Deep-learning for galaxy morphology and evolution

A review of on-going projects - questions / ideas

Marc Huertas-Company

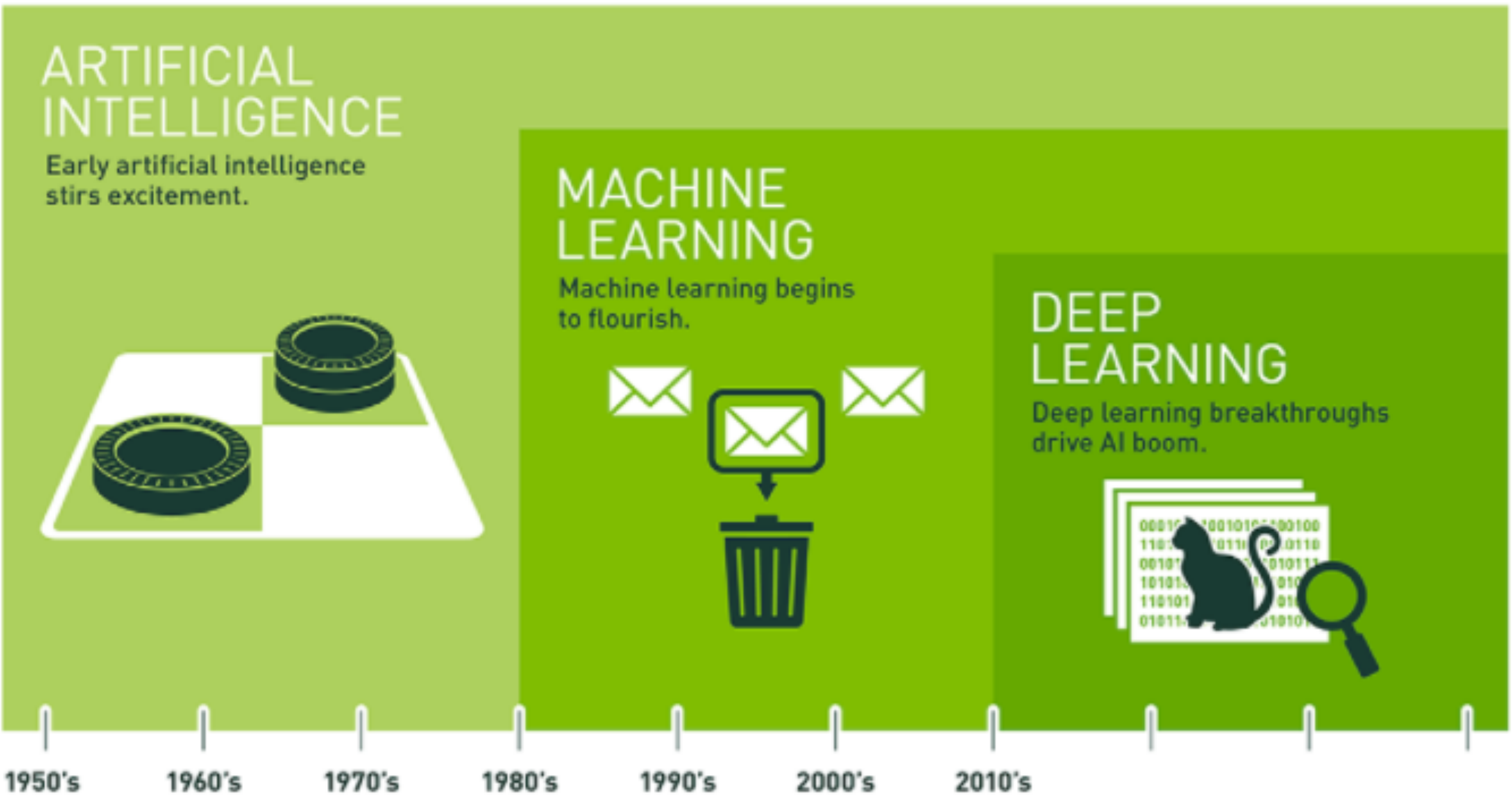
D. Koo, J. Primack, H. Dominguez-Sanchez, M. Bernardi, S. Faber, F. Caro, D. Tuccillo, C. Lee, B. Margalef-Bentabol, E. Decencière, S. Velasco-Forero, G. Cabrera-Vives….

Journées PNCG, November 17 2017

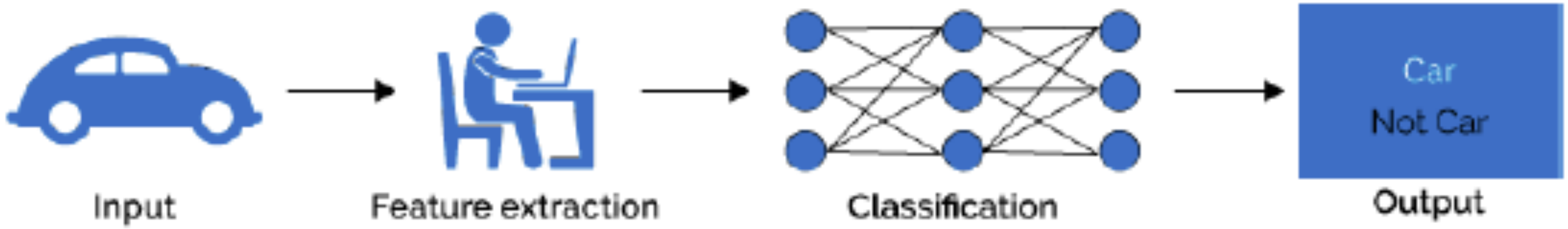
GOLDEN AGE FOR ARTIFICIAL INTELLIGENCE with the generalization of deep-learning since 2012

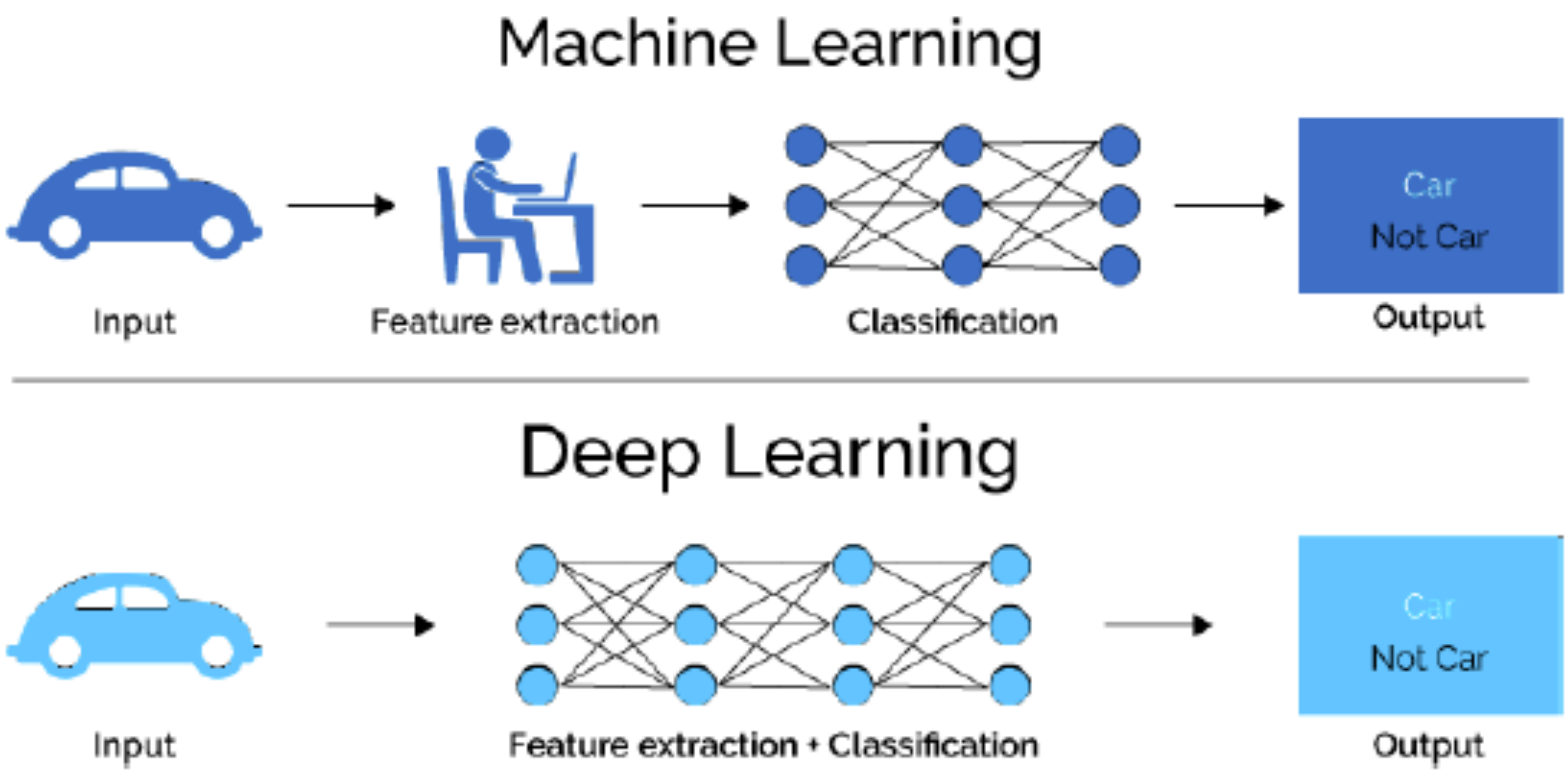
(Nature, 01/2016)

> **"Deep learning is killing every problem in AI."**



Machine Learning





The 3 "phases" of deep-learning:

skepticism, acceptance, frustration

GOLDEN AGE FOR ARTIFICIAL INTELLIGENCE
("Artificial Intelligence") with the generalization of deep-learning since 2012

QUESTIONS WE ARE FOCUSING ON :

How AI (i.e. deep-learning) can be used to understand galaxy formation?

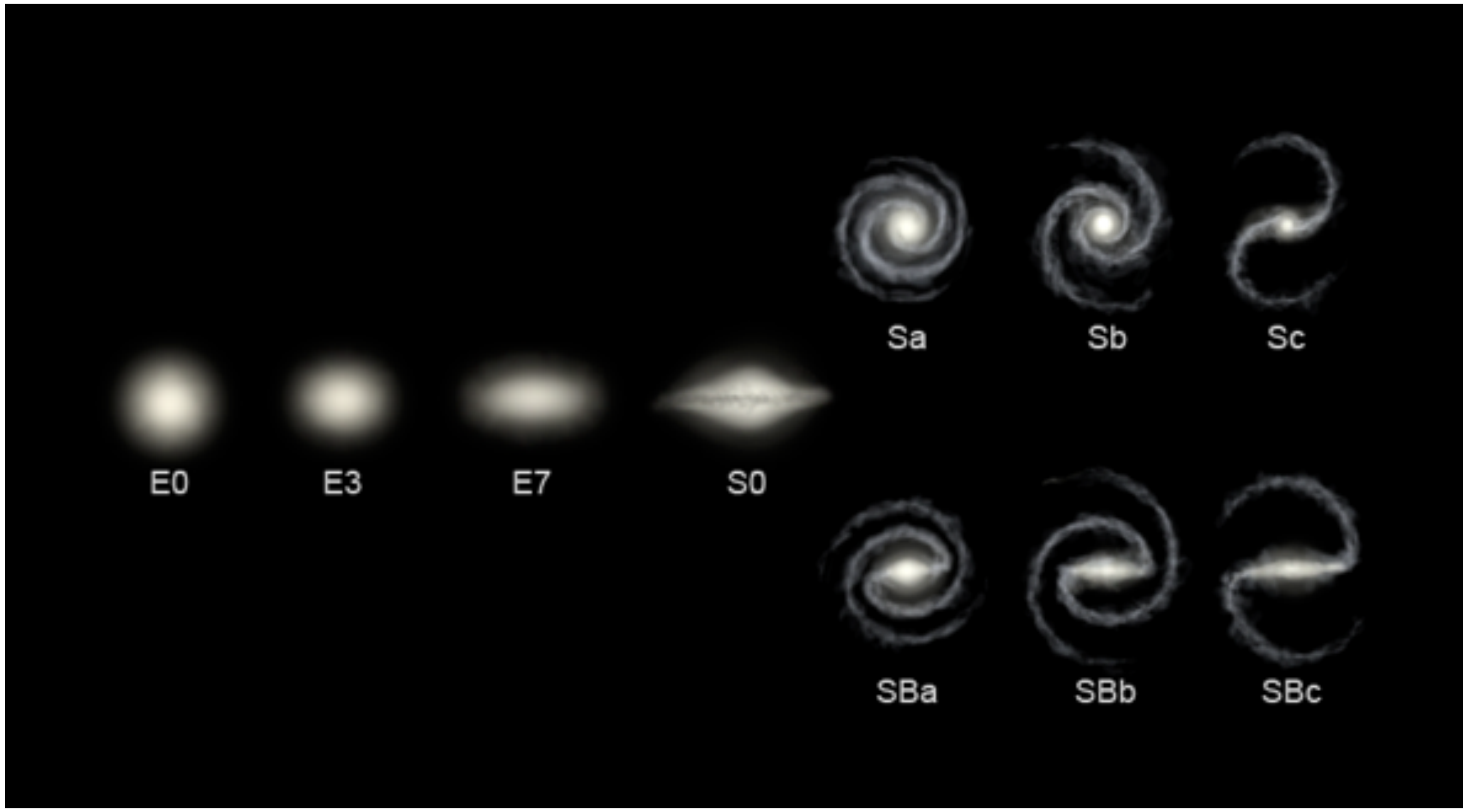
Can we do things with AI that we could not do before?

Can we learn something new about the physics of galaxies?

DL FOR FOR GALAXIES?

- **GROUP #1:** Time consuming tasks that humans do easily but classically challenging for computers - classification of objects
- **GROUP #2:** Efficient and fast quantitative measurements on large amount of (multi-lambda) data [photoz's, sizes, ellipiticities]
- **GROUP #3:** Find hidden new observables in the data, - Linking observations and theory
- **GROUP #4**: Finding the unknown?

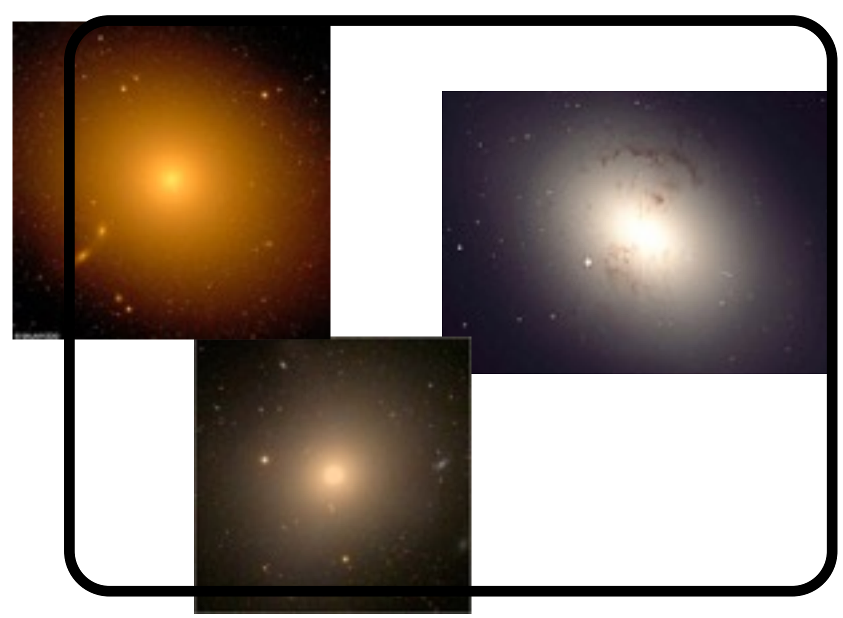
- **GROUP #1**: Time consuming tasks that humans do easily but classically challenging for computers - <u>classification of objects</u>



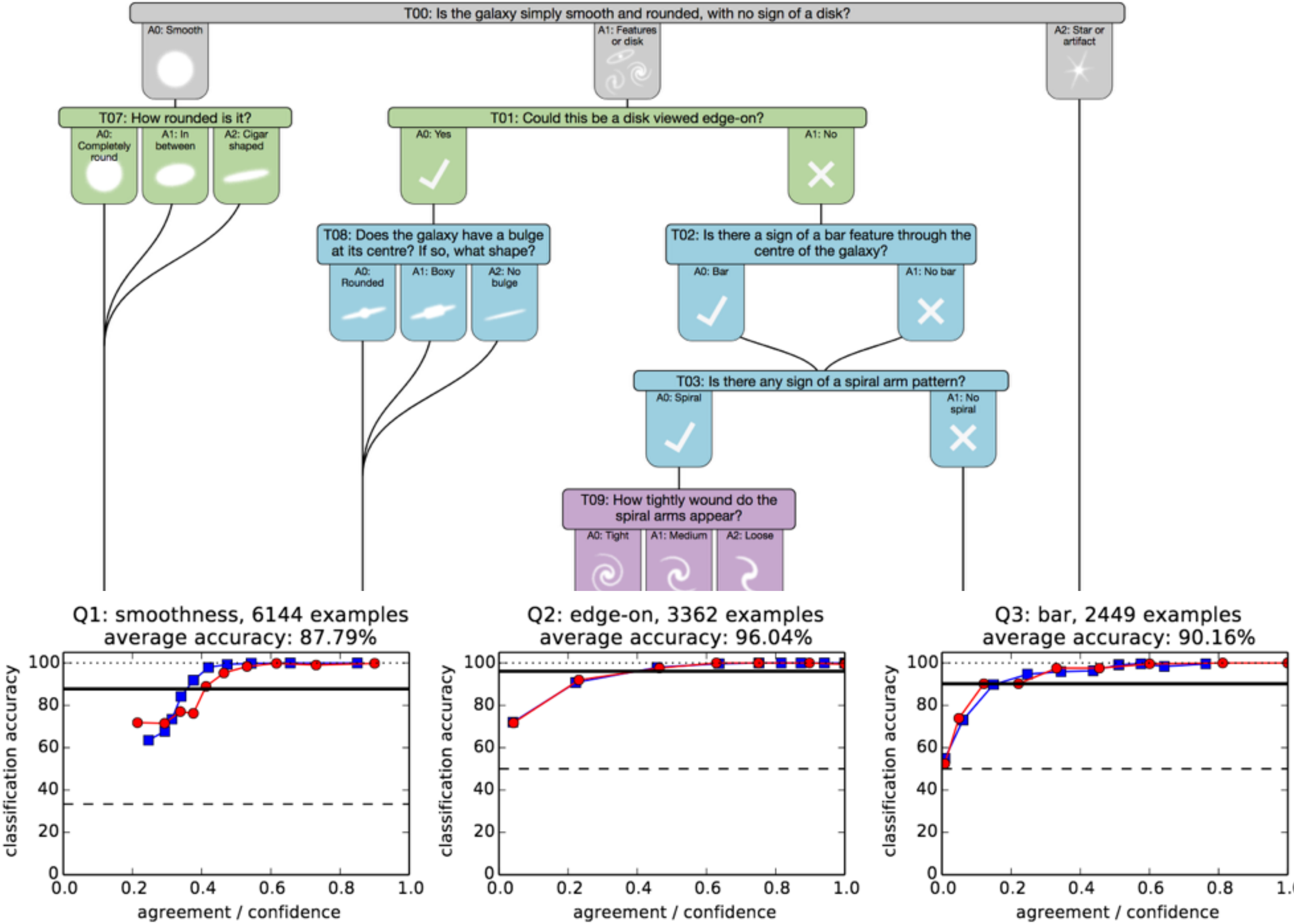
The Hubble Sequence

"Objects in the same box experienced the same physics"

BOX I

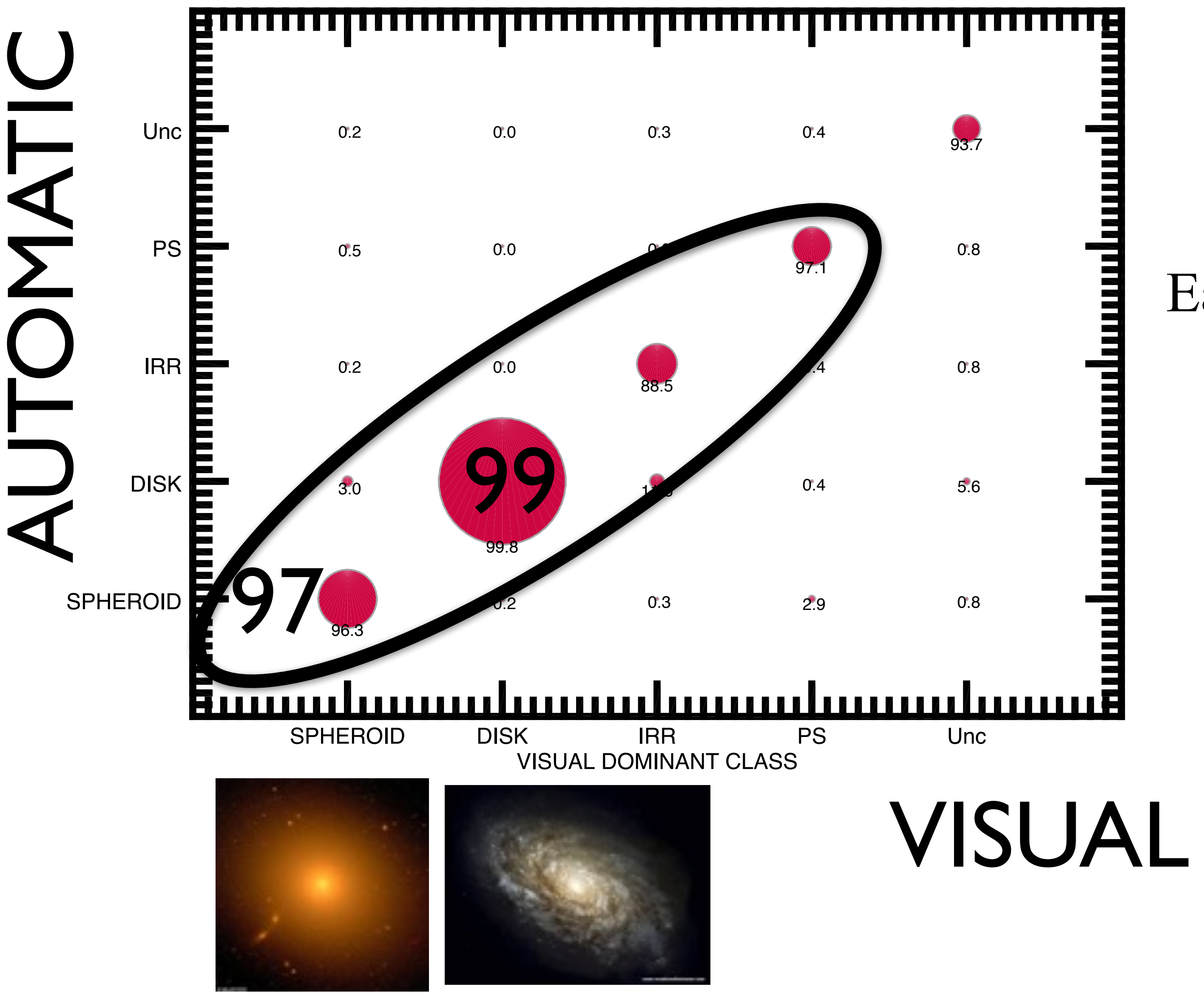


BOX II

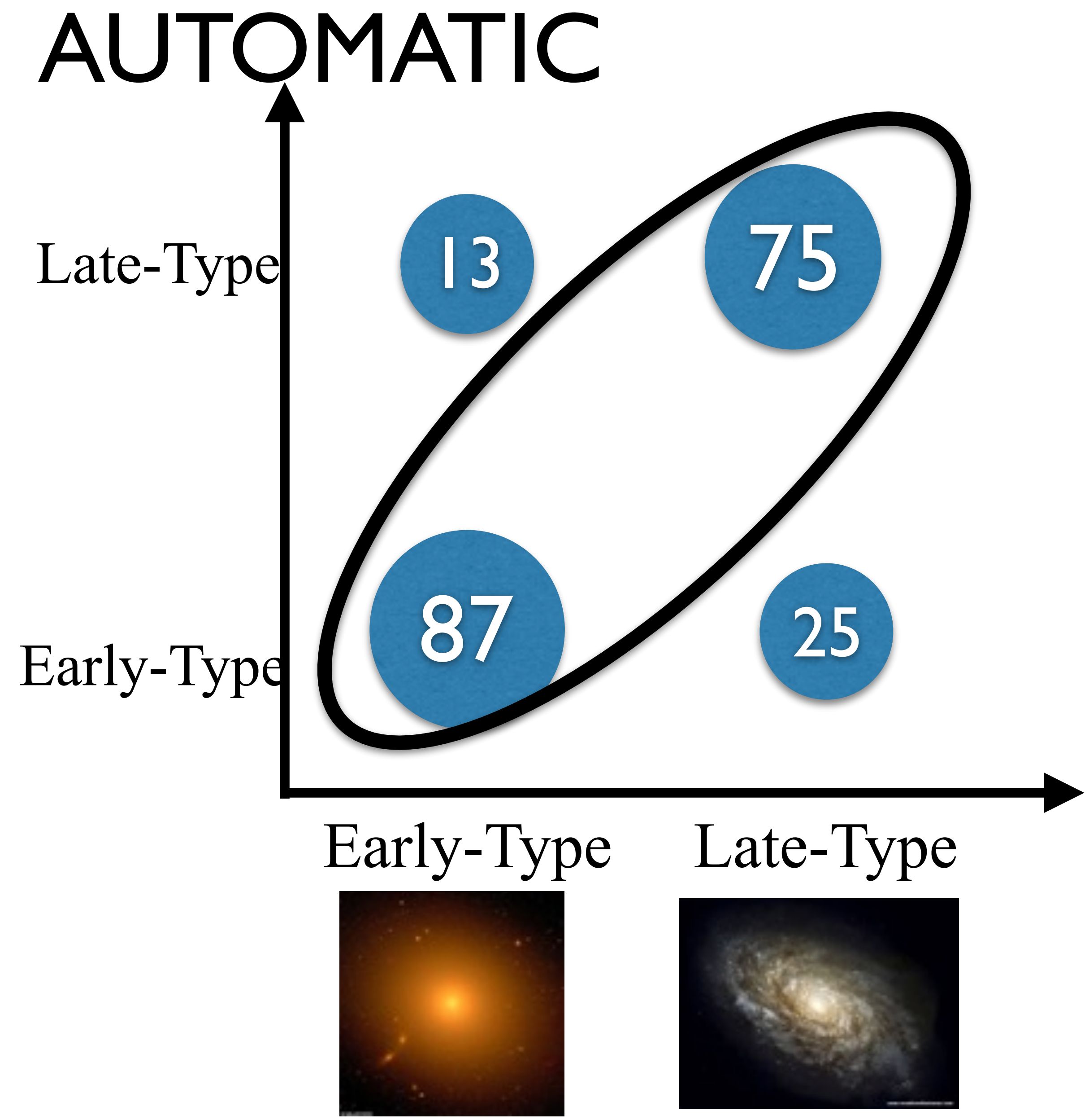

T00: Is the galaxy simply smooth and rounded, with no sign of a disk?
A0: Smooth
A1: Features or disk
A2: Star or artifact
T07: How rounded is it?
A0: Completely round
A1: In between
A2: Cigar shaped
T01: Could this be a disk viewed edge-on?
A0: Yes
A1: No
T08: Does the galaxy have a bulge at its centre? If so, what shape?
A0: Rounded
A1: Boxy
A2: No bulge
T02: Is there a sign of a bar feature through the centre of the galaxy?
A0: Bar
A1: No bar
T03: Is there any sign of a spiral arm pattern?
A0: Spiral
A1: No spiral
T09: How tightly wound do the spiral arms appear?
A0: Tight
A1: Medium
A2: Loose
Q1: smoothness, 6144 examples
average accuracy: 87.79%
Q2: edge-on, 3362 examples
average accuracy: 96.04%
Q3: bar, 2449 examples
average accuracy: 90.16%
classification accuracy
agreement / confidence
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100

Dieleman+15

HIGH-REDSHIFT MORPHOLOGIES (CANDELS)



CNNs

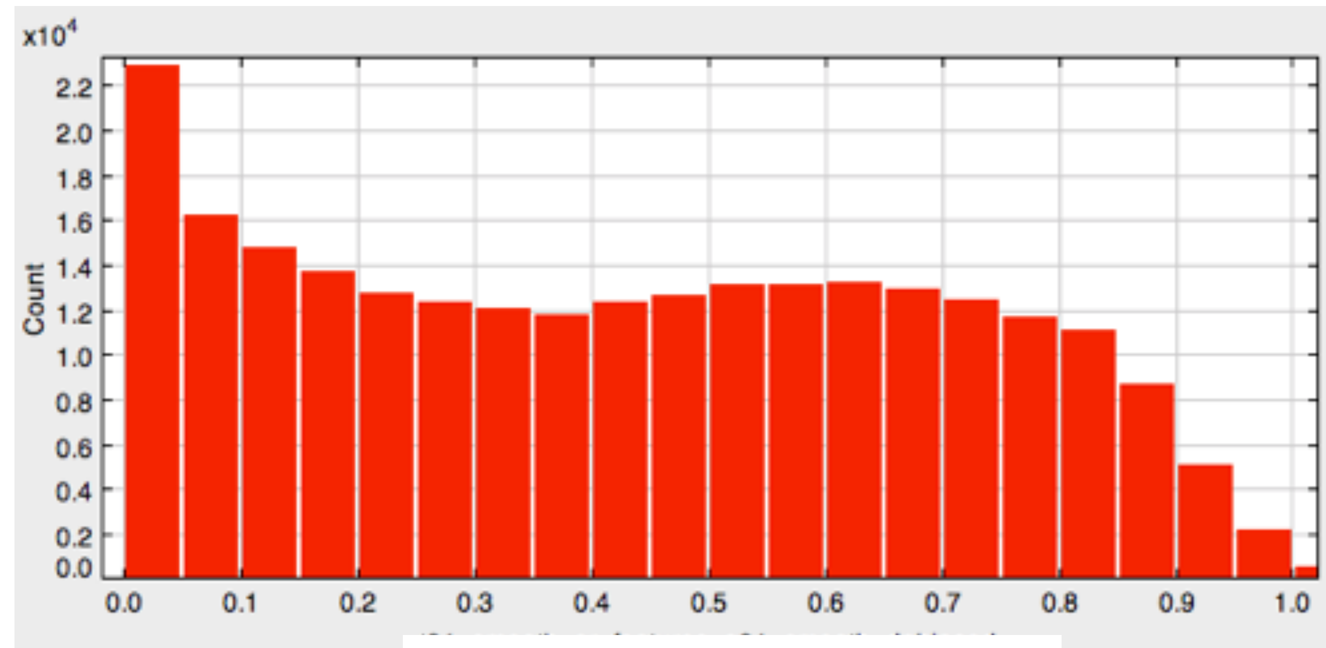


SVMs

MHC+15b

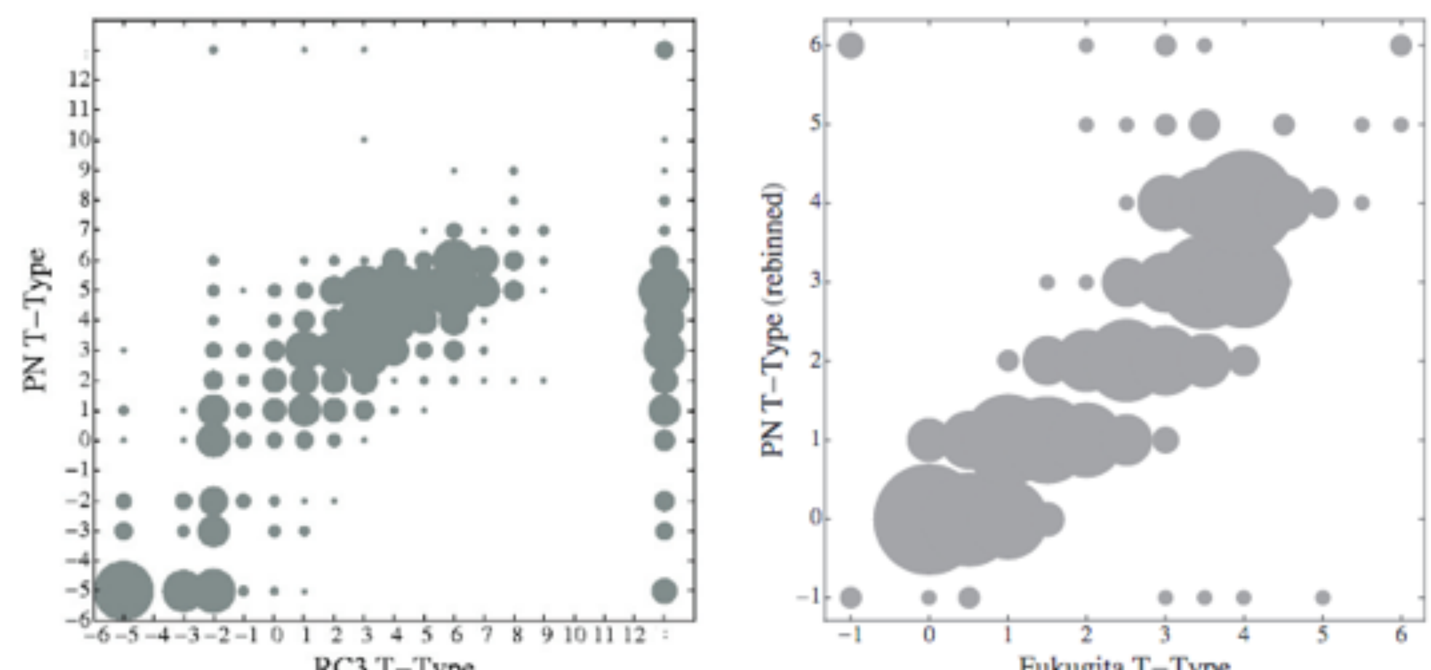
Can we improve human biases with machines?

