A Physically-Motivated Scheme for Matching Galaxies with Dark Matter Halos

Stephanie Tonnesen
Associate Research Scientist
CCA, Flatiron Institute

Kavli IPMU
June 30, 2021

arXiv:2102:13122

JP Ostriker (Princeton, Columbia, CCA)

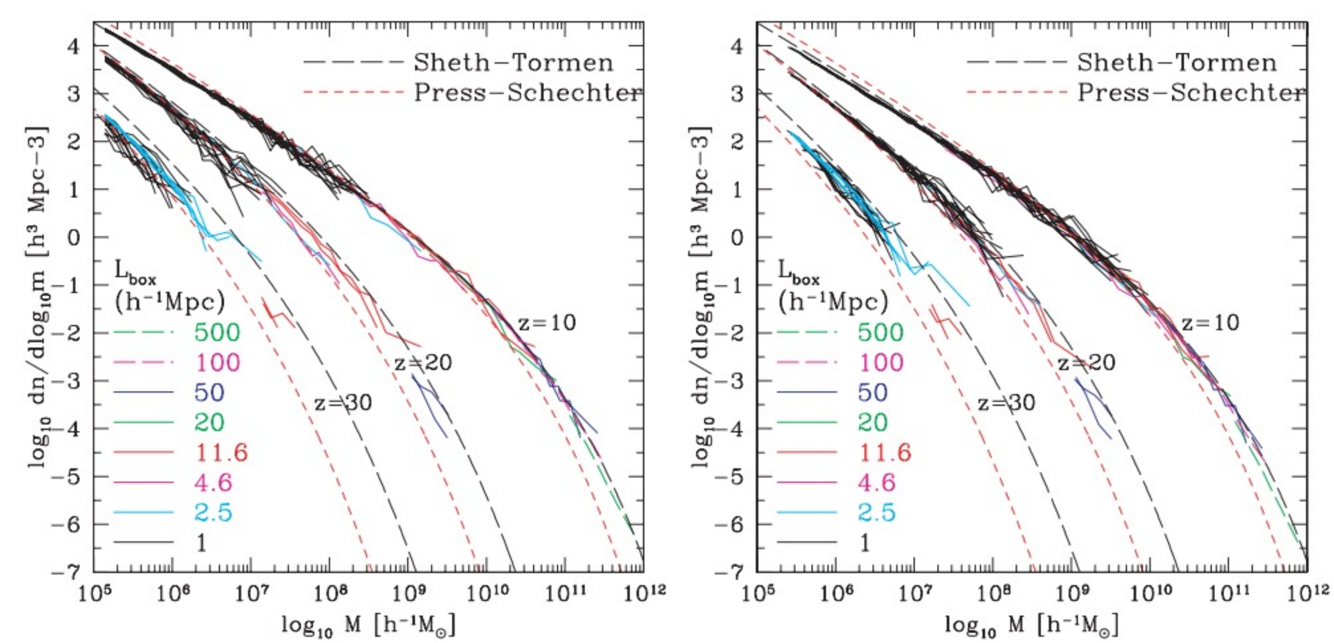
Dr. Tjitske Starkenburg (CIERA postdoc),
Claire Kopenhafer (Michigan State PhD student)

Dark Matter Halo Formation

Gaussian fluctuations in the dark matter density distribution collapse to form bound halos

Press-Schechter (1974) formalism describes the mass function of these halos

(see also
Sheth & Torman 1999;
Jenkins+ 2001;
Warren+ 2006...
many many more....)



(a) Raw simulation mass function   (b) Global mass function

Reed+ 2006

The Connection between Dark Matter and baryons in halos

Rees & Ostriker 1977; Silk 1977; Binney 1977; White and Rees 1978:

- Gas infalls and shocks at the virial radius to the virial temperature
- Slowly cools and infalls to form the dense central component of the galaxy

Dekel and Birnboim 2003; Kerěs et al 2005; van de Voort 2011; more...:

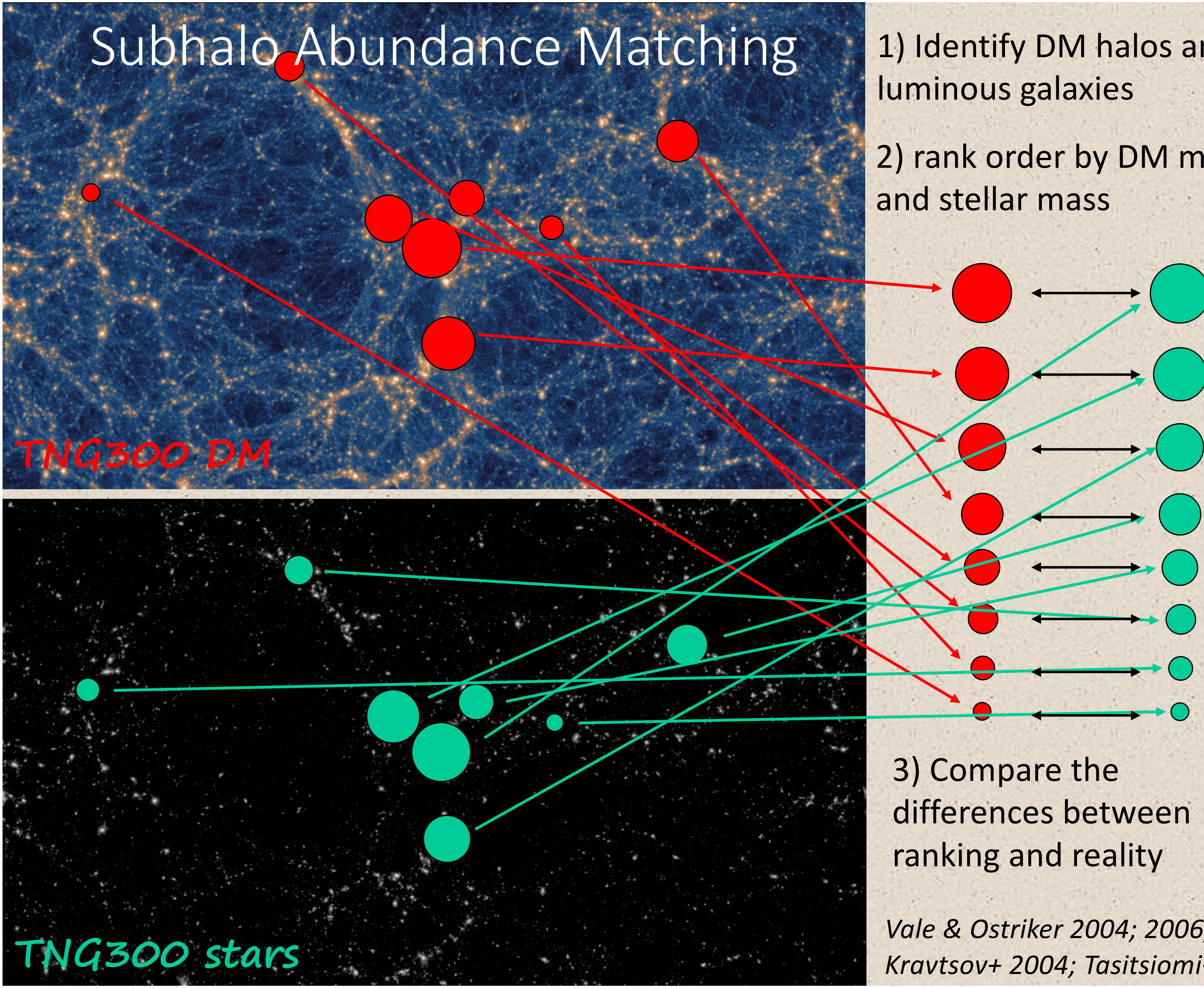
- Not all gas is shock-heated, and the fraction of shock-heated gas depends on the halo mass
- Cold mode accretion dominates at low redshifts in halos with masses below ~5×10^{11} M_{sun}

The Importance of the Galaxy-Halo Connection

- Deducing Cosmological Parameters
  - *clustering + halo occupation can constrain cosmologies*
- Distribution of Dark Matter
  - *Predict the amount of substructure given mass and concentration of halos*
- Physics of Galaxy Formation
  - *Which properties of dark matter halos influence the baryonic galaxy?*
  - *What is the effect of baryonic processes like feedback on galaxies?*

