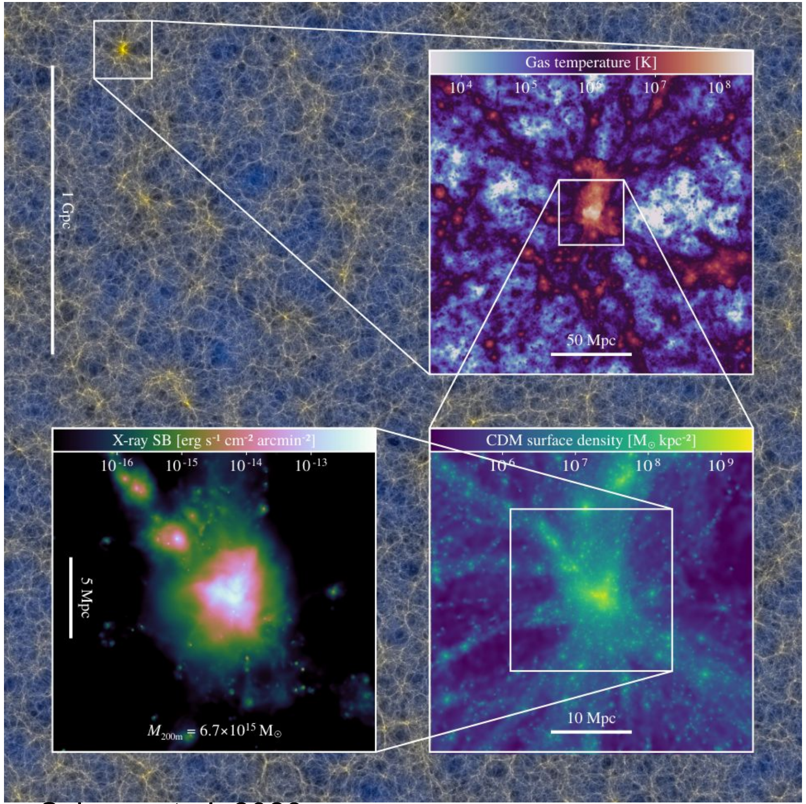
Machine Learning to improve hydrodynamic simulation resolution

Elliot Scott

Newcastle University

Background



Schaye et al. 2023

- Most matter is dark matter, which tends to clump together due to gravity
- Baryonic matter tends to reside in those clumps
- The properties (amount, density, etc.) of normal matter is related to the properties of the dark matter e.g. higher dark matter mass correlates with higher stellar mass

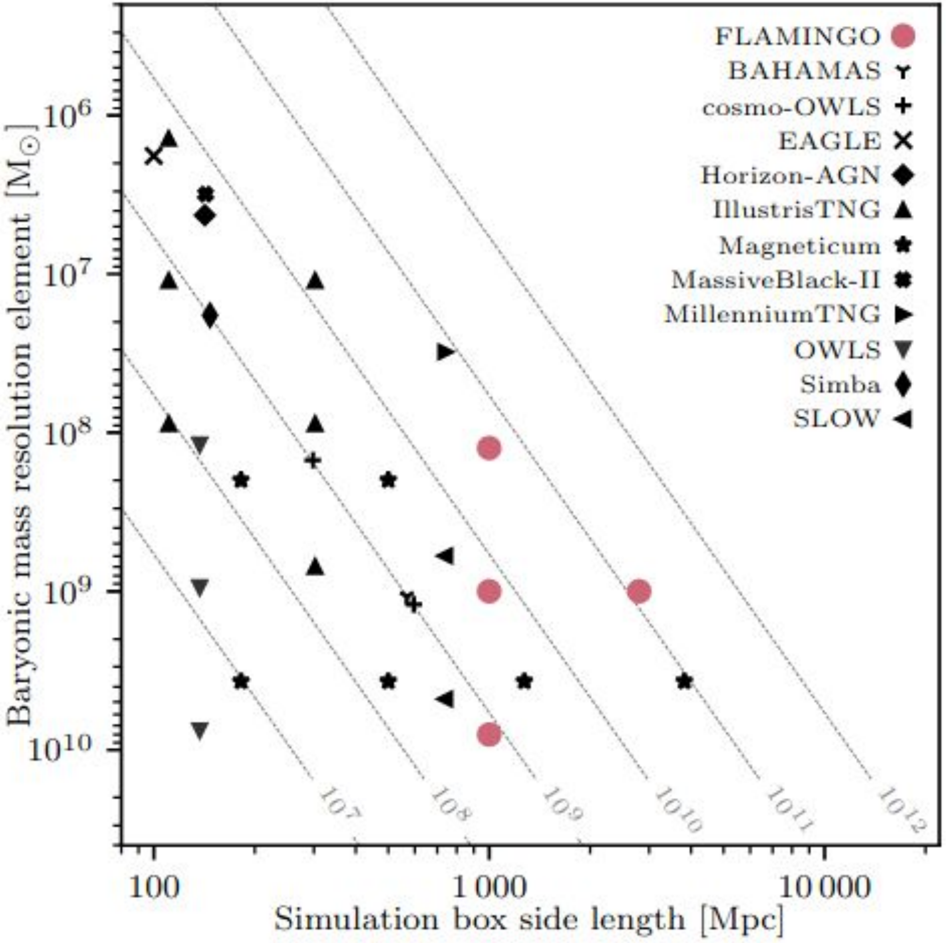
FLAMINGO

- Set of simulations, with both hydrodynamic (dark and normal matter) and dark-matter-only versions.
- Very large (up to 2.8Gpc) but not that high resolution
- Has been calibrated to match the stellar mass function and the gas mass fraction in observations

FLAMINGO Project; flamingo.strw.leidenuniv.nl

FLAMINGO

- FLAMINGO exceeds comparable simulations in terms of number of particles
- Lower resolution than comparable simulations but larger box size

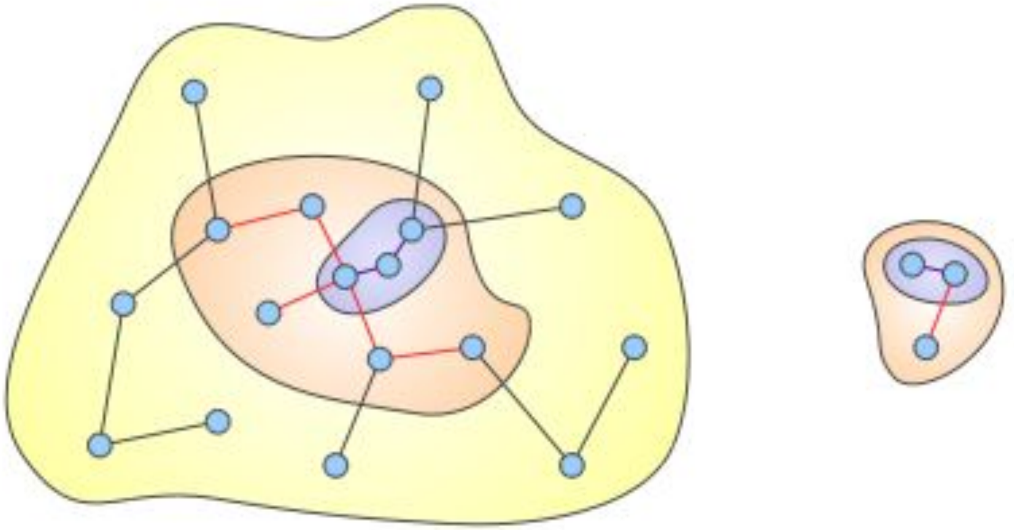


Schaye et al. 2023, MNRAS, 526, 4978

Particle Simulations to Halo Catalogues

Halos are identified using 6D Friends-of-Friends algorithm

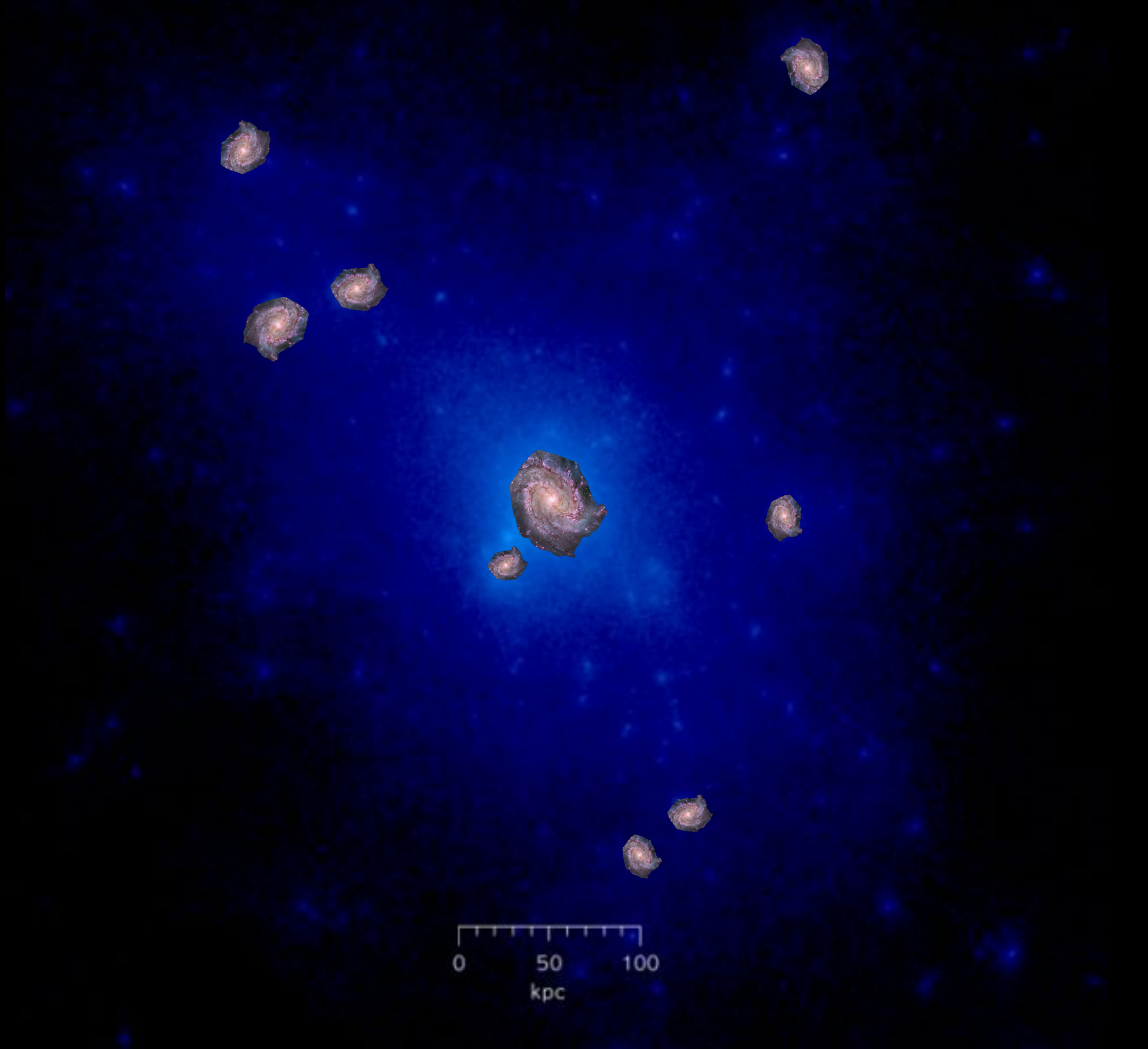
This identifies groups of particles which are similar in 6-dimensional phase space