GZOO



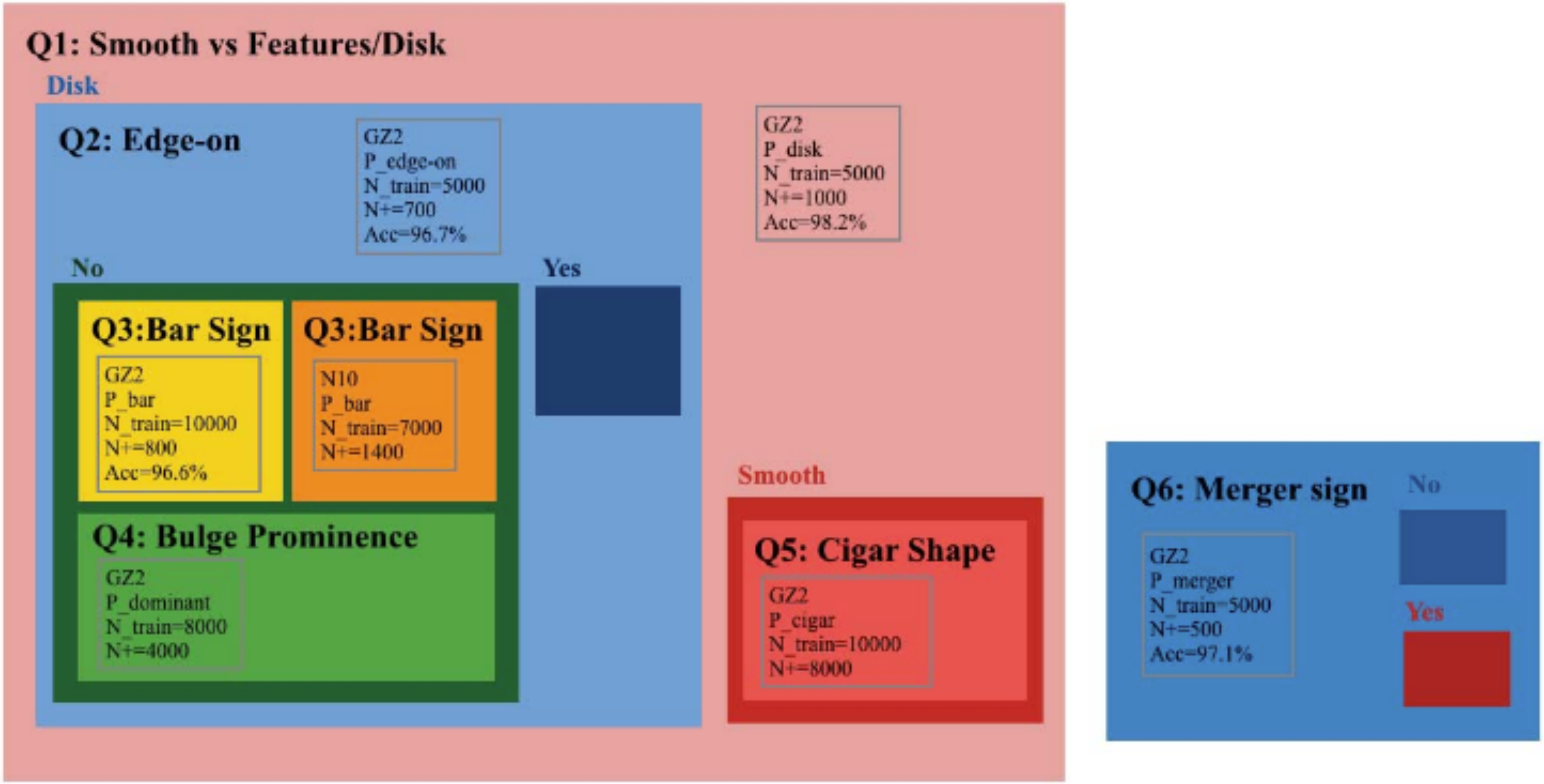
PROBABILITY

T-Type (astronomers)



Nair+10

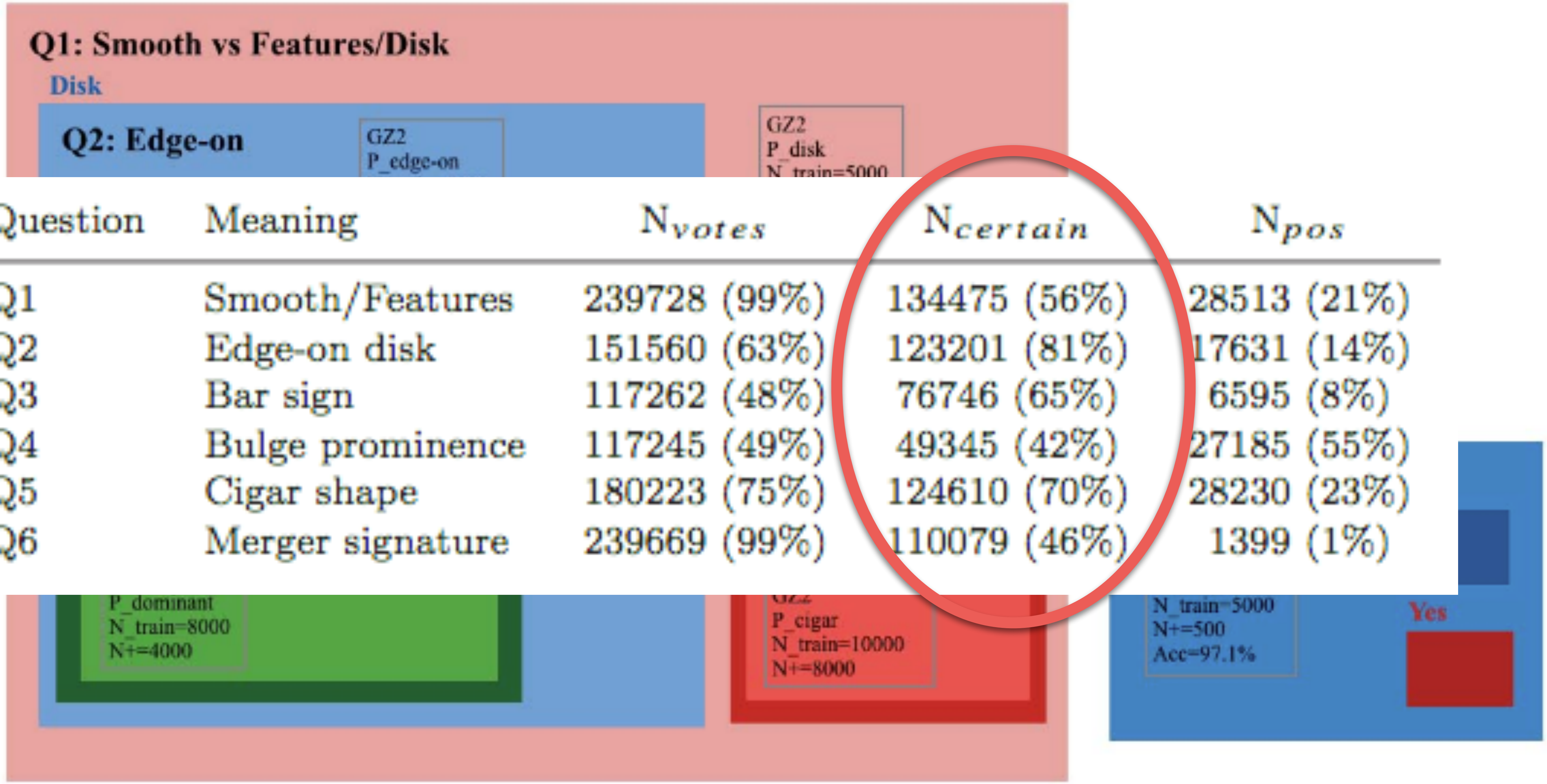
REVISITING THE SDSS MORPHOLOGY



Select only "safe" classifications for training [N>5, P>0.7]
Binary classification for each feature separately

Dominguez-Sanchez, MHC+17a

REVISITING THE SDSS MORPHOLOGY

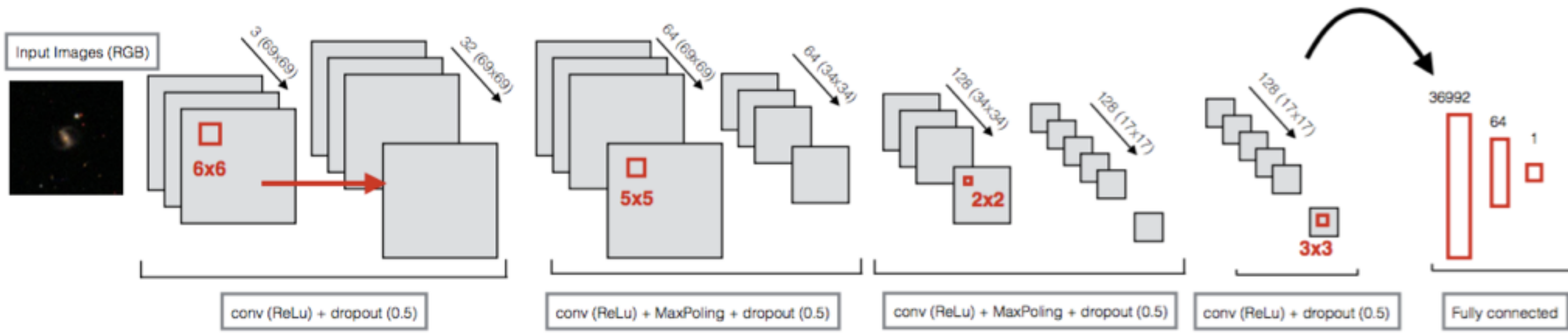


Question	Meaning	N_{votes}	$N_{certain}$	N_{pos}
Q1	Smooth/Features	239728 (99%)	134475 (56%)	28513 (21%)
Q2	Edge-on disk	151560 (63%)	123201 (81%)	17631 (14%)
Q3	Bar sign	117262 (48%)	76746 (65%)	6595 (8%)
Q4	Bulge prominence	117245 (49%)	49345 (42%)	27185 (55%)
Q5	Cigar shape	180223 (75%)	124610 (70%)	28230 (23%)
Q6	Merger signature	239669 (99%)	110079 (46%)	1399 (1%)

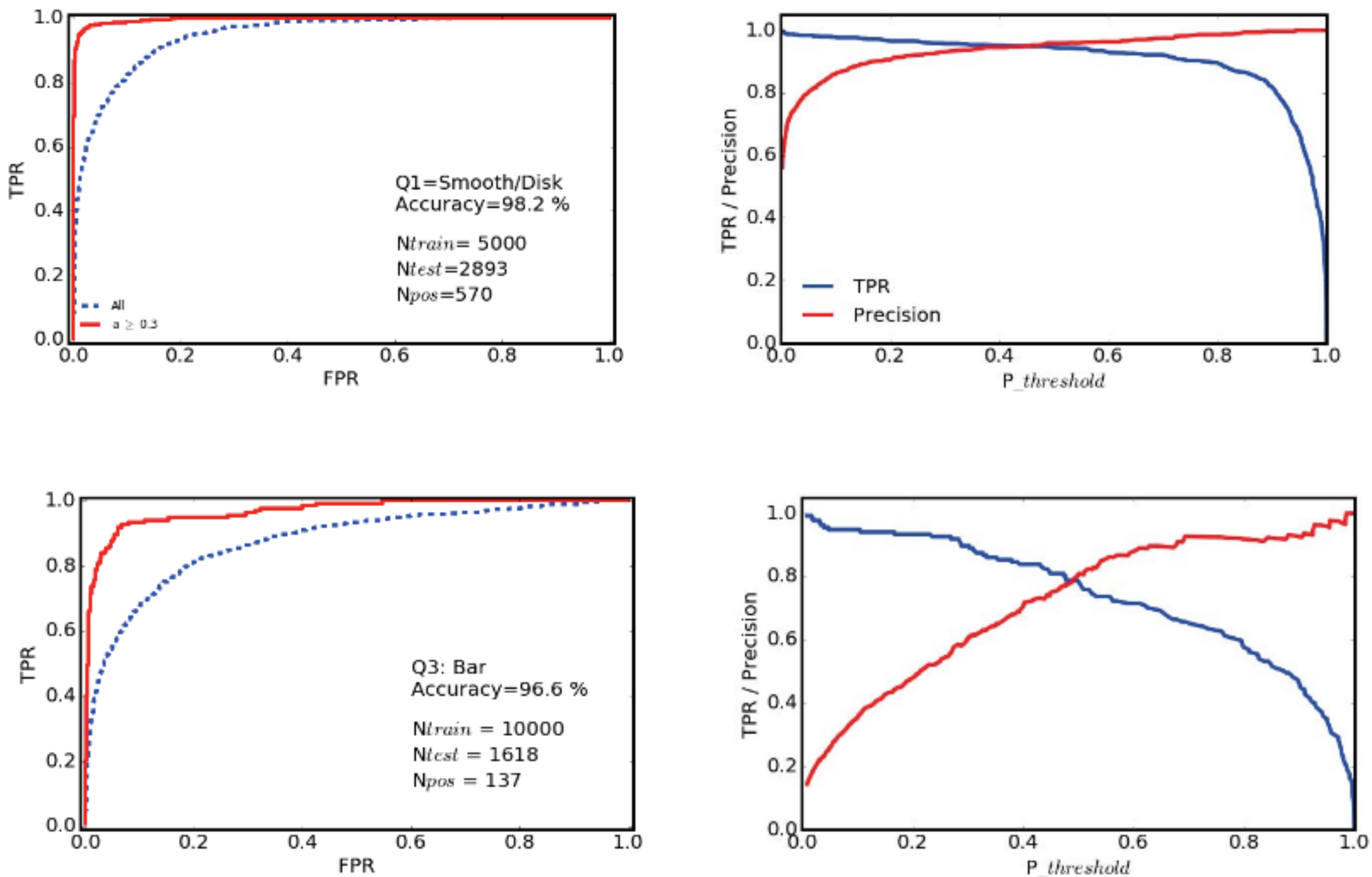
Select only "safe" classifications for training [N>5, P>0.7]
Binary classification for each feature separately

Dominguez-Sanchez, MHC+17a

VERY SIMPLE ARCHITECTURE

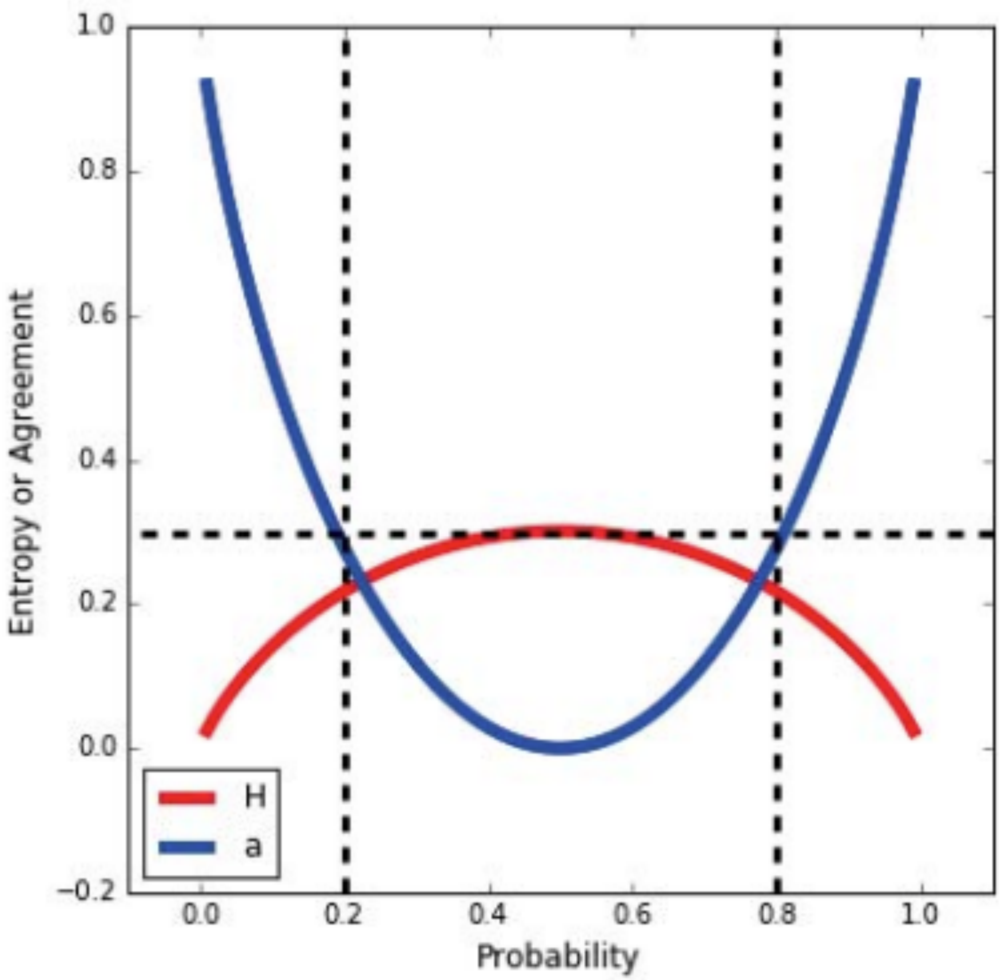


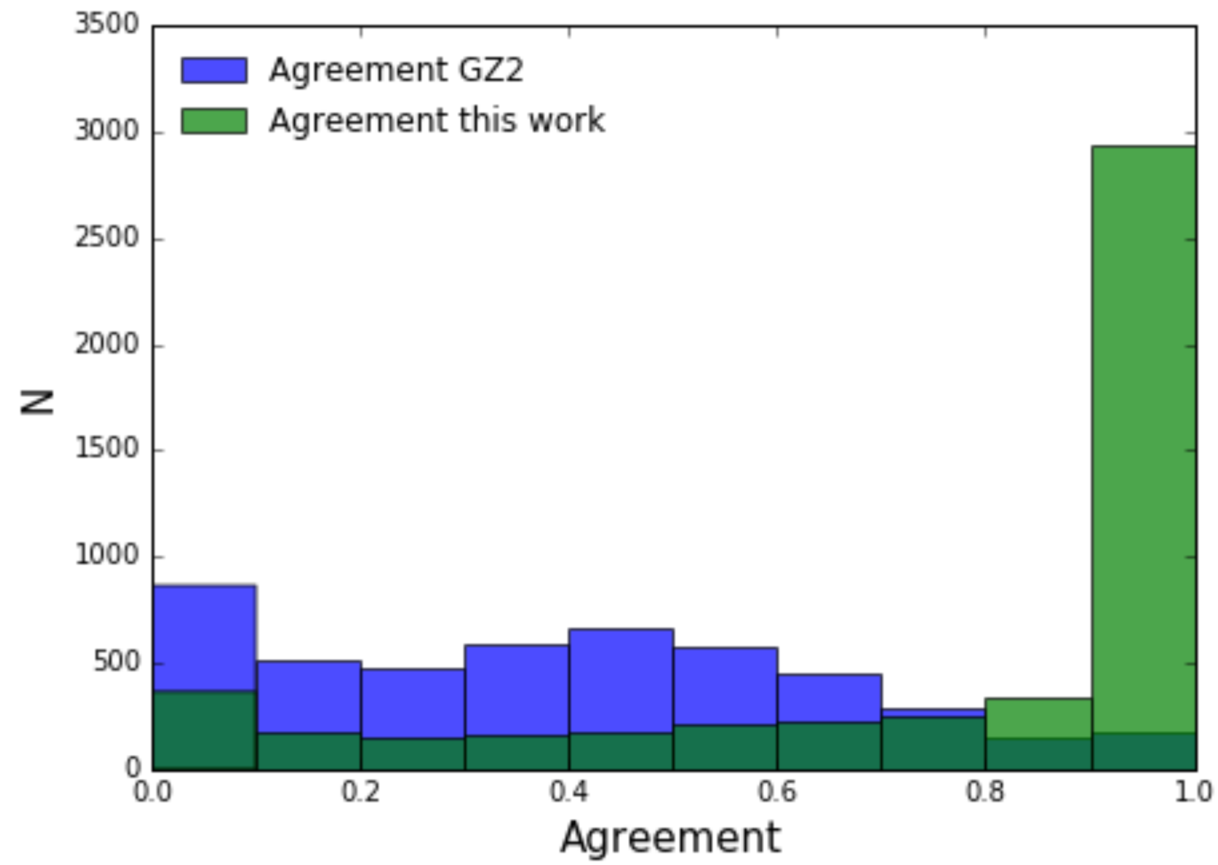
Dominguez-Sanchez, MHC+17a



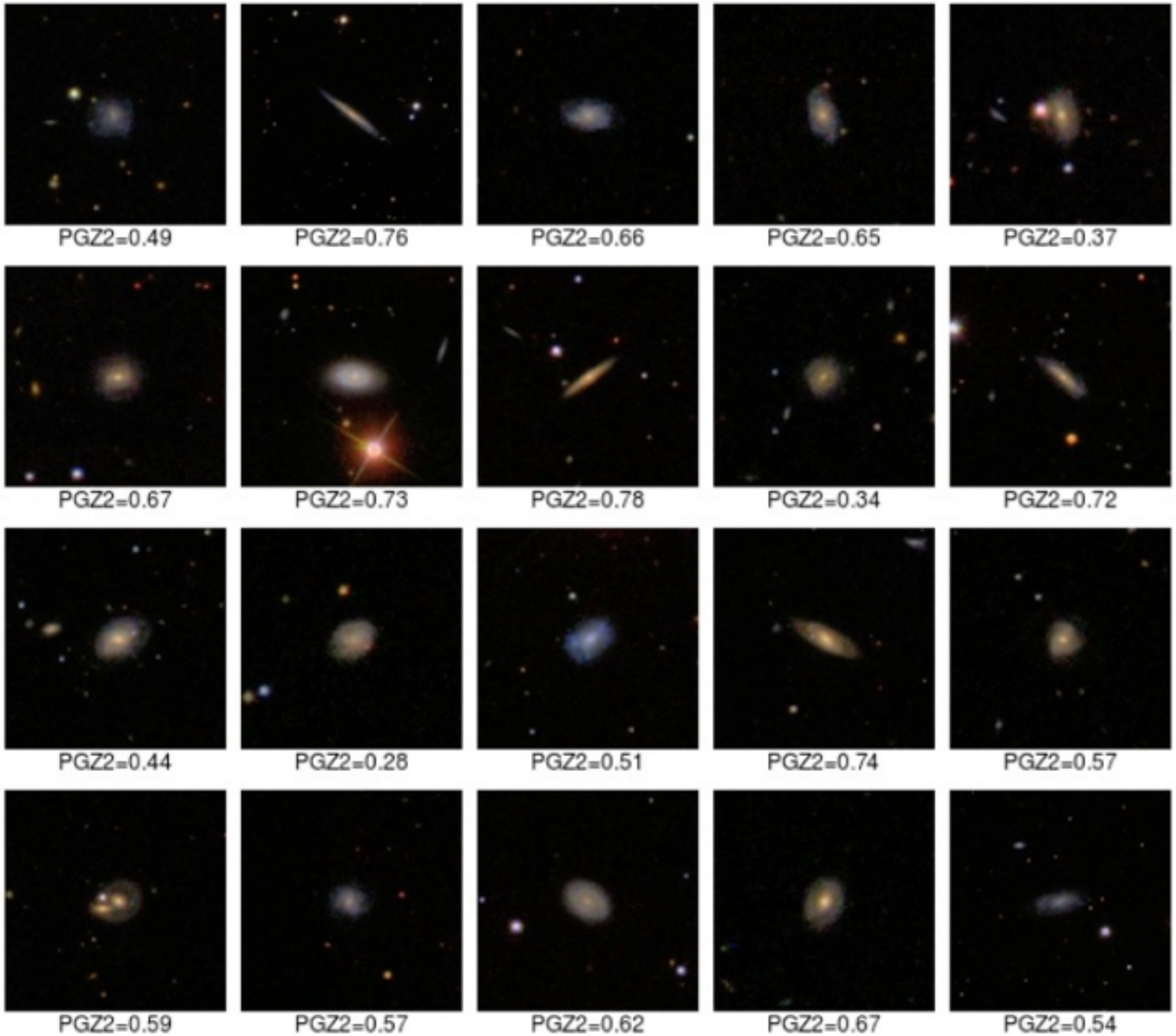
Dominguez-Sanchez, MHC+17a

Entropy of classification





Dominguez-Sanchez, MHC+17a



SECURE DISKS GALAXIES FOR DL - UNCLEAR FOR PEOPLE

Dominguez-Sanchez, MHC+17a

How robust to different datasets?
Do we always need a big training set?

DATA FROM NEW SURVEY

How robust to different datasets? Do we always need a big training set?

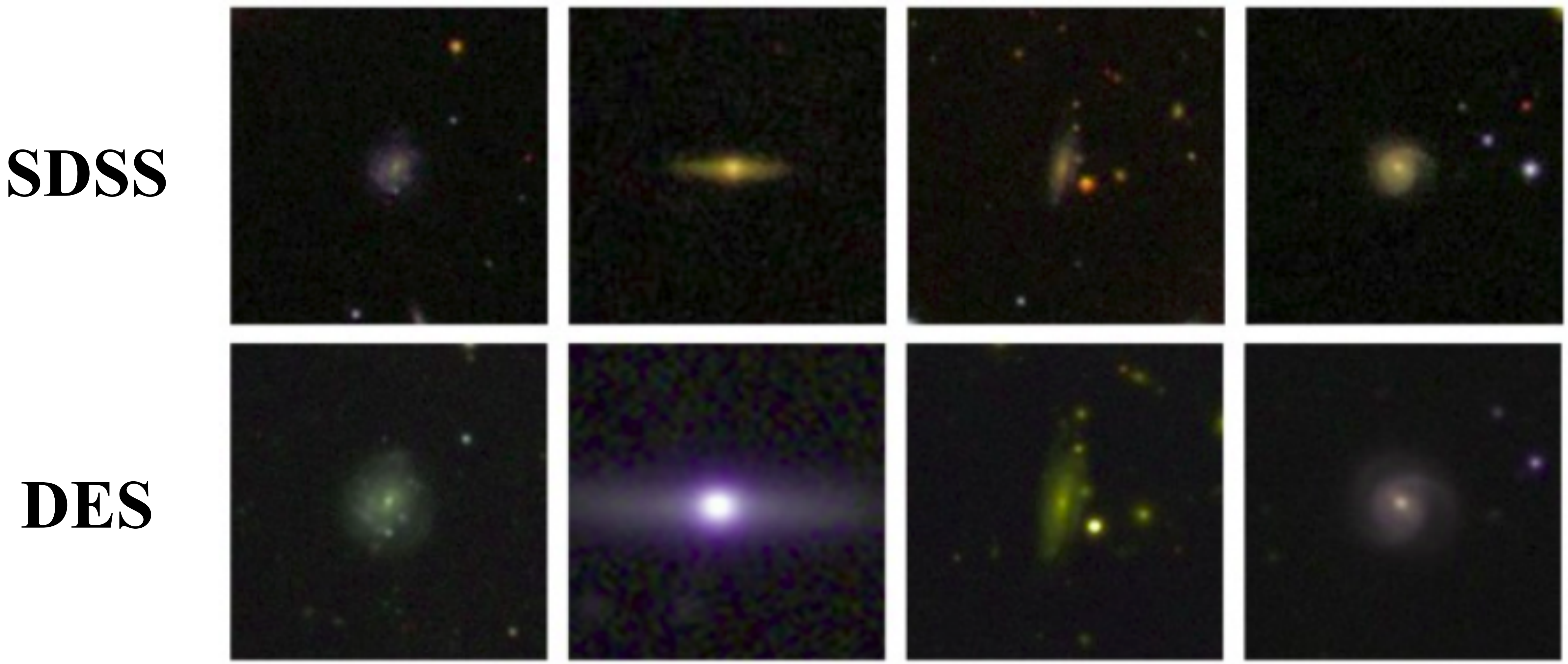
DEEP-LEARNING BASED MACHINE

"Improved" Galaxy ZOO like classifications for for the entire sample

Transfer knowledge?