Wechsler & Tinker 2018

Subhalo Abundance Matching



1) Identify DM halos and luminous galaxies

2) rank order by DM mass and stellar mass

3) Compare the differences between ranking and reality

Vale & Ostriker 2004; 2006; 2008; Kravtsov+ 2004; Tasitsiomi+ 2004

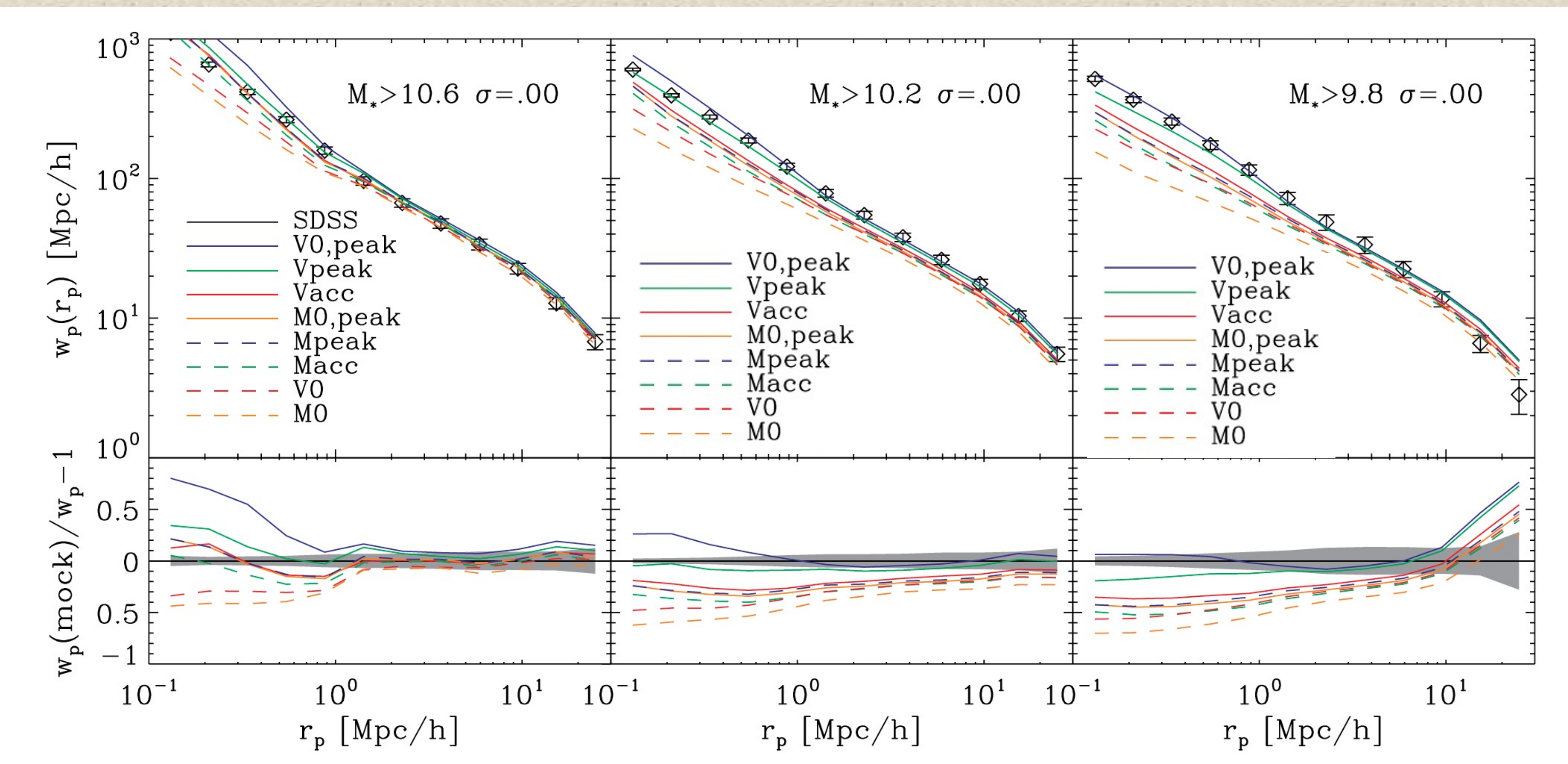
Which Halo Feature Should we Sort By?

- M_{DM}, current halo mass *(Vale & Ostriker)*
- M_{peak}, peak halo mass
- v_{max}, current maximum rotational velocity *(Kravtsov)*
- v_{peak}, peak maximum rotational velocity
- v_{relax}, v_{peak} while the halo fulfills a “relaxation” criterion *(Chaves-Montero+ 2016)*

Should we make different choices based on whether the galaxy is a central or satellite?

-v_{acc} *(Conroy+ 2006)*

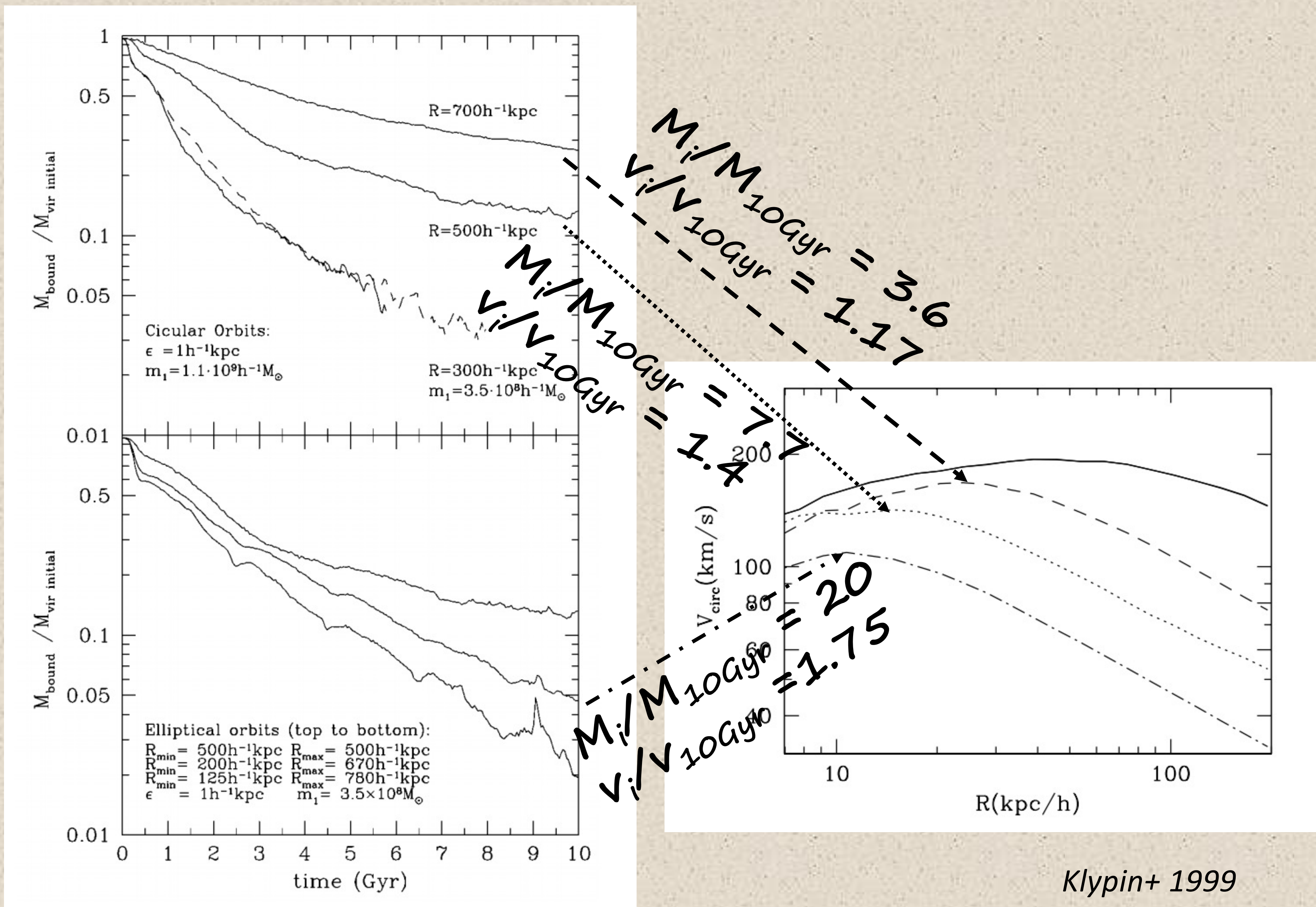
Comparing SHAMs with Observations:



Reddick+ 2013

2-point correlation function shows more discriminatory power at small scales, and may depend on the mass range considered

Why v_{max}? (I)



Klypin+ 1999

Arepo

- Moving mesh code (Springel 2010)
- Newtonian self-gravity
- Magnetohydrodynamical simulations
- TNG100: mbaryon 9.4 x 105 Msun/h
- TNG100: mDM 5.1 x 106 Msun/h

Cosmological parameters

- $\Omega_M = 1 - \Omega_\Lambda = 0.3089$, $\Omega_b = 0.0486$, $h = 0.6774$, $\sigma_8 = 0.8159$, $n = 0.9667$ (Planck 2015)
- TNG100 box size = 75 cMpc/h

Chemistry and microPhysics

- Primordial and metal-line radiative cooling inc self-shielding
- ionizing, redshift-dependent, spatially uniform background radiation field
- chemical enrichment from stellar pops (gas recycling), (SN Ia/II, AGB stars, and NS-NS mergers).
- Ideal MHD magnetic fields: small primordial seed field

Star formation, Black Holes, and feedback

- Stochastic SF in dense ISM gas above density threshold
- Evolution of stellar populations
- Stellar feedback: outflows from energy-driven kinetic wind scheme
- Seeding and growth of supermassive black holes
- BH feedback: 2 modes: high-accretion and low-accretion rates

Weinberger+ 2017; Pillepich+2018; Springel+ 2018; Naiman+ 2018; Nelson+ 2018, Marinacci+ 2018

SubHalo Abundance Matching in TNG100

Simulation Name	$L_{box}[Mpc]$	N_{DM}	m_{DM} $[M_\odot]$	m_{gas} $[M_\odot]$	N_{snap}	$N_{Subfind}(z=0)$
TNG100-1	110.7	1820^3	7.5×10^6	1.4×10^6	100	4371211
TNG100-1-Dark	110.7	1820^3	8.9×10^6	0	100	5012155

Selected galaxy – halo pairs that were well-resolved in both TNG100 and TNG100-Dark

- Required $M_* >= 10^9\ M_{sun}/h$ in TNG100
- Required $M_{DM} >= 10^{11}\ M_{sun}/h$ in TNG100-Dark