cosmosim.org

Painting Galaxies into Halos

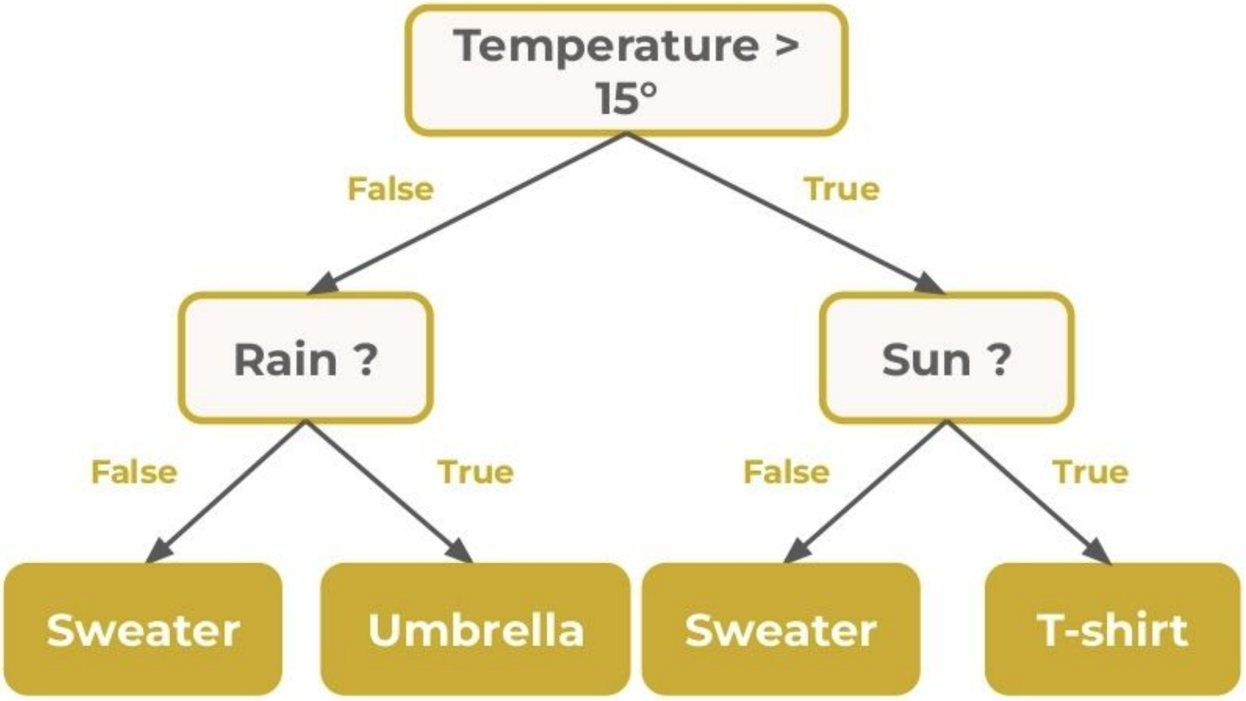
Painting Galaxies into Halos



Decision Trees

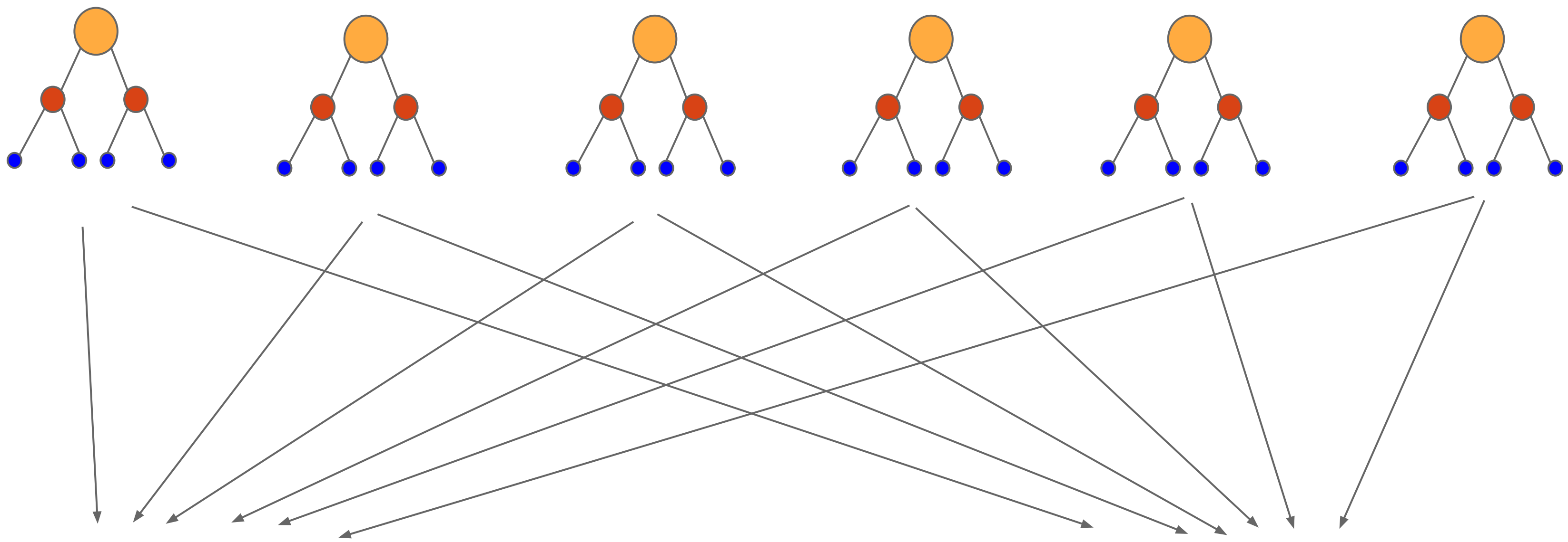
Decision trees are a way of categorising data

