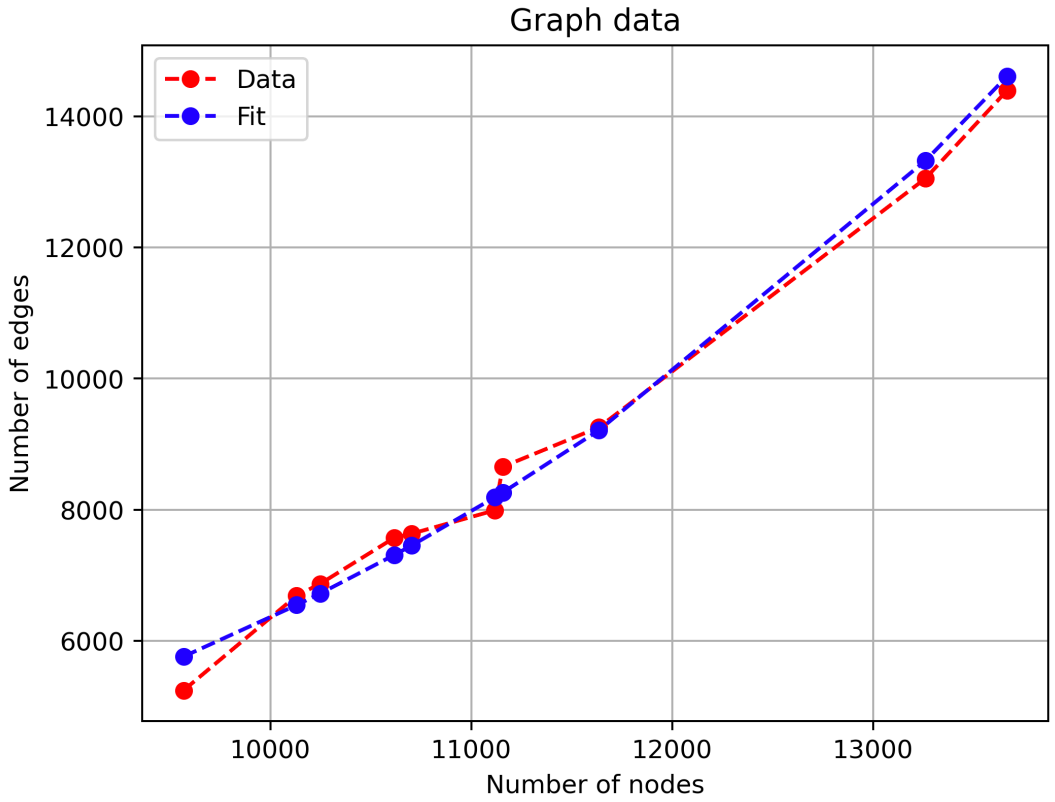
The Voronoi-Delaunay diagram shows the connectivity and density of galactic cells



Regions of smaller Voronoi cells may imply higher galactic density, potentially indicating dense cosmic structures



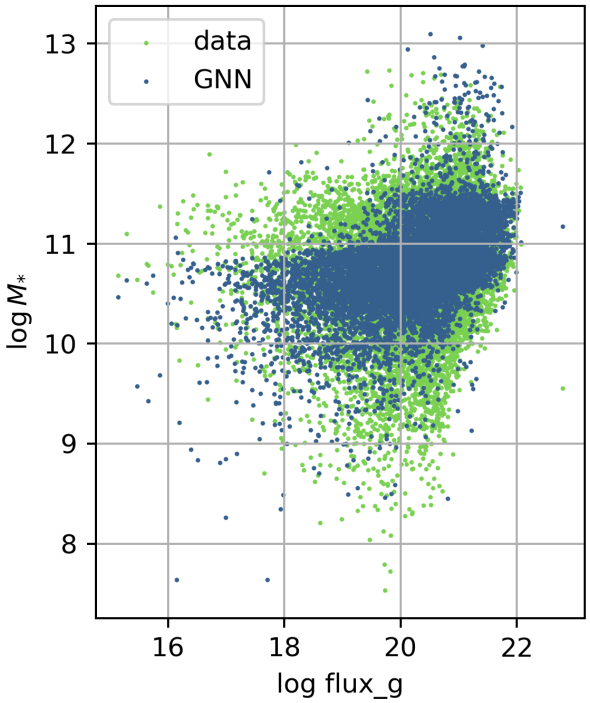
Rosette	*Nodes*	*Edgdes*
3	11157	7572
6	9568	5245
7	11635	9257
11	13667	13051
12	10617	6870
13	13260	14395
14	10704	7635
15	10130	8655
18	11117	7991
19	10248	6689