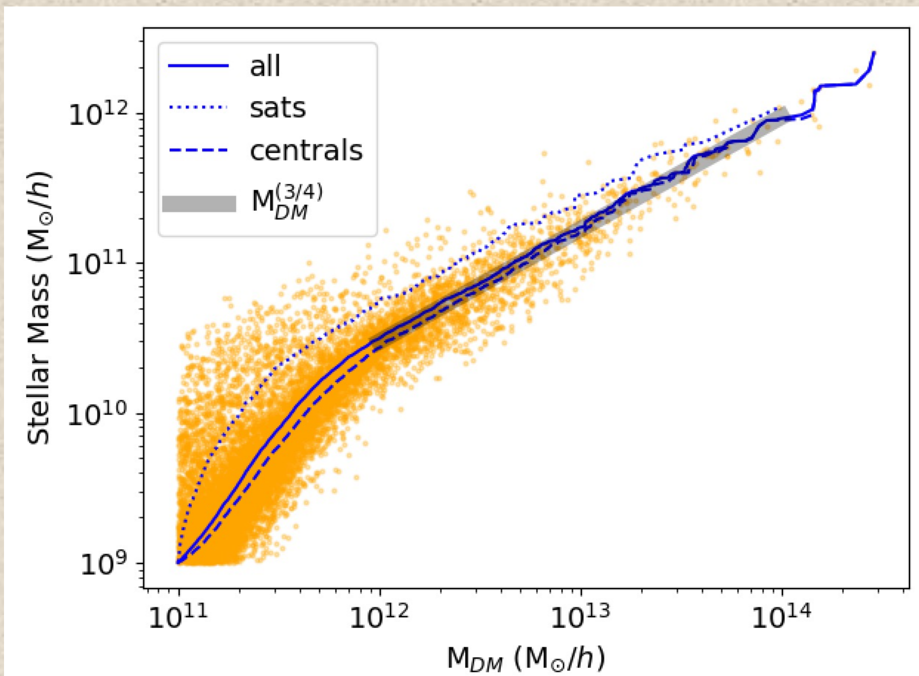
Halo Sample:

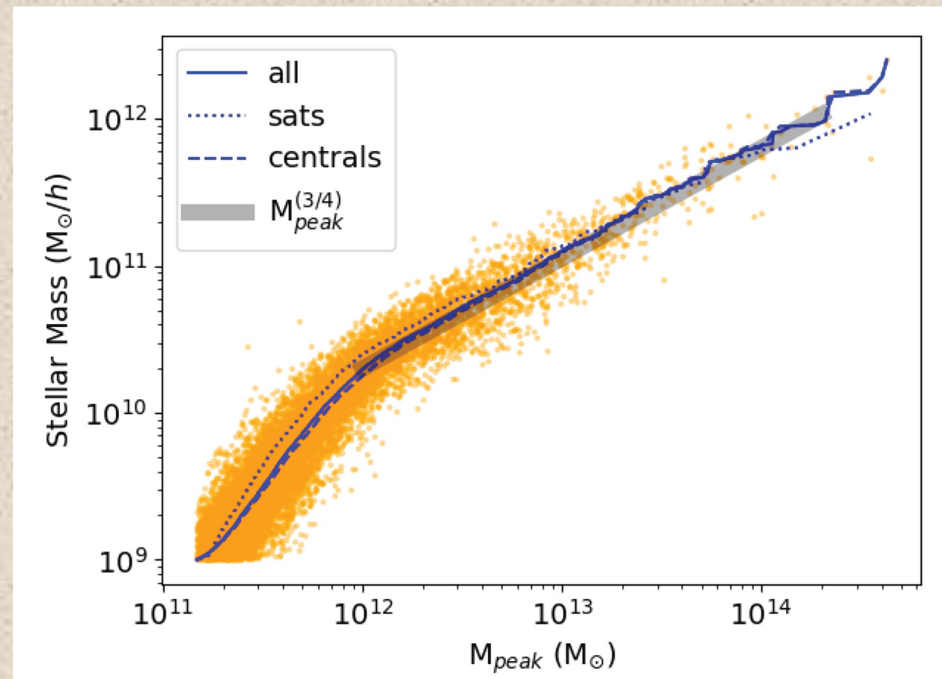
total: 11927

centrals: 9590

satellites: 2337

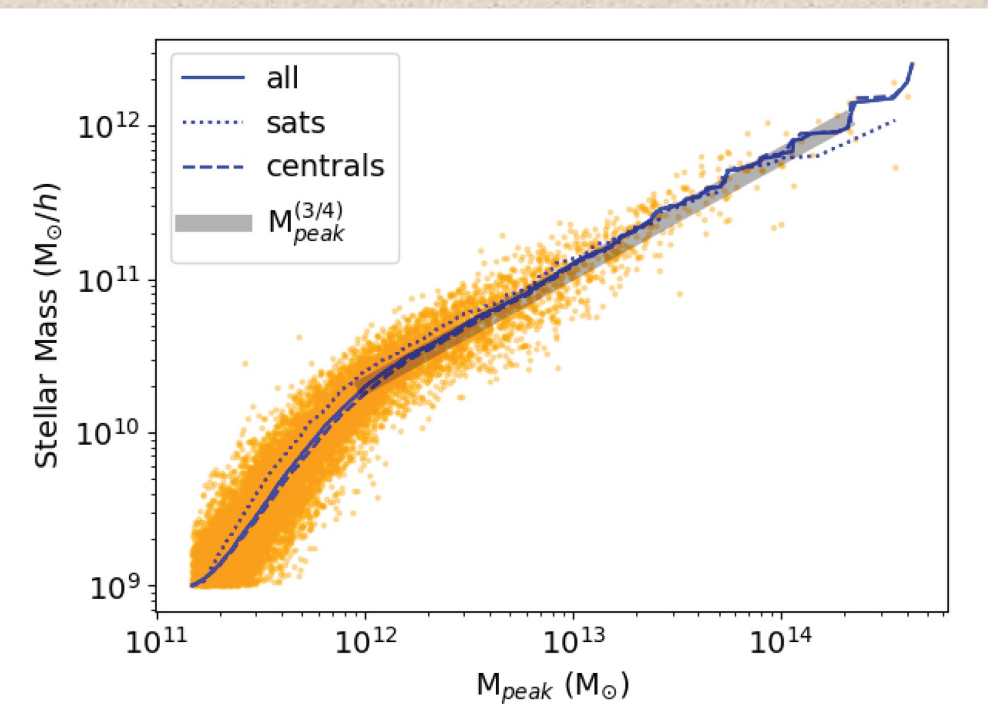
Finding the Best Sorting Feature (I)

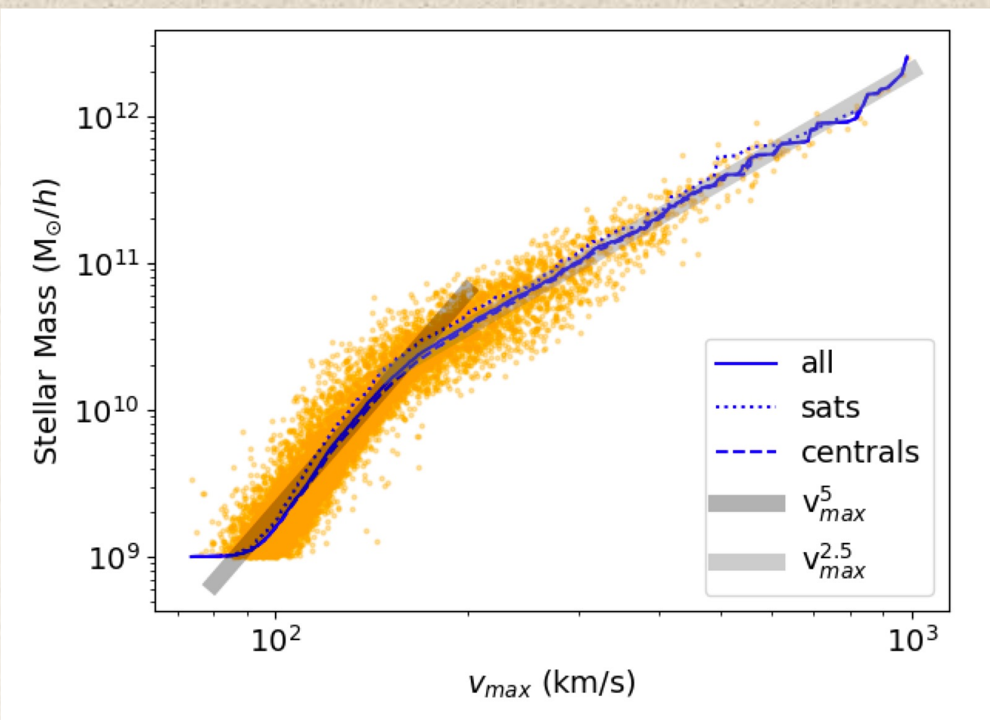




M_{peak} shows less scatter than M_{DM},
largely due to a reduction in the scatter for satellite galaxies
~The dependence on mass at high masses is the same~

Finding the Best Sorting Feature (II)





v_{max} shows even less scatter than M_{peak},
most clearly in the lower mass galaxies

Quantifying the Best Sorting Feature: Standard Set

$$Error \equiv \frac{\sum_N |\, log(M_{true}/M_{prediction}) \,|}{N}$$

Number of galaxies ($M_{DM} > 10^{11}$ $M_{\odot}$)	11927	9590	2337	11927	11927
Galaxy Sample	All	Centrals	Satellites	Mix	% Improvement
Rank Ordering using M_{DM}	0.198	0.130	0.279	0.159	–
Rank Ordering using M_{peak}	0.136	0.127	0.133	0.128	19
Rank Ordering using v_{max}	0.116	0.106	0.137	0.112	30

$$M_* \propto M_{peak} \frac{\Omega_b}{\Omega_d} \frac{t_{form}}{t_{cool,form}}$$

"monolithic collapse"
(Eggen+ 1962)

gravitational collapse
(Gunn & Gott 1972)

$$G < \rho > \equiv t_{form}^{-2}$$

$$t_{form} \propto \rho_{max}^{-\frac{1}{2}}.$$

radiative cooling

$$\Lambda(T_{max})\rho_{max}^2 \equiv \frac{\frac{3}{2}\rho_{max} k T_{max}}{t_{cool,form}}$$

$$t_{cool,form} \propto \rho_{max}^{-1} f^{-1}$$

where $f \equiv \Lambda(T_{max})/T_{max}$.