The split points and features to split on are optimised using the training data



inside-machinelearning.com

Random Forest



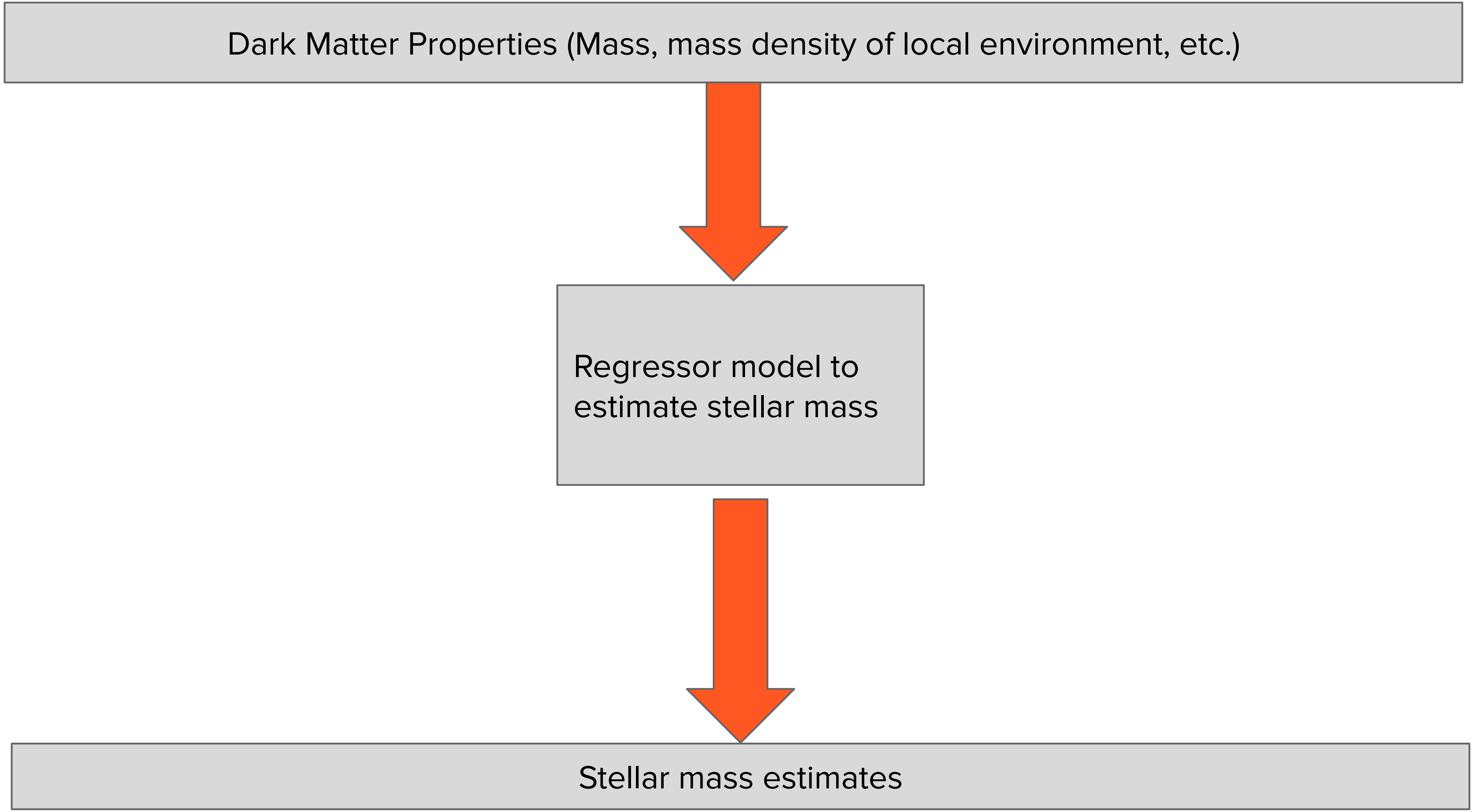
Regression:
Mean of all trees

Classification:
Modal class choice

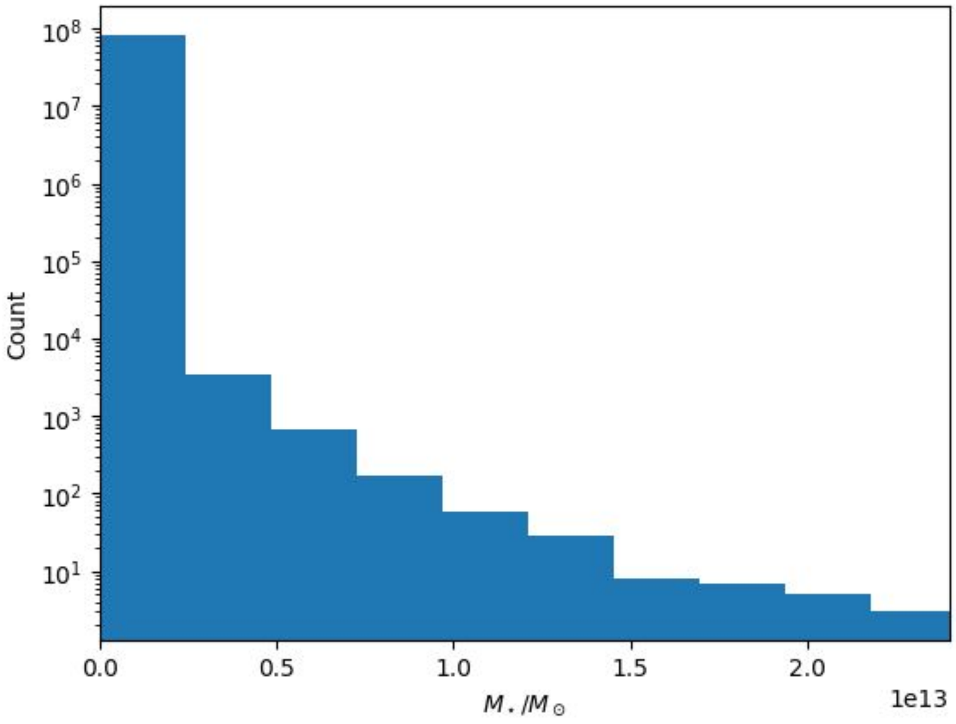
Model Architecture

Dark Matter Properties (Mass, mass density of local environment, etc.)