Human classifications from existing survey

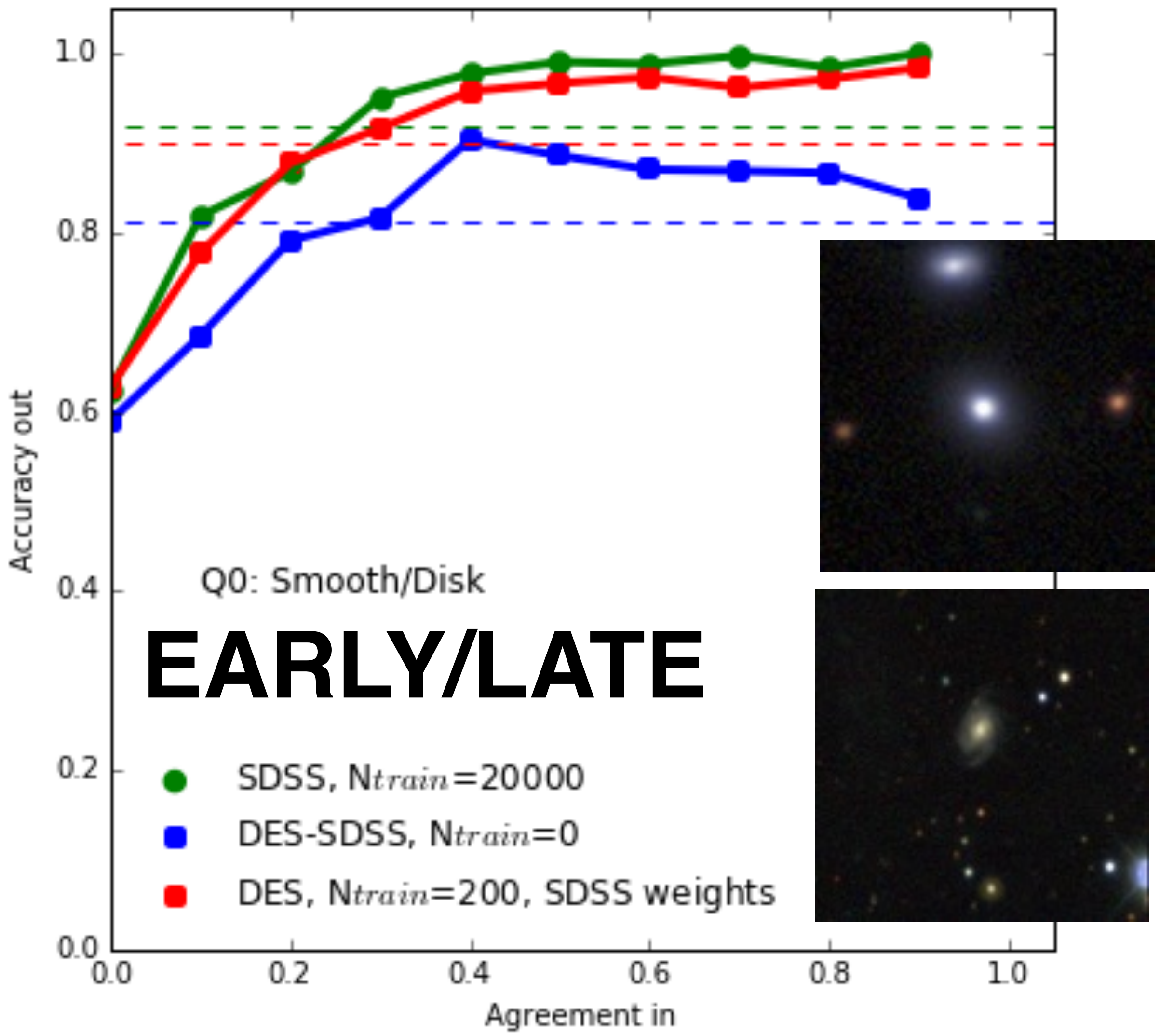
Dominguez-Sanchez, MHC+17b

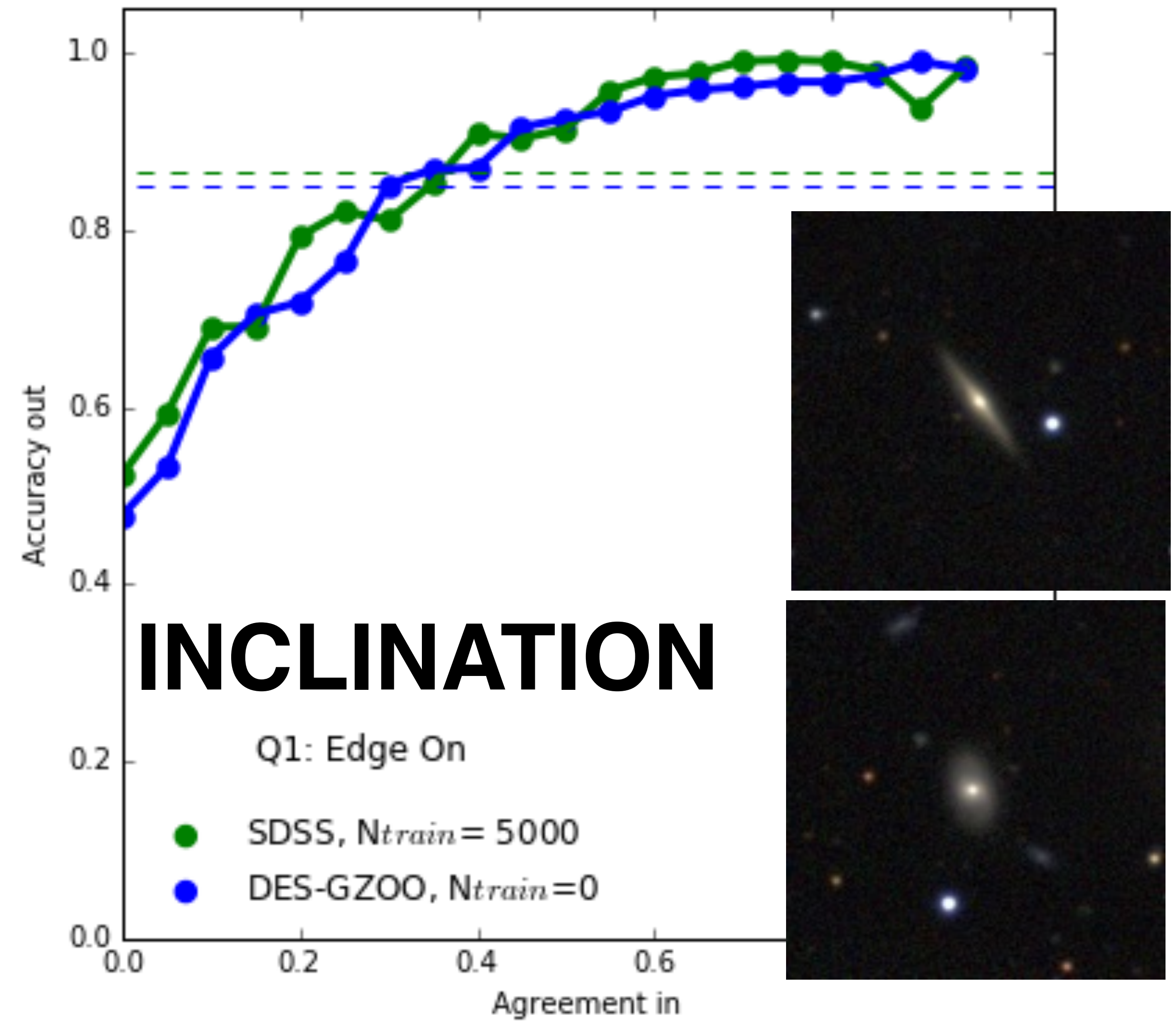


Dominguez-Sanchez, MHC+17b

Knowledge transfer from SDSS to DES

DOMINGUEZ-SANCHEZ, HUERTAS-COMPANY, BERNARDI et al. 17b

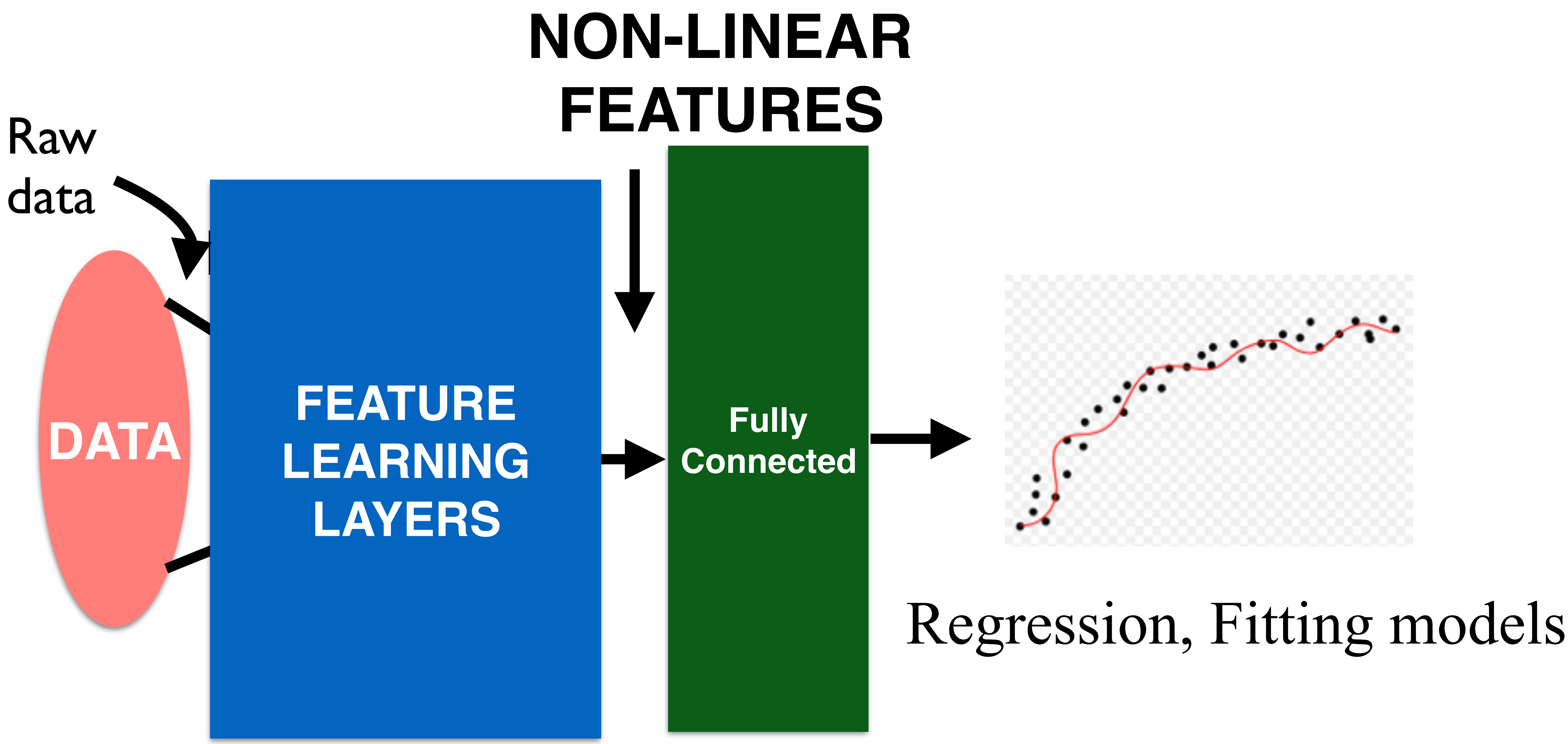




Only 200 (**1%!**) objects classified in DES are needed to reach an accuracy >90% if a machine trained on the SDSS is used

For some properties, i.e. EDGE-ON galaxies. No training at all is needed to go from SDSS to DES

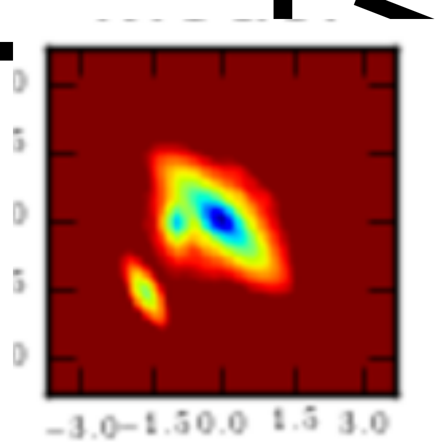
GROUP #2: Efficient and fast quantitative measurements on large amount of (multi-lambda) data [photoz's, sizes, ellipiticities]



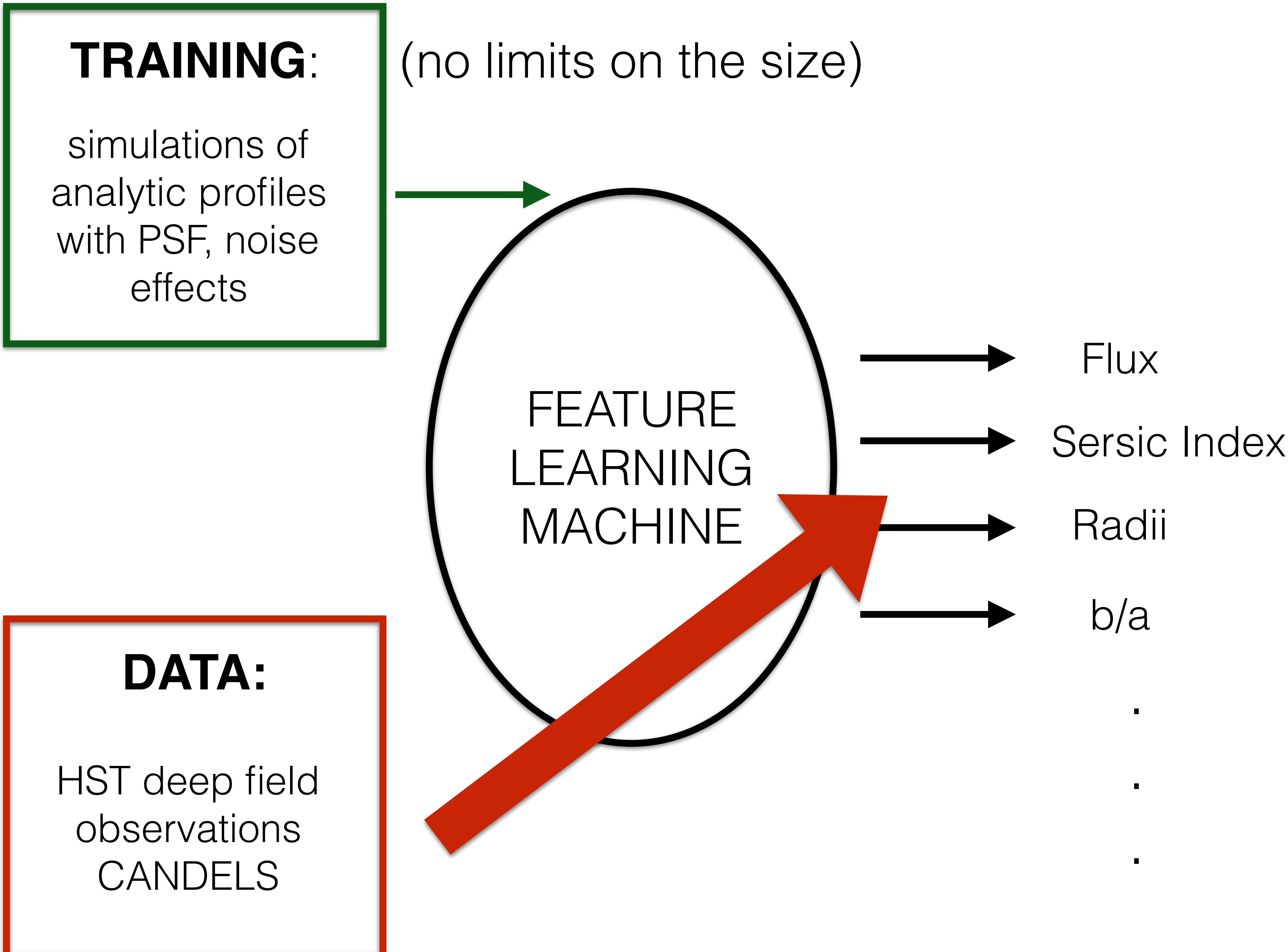
MORPHOMETRICS

flux, size, axis-ratio, PA,
Sersic index

[PSF corrected]



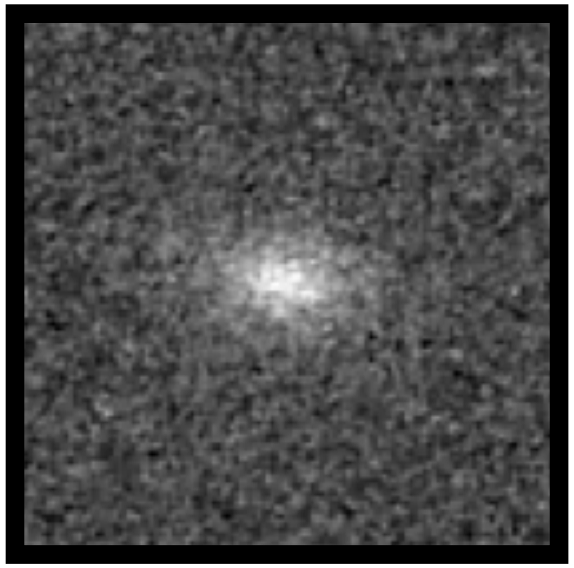
MODEL → **MODEL*PSF** → **MINIMIZATION**

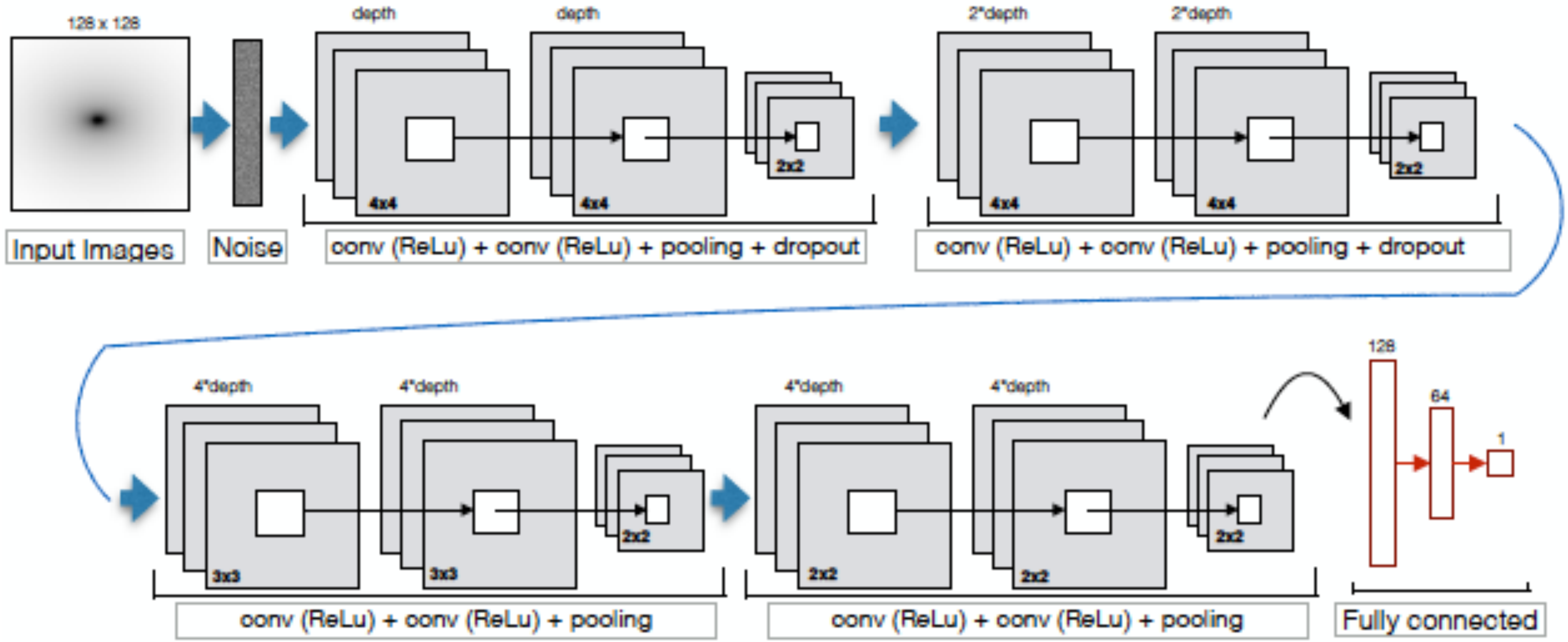


TRAINING:
simulations of analytic profiles with PSF, noise effects
(no limits on the size)
FEATURE LEARNING MACHINE
Flux
Sersic Index
Radii
b/a
.
.
.
DATA:
HST deep field observations CANDELS

Standard analytic profiles

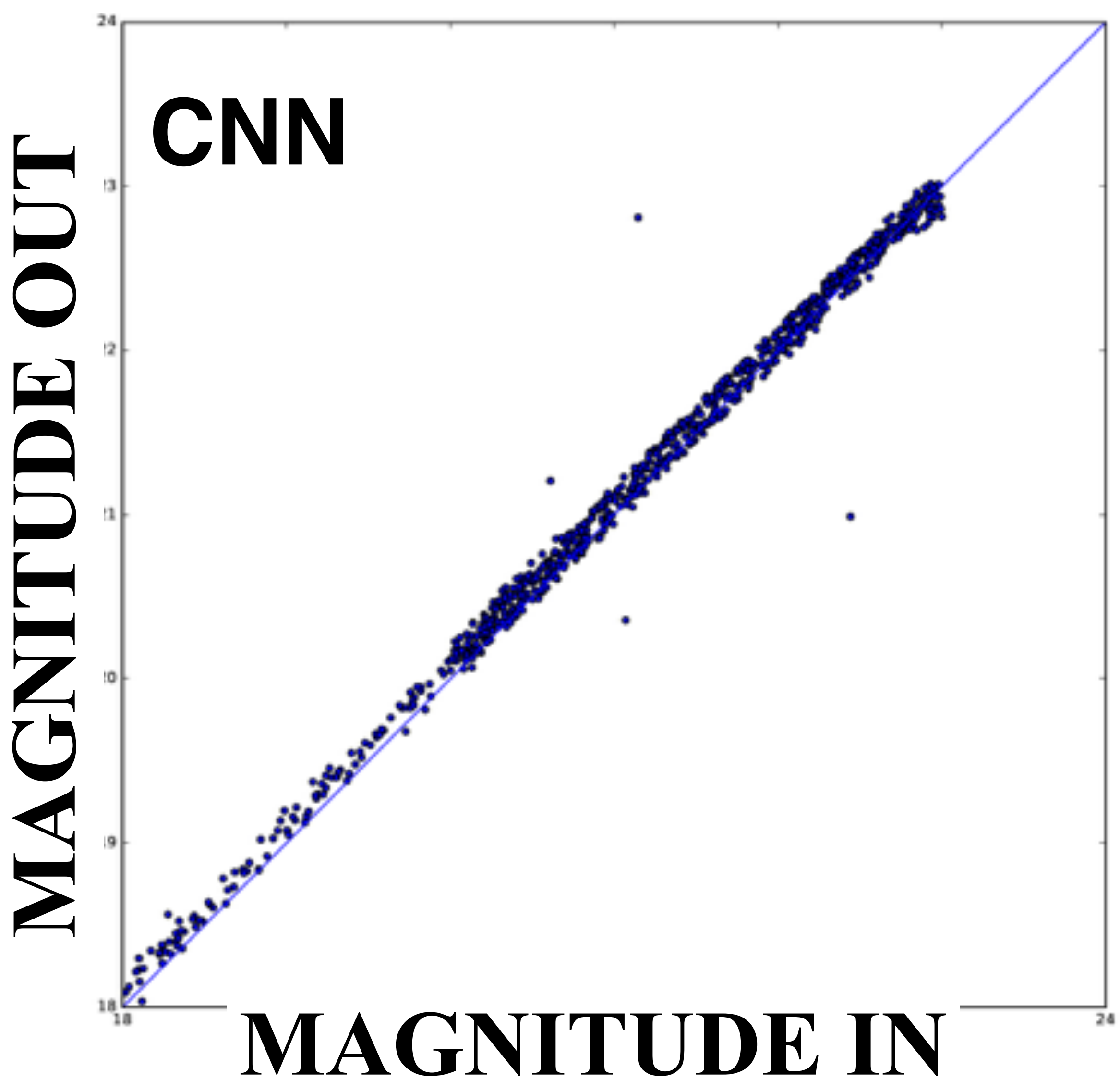
- 100.000-300.000 galaxies [GALSIM]
  - Real HST background added + PSF (F160)
  - Random distribution of parameters (uniform):
    - 18<Mag<24, 0<BT<1, <Nb<, Nd=1, 0.2<log(rb)<1.3, 0.2<log(rd)<1.5, 0.05<eb<0.95, 0.05<ed<0.95, 0<PA<180
  - 64*64 stamps
  - **FULLY IDEALISTIC -**
    NO COMPANIONS
    NO IRREGULARS
    NO CLUMPY!

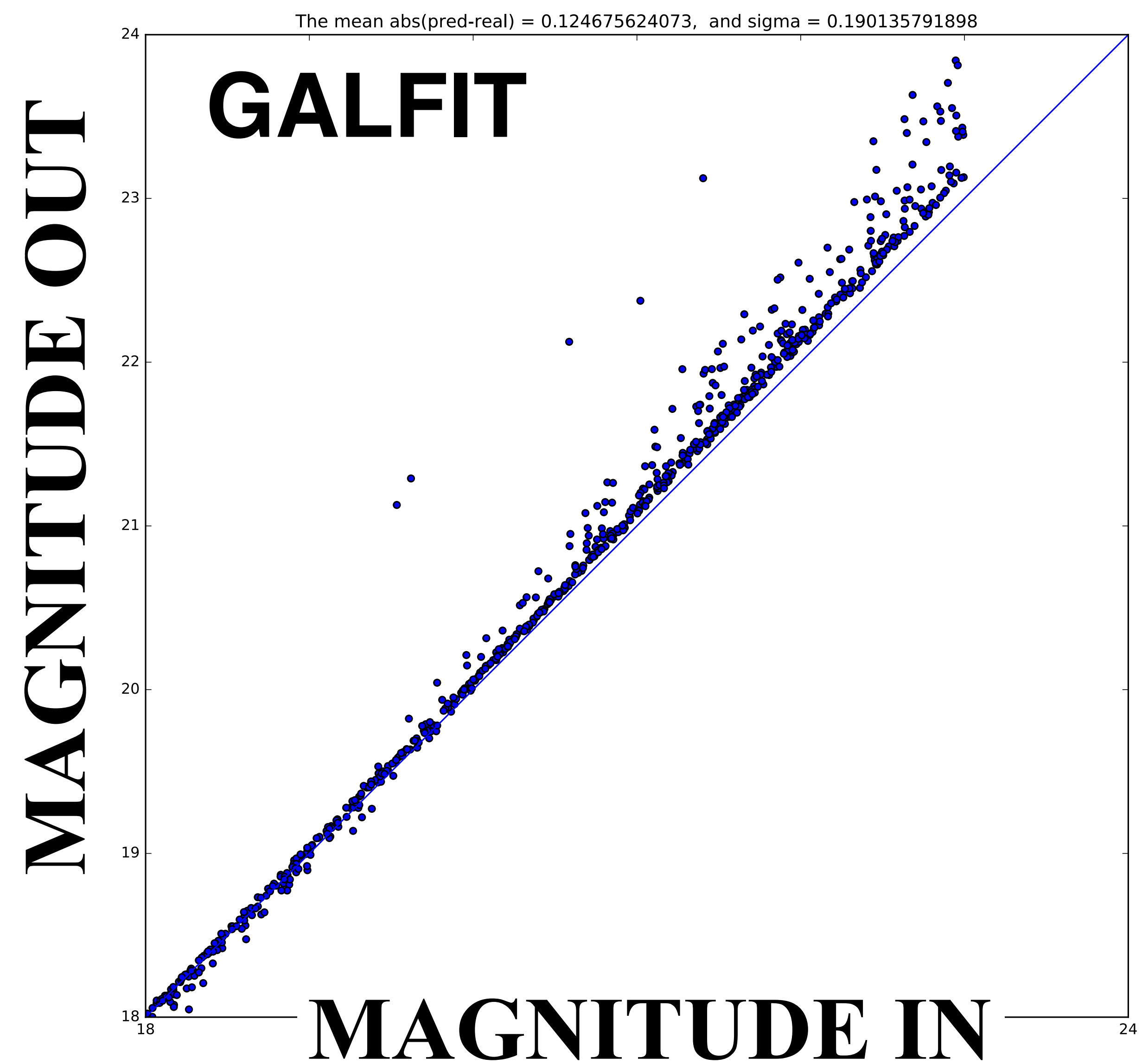



128 x 128
Input Images
Noise
depth
depth
4x4
4x4
2x2
conv (ReLu) + conv (ReLu) + pooling + dropout
2*depth
2*depth
4x4
4x4
2x2
conv (ReLu) + conv (ReLu) + pooling + dropout
4*depth
4*depth
3x3
3x3
2x2
conv (ReLu) + conv (ReLu) + pooling
4*depth
4*depth
2x2
2x2
2x2
conv (ReLu) + conv (ReLu) + pooling
128
64
1
Fully connected

ON SIMULATIONS

TUCCILLO, HUERTAS-COMPANY et al. 17

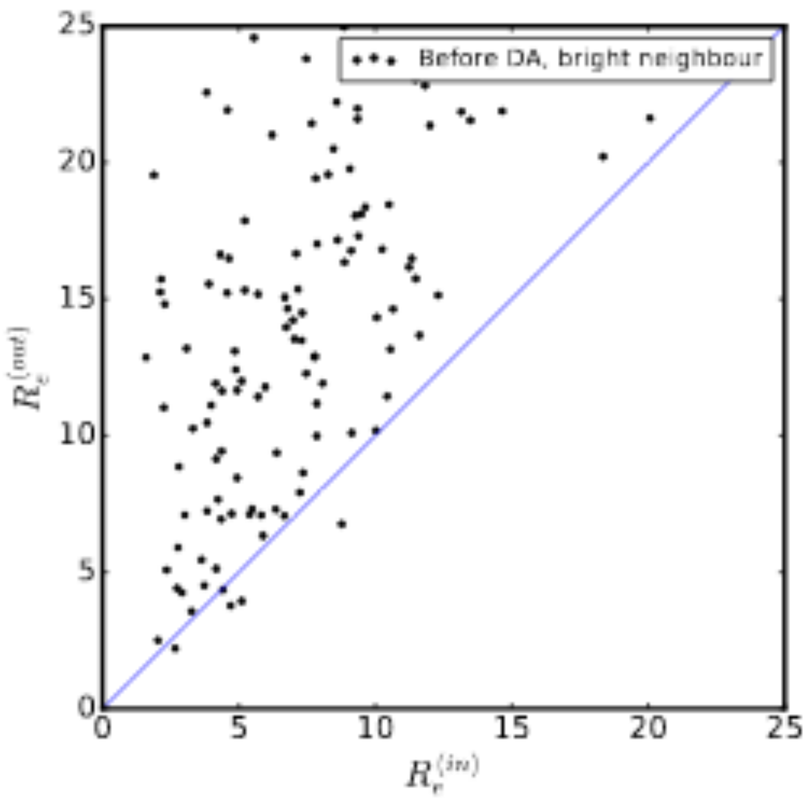




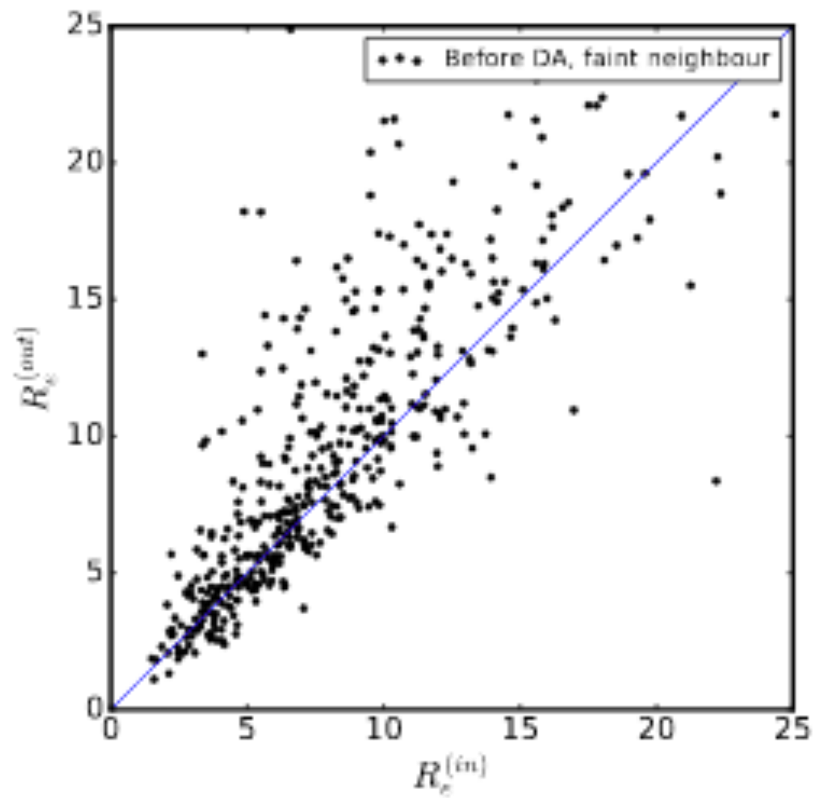
VERY SIMILAR RESULTS ON THE SAME SIMULATIONS, BUT CNNs are several orders of magnitude faster [**3.5 hrs vs. <1 sec for ~1000 objects**]