halo density
and velocity

$$M_* \propto M_{peak}\rho_{max}^{\frac{1}{2}} f \propto \left(\frac{M_{max}}{r_{max}}\right)^{\frac{3}{2}} f \propto v_{max}^3 f$$

$$\rho_{max} \equiv \frac{M_{max}}{\frac{4}{3}\pi r_{max}^3}$$

$$v_{max}^2 = \frac{GM_{max}}{r_{max}}.$$

Quantifying the Best Sorting Feature: High Mass

$$Error \equiv \frac{\sum_{N} | \, log(M_{true}/M_{prediction}) \, |}{N}$$

Number of galaxies ($M_{DM} > 10^{12}$ $M_{\odot}$)	1659	1463	196	1659
Galaxy Sample	All	Centrals	Satellites	Mix
Rank Ordering using M_{DM}	0.132	0.118	0.181	0.126
Rank Ordering using M_{peak}	0.120	0.118	0.122	0.119
Rank Ordering using vmax	0.124	0.123	0.129	0.124

Quantifying the Best Sorting Feature: Combination

$$\phi \equiv v_{norm} + m_{norm}$$

$$v_{norm} \equiv v_{max}/v_{max,12.7}$$

$$m_{norm} \equiv M_{peak}/10^{12.7}$$

Number of galaxies ($M_{DM} > 10^{11}\ M_{\odot}$)	11927	9590	2337	11927	11927
Galaxy Sample	All	Centrals	Satellites	Mix	% Improvement
Rank Ordering using M_{DM}	0.198	0.130	0.279	0.159	–
Rank Ordering using M_{peak}	0.136	0.127	0.133	0.128	19
Rank Ordering using v_{max}	0.116	0.106	0.137	0.112	30
Rank Ordering using $\phi \equiv v_{norm} + m_{norm}$	0.111	0.101	0.119	0.105	34

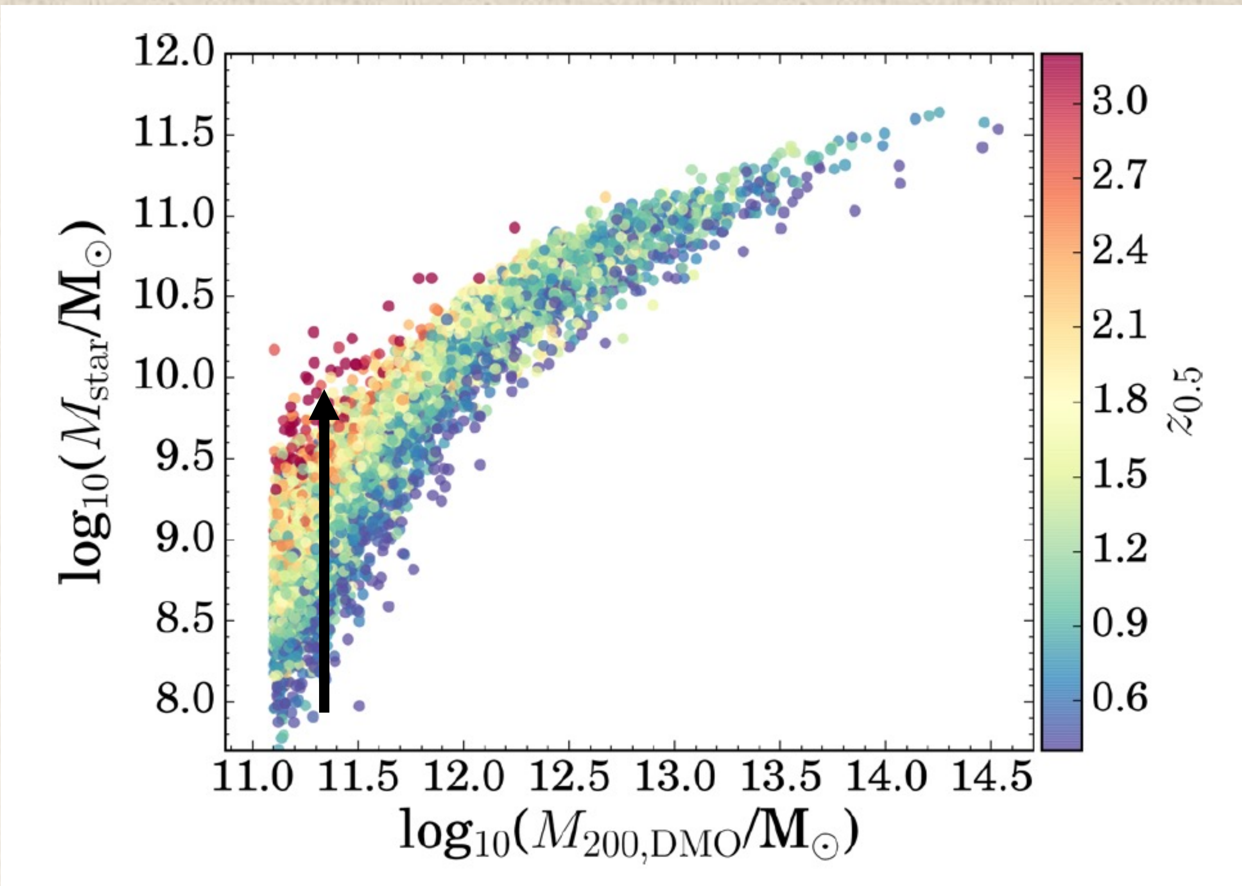
Lehmann+ (2017) used a similar "composite" feature for abundance matching:

$$v_{\alpha} = v_{\rm vir}\left(\frac{v_{\rm max}}{v_{\rm vir}}\right)^{\alpha},$$

Improvements with Secondary Features

- **formation time**
- **halo concentration**
- **local environmental density**

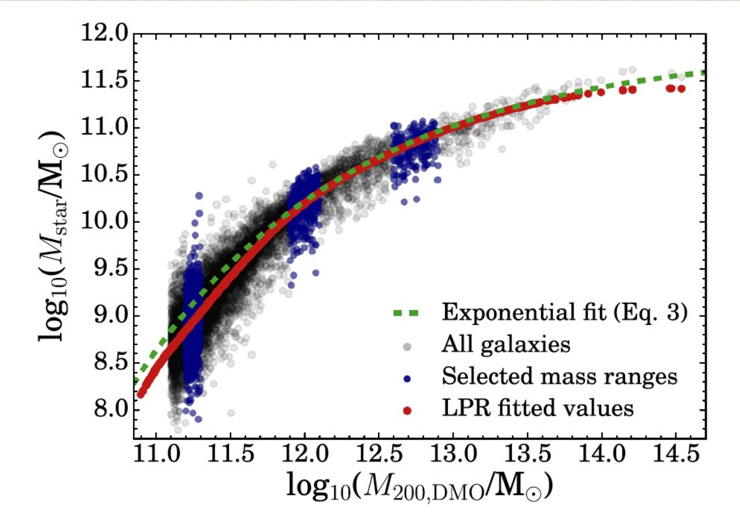
Formation time



Early formation time
➔ higher M_*

The halo is more massive at early times, when there is more gas to accrete and form stars

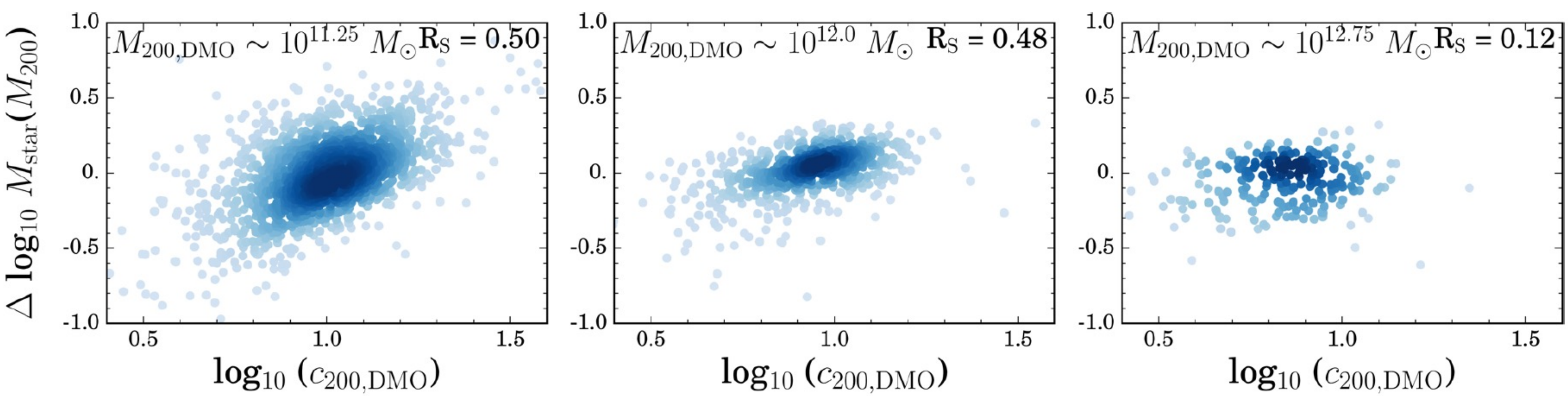
Matthee+ 2017



Matthee+ 2017

Concentration

$M_* - M_{200,\,DM}$ relation



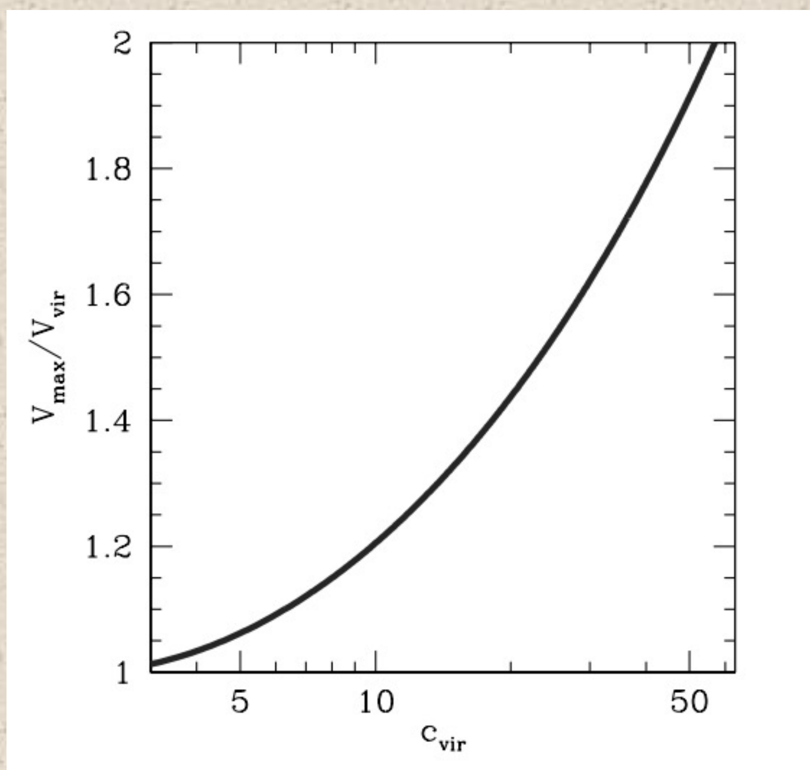
Distance from the $M_* - M_{200,\,DM}$ relation as a function of concentration