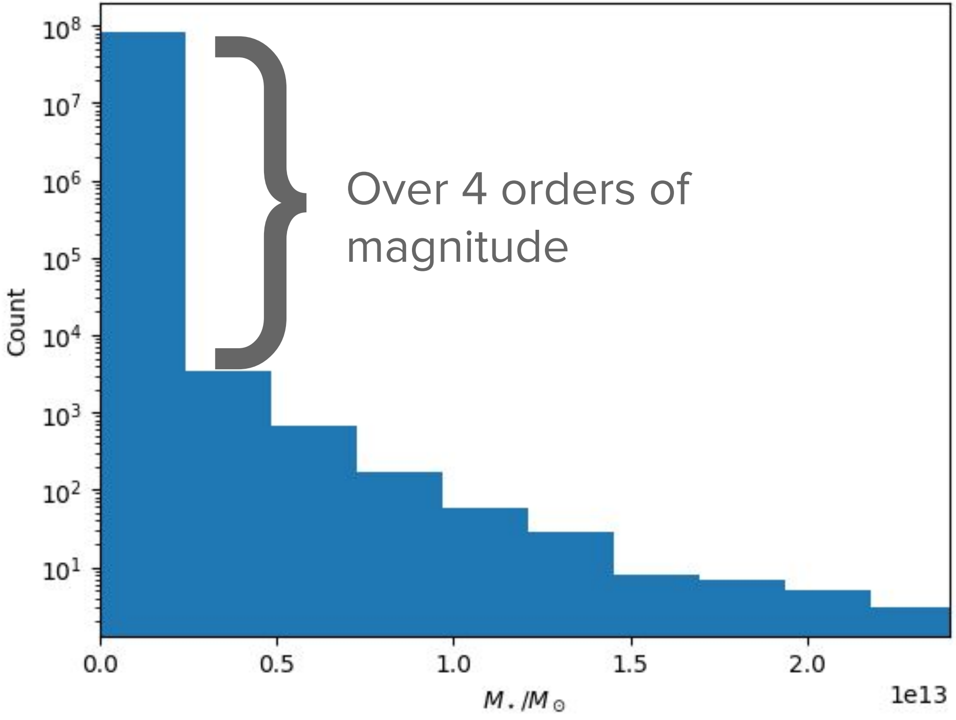
Naive Model Architecture



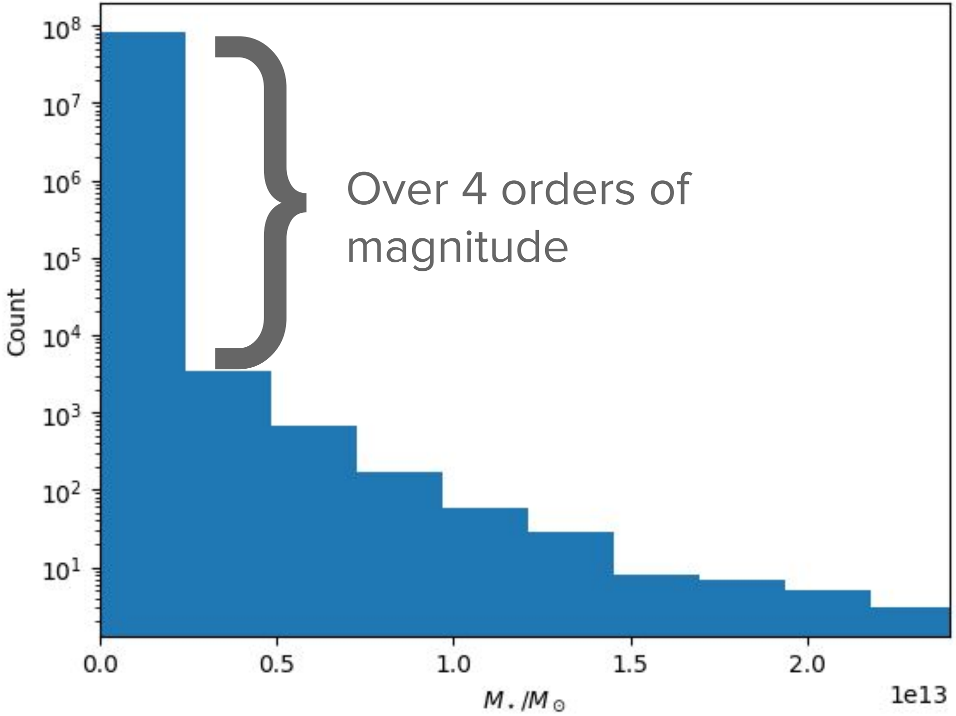
Stellar Mass Function



Stellar Mass Function

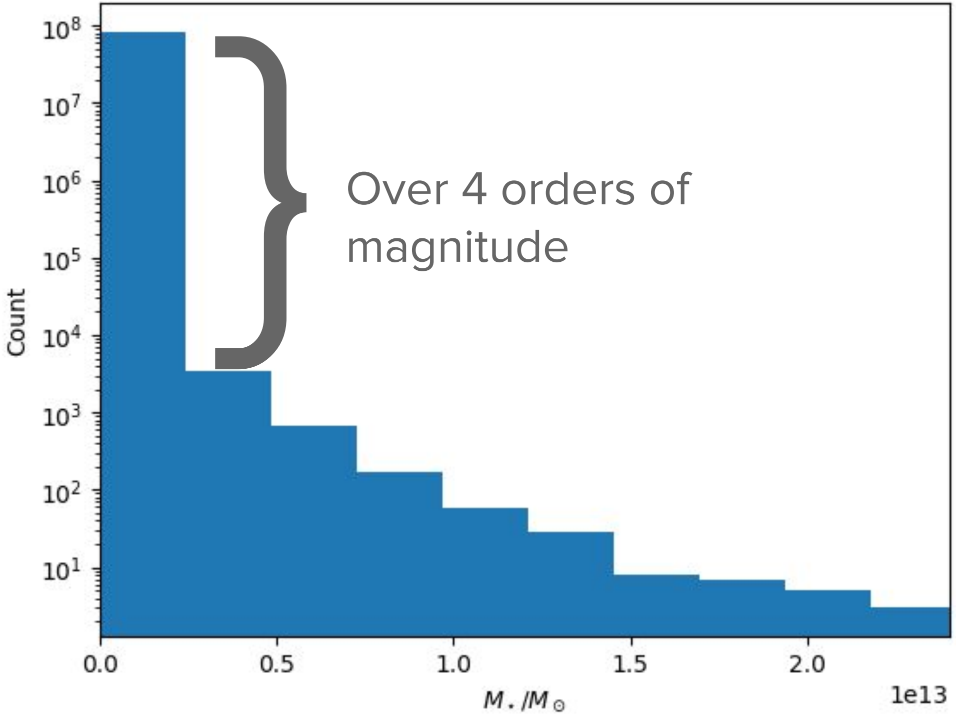


Stellar Mass Function



1. The subhalo would not be expected to have any stellar mass

Stellar Mass Function



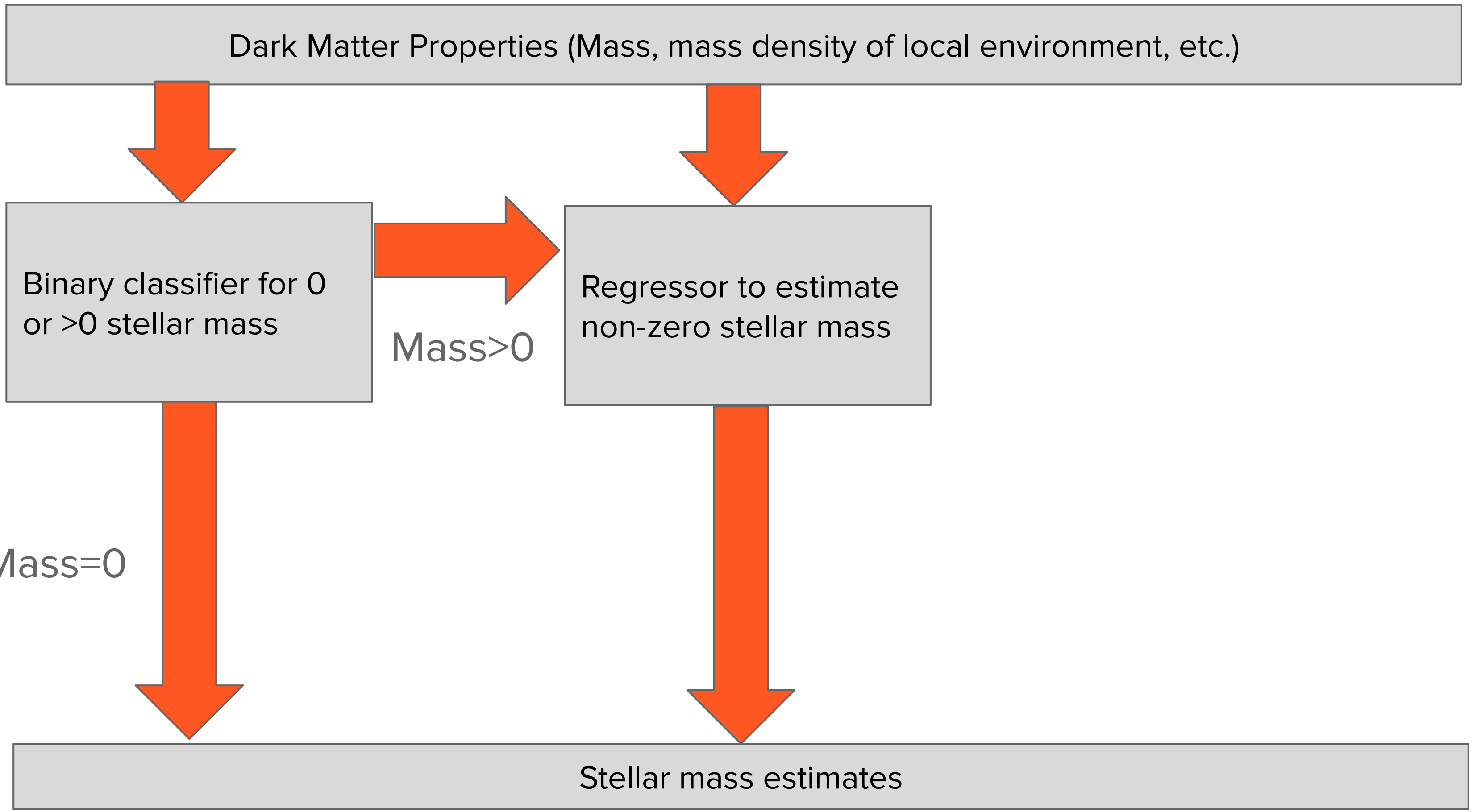
1. The subhalo would not be expected to have any stellar mass
2. The subhalo would have stellar mass but an amount less than the resolution limit of the simulations

Less Naive Model Architecture

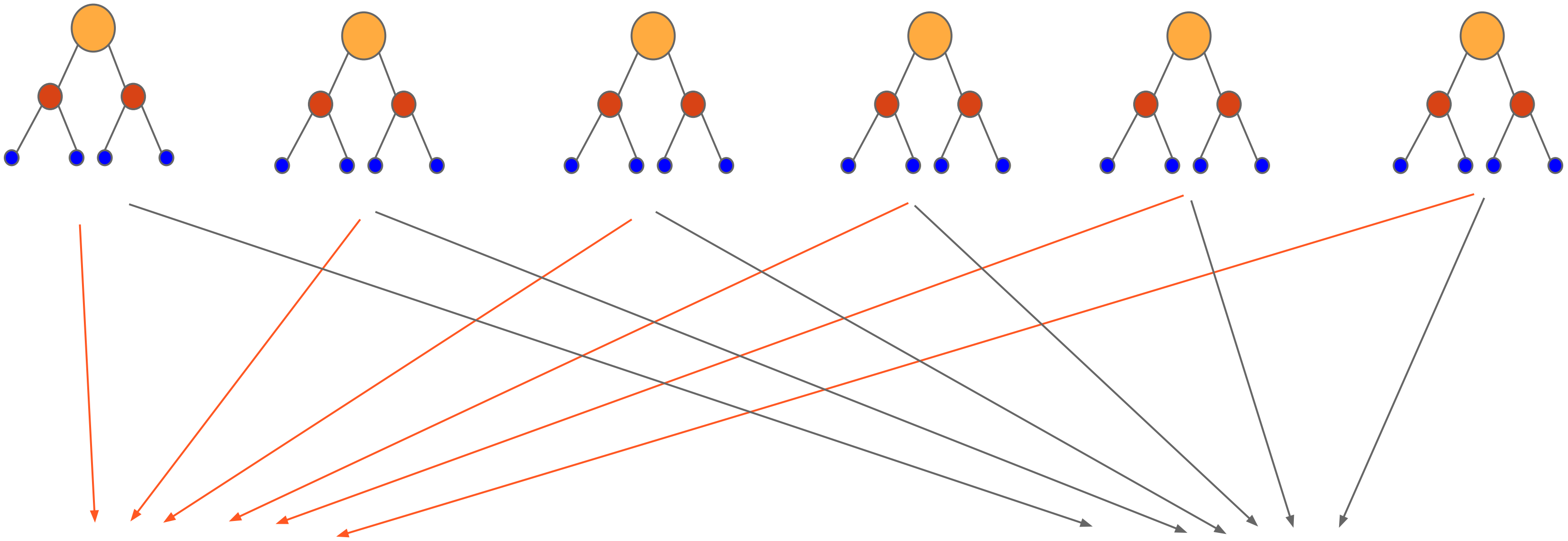
Dark Matter Properties (Mass, mass density of local environment, etc.)