MORPHOMETRY OF REAL GALAXIES TRAINED ON ANALYTIC PROFILES

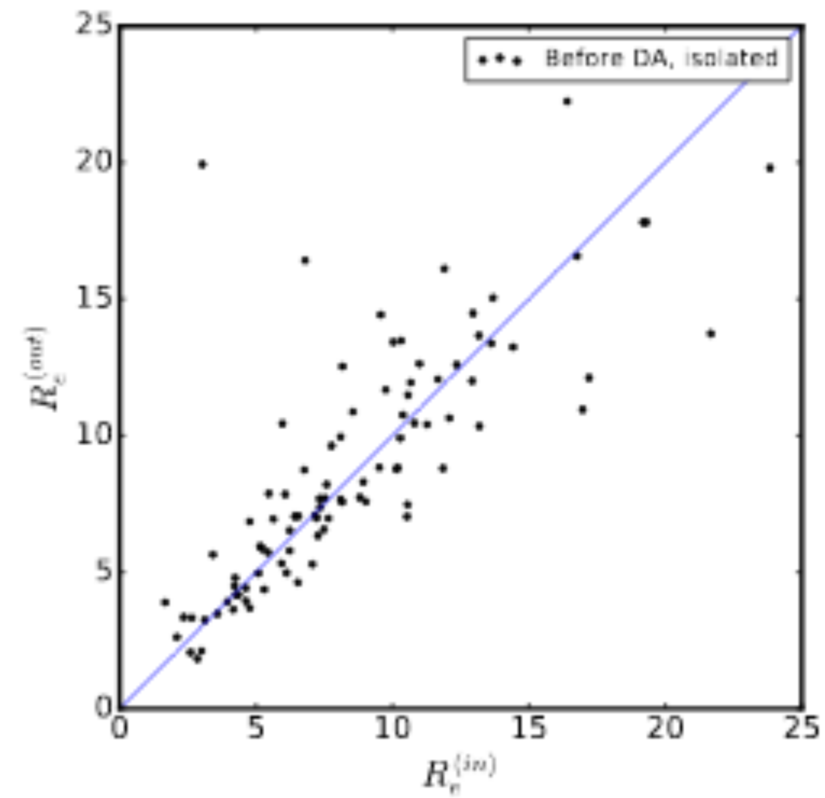
BRIGHT NEIGH. **FAINT NEIGH.** **ISOLATED**



(c) BDA bright neigbours



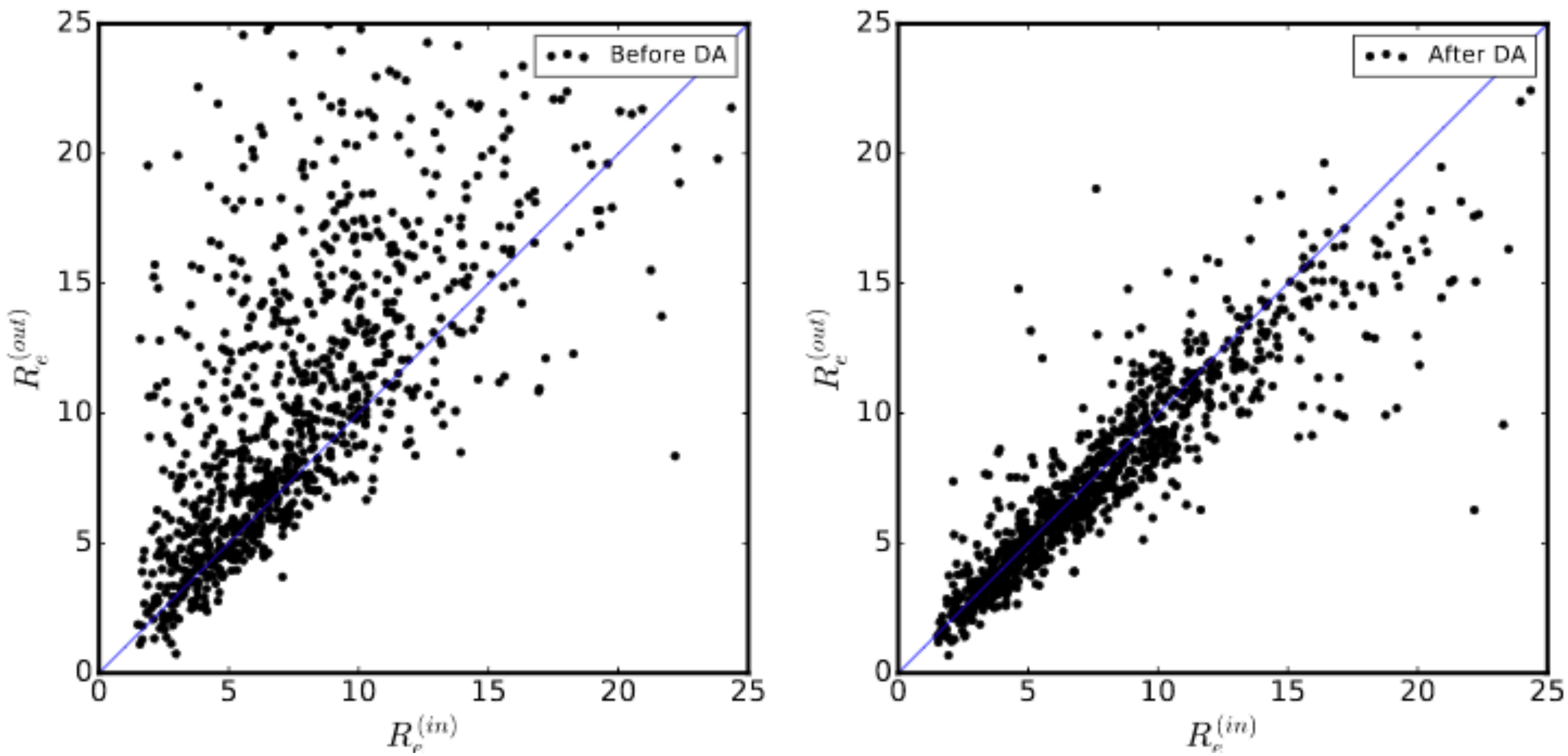
(d) BDA faint neigbours



(e) BDA isolated galaxies

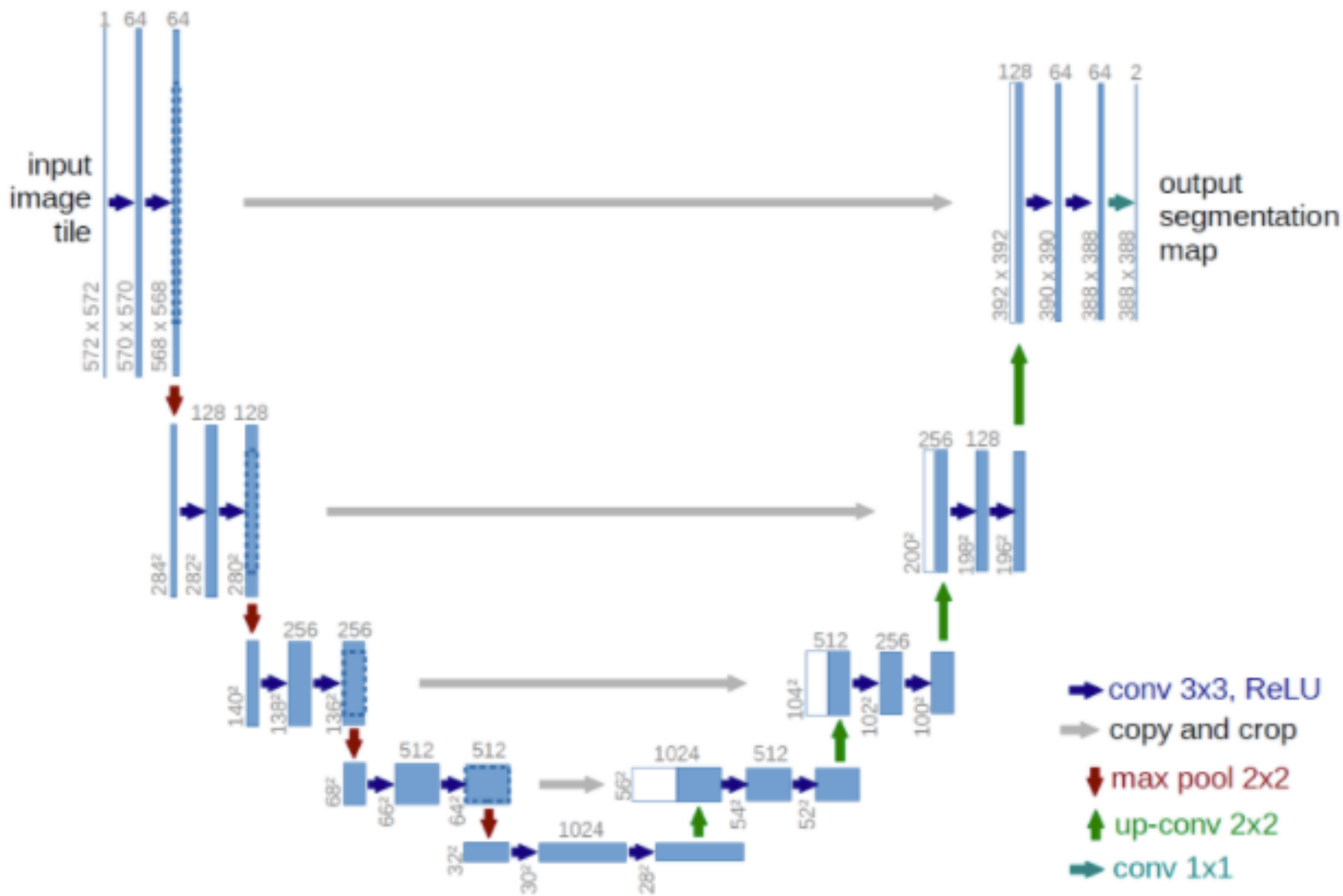
TUCCILLO, HUERTAS-COMPANY et al. 17

DOMAIN ADAPTATION: 0.1% OF "REAL" GALAXIES



TUCCILLO, HUERTAS-COMPANY et al. 17

Coming soon: U-net for bulge/disc decompositions…

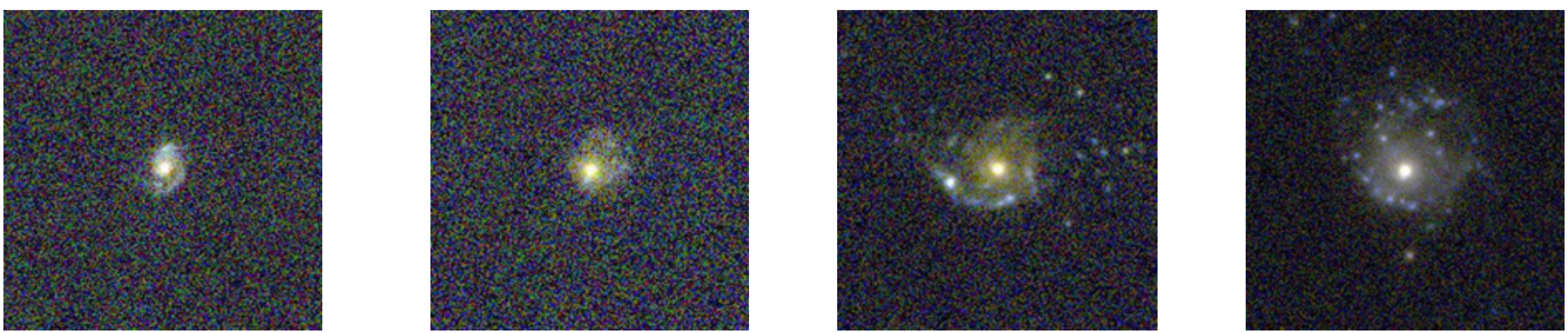


VELA hydrodynamic simulations

Ceverino+15

35 high res (~20pc) zoom-in simulations
hydroART
radiative and supernovae feedback
stops at z=1 - Mh=10^11-2.10^12

mock images [sunrise]
Tstep ~ 200Myrs
10 projections
HST like



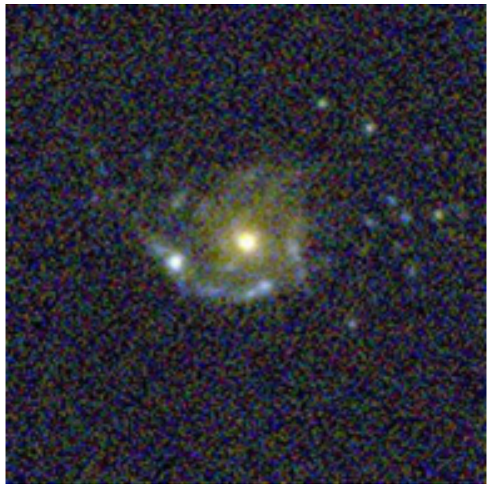
[G. Snyder, J. Lotz]

Margalef, MHC in prep.

CANDELS (observations)

Re (2D, light)

DEEPLEGATO

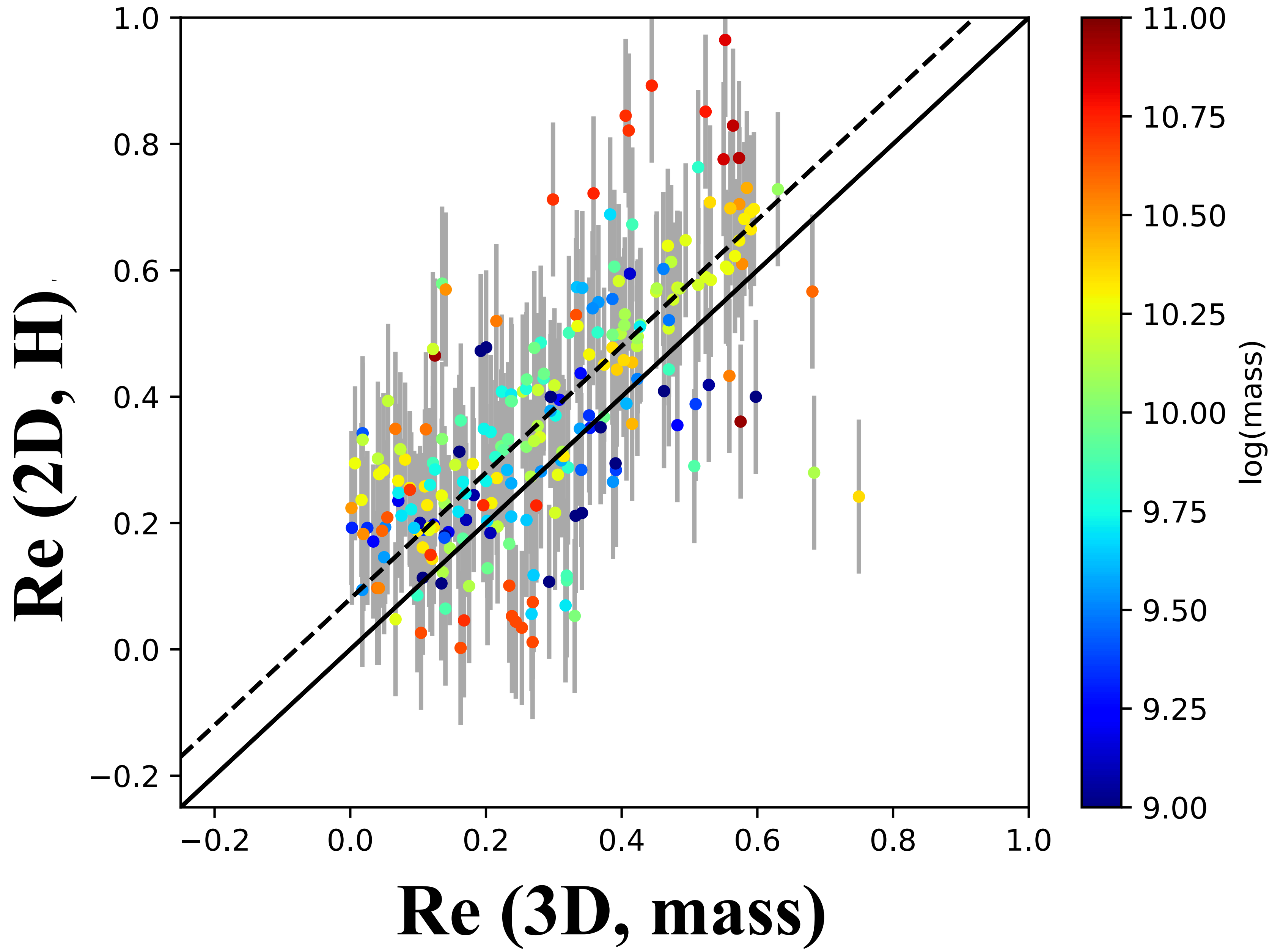


VELA (simulation)

Re (2D, light)

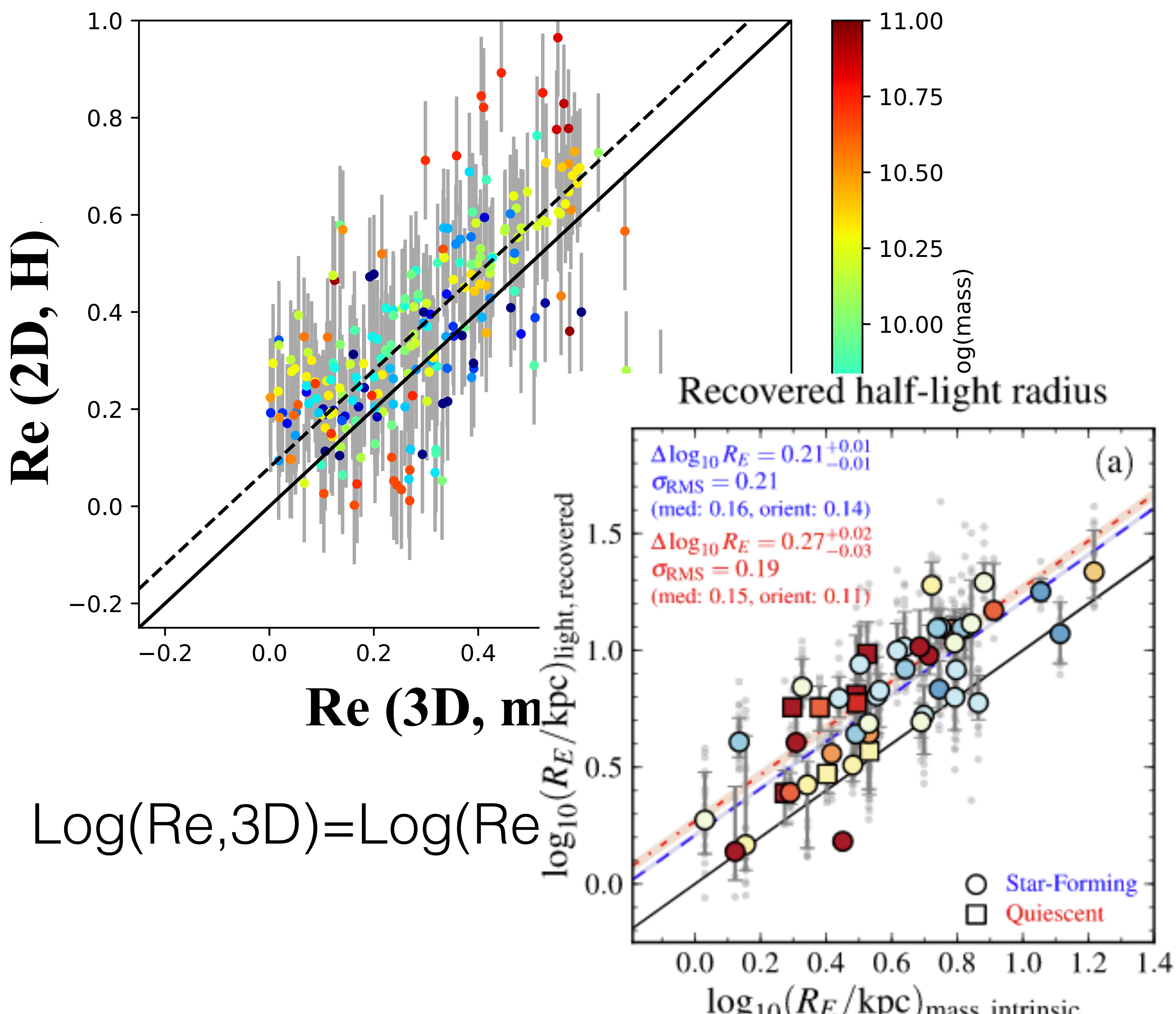
SIMULATION METADATA

Re (3D, mass,<0.1Rvir)



Log(Re,3D)=Log(Re,2D)*0.8+0.13

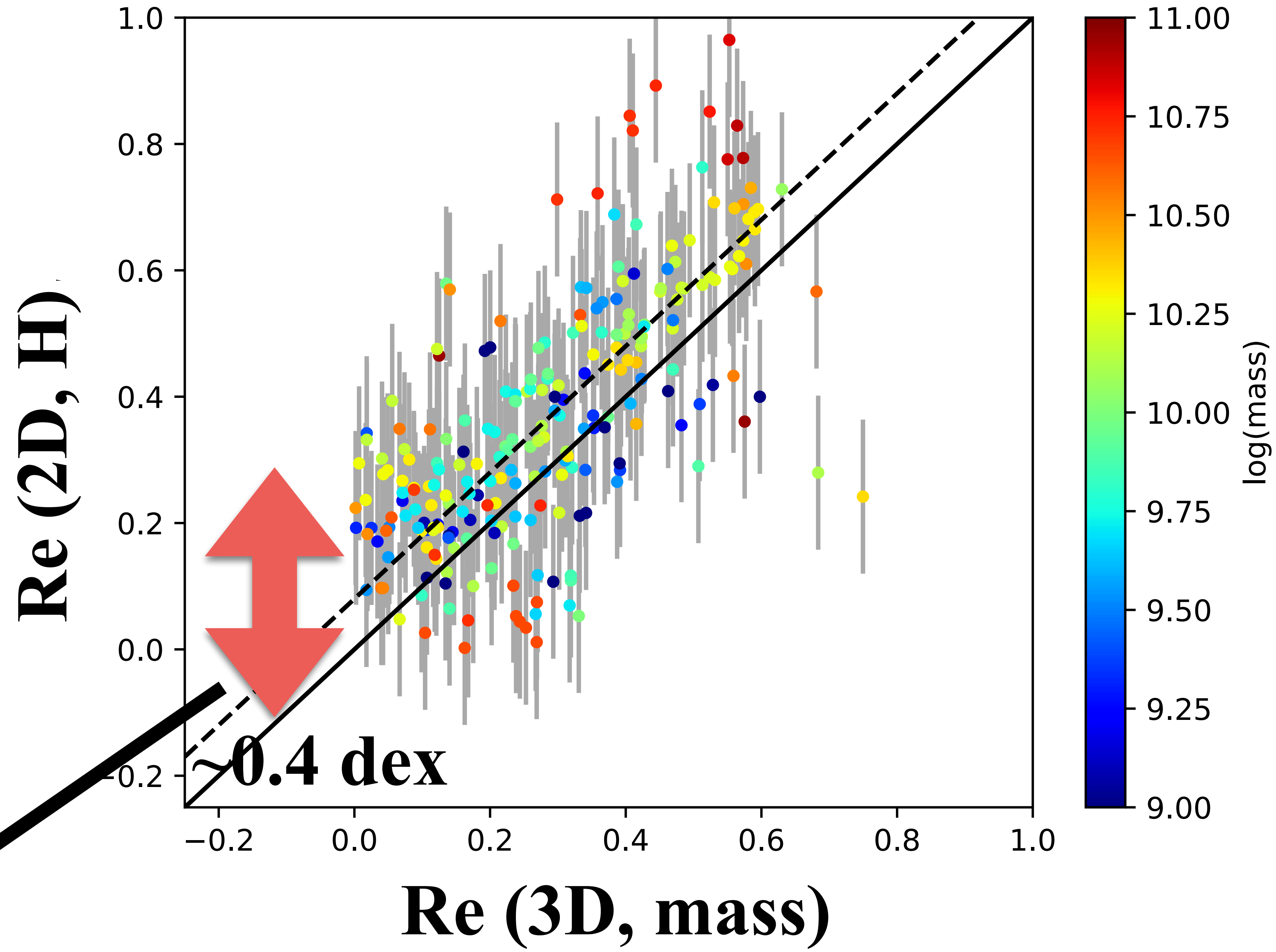
Margalef, MHC in prep.



Log(Re,3D)=Log(Re

Margalef, MHC in prep.

Price+17

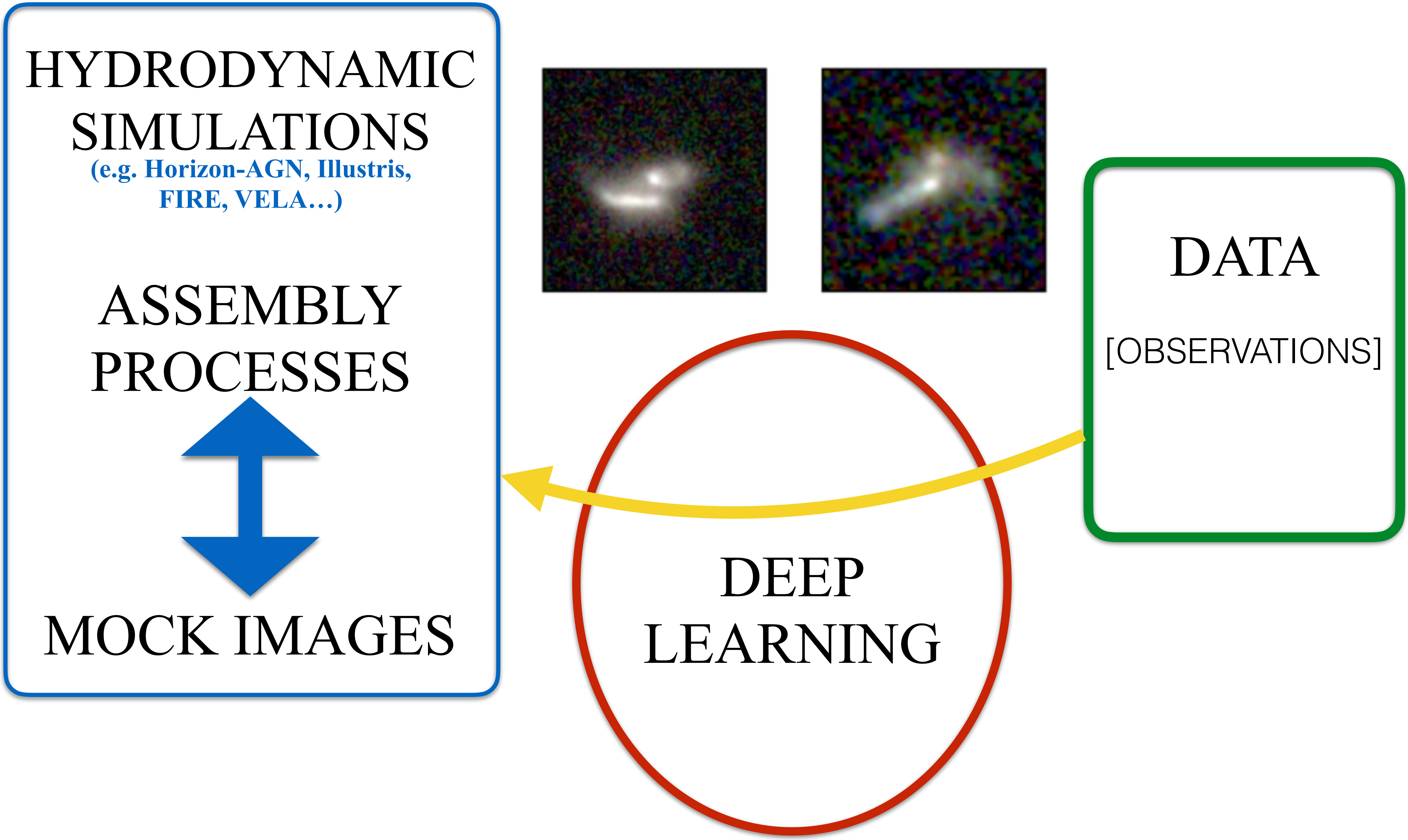


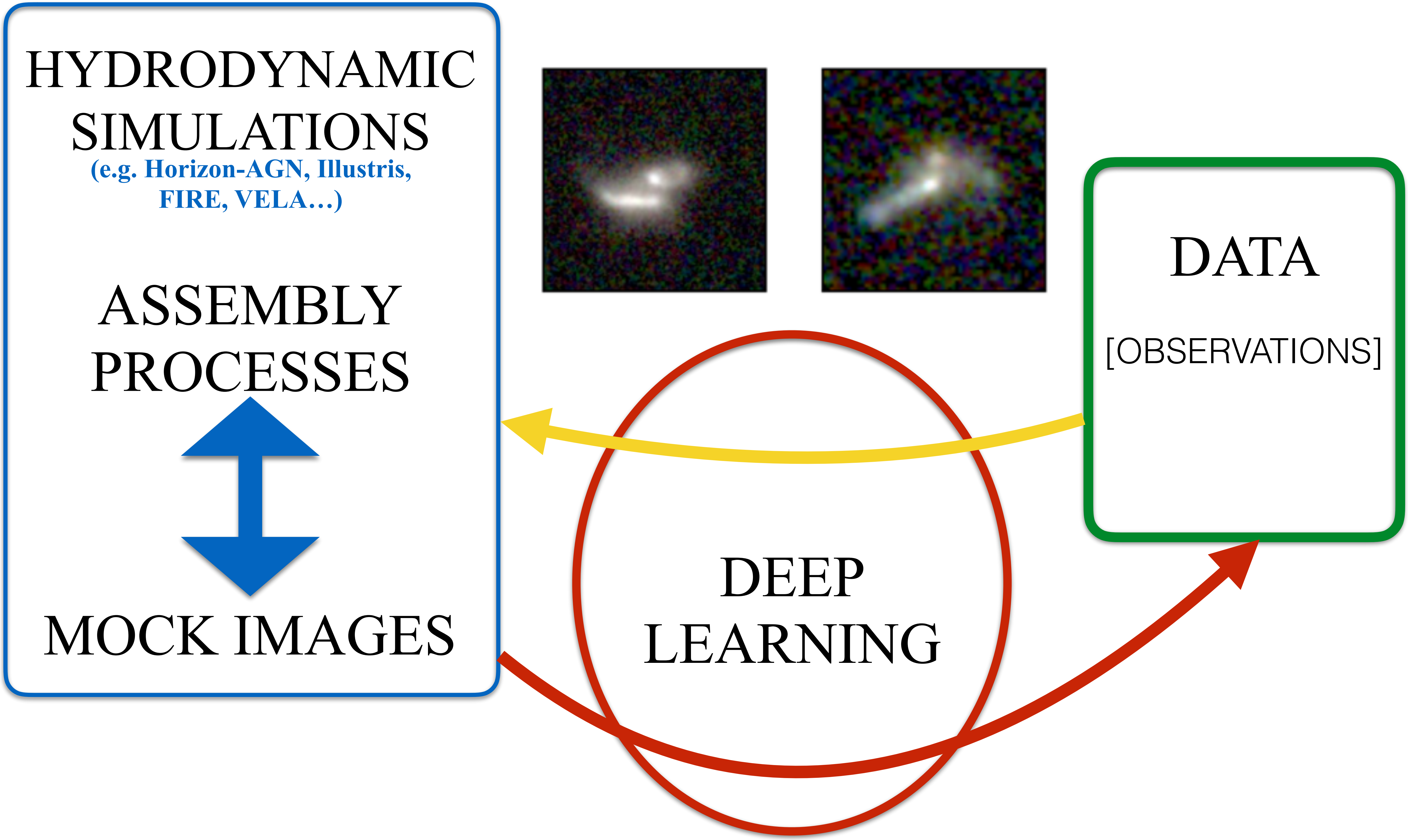
reduce scatter with DL?