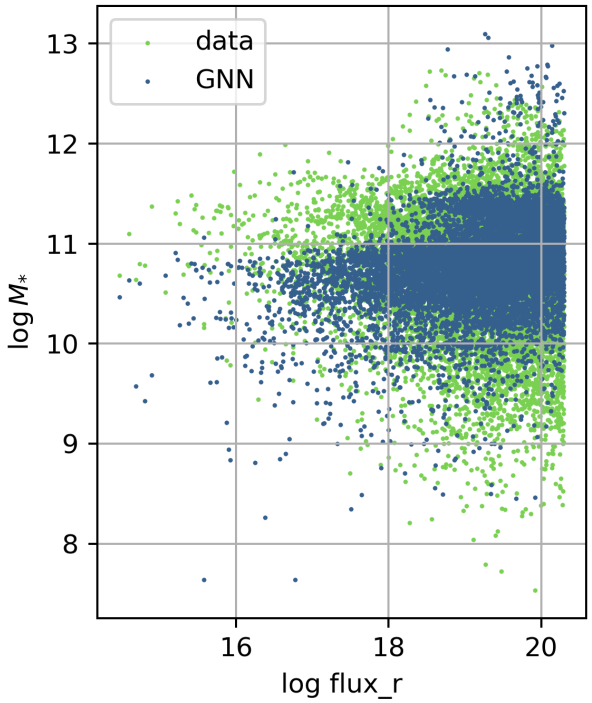


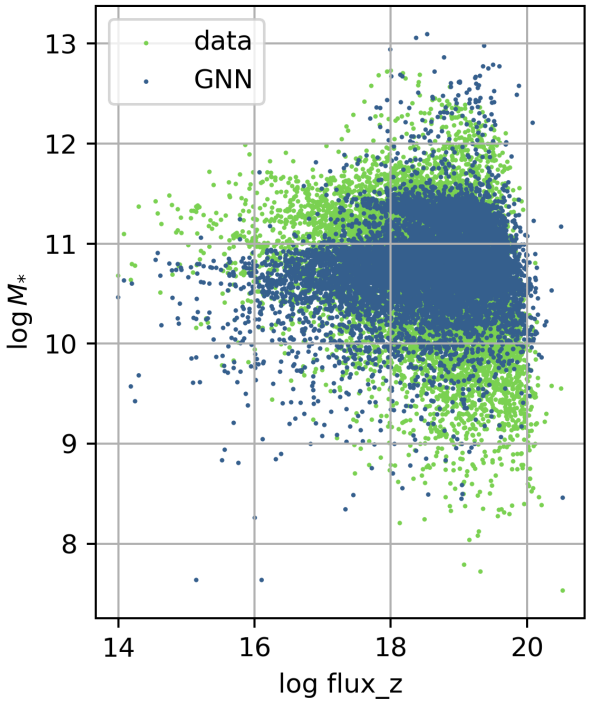
Our GNN model

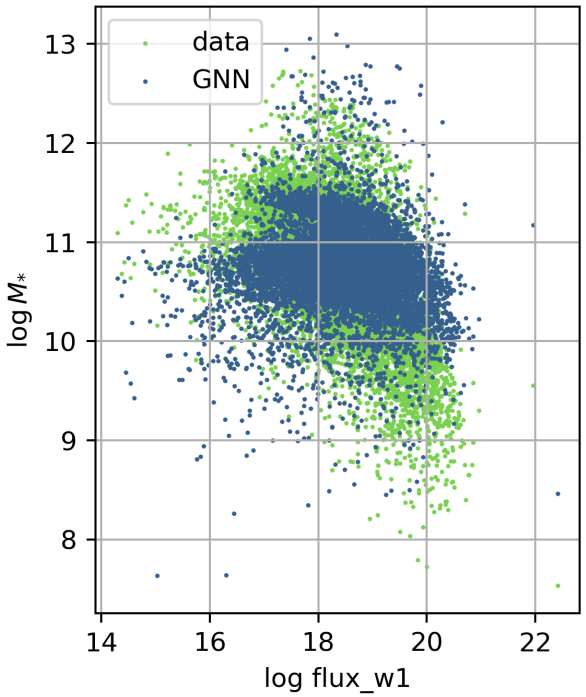
Some training results on a pretty basic model