Binary classifier for 0 or >0 stellar mass

Less Naive Model Architecture



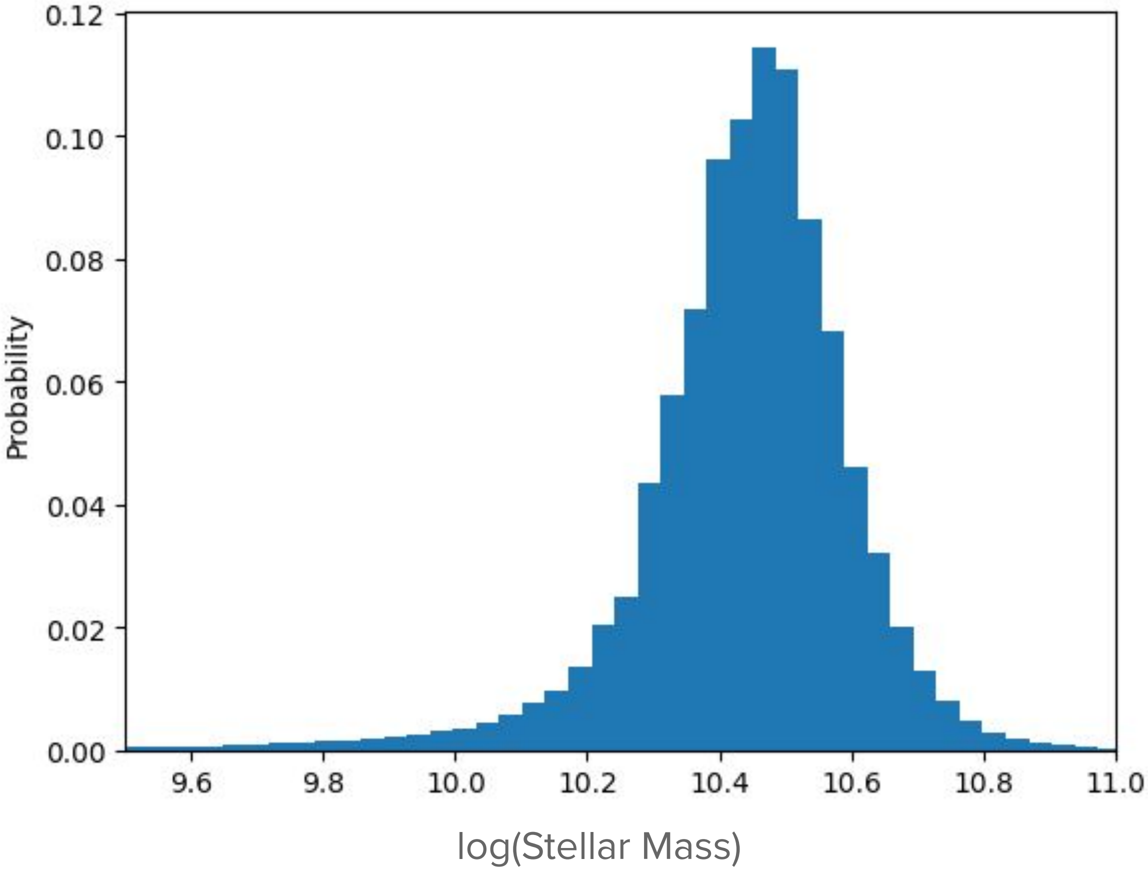
Random Forest Regression



Regression:
Mean of all trees

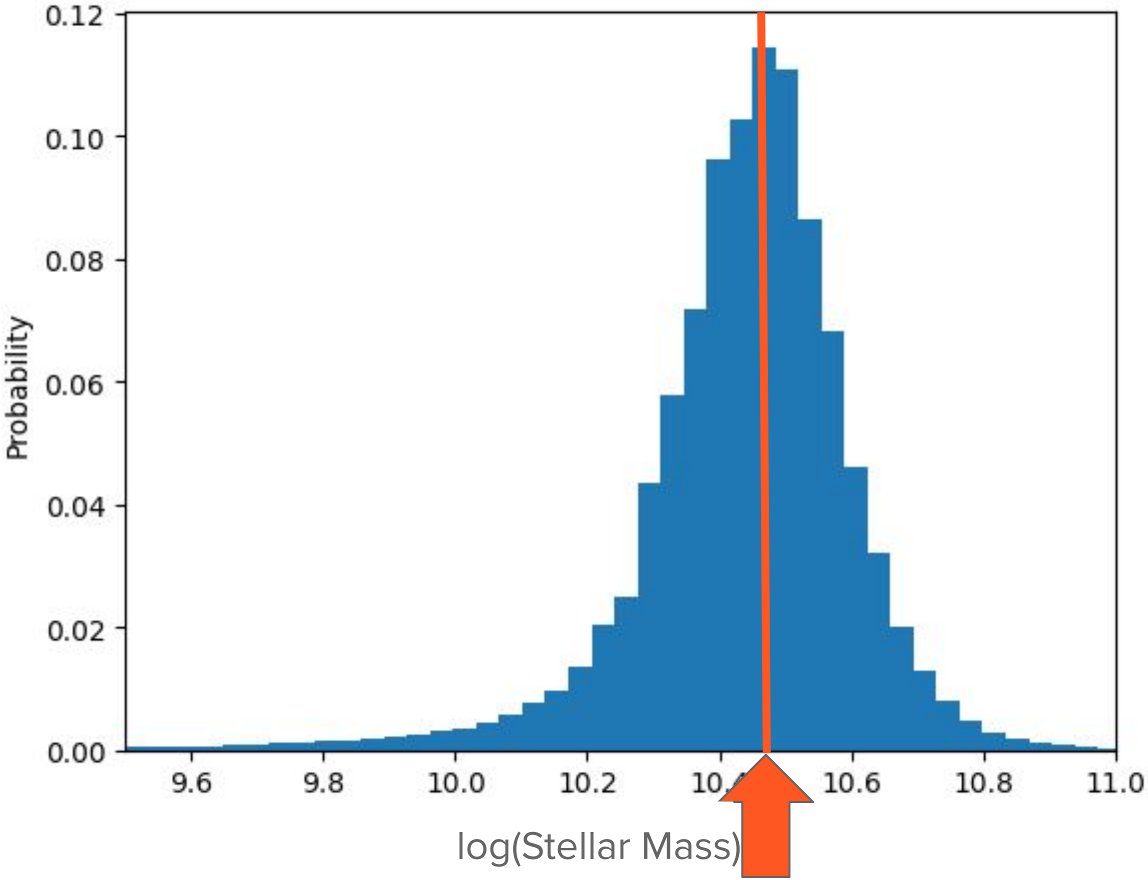
Classification:
Modal class choice

Random Forest Regression

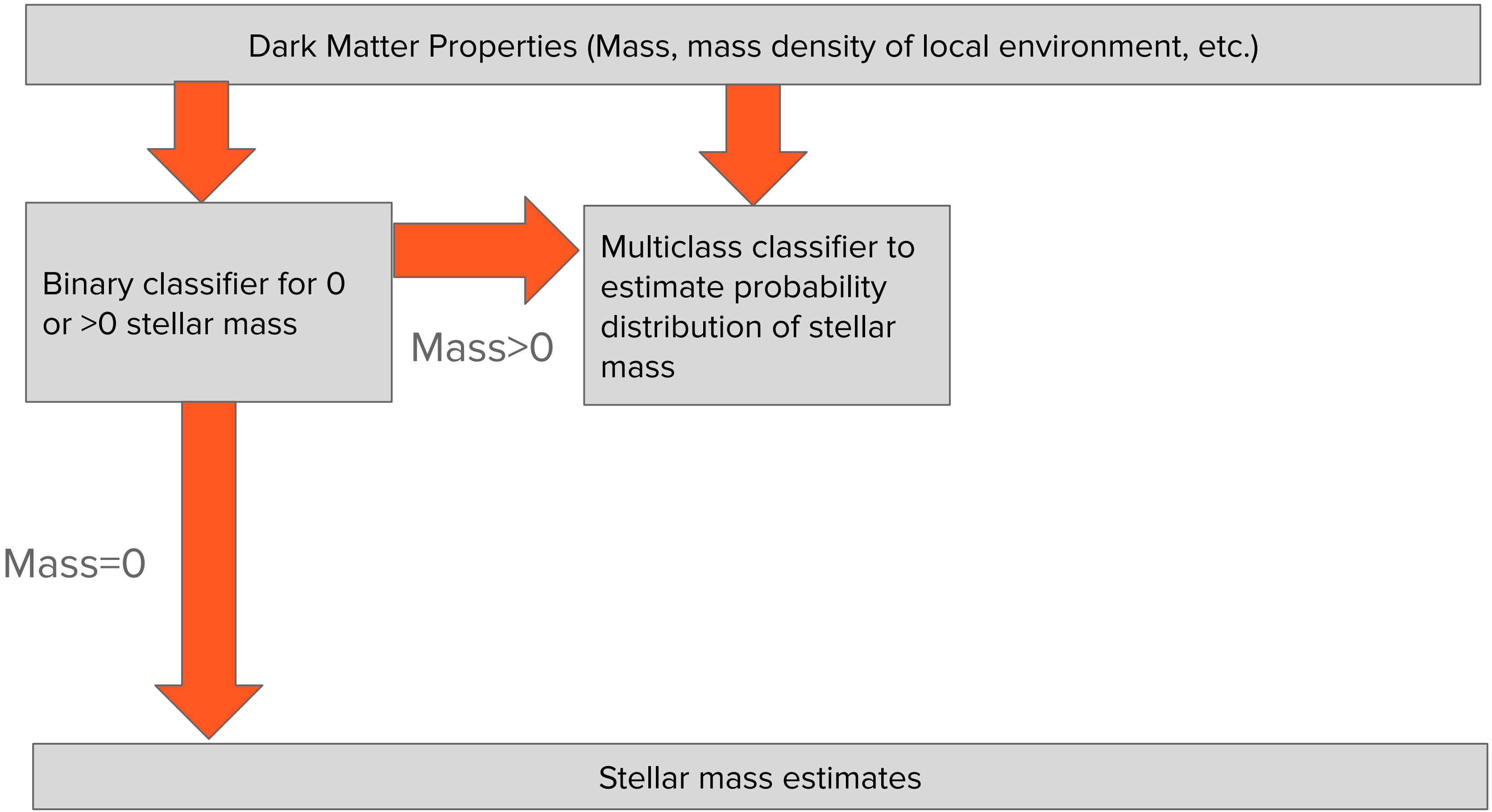


Random Forest Regression

The mean of the estimates of the decision trees will give an estimate of the mean of the probability distribution



Least Naive Model Architecture



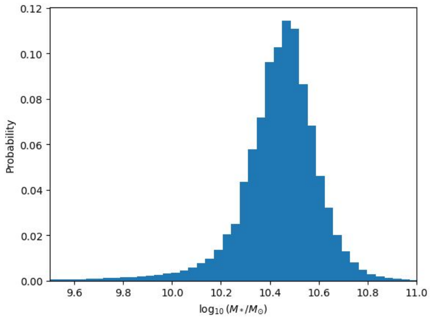
Least Naive Model Architecture

Dark Matter Properties (Mass, mass density of local environment, etc.)

Binary classifier for 0 or >0 stellar mass

Mass>0

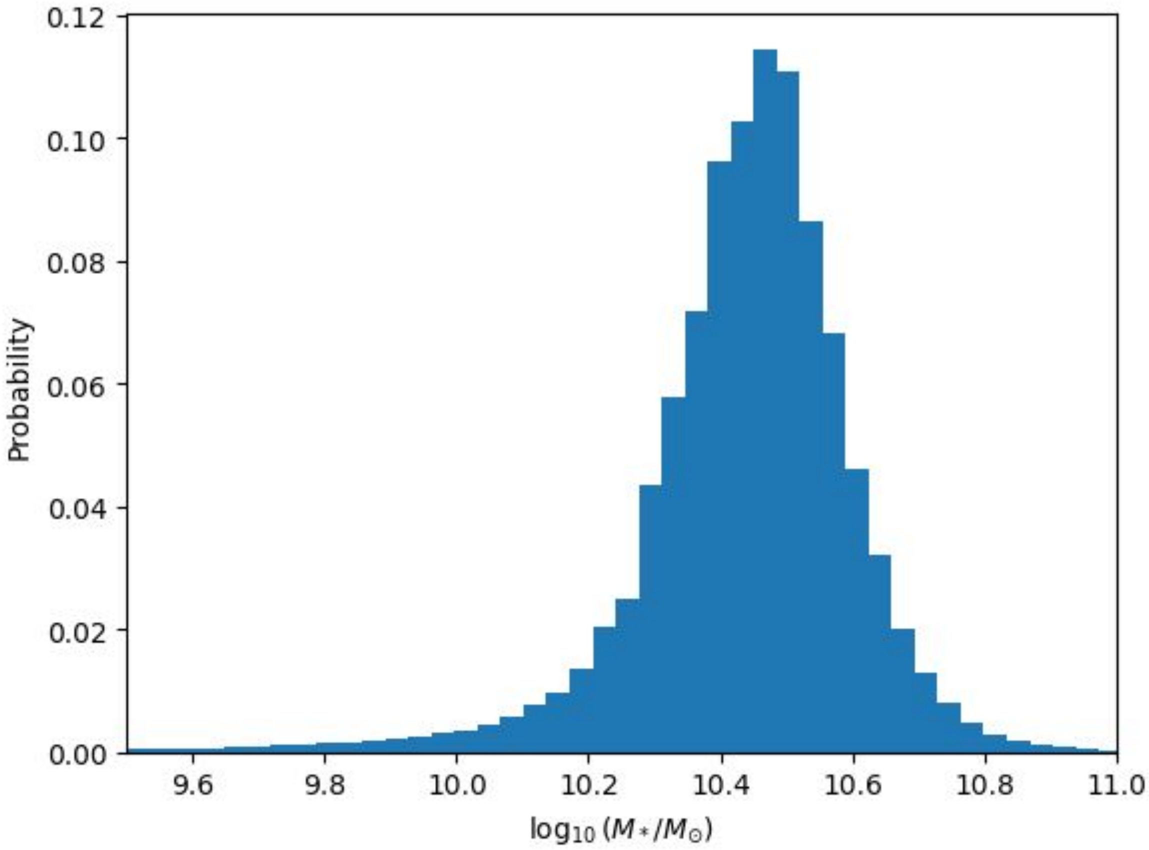
Multiclass classifier to estimate probability distribution of stellar mass