Log(Re,3D)=Log(Re,2D)*0.8+0.13

Margalef, MHC in prep.

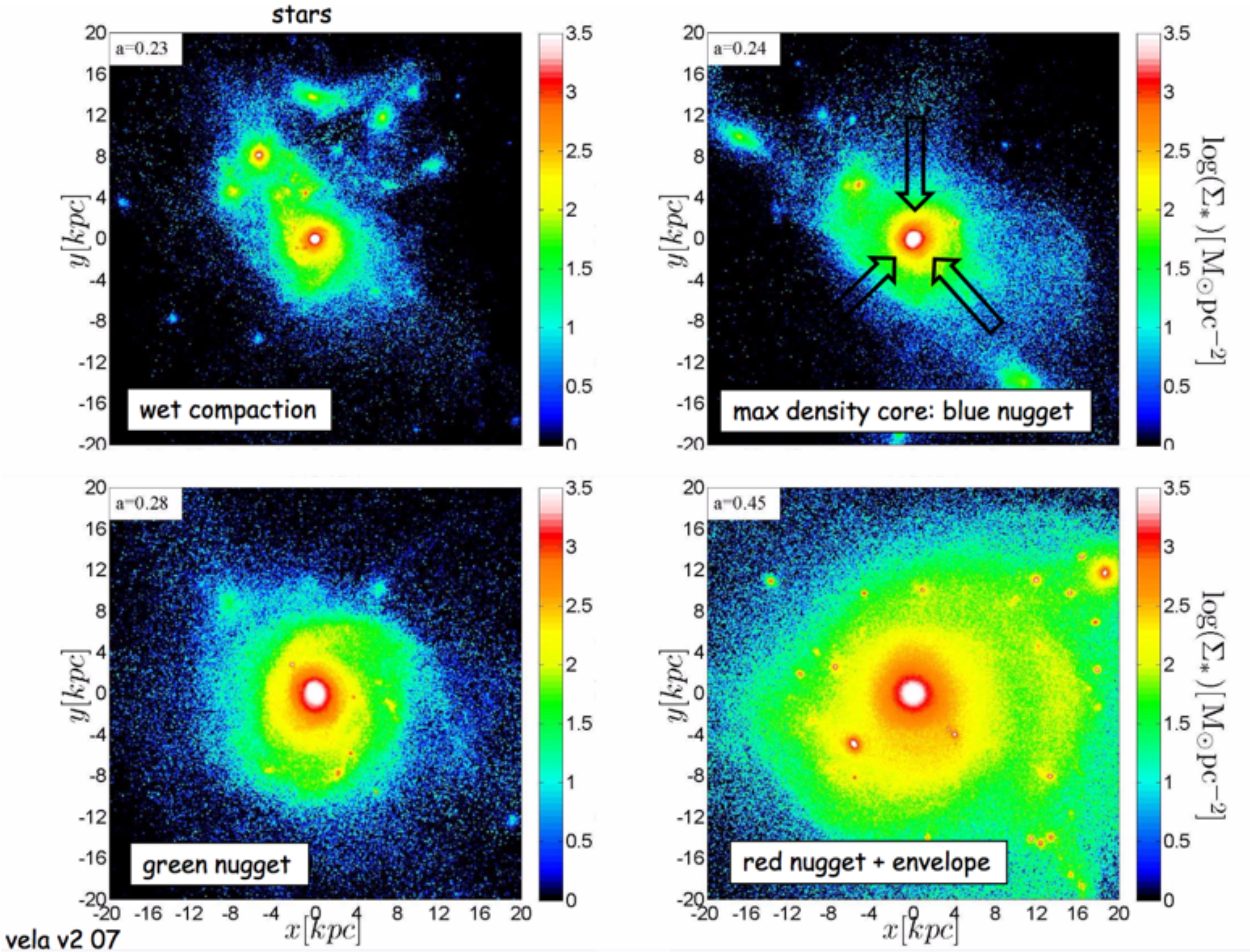
- **GROUP #3:** Find new hidden observables in the data, - <u>Linking observations and theory</u>


HYDRODYNAMIC SIMULATIONS
(e.g. Horizon-AGN, Illustris, FIRE, VELA…)
ASSEMBLY PROCESSES
MOCK IMAGES
DATA
[OBSERVATIONS]
DEEP LEARNING


HYDRODYNAMIC SIMULATIONS
(e.g. Horizon-AGN, Illustris, FIRE, VELA…)
ASSEMBLY PROCESSES
MOCK IMAGES
DEEP LEARNING
DATA
[OBSERVATIONS]

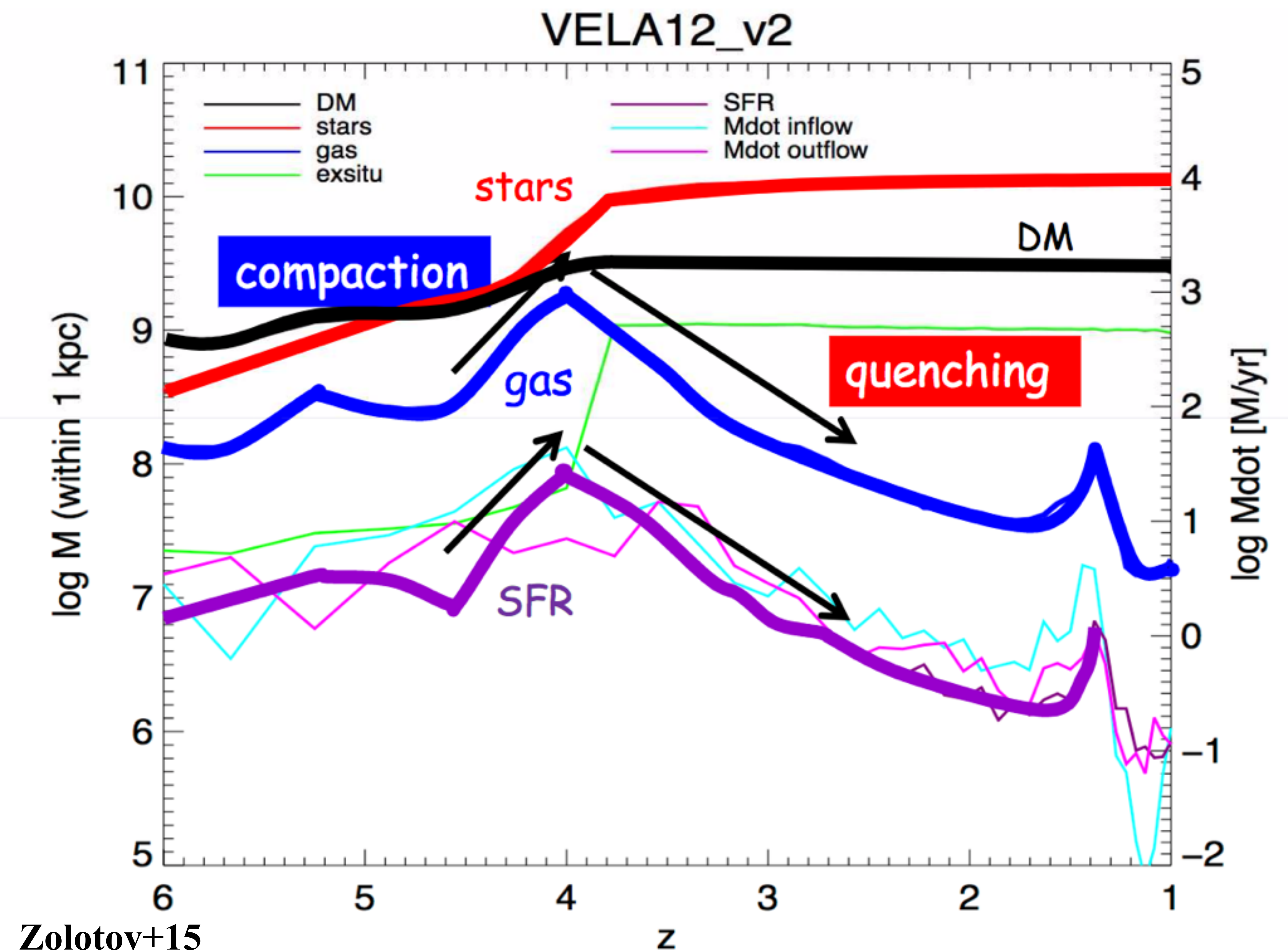
Is "wet compaction" a common mechanism for bulge formation?

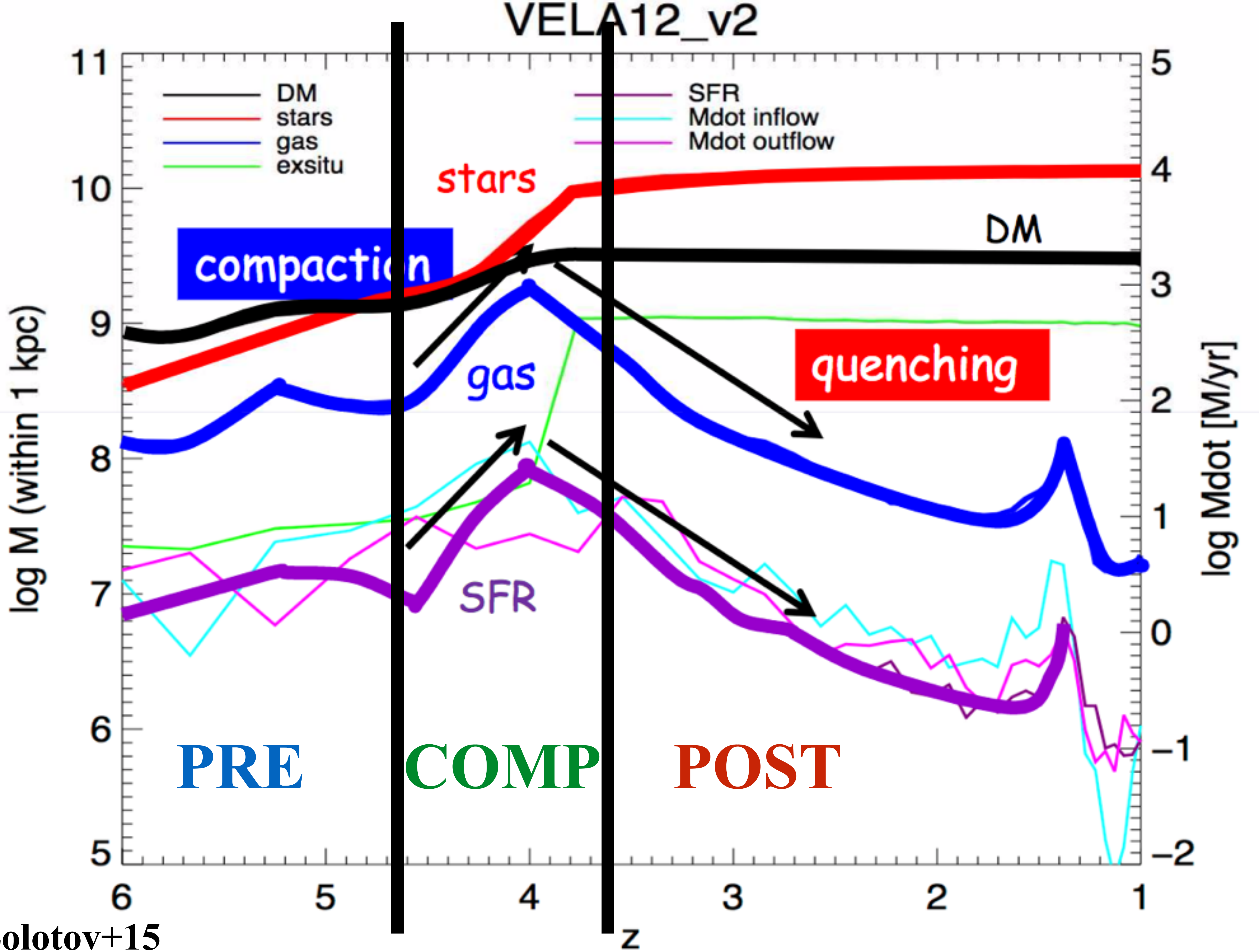
Is "wet compaction" a common mechanism for bulge formation?



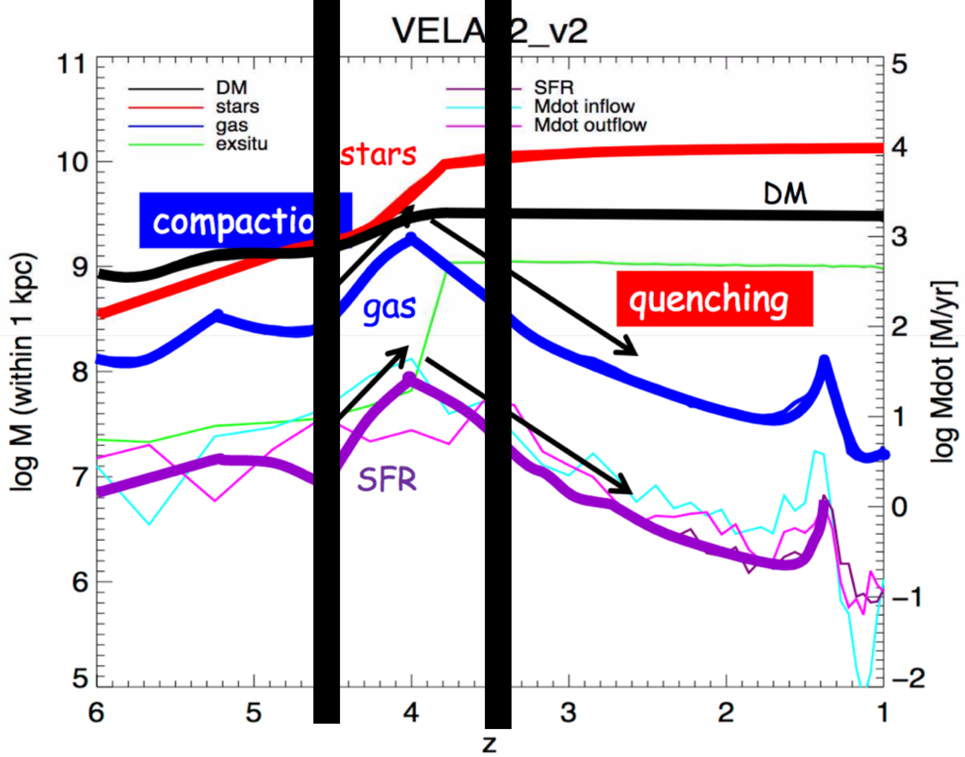
Courtesy of A. Dekel

Ceverino+15
Zolotov+15
Tacchella+17


VELA12_v2
DM
stars
gas
exsitu
SFR
Mdot inflow
Mdot outflow
stars
DM
compaction
gas
quenching
SFR
log M (within 1 kpc)
log Mdot [M/yr]
z
11
10
9
8
7
6
5
5
4
3
2
1
0
−1
−2
6
5
4
3
2
1
Zolotov+15

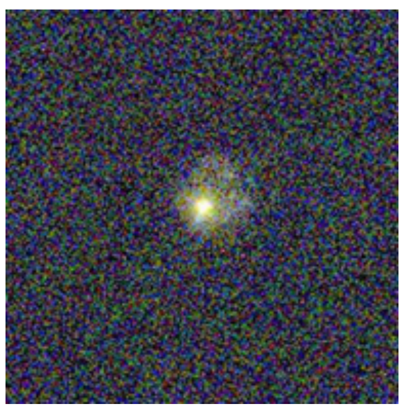

VELA12_v2
DM
stars
gas
exsitu
SFR
Mdot inflow
Mdot outflow
stars
DM
compaction
gas
quenching
SFR
PRE
COMP
POST
log M (within 1 kpc)
log Mdot [M/yr]
z
11
10
9
8
7
6
5
5
4
3
2
1
0
−1
−2
6
5
4
3
2
1

Zolotov+15



"WET COMPACTION"

[simulation metadata]



MOCK HST IMAGES (CANDELS filters)

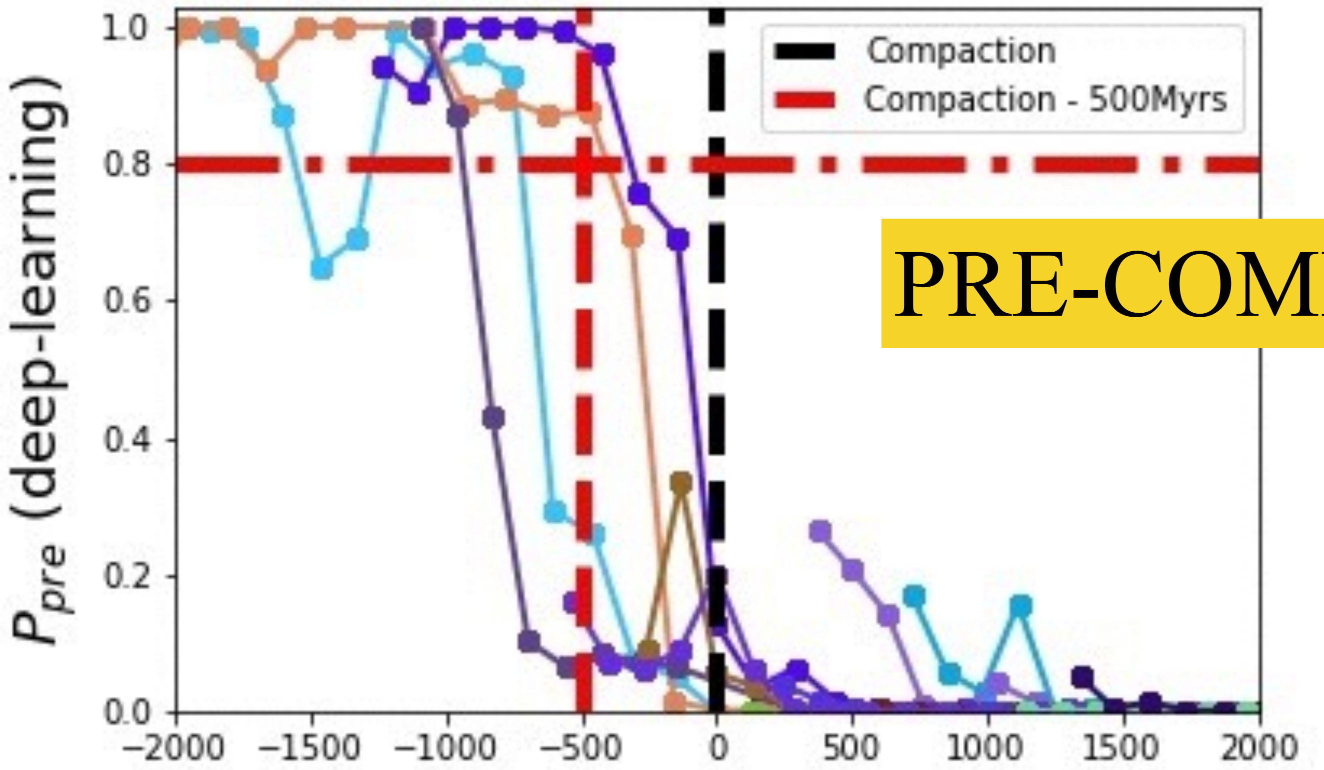
ANY SIGNATURE OF "COMPACTION" IN THE STELLAR DISTRIBUTION?

DEEP-NET

Pre-Compaction

Compaction

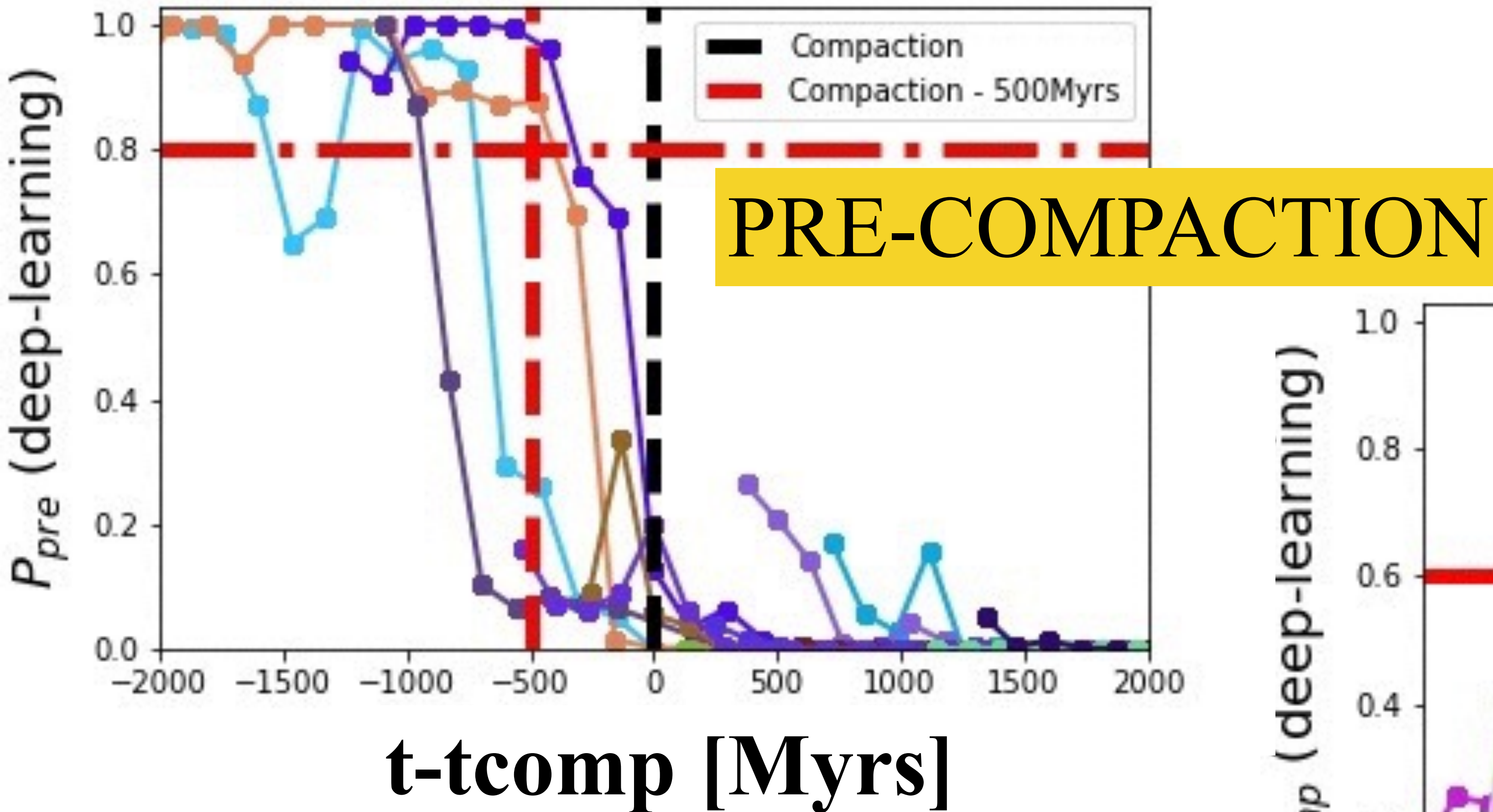
Post-Compaction

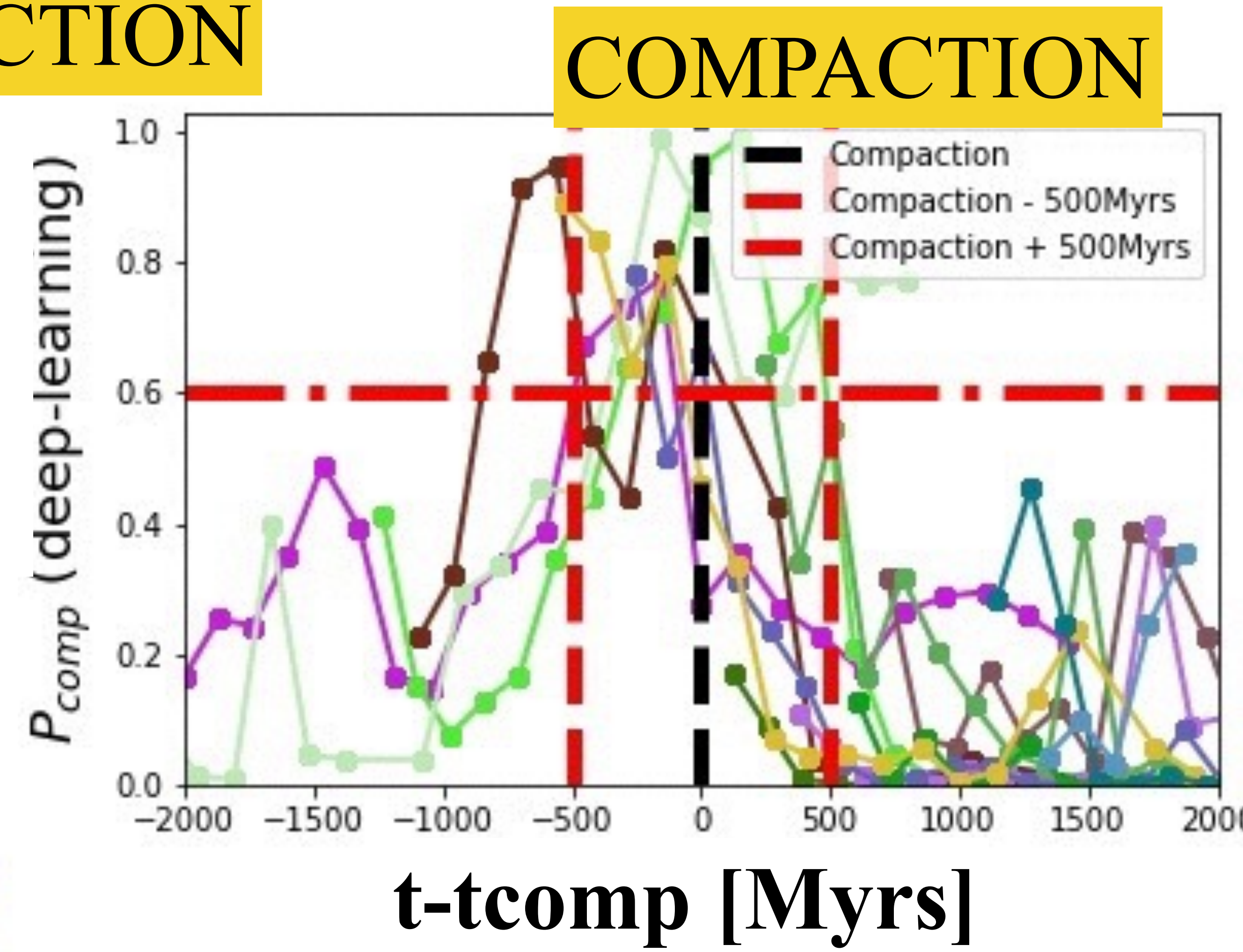


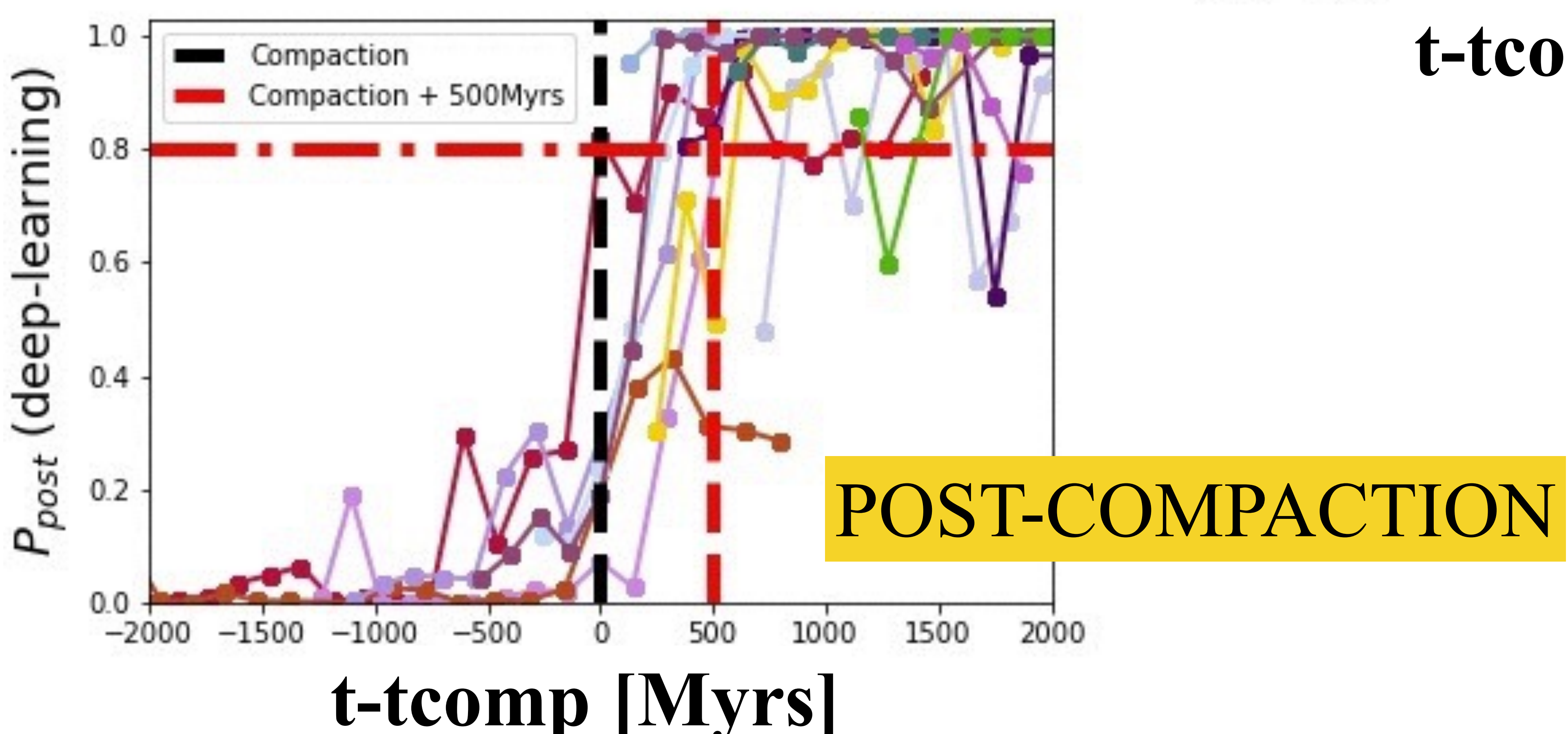
PRE-COMPACTION

t-tcomp [Myrs]