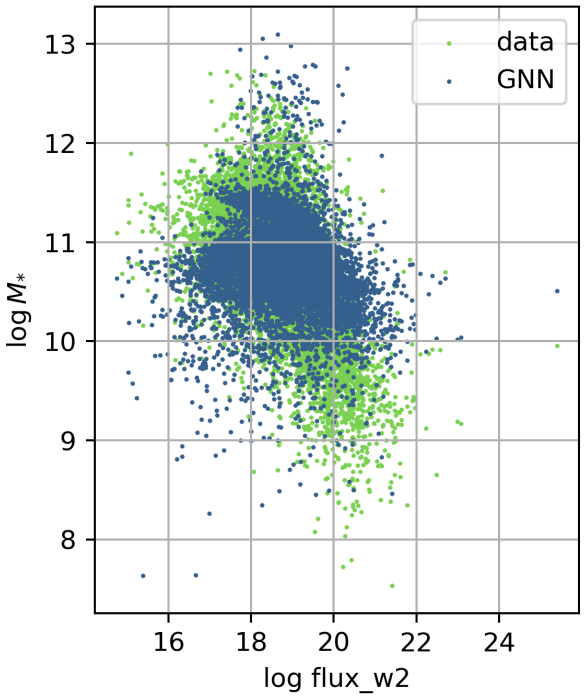
$\boldsymbol{Rosette\ number}$	$\boldsymbol{mse}$
$3\ (train)$	0.2661

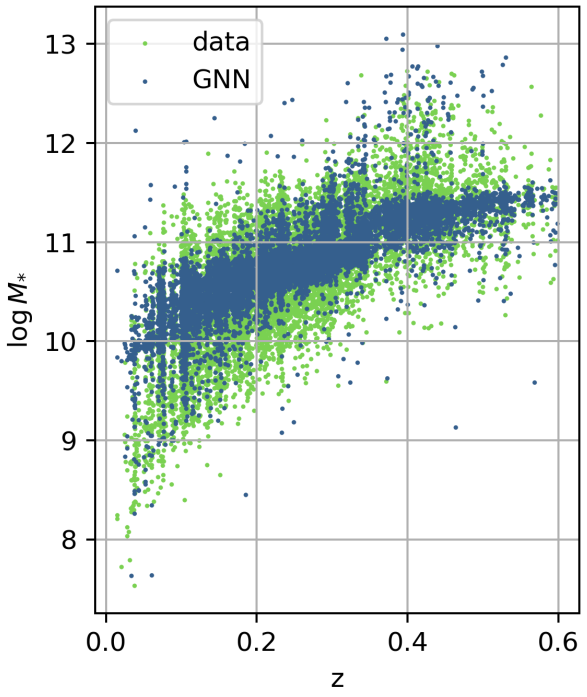

data
GNN
log M_*
log flux_g
13
12
11
10
9
8
16
18
20
22


data
GNN
log M_*
log flux_r
13
12
11
10
9
8
16
18
20


data
GNN
log M_*
log flux_z
13
12
11
10
9
8
14
16
18
20


data
GNN
log M_*
log flux_w1
13
12
11
10
9
8
14
16
18
20
22


data
GNN
log M_*
log flux_w2
13
12
11
10
9
8
16
18
20
22
24


data
GNN
log M_*
z
13
12
11
10
9
8
0.0
0.2
0.4
0.6

Conclussions

The success of the initial simple model has led to endeavors to enhance and develop a more robust neural network, with the aim of improving predictions to potentially surpass the performance of other models.

Our goal is to discern the effectiveness of the GNN model in harnessing spatial relationships encoded in cosmic graphs to enhance precision.

These findings set the stage for a comparative analysis between the Linear Regression, Random Forest, and Graph Neural Network models. Given the inherent graph structure of the galactic dataset, the GNN is anticipated to capture intricate relationships and patterns that may be overlooked by traditional models.

The forthcoming evaluation will center on exploring potentially novel graph-based metrics to assess the performance of predictive models for stellar mass estimation.

Adding a layer of novelty to this research is the utilization of observational data from DESI. In contrast to prior studies relying on simulated data, our approach delivers a more realistic and applicable estimation of stellar masses, providing a substantial advancement in the field of computational astrophysics

References

[1] DESI. (2023). DESI Data Early Data Release (EDR). https://data.desi.lbl.gov/doc/releases/edr/

[2] Pablo Villanueva-Domingo and Francisco Villaescusa-Navarro. Learning cosmology and clustering with cosmic graphs. The Astrophysical Journal, 937(2):115, October 2022.

[3] Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/user_guide.html

[4] Wilson, R. (1996), Introduction to Graph Theory. https://www.maths.ed.ac.uk/~v1ranick/papers/wilsongraph.pdf