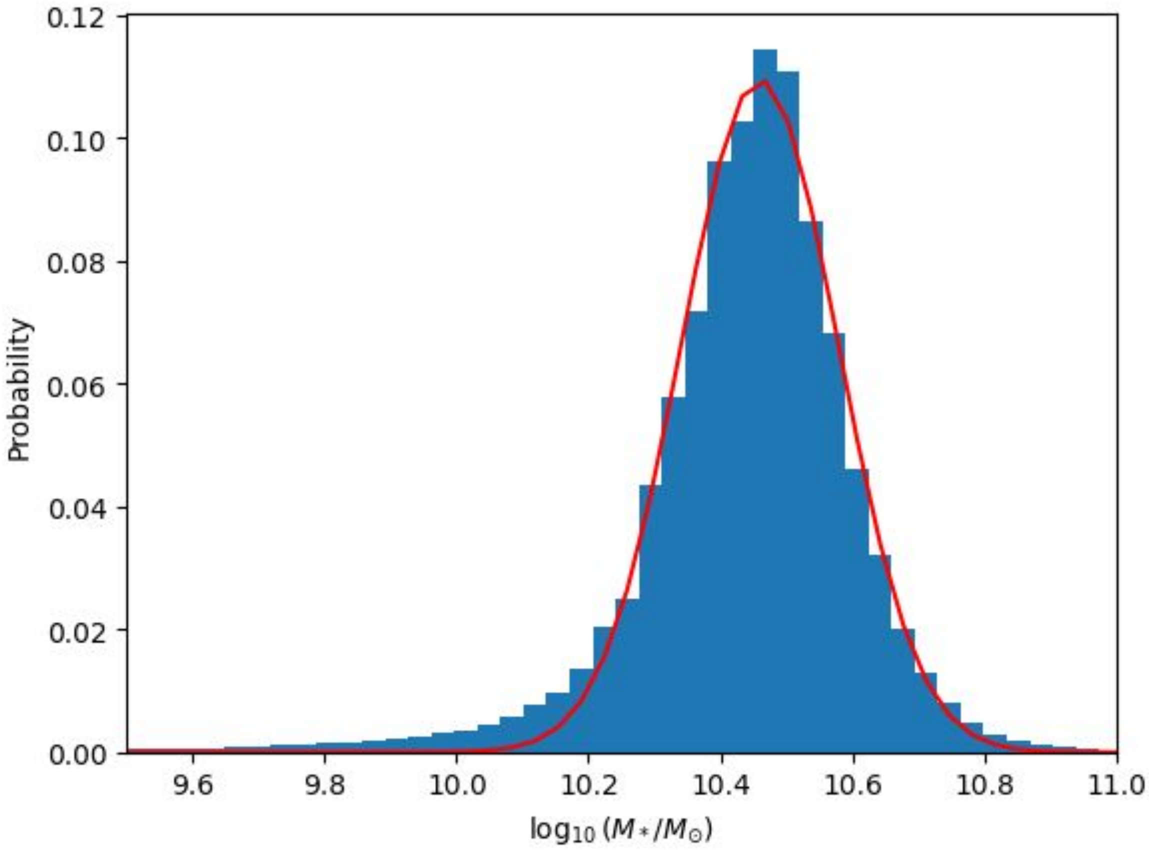
Mass=0

Stellar mass estimates

Model Architecture



Model Architecture



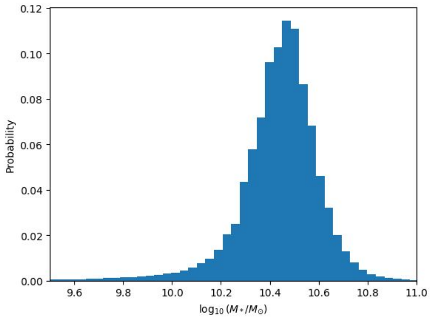
Final Model Architecture

Dark Matter Properties (Mass, mass density of local environment, etc.)

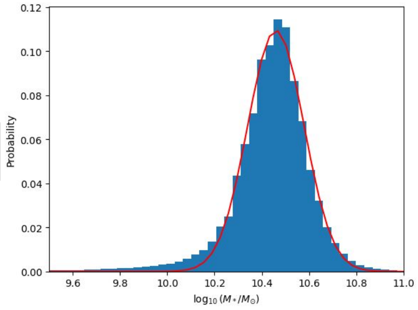
Binary classifier for 0 or >0 stellar mass