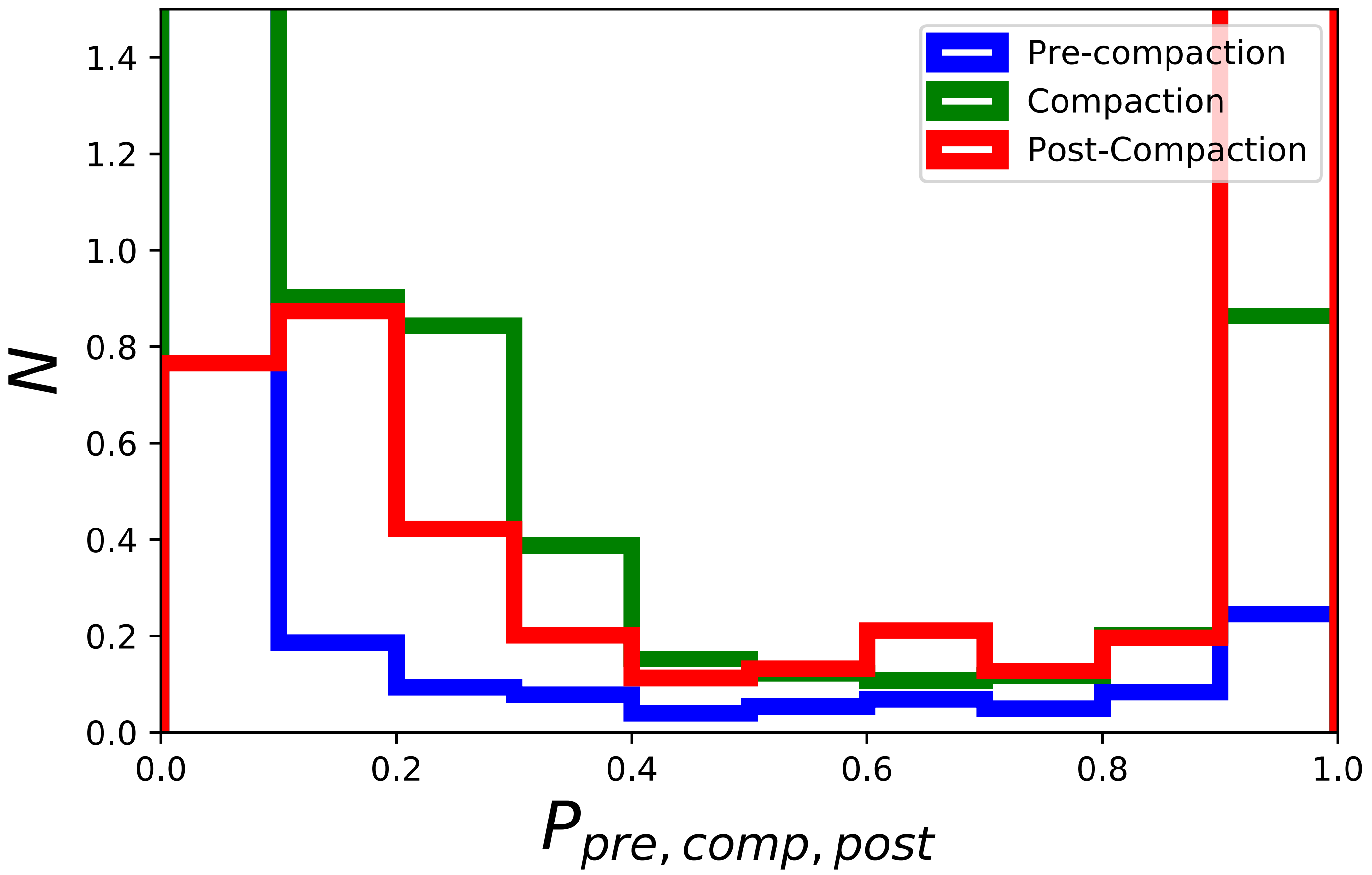
MHC+18 (in prep)





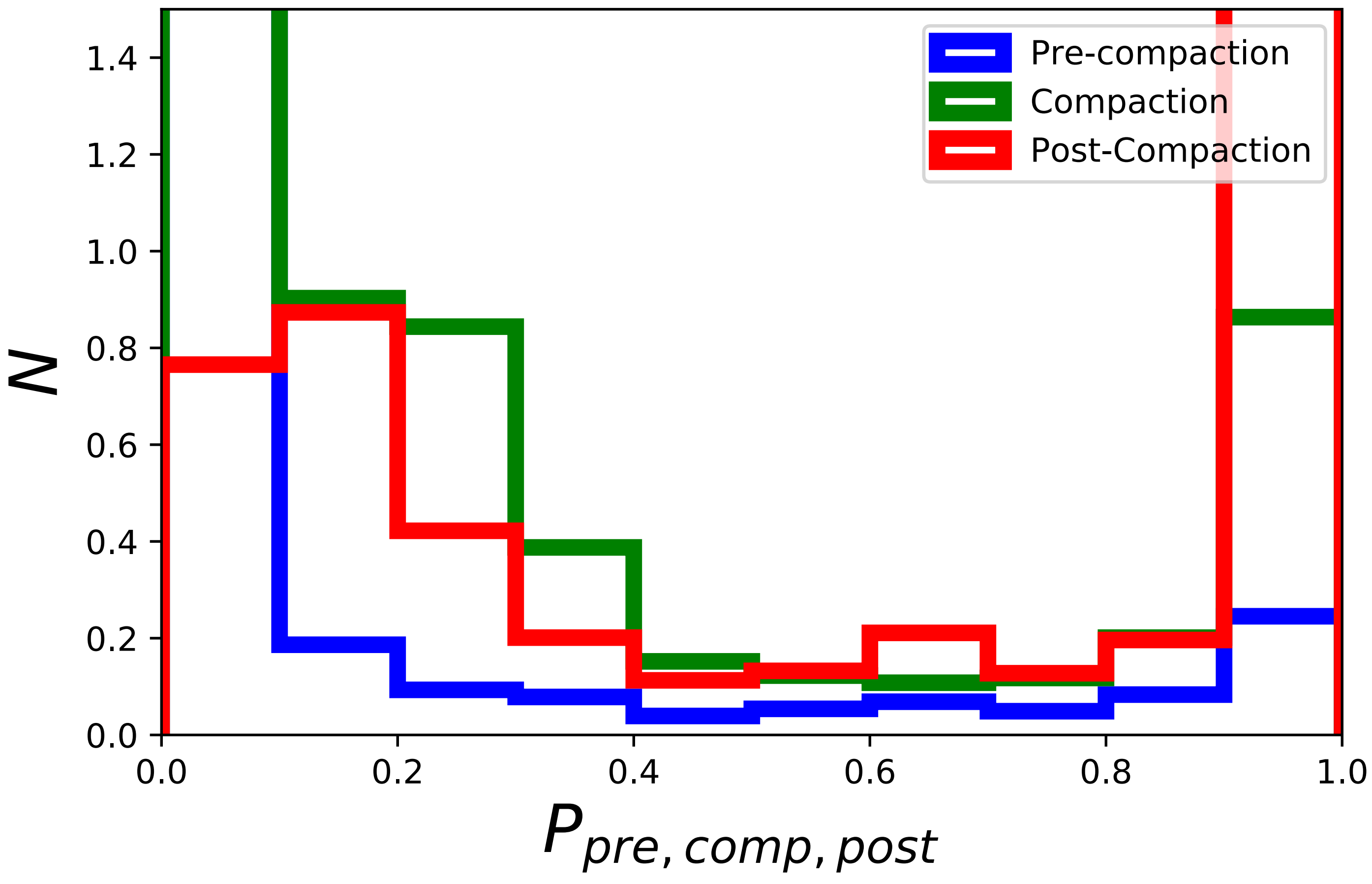


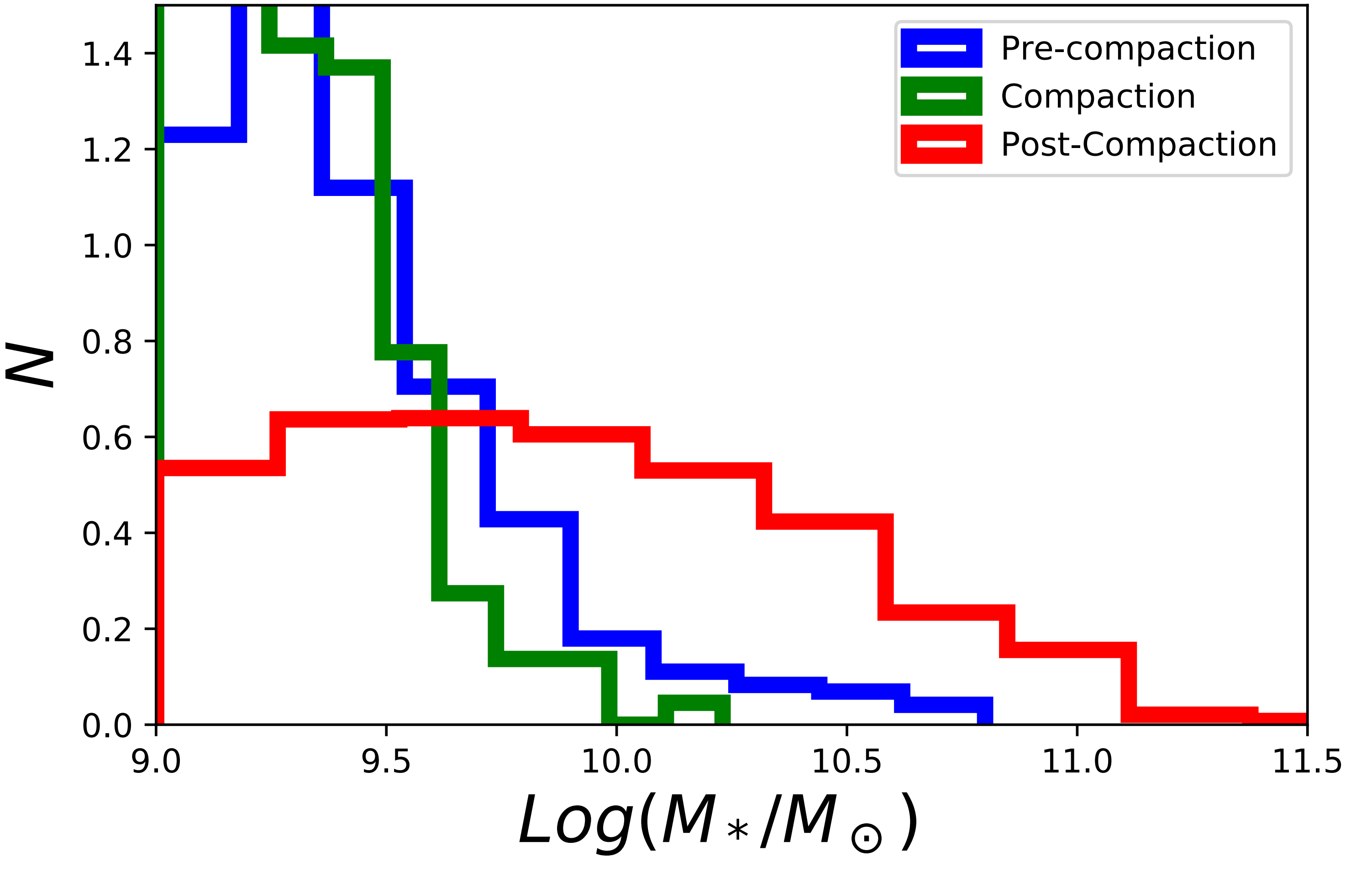
MHC+18 (in prep)



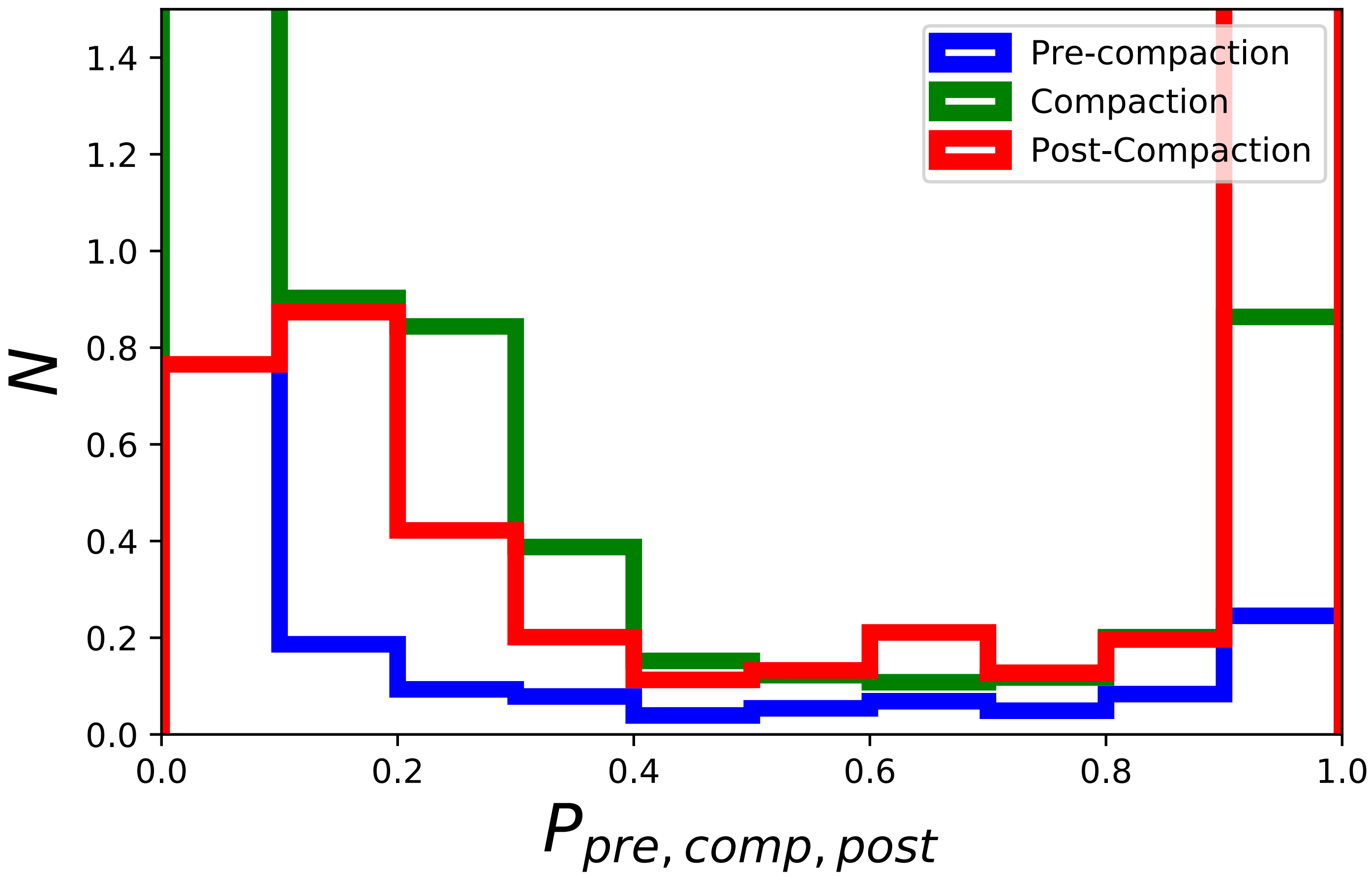
WHEN APPLIED TO REAL CANDELS DATA

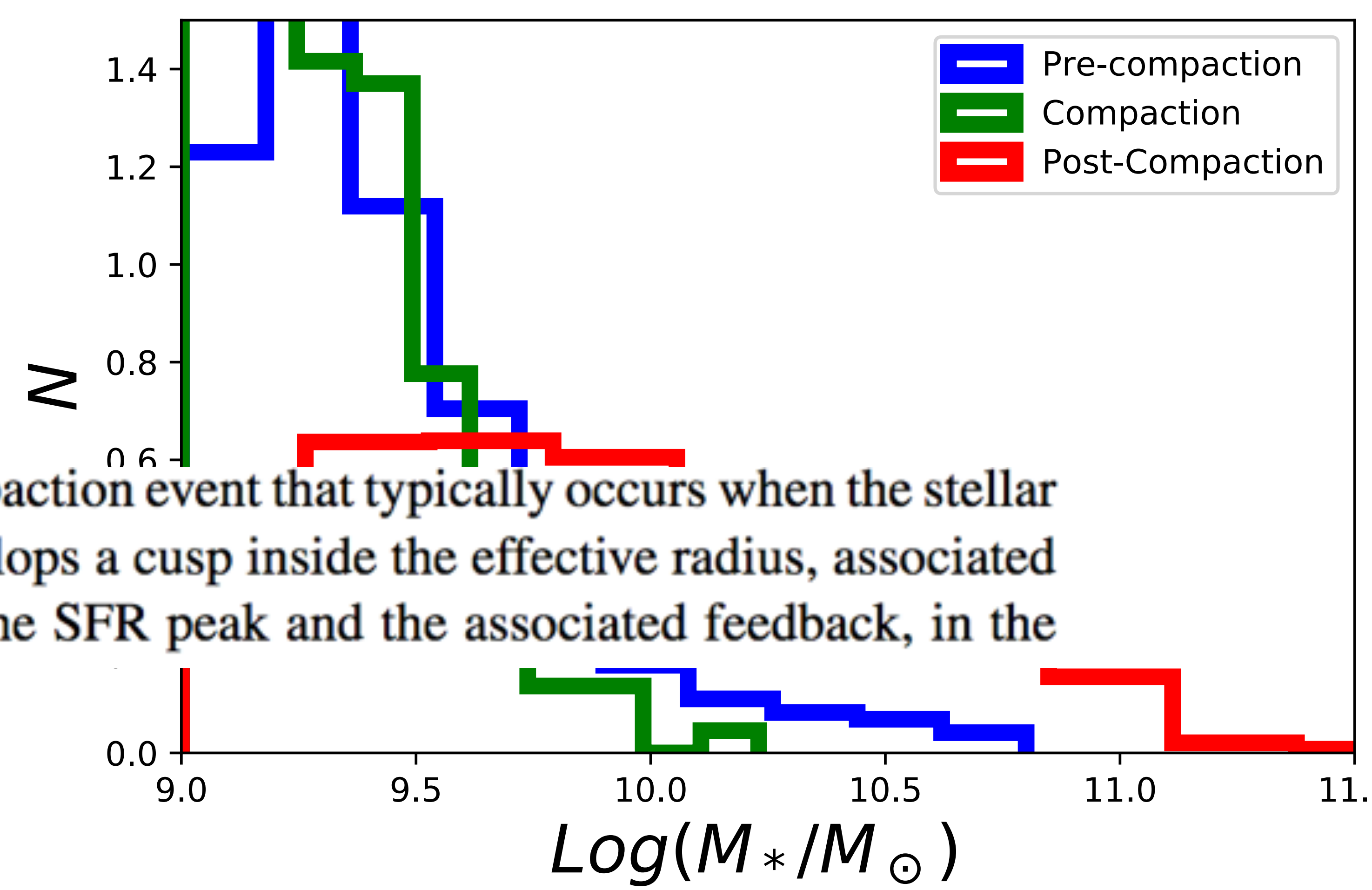
WHEN APPLIED TO REAL CANDELS DATA





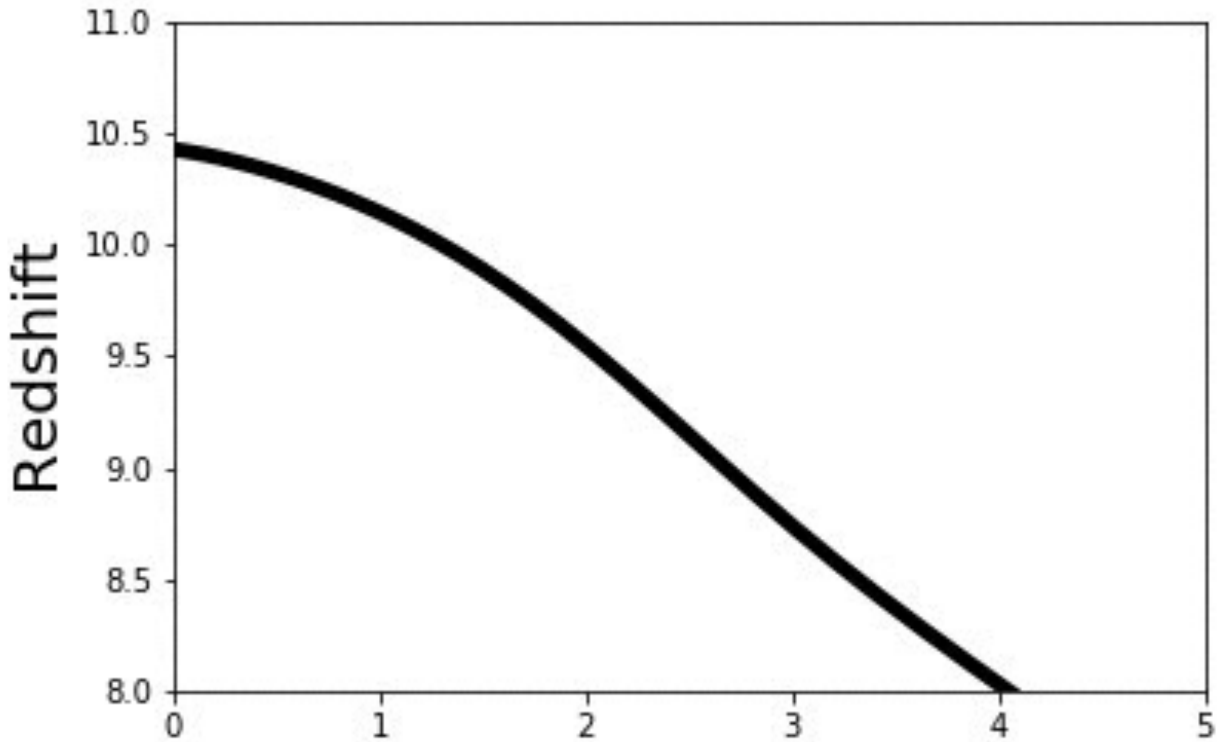
WHEN APPLIED TO REAL CANDELS DATA





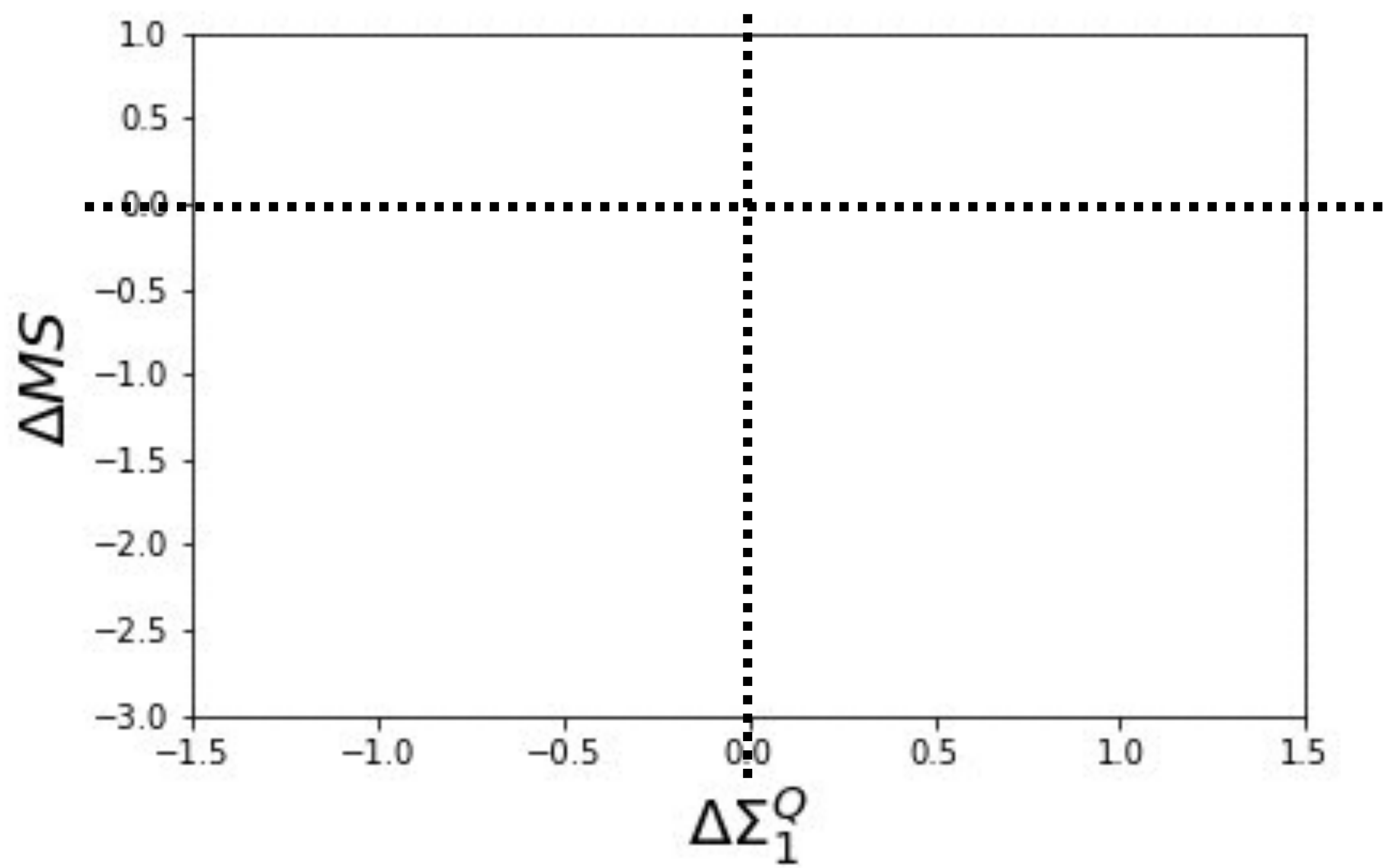
their evolution phases. Following a wet compaction event that typically occurs when the stellar mass is $\sim 10^{9.5}$ $M_\odot$ at $z \sim 2$–4, the gas develops a cusp inside the effective radius, associated with a peak in star formation rate (SFR). The SFR peak and the associated feedback, in the

Tacchella+16

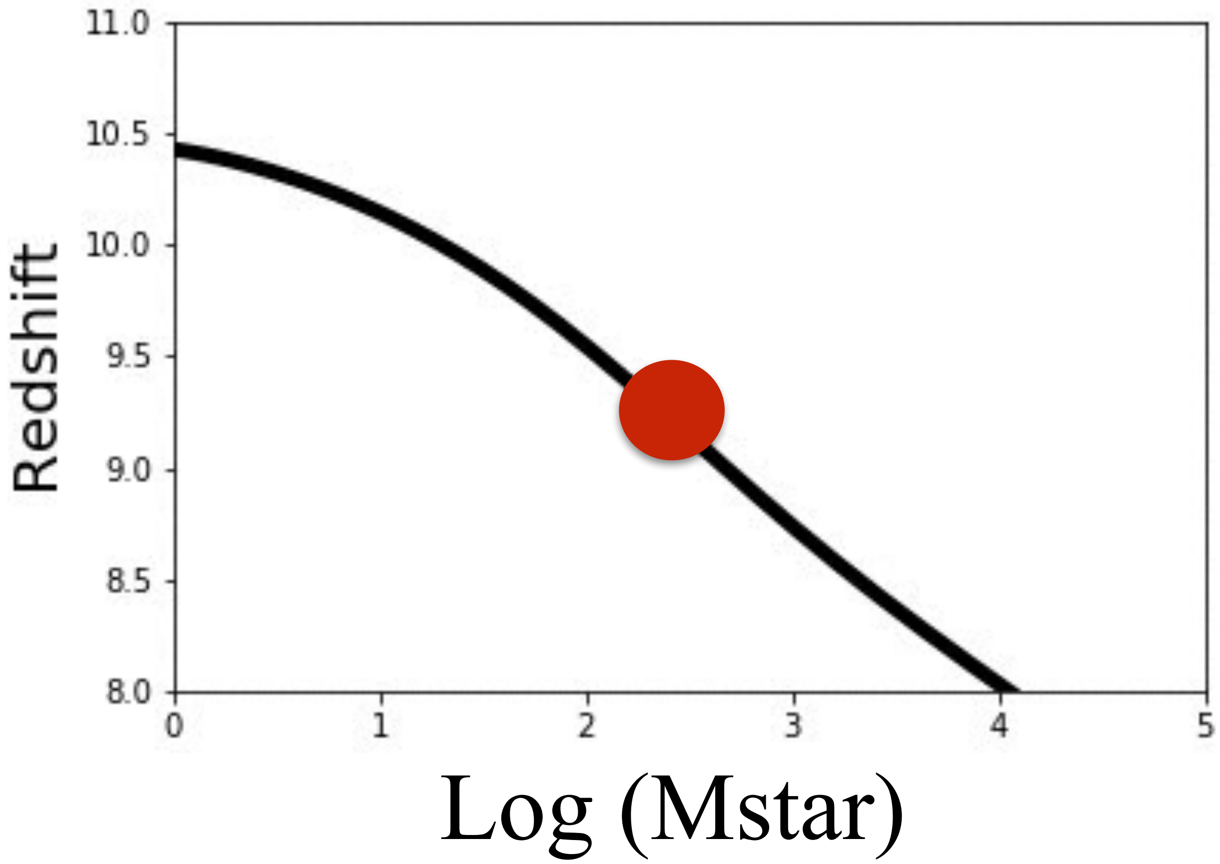


Log (Mstar)