Higher concentration ➔ higher M_*

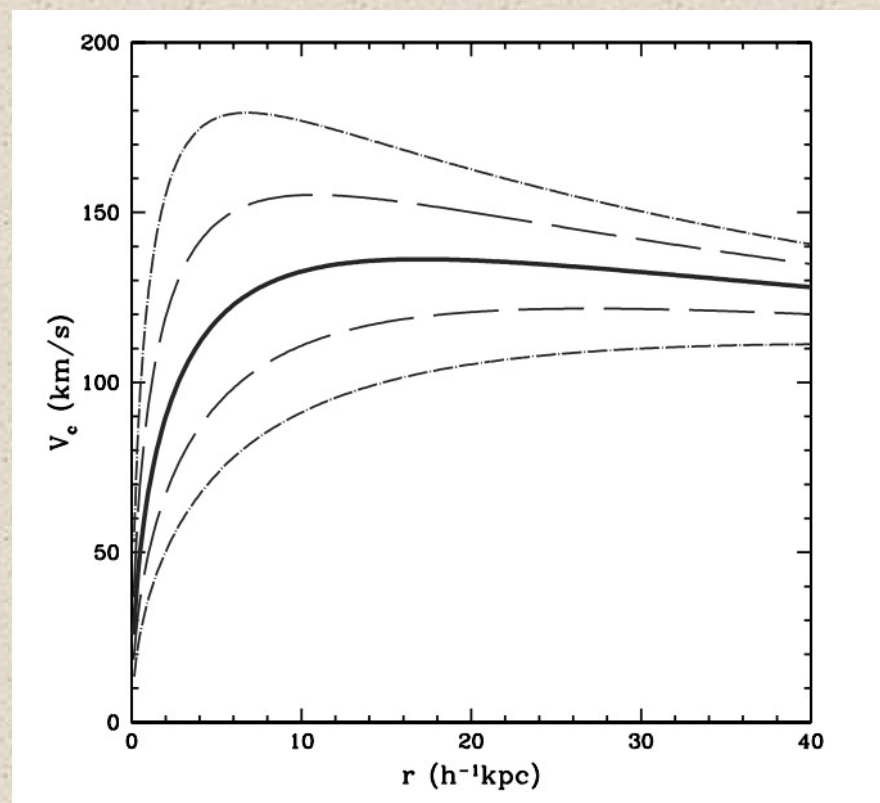
Steeper slope at lower mass

Why v_{max}? (II)

v_{max} includes a dependence on concentration



v_{max}-concentration relation

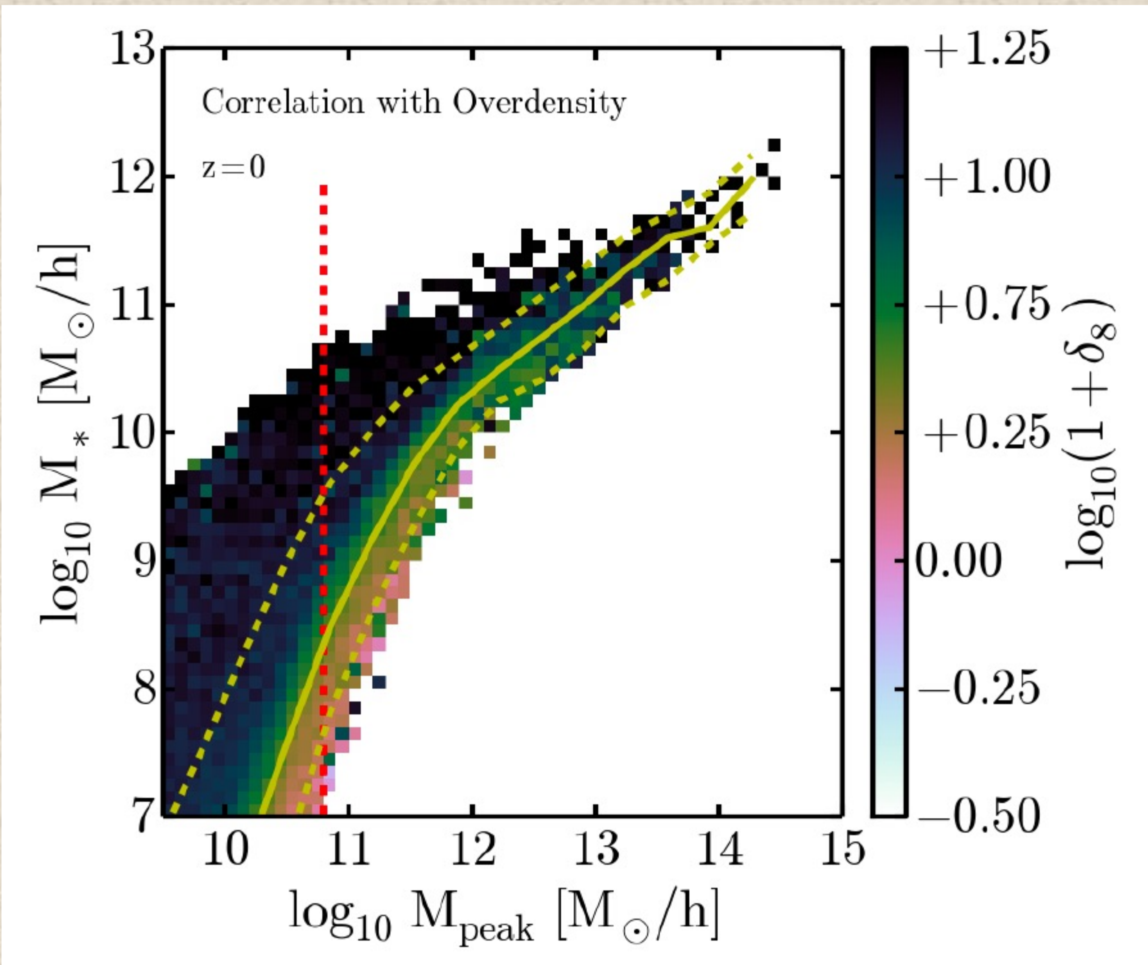


Different rotation curves from varying concentration

Bullock+ 2001

(also Klypin+ 2011)

Environment

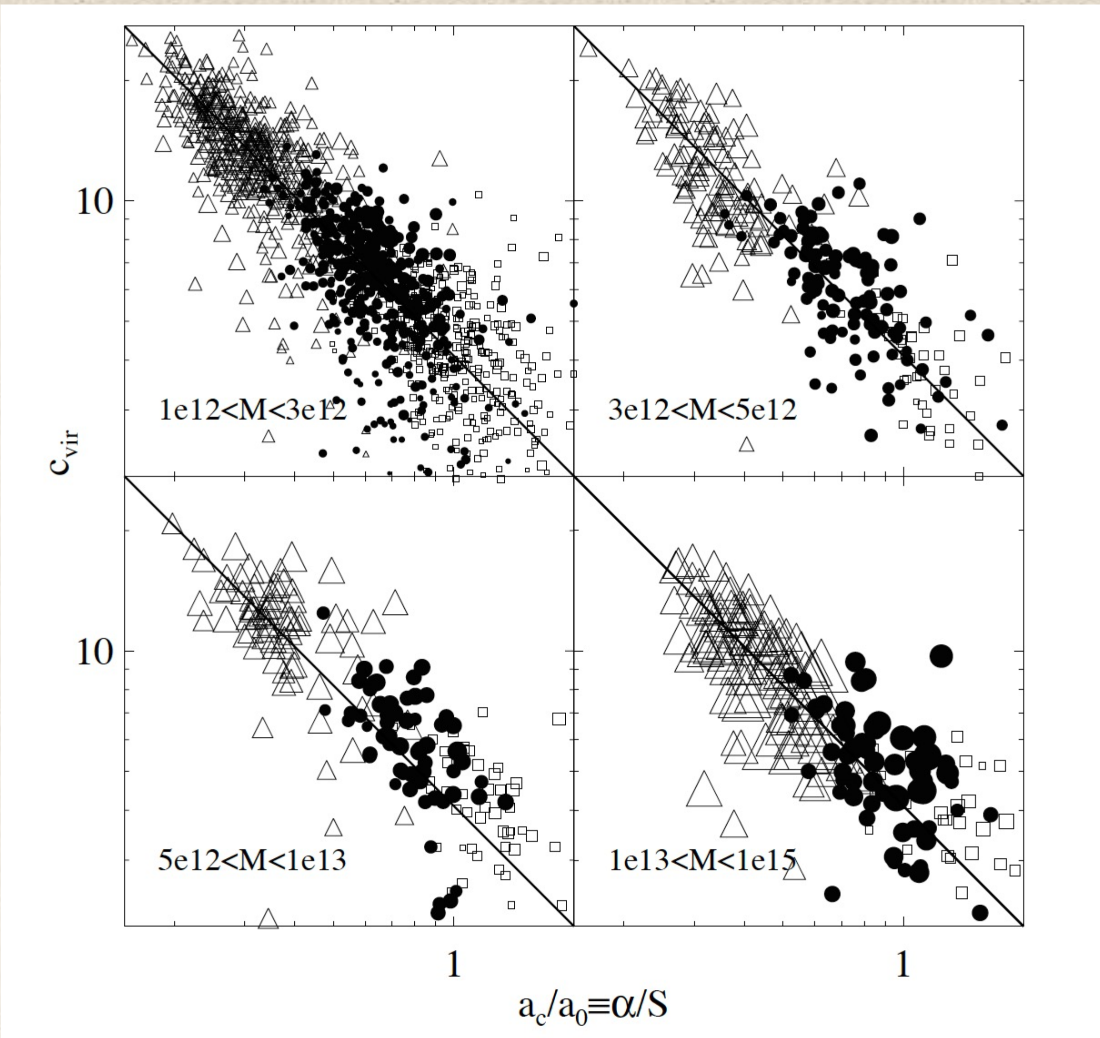


High local density
➔ higher M_*

Assembly Bias
(Gao et al. 2005)