Mass>0

Multiclass classifier to estimate probability distribution of stellar mass



Fit continuous gaussian distribution which can be sampled from

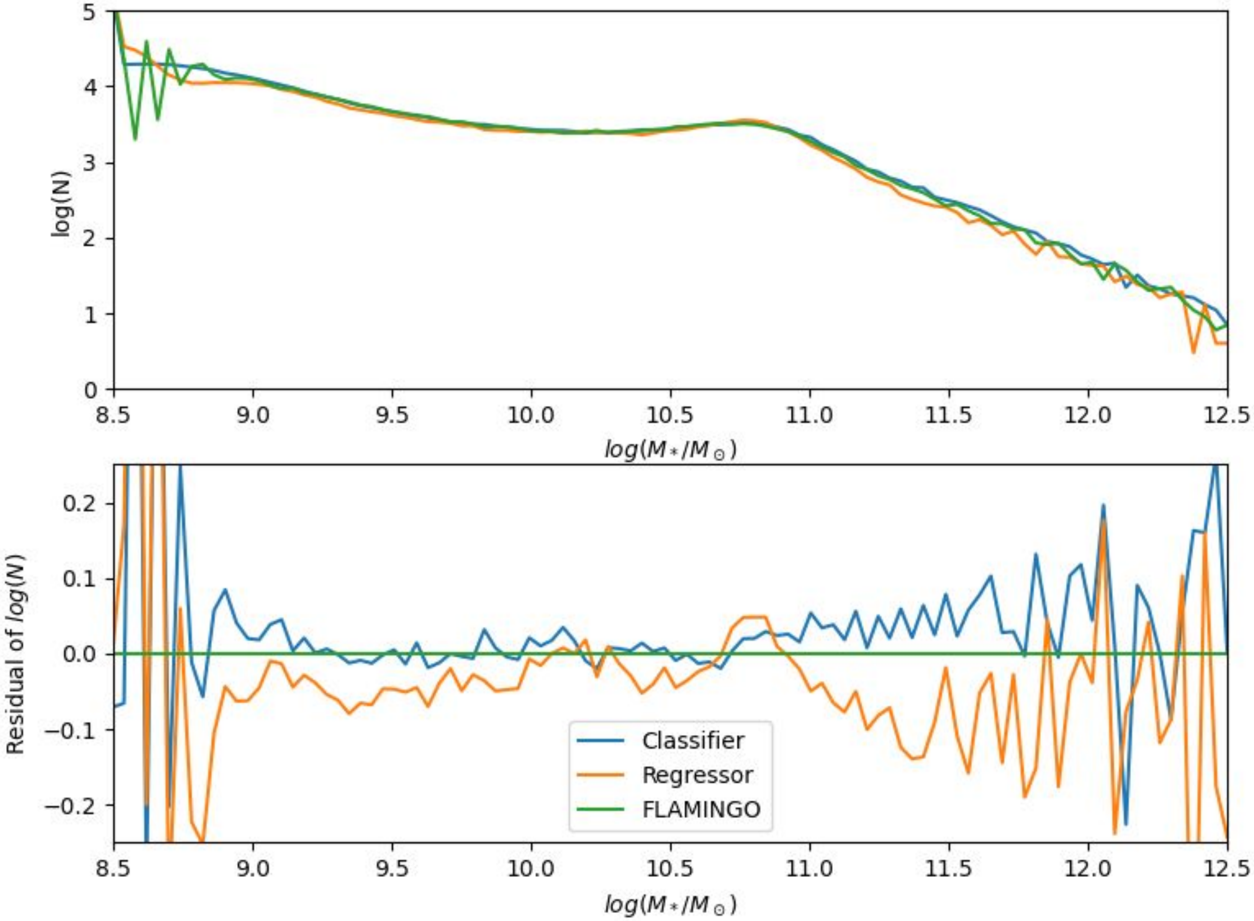


Mass=0

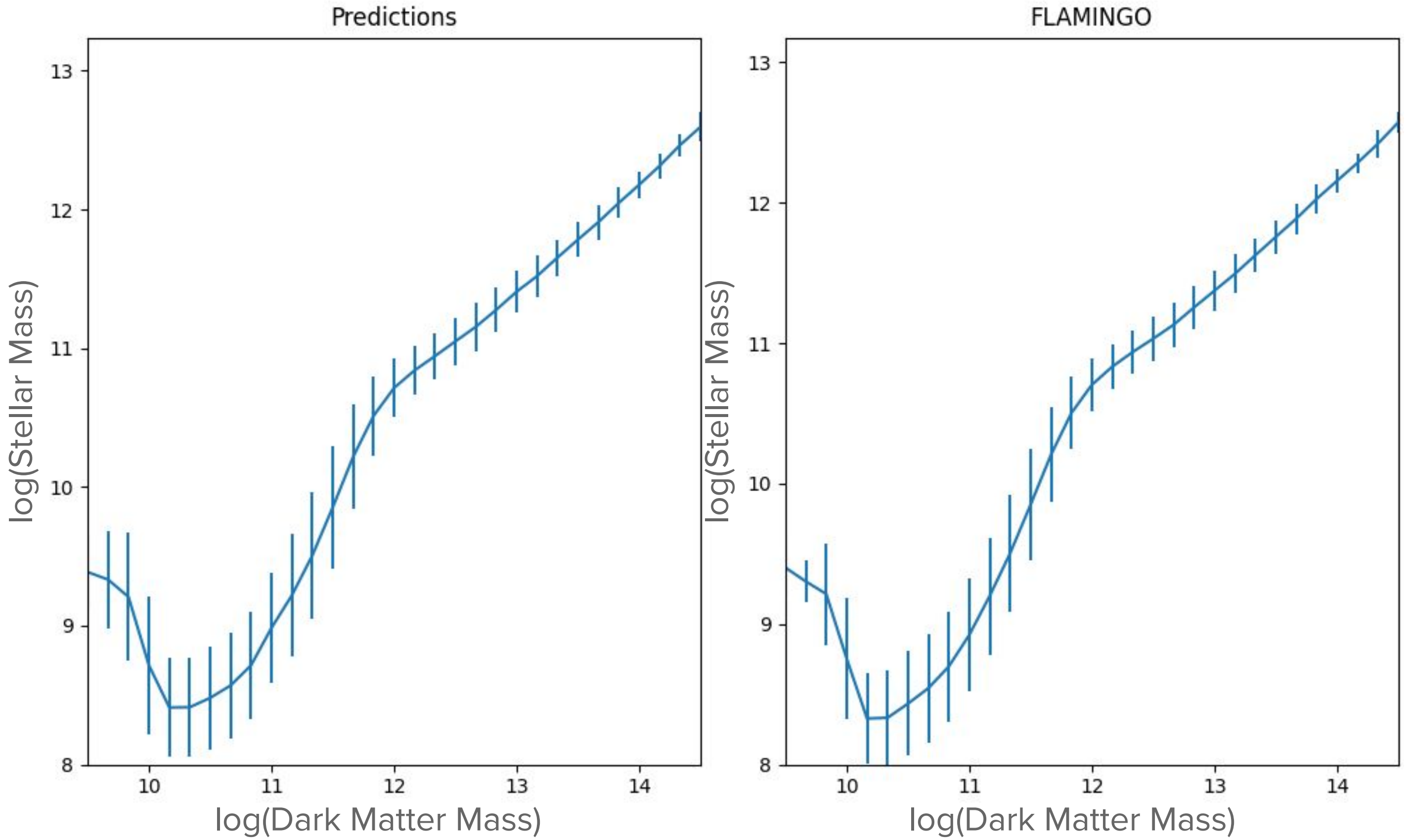
Stellar mass estimates

Stellar Mass Function

Both models show good agreement in regions where there are a large number of well resolved galaxies

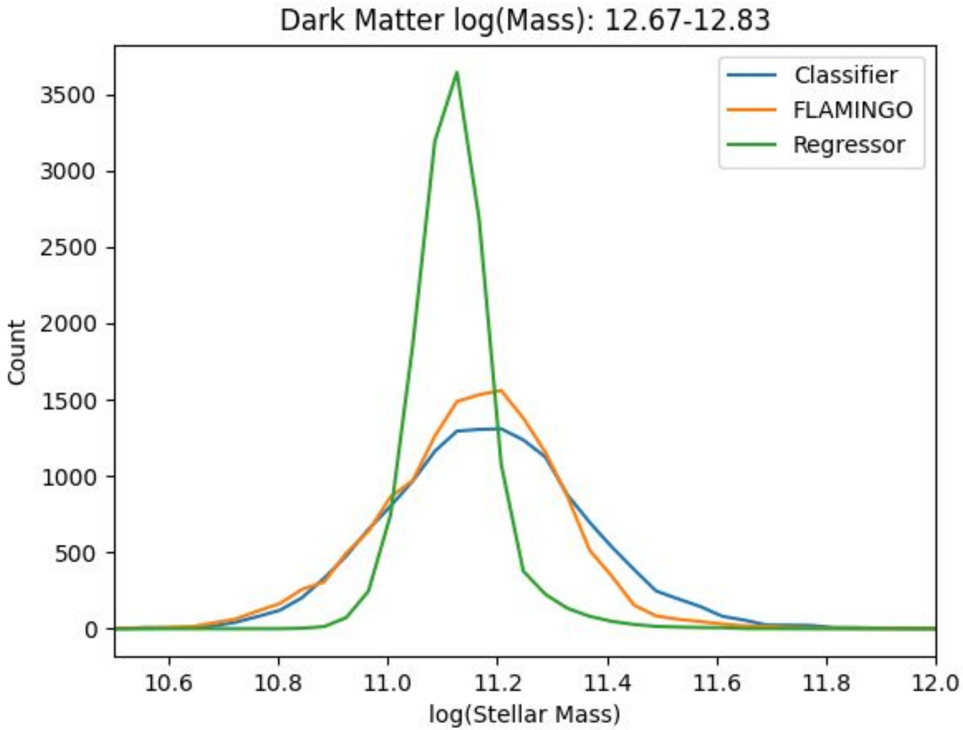


Conditional Stellar Mass Function



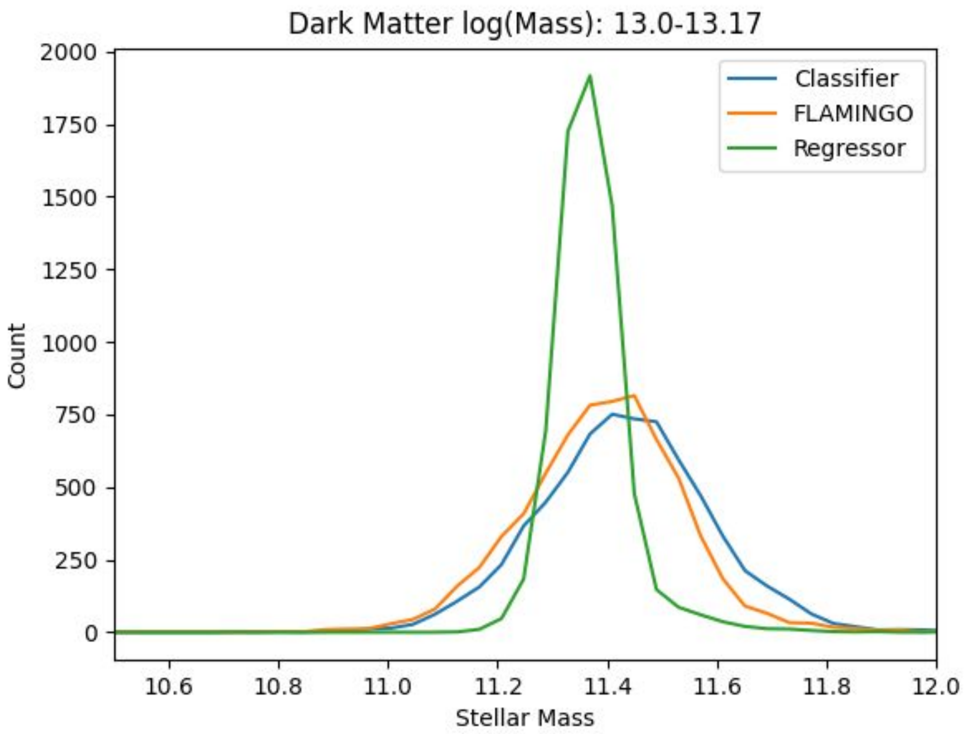
Conditional Stellar Mass Function

Crucially, the classifier model has much closer standard deviations than the regressor

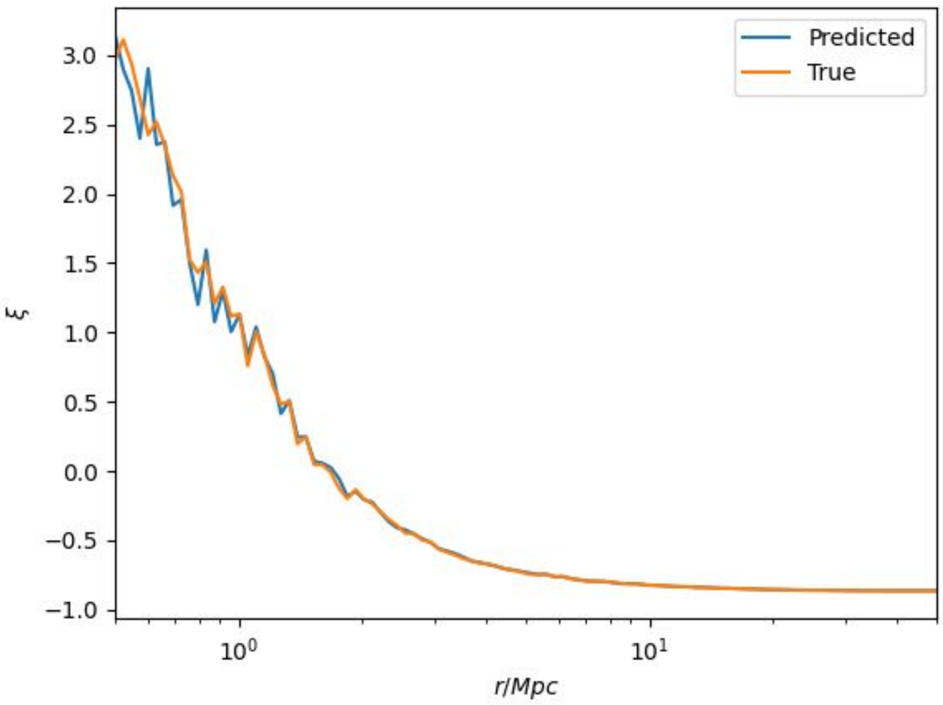


Conditional Stellar Mass Function

Crucially, the classifier model has much closer standard deviations than the regressor

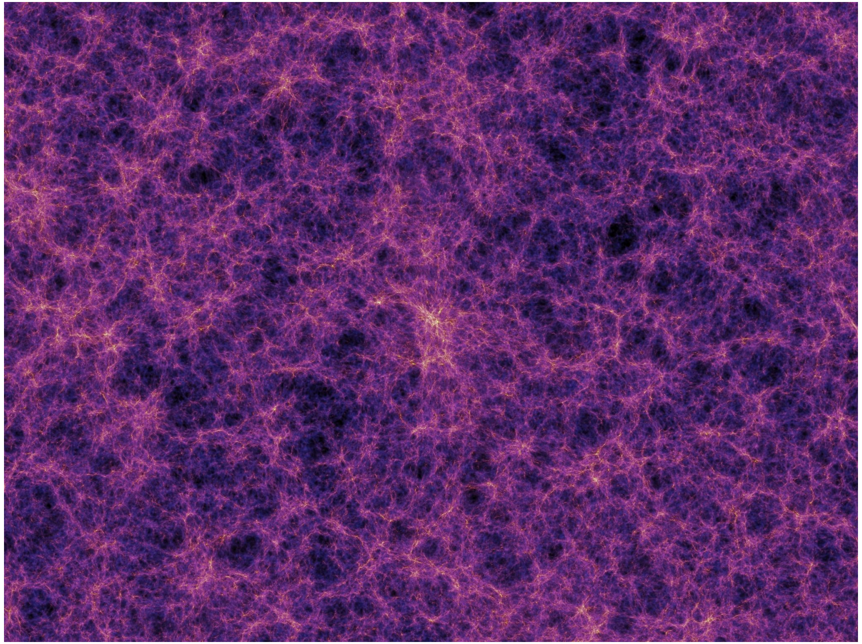


NN-Correlation Function



Similarity in other properties not directly predicted such as the NN-Correlation function