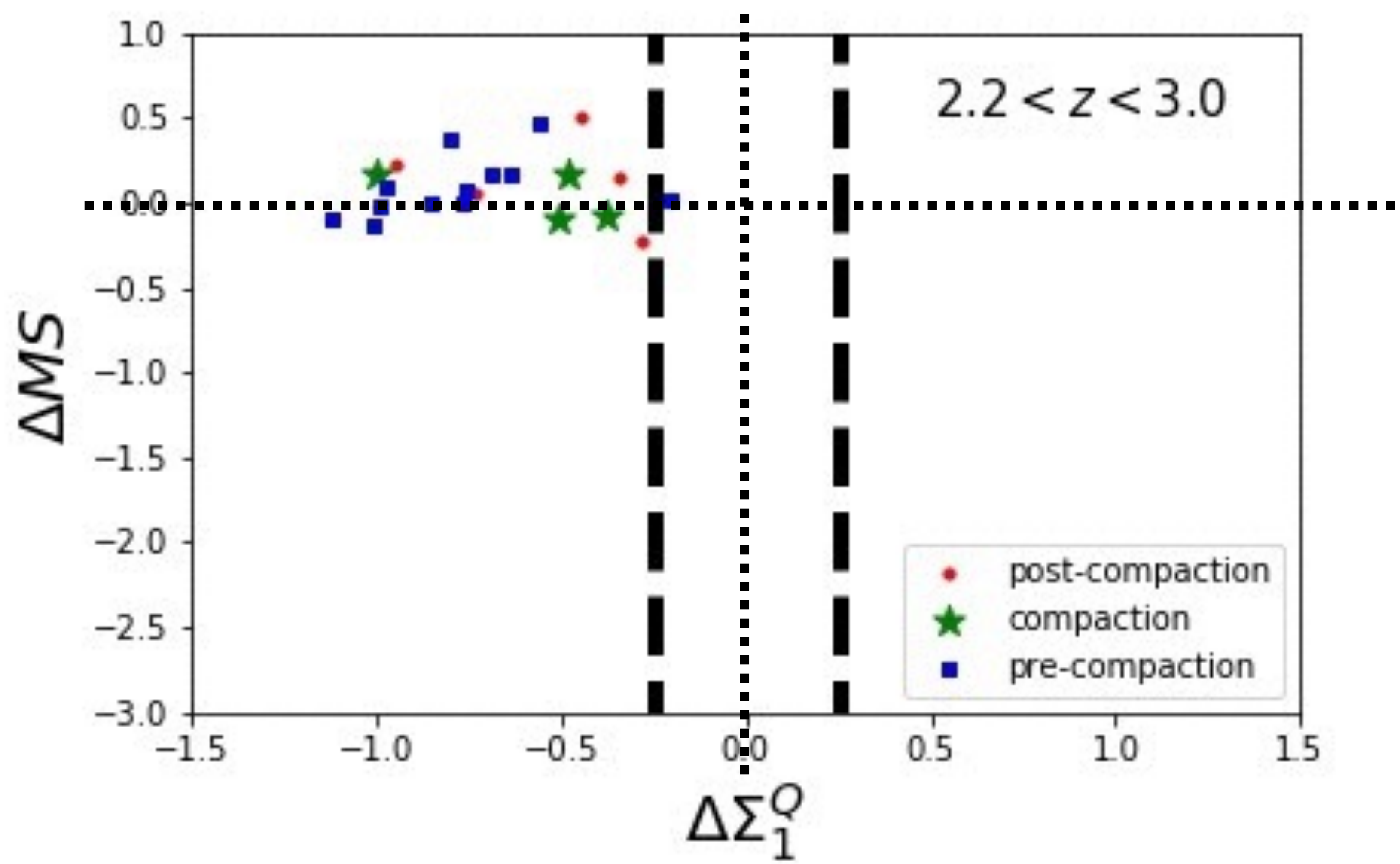
abundance matching
[Rodriguez-Puebla+17]



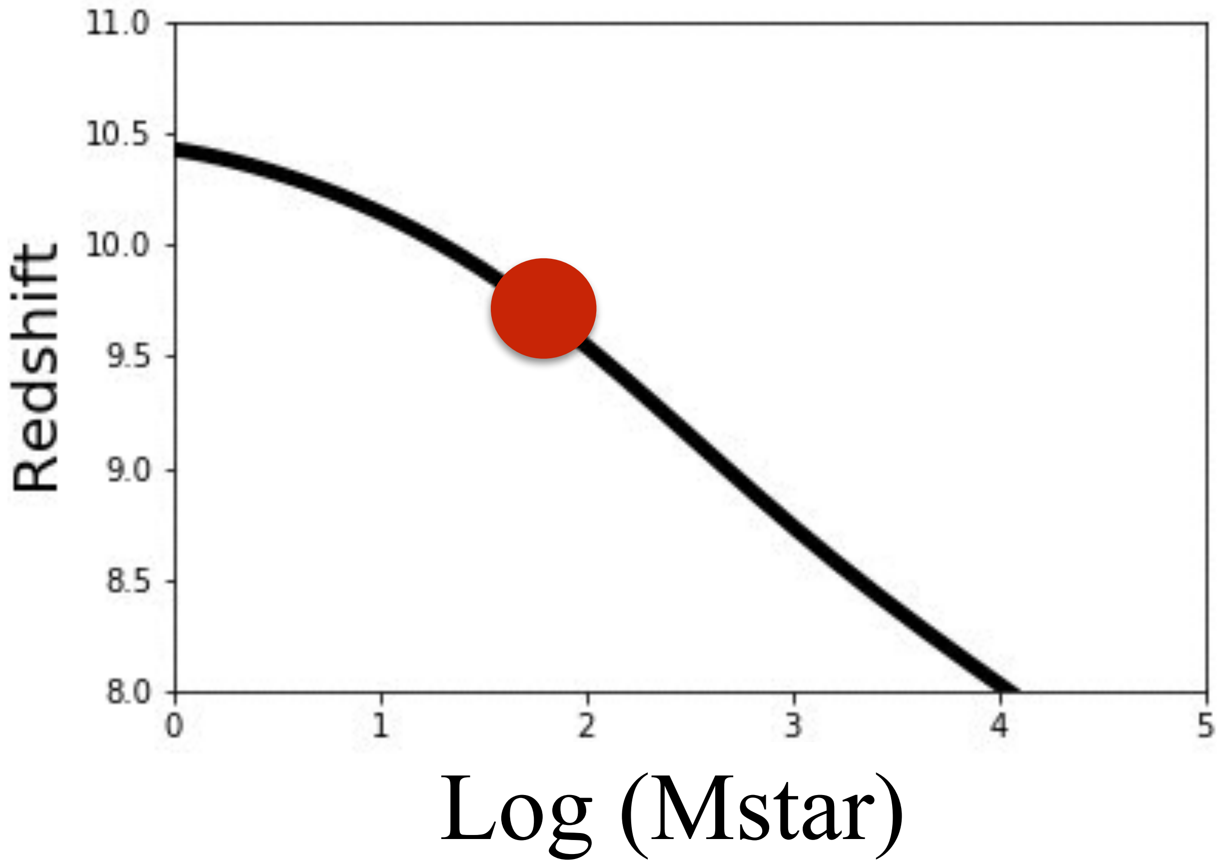
MHC+18 (in prep)



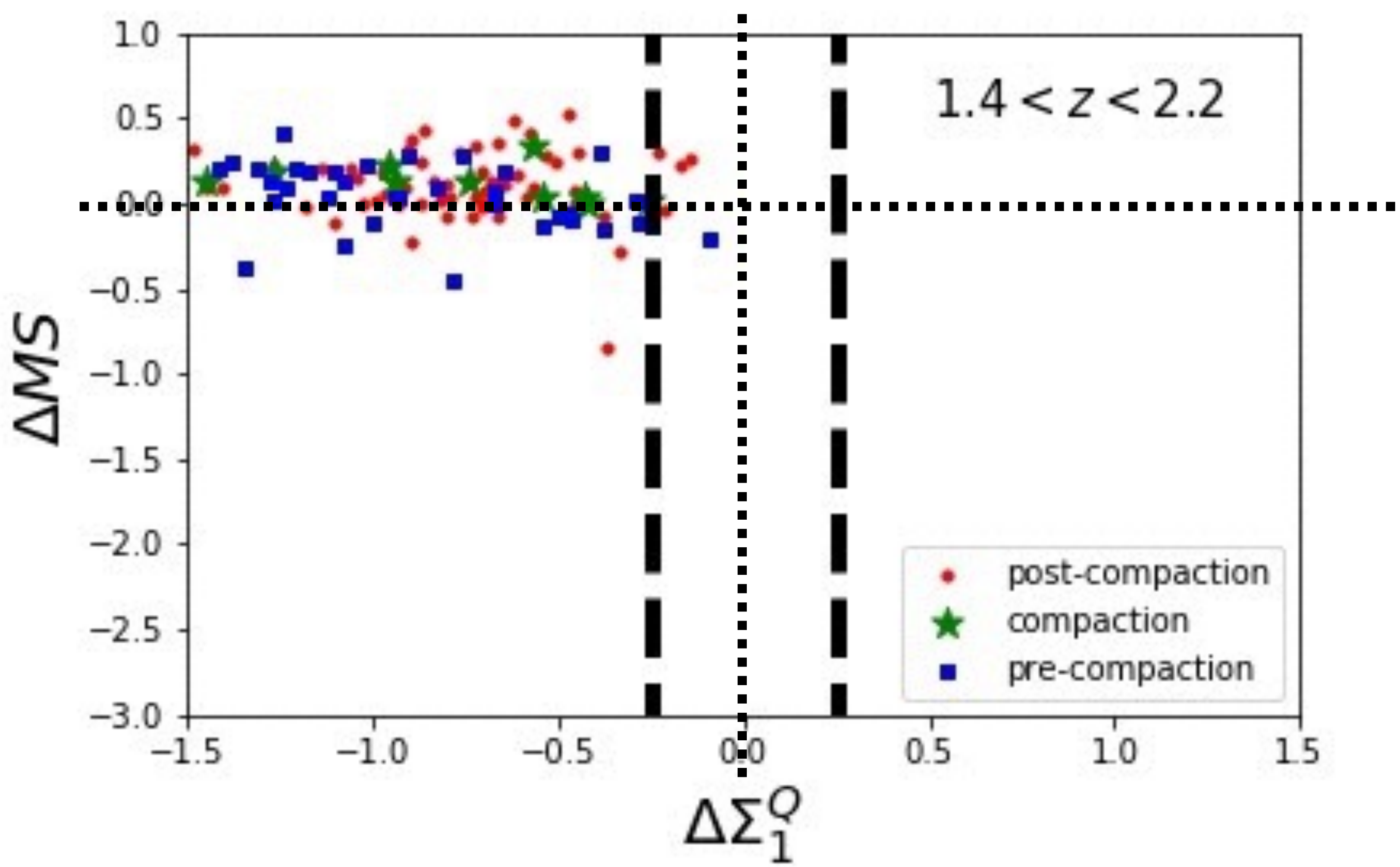
abundance matching [Rodriguez-Puebla+17]



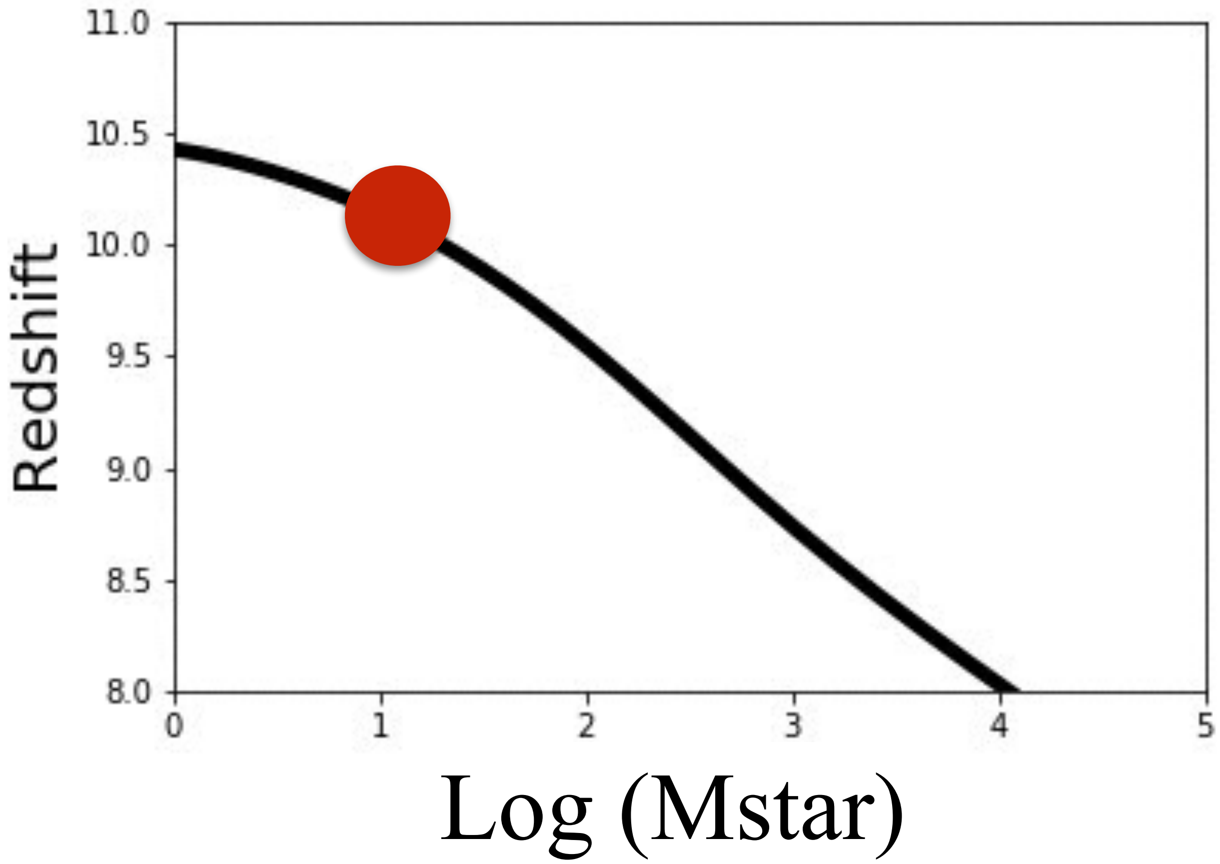
MHC+18 (in prep)



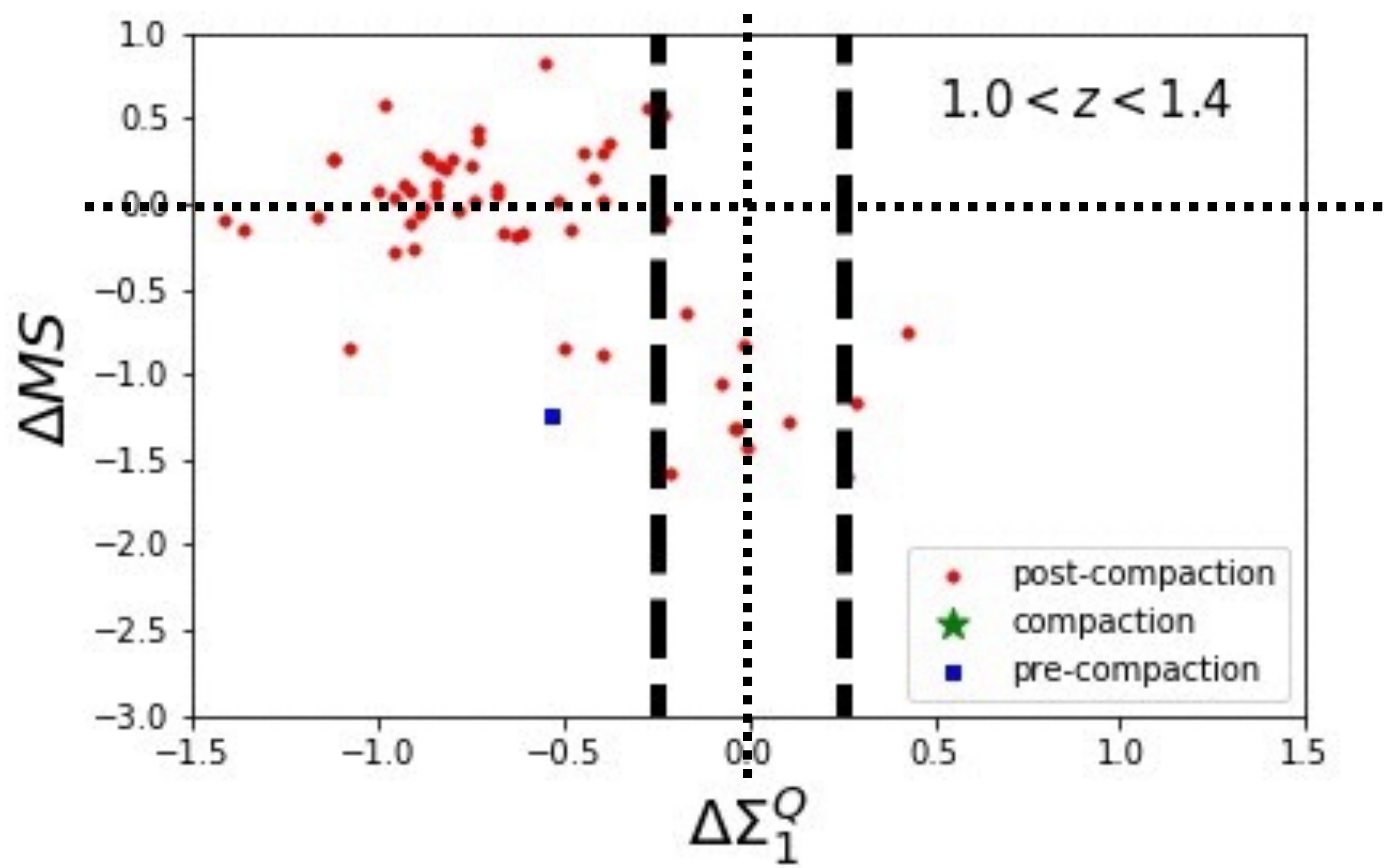
abundance matching
[Rodriguez-Puebla+17]



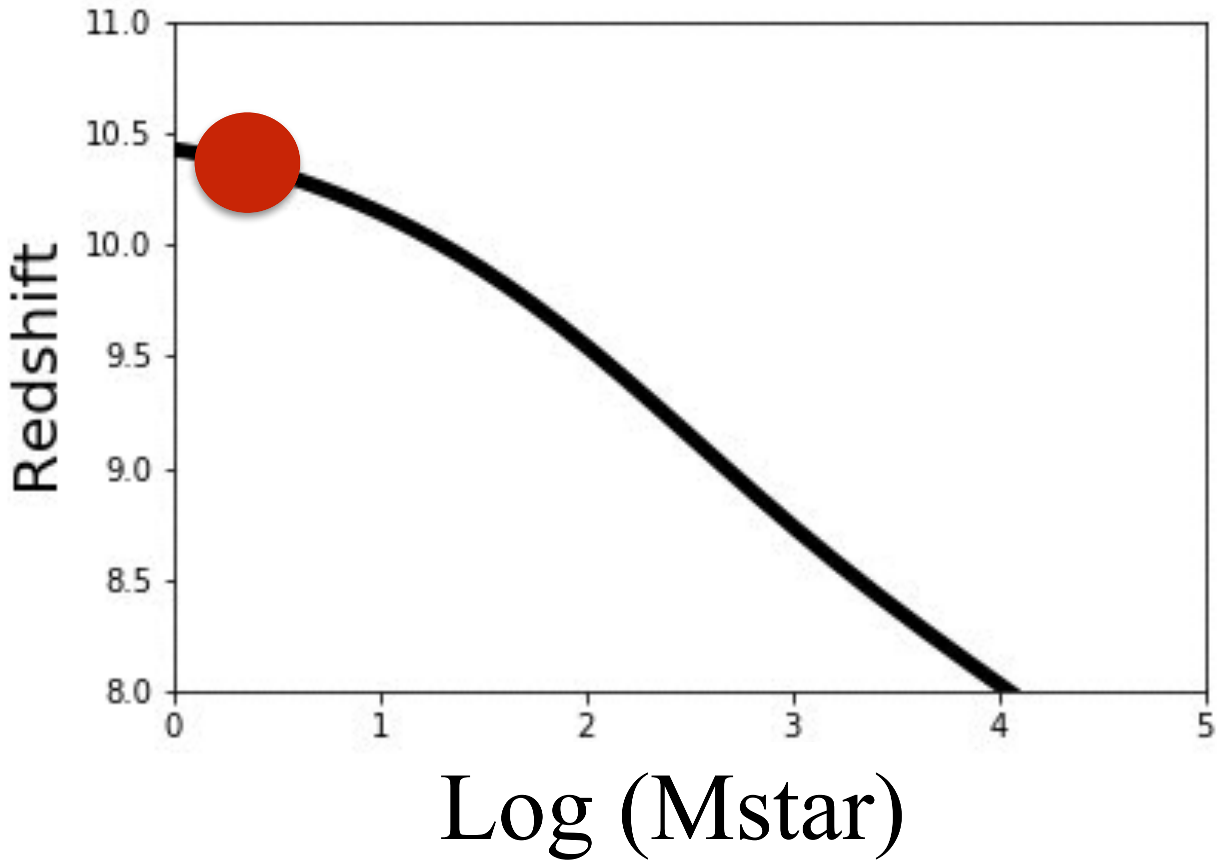
MHC+18 (in prep)



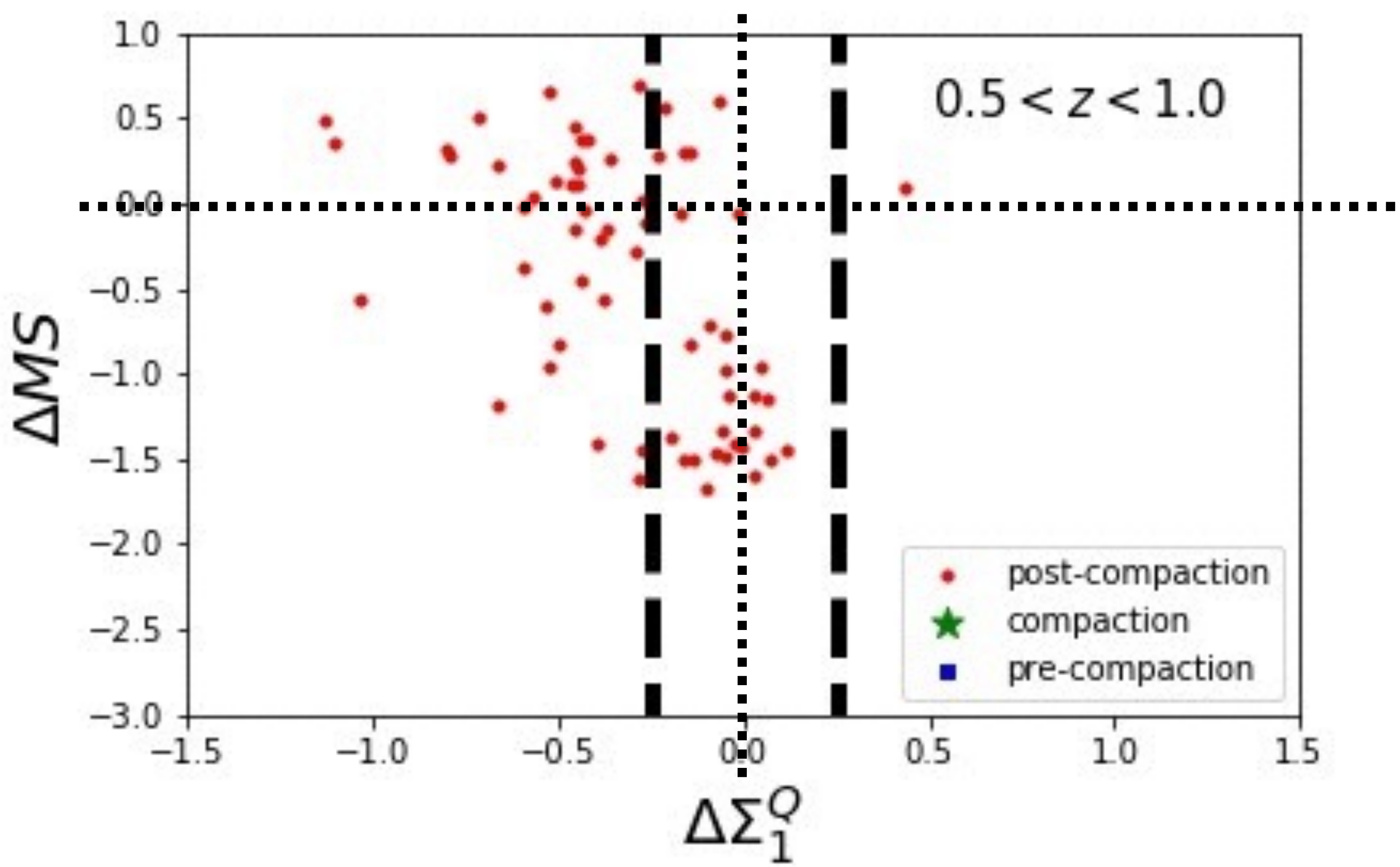
abundance matching [Rodriguez-Puebla+17]



MHC+18 (in prep)

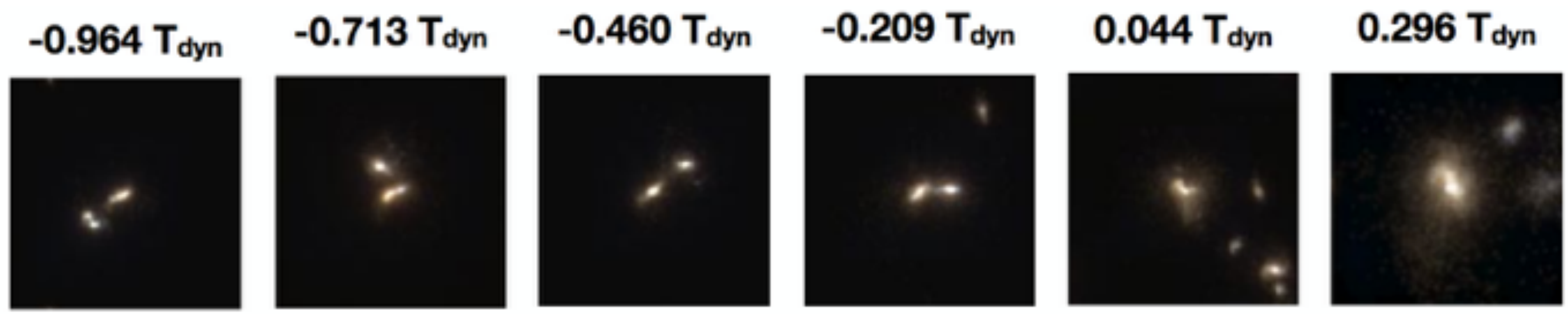


abundance matching [Rodriguez-Puebla+17]



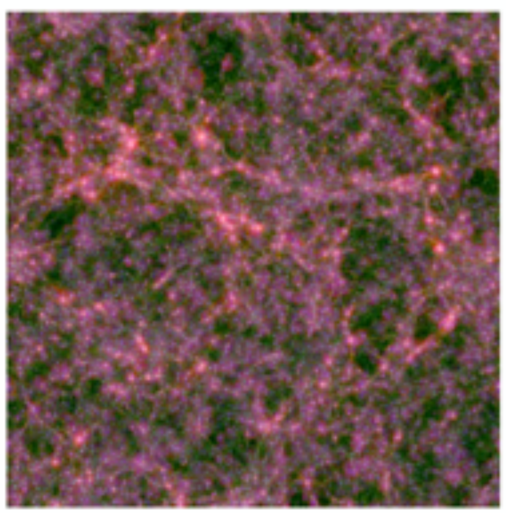
MHC+18 (in prep)

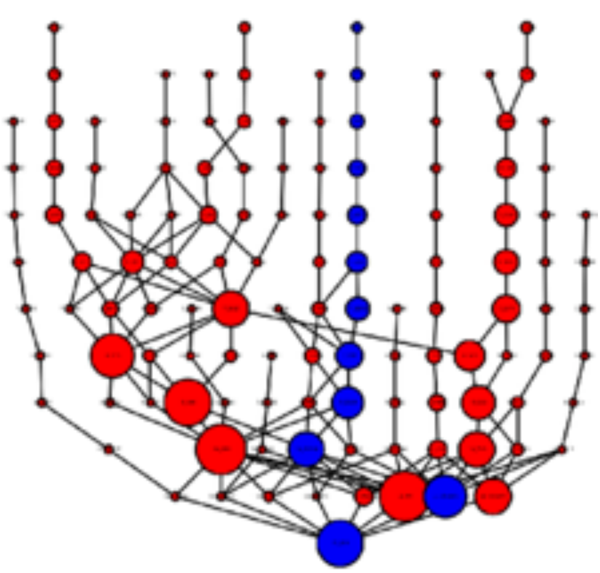
Dissecting mergers with DL



Horizon AGN hydro sim

[Dubois+14]



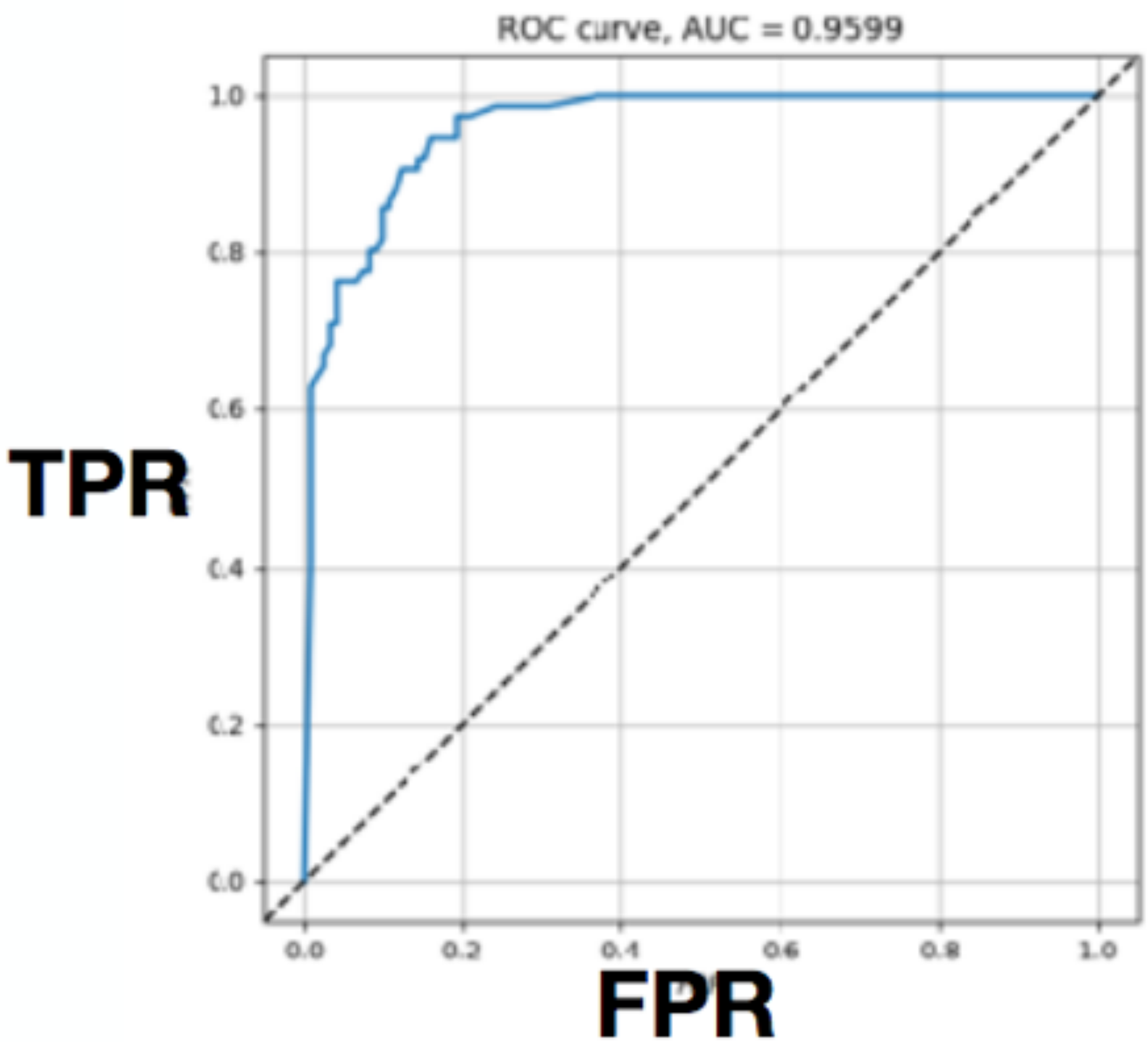


Mock images → Deep Learning →

Merger properties:

- mass-ratio
- stage
- timescale

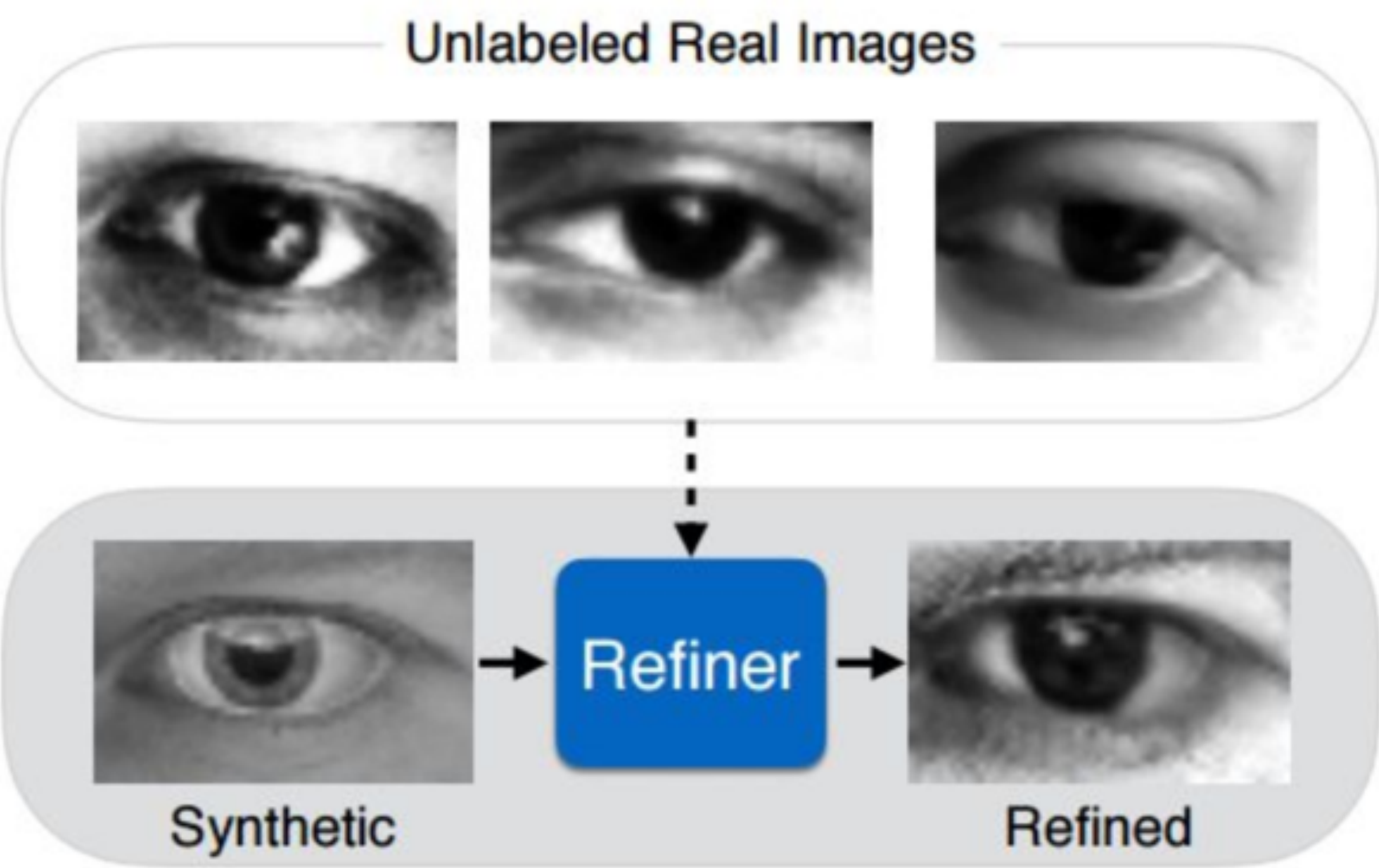
MERGER PHASE



Accuracy over validation set: ~ 96%

- **GROUP #4**: Finding the unknown?