Martizzi+ 2020

Concentration and Formation Time

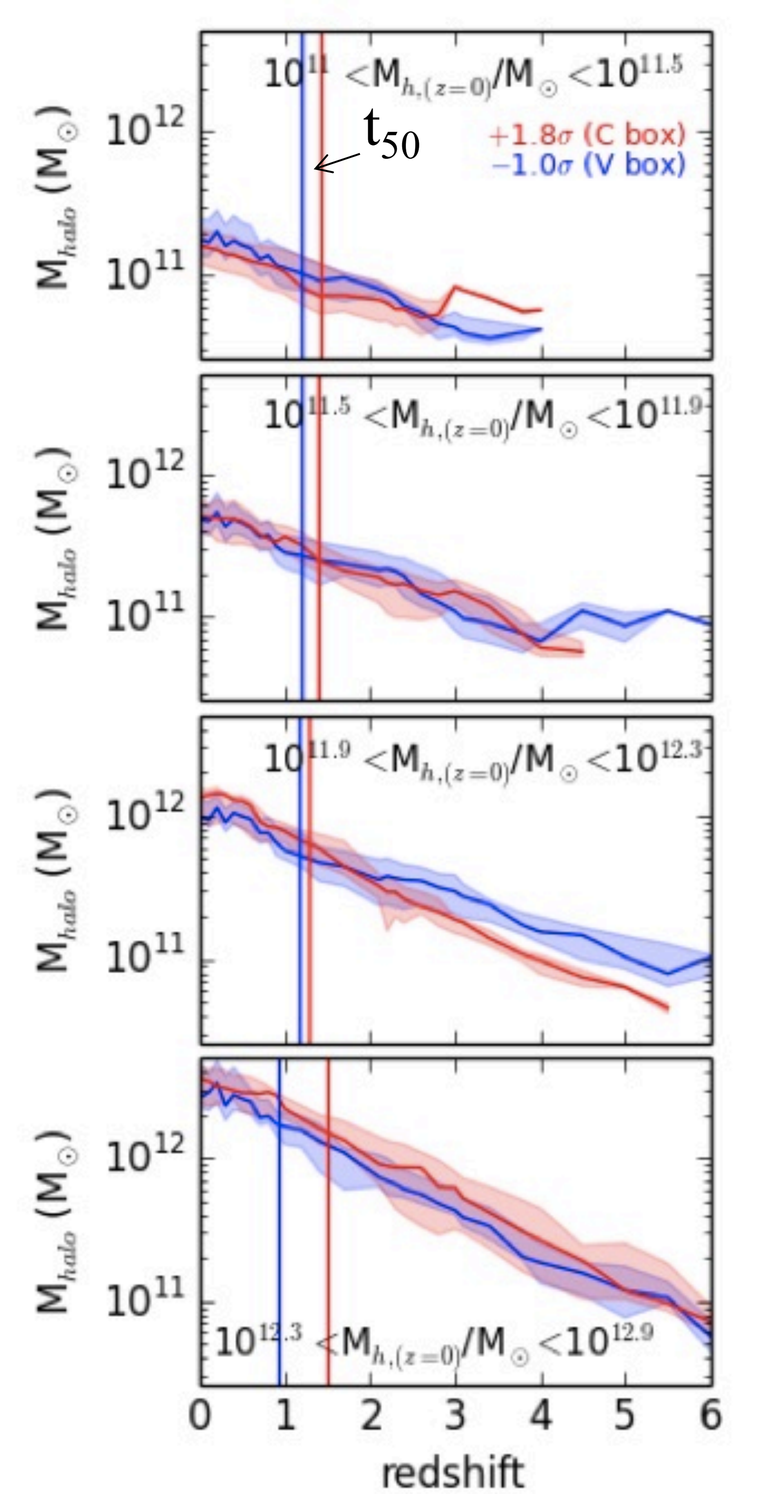


Wechsler+ 2002

higher concentration
➔ earlier formation time

Early formation times, when the density of the universe is higher, results in higher concentration halos

see also NFW+ 1997; Bullock+ 2001;



Environment and Formation Time

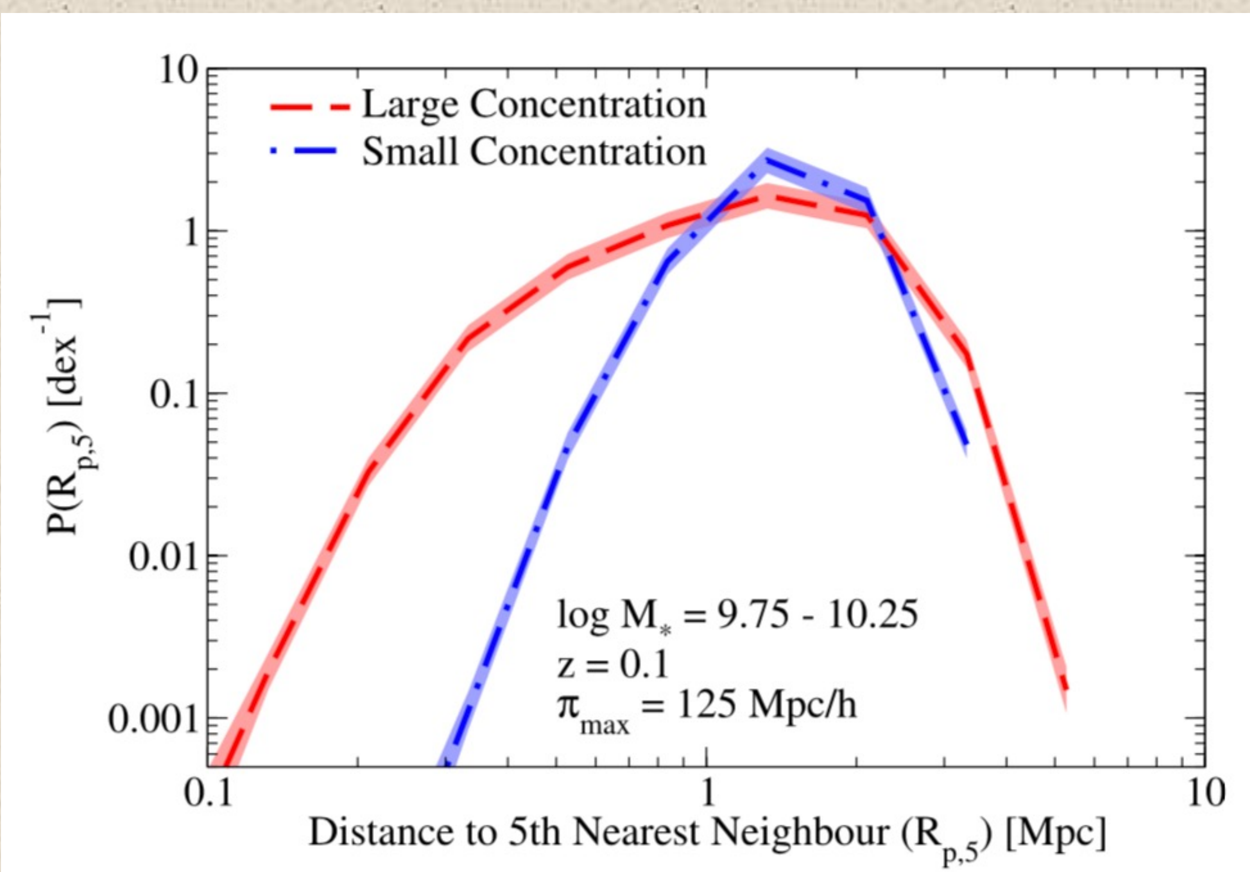
Underdensity: 31 × 31 × 35 h^{-3} Mpc^3
-1.0σ fluctuation