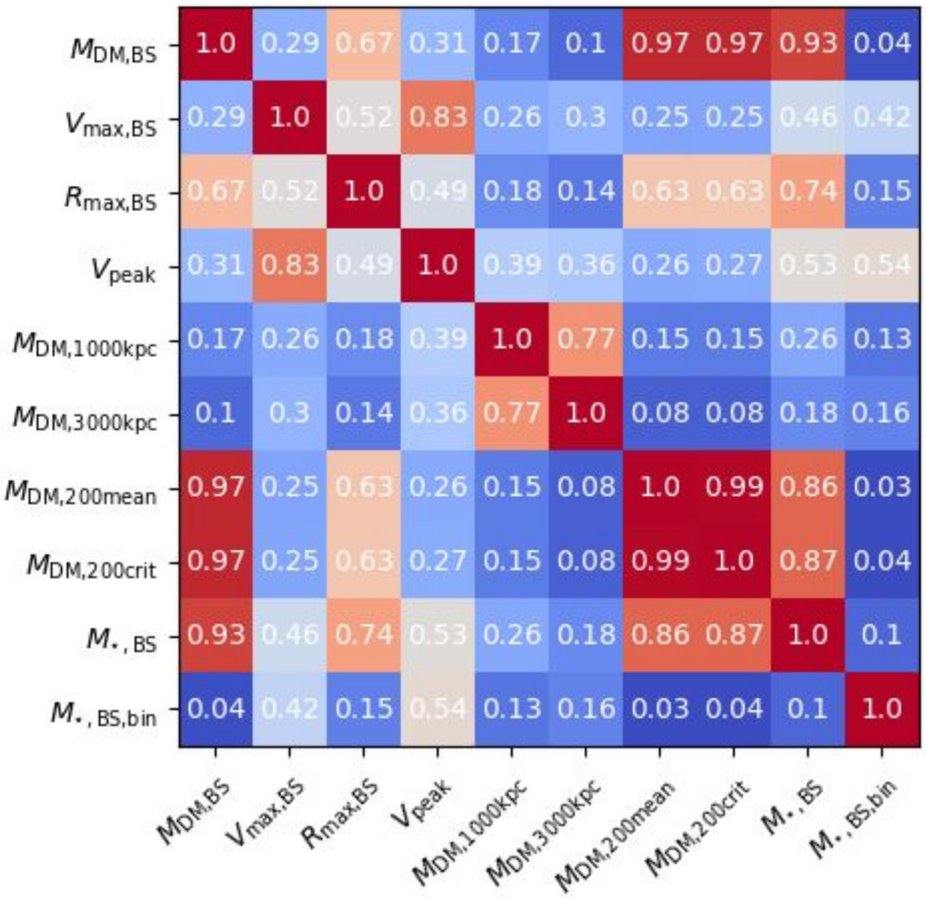
Conclusions



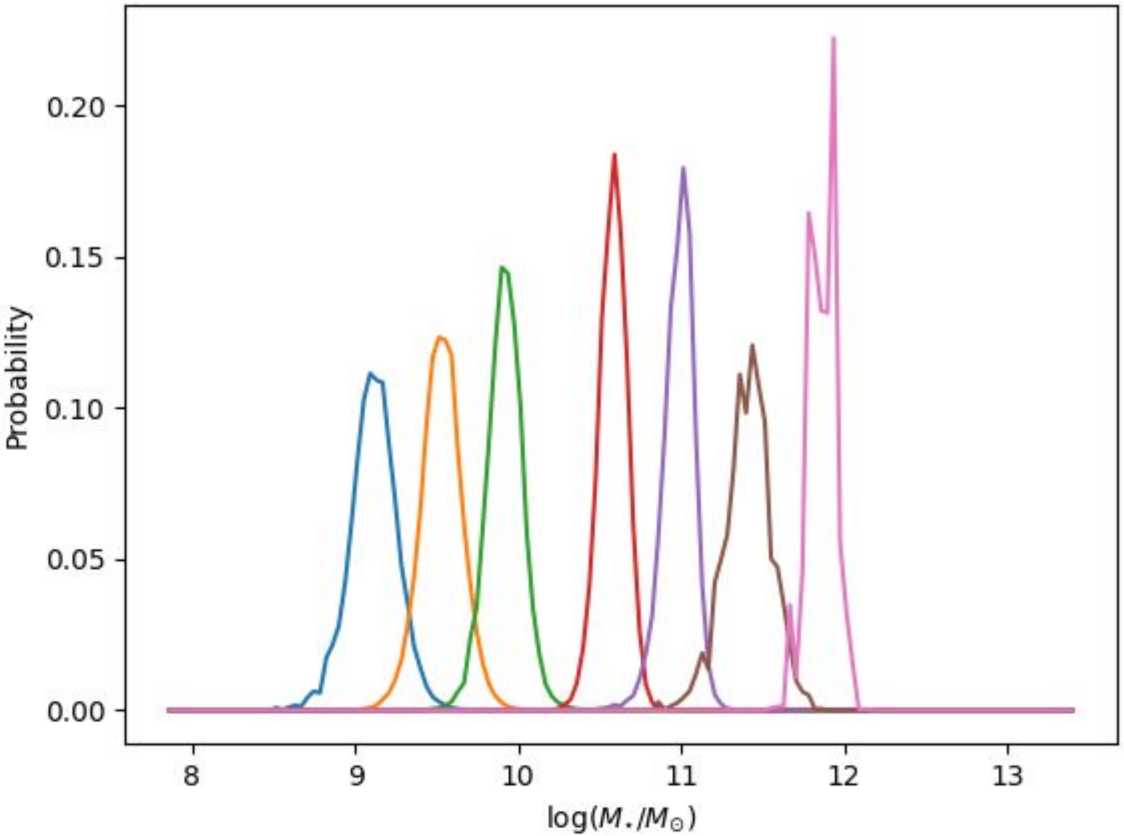
ESA/Hubble

- Machine learning provides an effective method for modelling the galaxy-halo connection
- Bulk properties such as the stellar mass function and NN-Correlation are reproduced effectively
- Estimating the posterior stellar mass distribution and sampling from that helps to preserve the morphology of the conditional stellar mass function

Additional Slides and Contextless Plots



$M_{\mathrm{DM,BS}}$ 1.0 0.29 0.67 0.31 0.17 0.1 0.97 0.97 0.93 0.04
$V_{\mathrm{max,BS}}$ 0.29 1.0 0.52 0.83 0.26 0.3 0.25 0.25 0.46 0.42
$R_{\mathrm{max,BS}}$ 0.67 0.52 1.0 0.49 0.18 0.14 0.63 0.63 0.74 0.15
V_{peak} 0.31 0.83 0.49 1.0 0.39 0.36 0.26 0.27 0.53 0.54
$M_{\mathrm{DM,1000kpc}}$ 0.17 0.26 0.18 0.39 1.0 0.77 0.15 0.15 0.26 0.13
$M_{\mathrm{DM,3000kpc}}$ 0.1 0.3 0.14 0.36 0.77 1.0 0.08 0.08 0.18 0.16
$M_{\mathrm{DM,200mean}}$ 0.97 0.25 0.63 0.26 0.15 0.08 1.0 0.99 0.86 0.03
$M_{\mathrm{DM,200crit}}$ 0.97 0.25 0.63 0.27 0.15 0.08 0.99 1.0 0.87 0.04
$M_{\star,\mathrm{BS}}$ 0.93 0.46 0.74 0.53 0.26 0.18 0.86 0.87 1.0 0.1
$M_{\star,\mathrm{BS,bin}}$ 0.04 0.42 0.15 0.54 0.13 0.16 0.03 0.04 0.1 1.0
$M_{\mathrm{DM,BS}}$ $V_{\mathrm{max,BS}}$ $R_{\mathrm{max,BS}}$ V_{peak} $M_{\mathrm{DM,1000kpc}}$ $M_{\mathrm{DM,3000kpc}}$ $M_{\mathrm{DM,200mean}}$ $M_{\mathrm{DM,200crit}}$ $M_{\star,\mathrm{BS}}$ $M_{\star,\mathrm{BS,bin}}$



0.20
0.15
0.10
0.05
0.00
Probability
8
9
10
11
12
13
$\log(M_\star/M_\odot)$

Parameter	Possible Values	Optimal Value
Number of Features per Tree	2,3,4,5,6	6
Minimum Samples to Split	5,10,20	5
Maximum Depth of Tree	5,10,15	15

	Has Stellar Mass	Does Not Have Stellar Mass
Predicted Stellar Mass	45715278	957289
Predicted No Stellar Mass	1391542	30298637

$$\Delta_\mu = \sqrt{\sum_i n_i(\mu_{i,\mathrm{true}} - \mu_{i,\mathrm{pred}})^2} \qquad (1)$$

$$\Delta_\sigma = \sqrt{\sum_i n_i(\sigma_{i,\mathrm{true}} - \sigma_{i,\mathrm{pred}})^2} \qquad (2)$$

Parameter	Possible Values	Optimal Value
Number of Features per Tree	2,3,4,5,6	6
Minimum Samples to Split	5,10,20	10
Maximum Depth of Tree	5,10,15	15
Number of Stellar Mass Bins	10,30,50,70,90,110,130,150,170,190,210,230	150

Binary Classifier

Input Feature	Importance
M_{DM}	0.08406035
V_{max}	0.13010774
R_{max}	0.00568454
$M_{\mathrm{DM,1Mpc}}$	0.01735433
$M_{\mathrm{DM,3Mpc}}$	0.02279006
V_{peak}	0.74000297

Secondary Classifier

Input Feature	Importance
M_{DM}	0.16416795
V_{max}	0.29223963
R_{max}	0.0171641
$M_{\mathrm{DM,1Mpc}}$	0.02592992
$M_{\mathrm{DM,3Mpc}}$	0.02288563
V_{peak}	0.47761276

Binary Classifier

Metric	Value
Training Accuracy	0.975
Test Accuracy	0.970
Test Precision	0.979
Test Recall	0.970
Test F1 Score	0.974

Secondary Classifier

Metric	Value
Training Accuracy	0.245
Test Accuracy	0.193