Overdensity: 21 × 24 × 20 h^{-3} Mpc^3
+1.8σ fluctuation

higher density environment
➔ earlier formation time

Tonnesen & Cen 2015

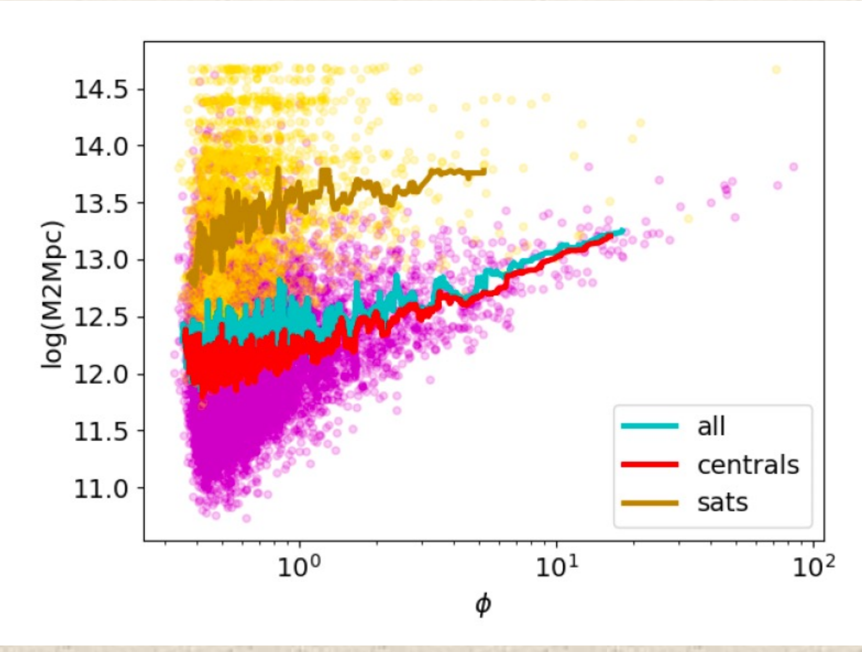
Environment and Concentration



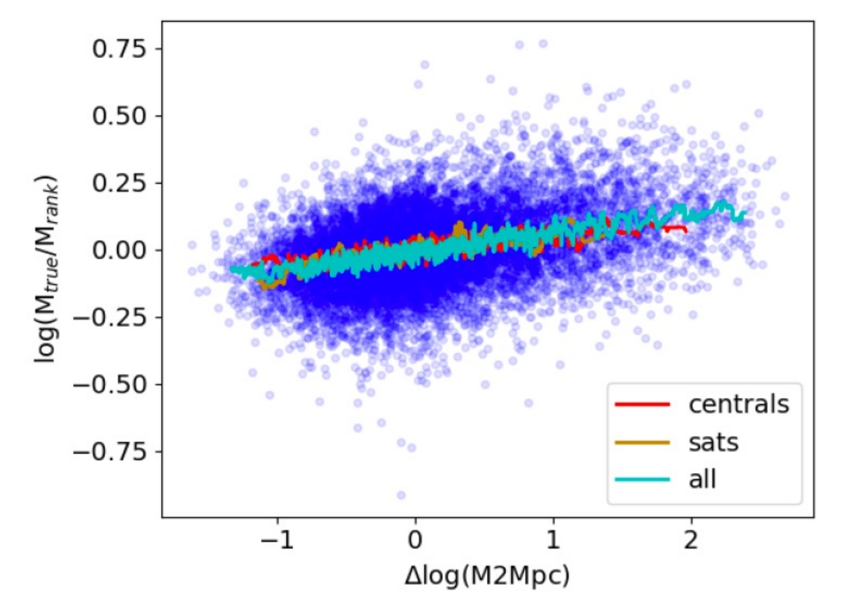
higher concentration
➔ higher local density

Behroozi+ 2020

Improving the fit in TNG:



1) Plot the secondary feature as a function of ϕ



2) Find Mtrue/Mrank as a function of Δlog(feature)

3) Solve for the new predicted M*

$$log(M_{*,pred}) = log(M_{*,rank}) + \alpha \Delta log(feature)^2 + \beta \Delta log(feature) + \gamma$$

Quantifying Improvement

	Number of galaxies ($M_{DM} > 10^{11}\ M_{\odot}$)	11927	9590	2337	11927	11927
	Galaxy Sample	All	Centrals	Satellites	Mix	% Improvement
Mass proxies	$\phi + v_{disp}$	0.112	0.102	0.117	0.105	0
	$\phi + v_{max}$	0.111	0.101	0.117	0.104	1
	$\phi + M_{DM}$	0.105	0.101	0.110	0.103	2
	$\phi + M_{peak}$	0.111	0.101	0.117	0.104	1
halo size	$\phi + r_{max}$	0.111	0.101	0.118	0.105	0
	$\phi + r_{DM}$	0.105	0.100	0.114	0.103	2
concentration	$\phi + c_v$	0.111	0.101	0.118	0.105	0
	$\phi + c_h$	0.109	0.101	0.117	0.104	1
	$\phi + c_r$	0.109	0.101	0.116	0.104	1
formation time	$\phi + t_{peak}$	0.105	0.101	0.110	0.103	2
	$\phi + t_{50}$	0.106	0.099	0.116	0.102	3
	$\phi + t_{85}$	0.104	0.099	0.111	0.101	4
environment	$\phi + M_{DM,r<1Mpc}$	0.104	0.100	0.115	0.103	2
	$\phi + M_{DM,r<2Mpc}$	0.103	0.099	0.113	0.102	3
	$\phi + M_{DM,r<5Mpc}$	0.105	0.099	0.115	0.102	3
	$\phi + M_{DM,r<8Mpc}$	0.107	0.100	0.116	0.103	2
	$\phi + M_{DM,r<15Mpc}$	0.109	0.100	0.117	0.104	1
ranking	Rank Ordering using $\phi \equiv v_{norm} + m_{norm}$	0.111	0.101	0.119	0.105	

Throwing it all together

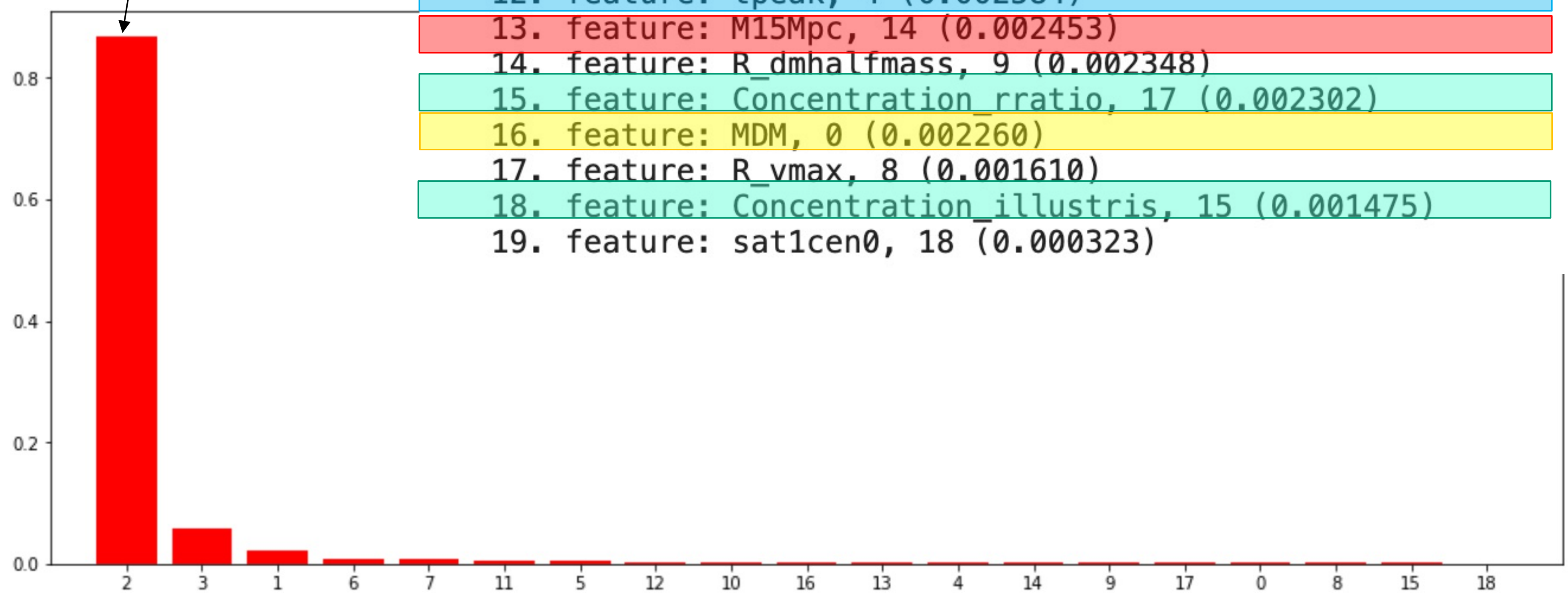
```
#Auth[illegible]: Viviana Acquaviva
#Licen[illegible]BSD but really[illegible]ld be TBD - just be nice.

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import sklearn
import time
from scipy import stats

from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, cross_val_predict
from sklearn.model_selection import KFold, StratifiedKFold
from sklearn.model_selection import GridSearchCV
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import scale
from sklearn.utils import shuffle
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
from sklearn.feature_selection import SelectFromModel
```

Pedregosa+ 2011

Feature Ranking

ϕ

```
Feature ranking:
1. feature: vnormMnorm, 2 (0.867145)
2. feature: Mpeak, 3 (0.056724)
3. feature: vmax, 1 (0.022002)
4. feature: t85, 6 (0.008597)
5. feature: vdisp, 7 (0.006519)
6. feature: M2Mpc, 11 (0.005737)
7. feature: t50, 5 (0.005414)
8. feature: M5Mpc, 12 (0.003322)
9. feature: M1Mpc, 10 (0.003297)
10. feature: Concentration_vratio, 16 (0.003278)
11. feature: M8Mpc, 13 (0.002611)
12. feature: tpeak, 4 (0.002584)
13. feature: M15Mpc, 14 (0.002453)
14. feature: R_dmhalfmass, 9 (0.002348)
15. feature: Concentration_rratio, 17 (0.002302)
16. feature: MDM, 0 (0.002260)
17. feature: R_vmax, 8 (0.001610)
18. feature: Concentration_illustris, 15 (0.001475)
19. feature: sat1cen0, 18 (0.000323)
```



Random Forest Regression

```
cvmethod = KFold(n_splits=5, shuffle = True)

parameters = {'max_depth':[10,14,20], \
              'max_features': [3,4,6,8,9,10,12,14,15,16,17,18,19], 'n_estimators':[50,100,200]}

nmodels = np.product([len(el) for el in parameters.values()])

gmodel = GridSearchCV(RandomForestRegressor(), parameters, cv = cvmethod, \
                     scoring = 'neg_mean_absolute_error', \
    verbose = 1, n_jobs = 4,return_train_score=True)
start = time.time()
gmodel.fit(normalized_X, y)
stop = time.time()
print('Best params, best score:', "{:.4f}".format(gmodel.best_score_), gmodel.best_params_),
print('Time per model (s):', "{:.4f}".format((stop-start)/float(nmodels*4)))
```

```
Fitting 5 folds for each of 117 candidates, totalling 585 fits
[Parallel(n_jobs=4)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:   41.0s
[Parallel(n_jobs=4)]: Done 192 tasks      | elapsed:  7.0min
[Parallel(n_jobs=4)]: Done 442 tasks      | elapsed: 18.9min
[Parallel(n_jobs=4)]: Done 585 out of 585 | elapsed: 28.2min finished
Best params, best score: -0.0920 {'max_depth': 20, 'max_features': 10, 'n_estimators': 100}
Time per model (s): 3.6320
```

7% improvement from 3 features…

Best Score uses 10 features…..

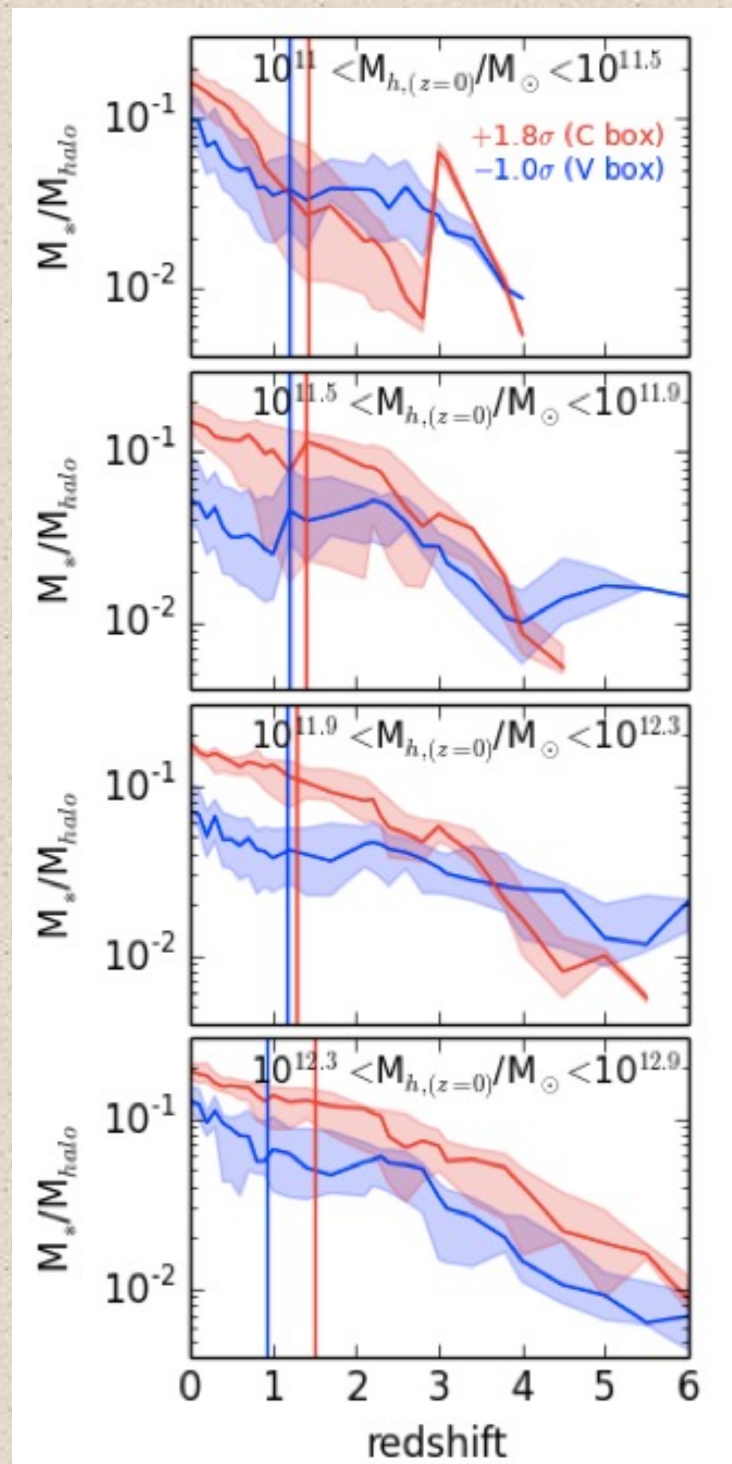
RFR does not require 10 features for a low error

	params	mean_test_score	std_test_score
71	{'max_depth': 14, 'max_features': 3, 'n_estimators': 200}	-0.092174	0.000666
47	{'max_depth': 10, 'max_features': 5, 'n_estimators': 200}	-0.092233	0.000827
43	{'max_depth': 10, 'max_features': 4, 'n_estimators': 200}	-0.092249	0.000832
67	{'max_depth': 14, 'max_features': 2, 'n_estimators': 200}	-0.092255	0.000717
39	{'max_depth': 10, 'max_features': 3, 'n_estimators': 200}	-0.092256	0.000767
75	{'max_depth': 14, 'max_features': 4, 'n_estimators': 200}	-0.092263	0.000815
42	{'max_depth': 10, 'max_features': 4, 'n_estimators': 100}	-0.092277	0.000716
66	{'max_depth': 14, 'max_features': 2, 'n_estimators': 100}	-0.092285	0.000449

perhaps there are several similarly relevant predictors...

But what about the SHMR in Different Environments?

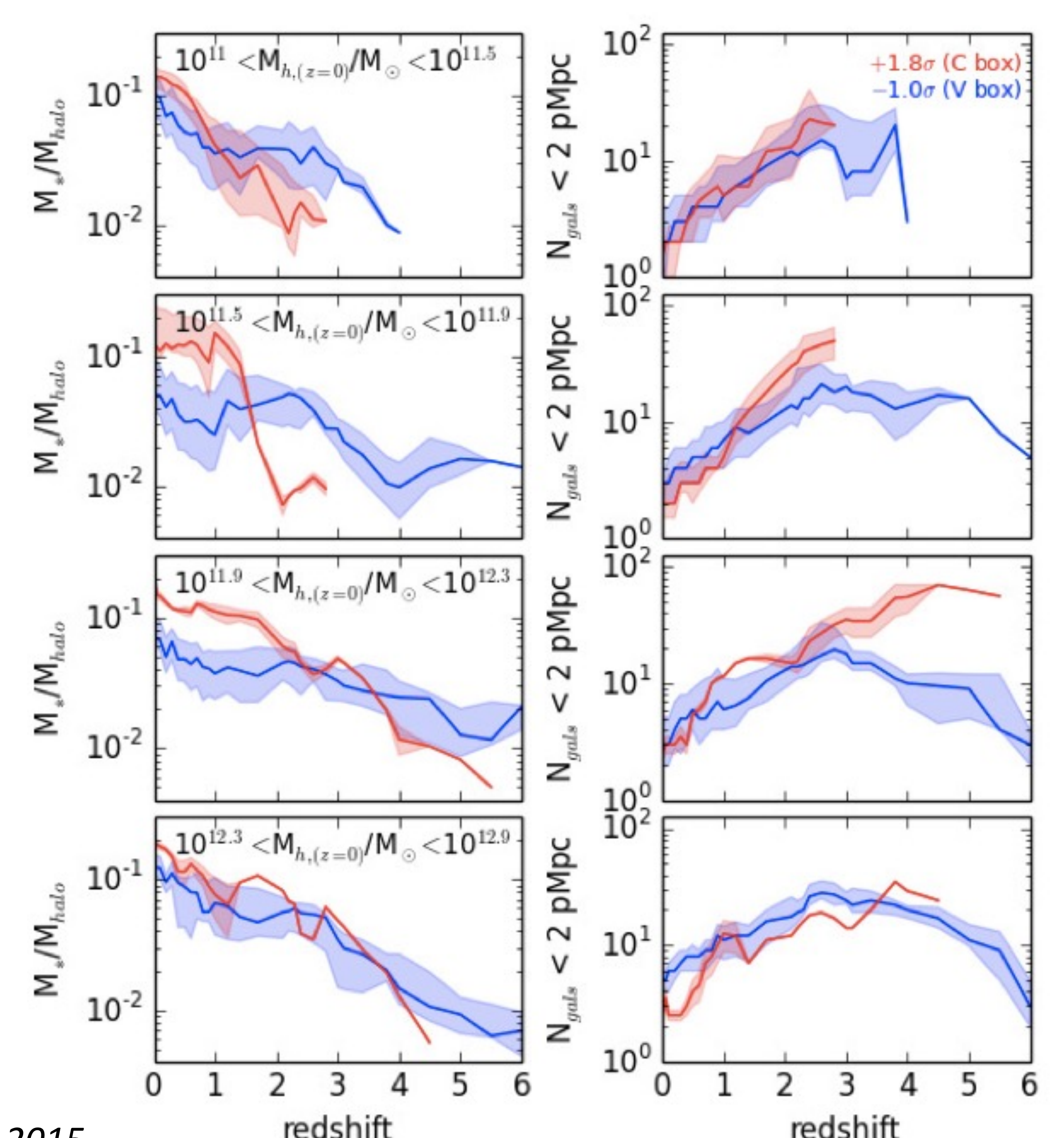
M_*/M_{halo} is larger in the large-scale overdensity



Tonnesen & Cen 2015

What about the large-scale environment?

Only select galaxies from the overdensity that have fewer than 3 galaxies within 2 physical Mpc at z=0. Therefore the "local galaxy density" is lower in the large-scale overdensity



Tonnesen & Cen 2015

Summary

- Scatter in the $M_* - M_{DM}$ relation can be dramatically reduced by ranking with v_{max}
- We further reduce scatter by ranking with a parameter that depends on v_{max} at low mass and M_{peak} at high mass (our ϕ)
- Secondary parameters based on formation time and local density gave the most improvement on standard ranking
- Correcting using secondary parameters—even a lot of them—does not substantially reduce scatter

- *Consider v_{peak} (or v_{relax})*
- *Consider local environment at halo formation time*
- *Test the impact of feedback*