[ANR project submitted]



GANs for outlier detection….

Shrivastava+16

DL FOR FOR GALAXIES?

- **GROUP #1:** Time consuming tasks that humans do easily but classically challenging for computers - classification of objects
- **GROUP #2:** Efficient and fast quantitative measurements on large amount of (multi-lambda) data [photoz's, sizes, ellipiticities]
- **GROUP #3:** Find hidden unknown correlations in the data, - Linking observations and theory
- **GROUP #4**: Finding the unknown?

Group #1: Classification of large datasets

- **Requires a huge volume** of “labeled” data to train. Human intervention is necessary anyway…

DATA FROM NEW SURVEY

- **Requires a huge volume** of “labeled” data to train. Human intervention is necessary anyway.

DEEP-LEARNING BASED MACHINE

Galaxy ZOO like classifications for for the entire sample

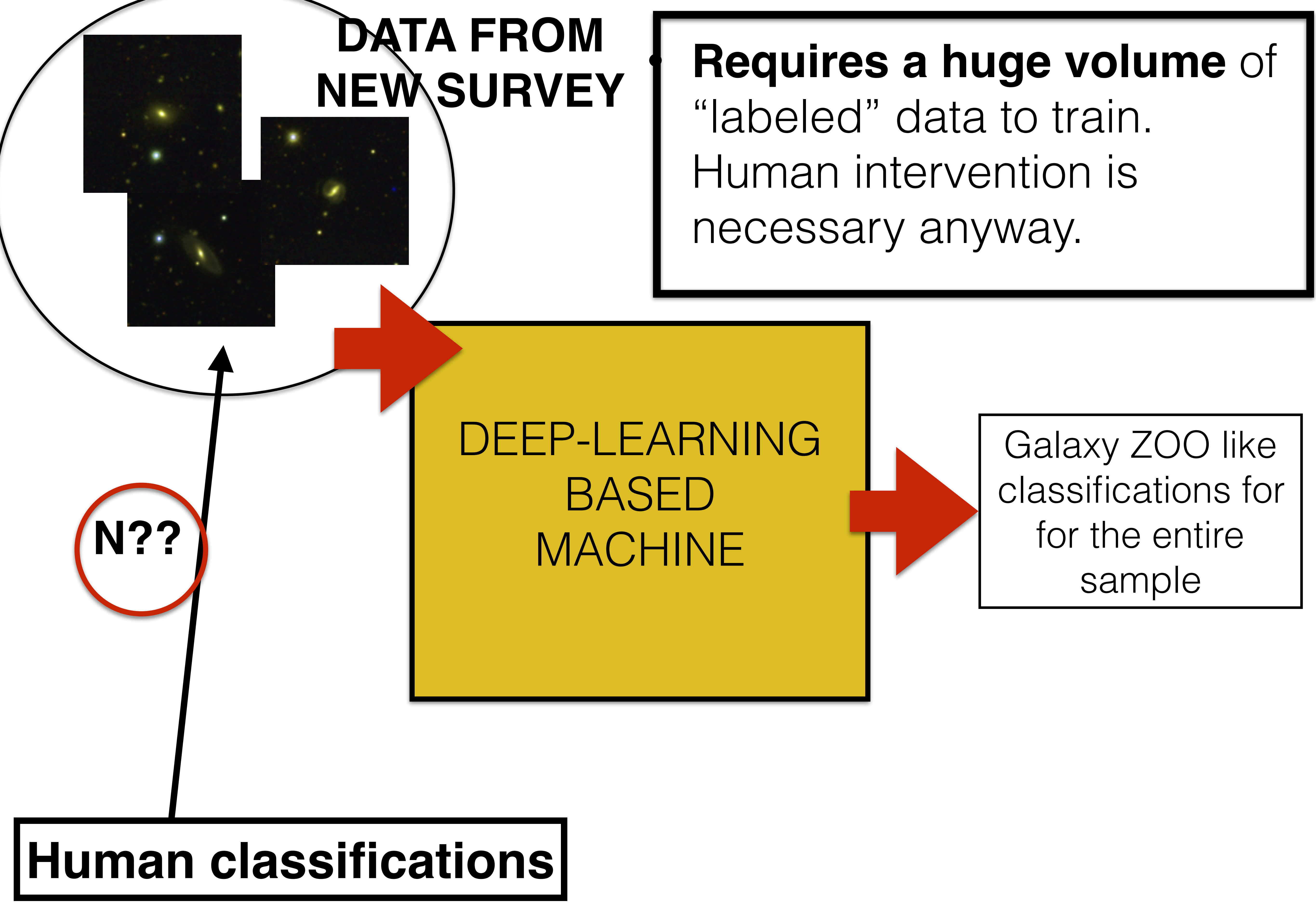
DATA FROM
NEW SURVEY

- **Requires a huge volume** of "labeled" data to train. Human intervention is necessary anyway.

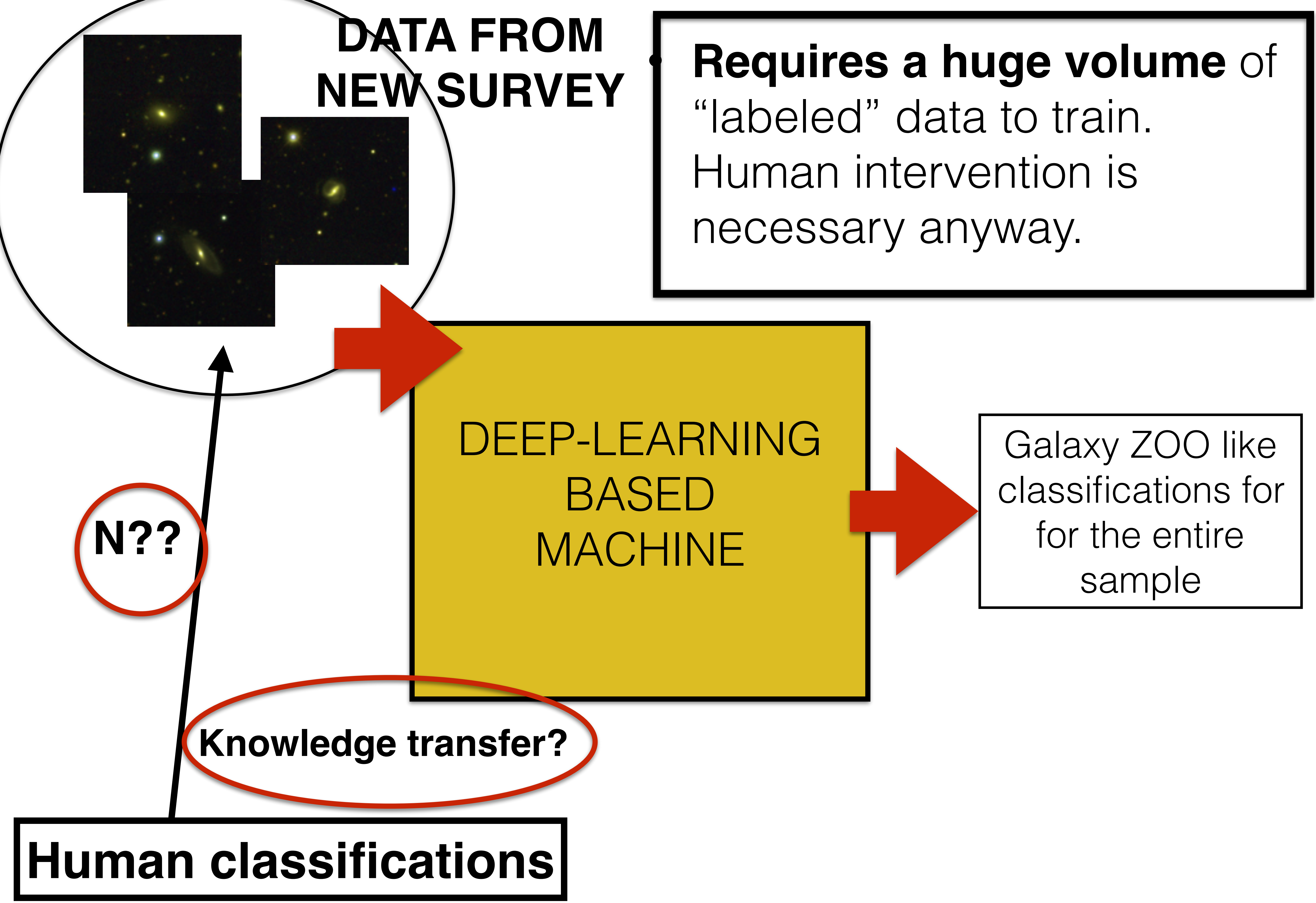
DEEP-LEARNING
BASED
MACHINE

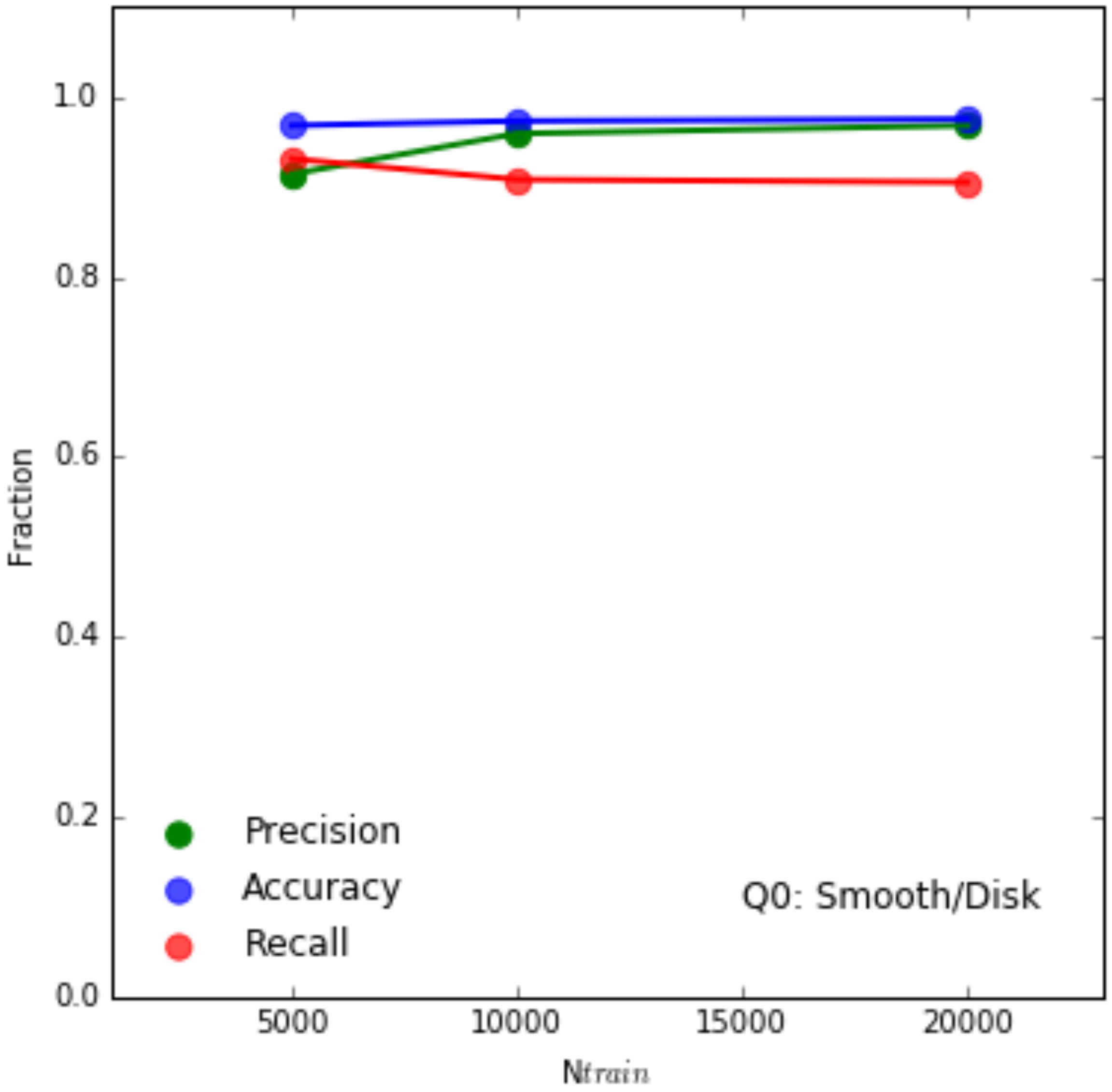
Galaxy ZOO like classifications for for the entire sample

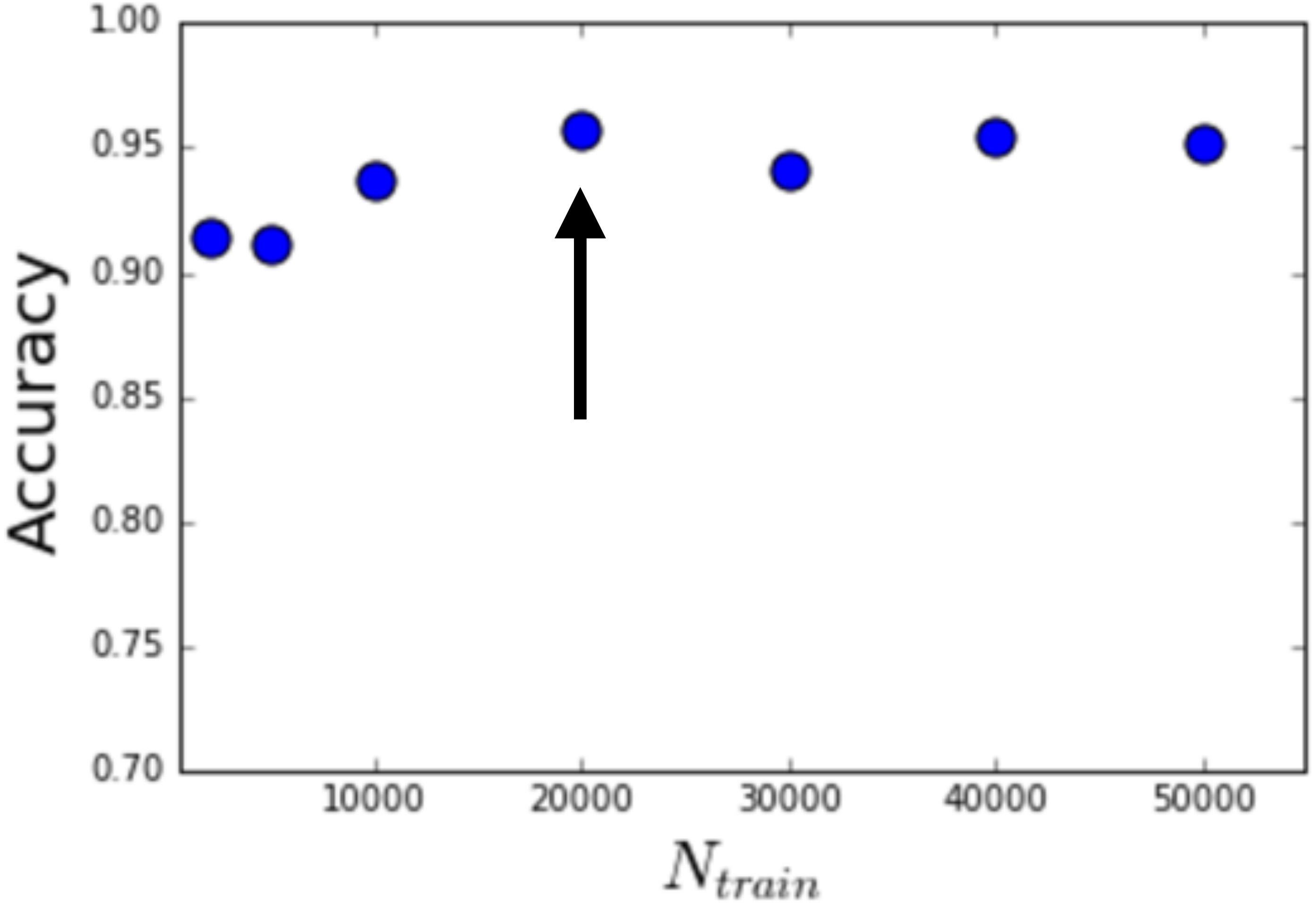
Human classifications

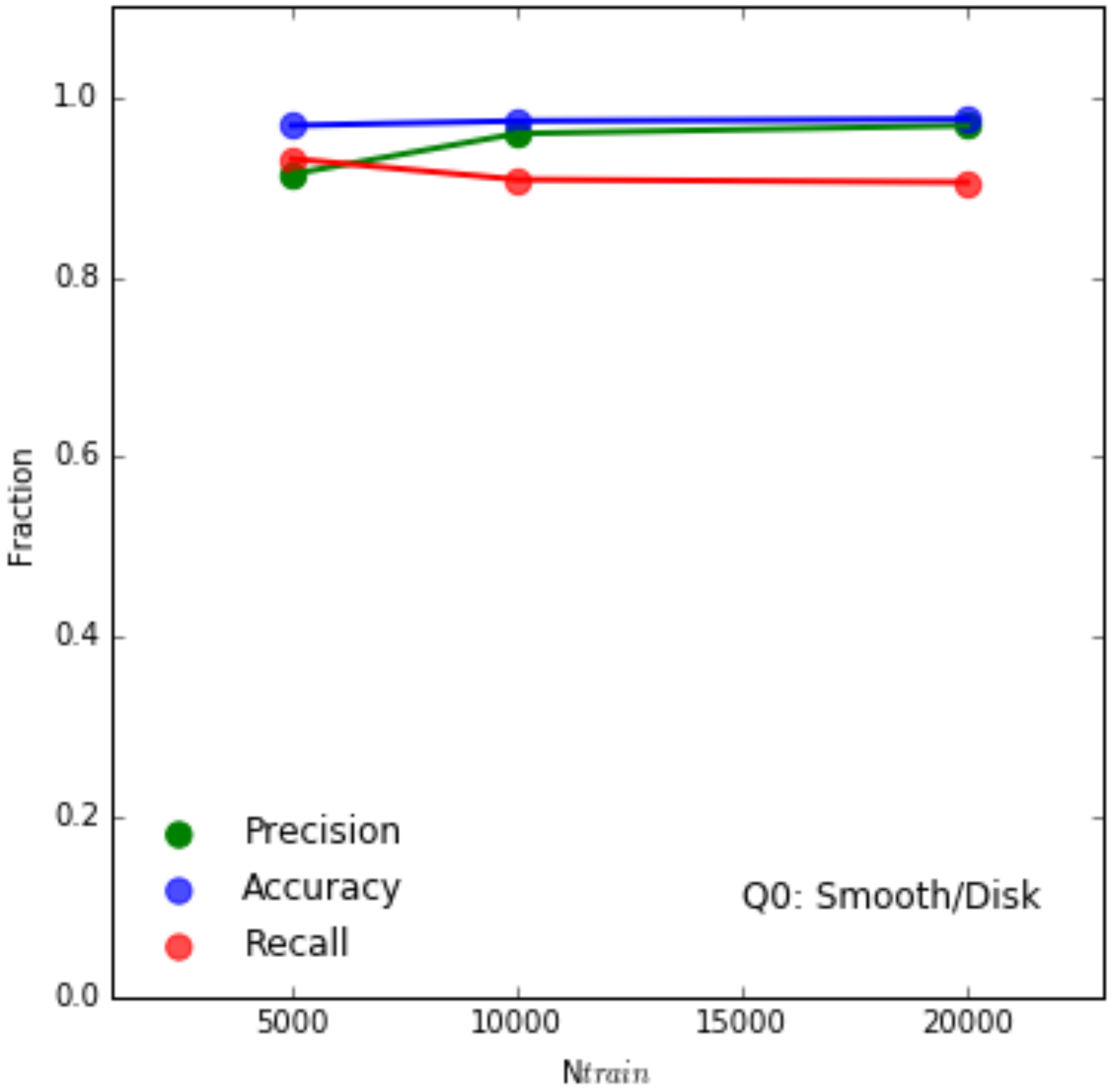


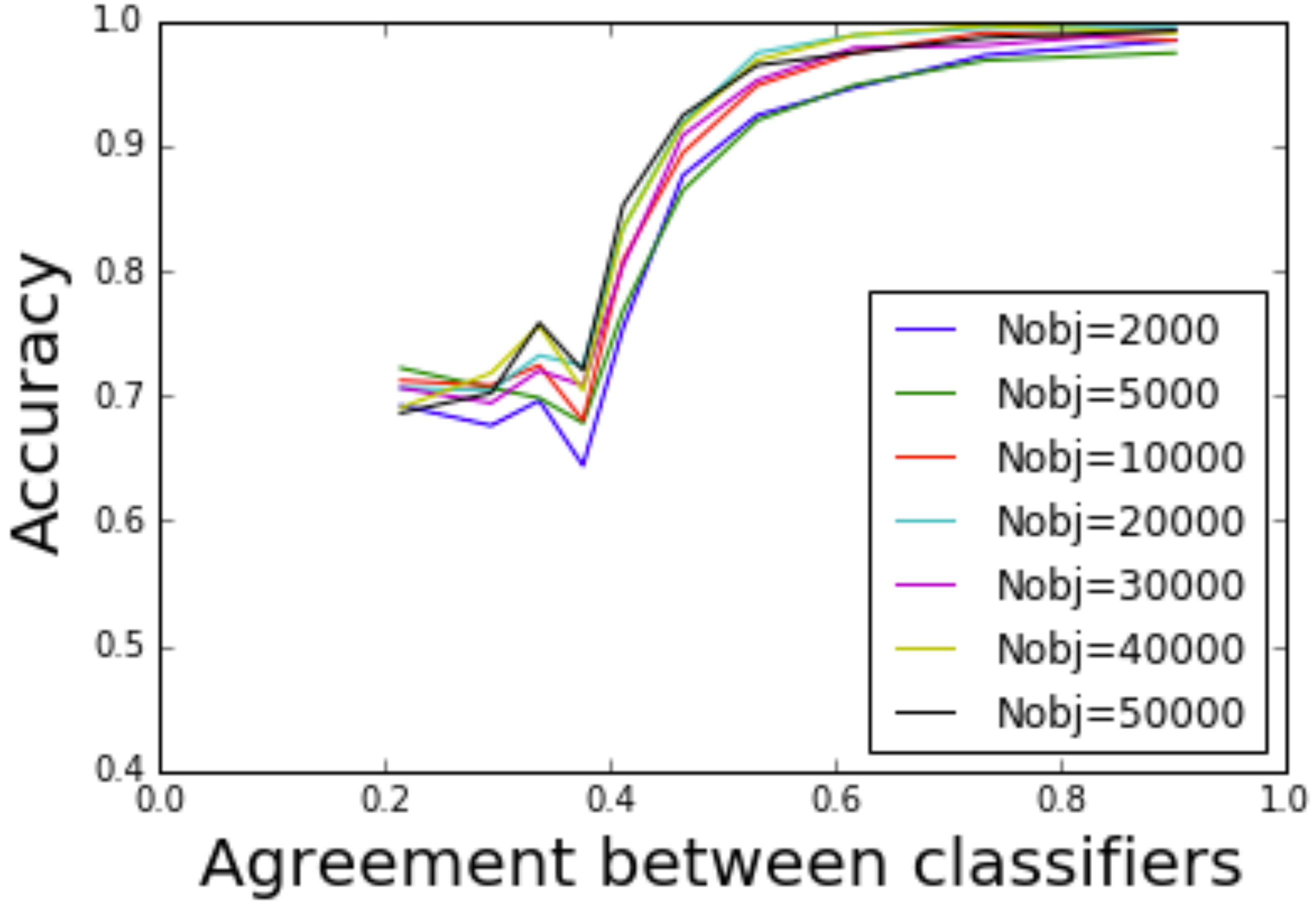
DATA FROM NEW SURVEY
• **Requires a huge volume** of "labeled" data to train. Human intervention is necessary anyway.
DEEP-LEARNING BASED MACHINE
Galaxy ZOO like classifications for for the entire sample
N??
Human classifications


DATA FROM NEW SURVEY
Requires a huge volume of "labeled" data to train. Human intervention is necessary anyway.
DEEP-LEARNING BASED MACHINE
Galaxy ZOO like classifications for for the entire sample
N??
Knowledge transfer?
Human classifications

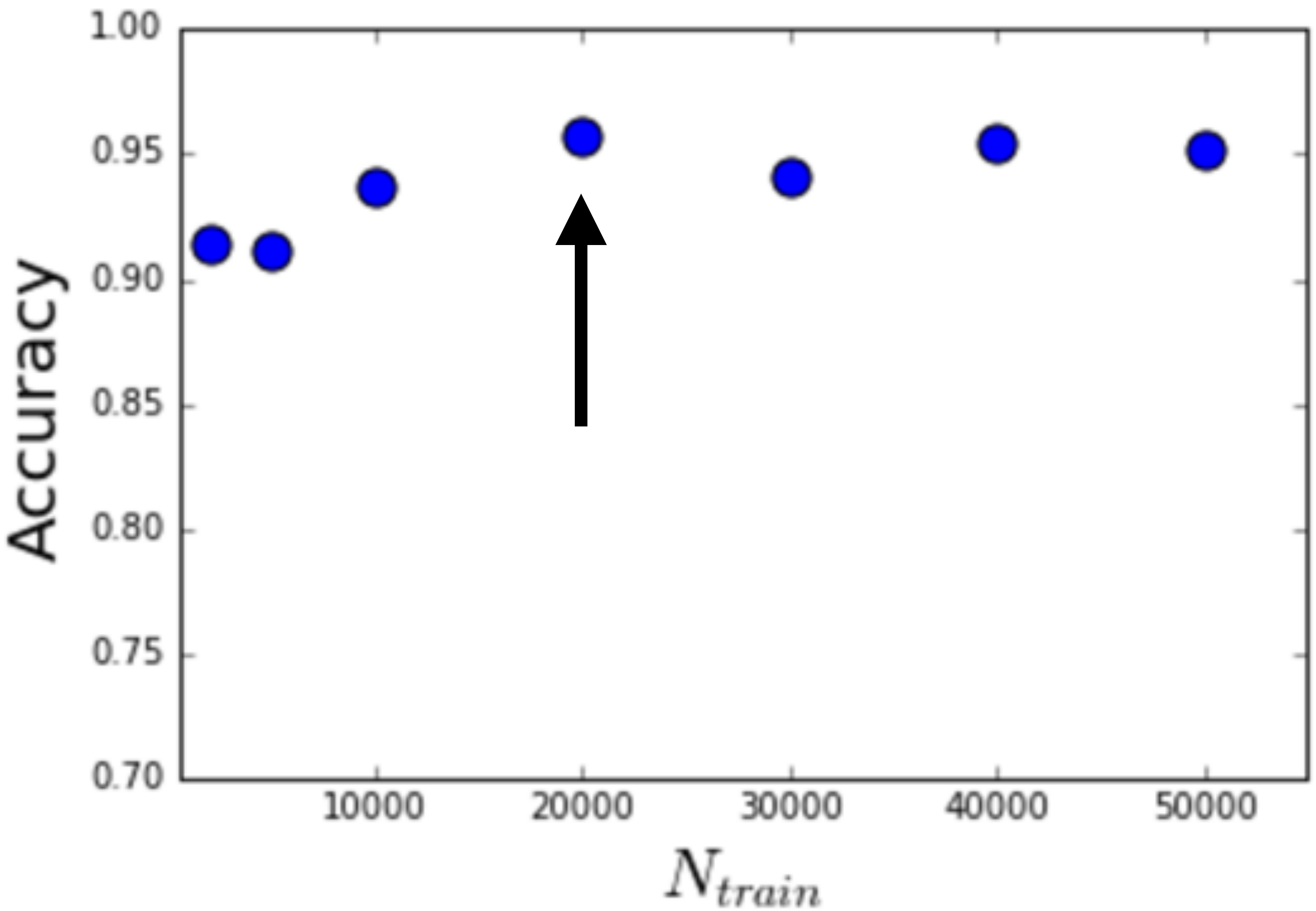

1.0
0.8
0.6
0.4
0.2
0.0
Fraction
5000
10000
15000
20000
N_{train}
Precision
Accuracy
Recall
Q0: Smooth/Disk


1.00
0.95
0.90
0.85
0.80
0.75
0.70
Accuracy
10000
20000
30000
40000
50000
N_{train}

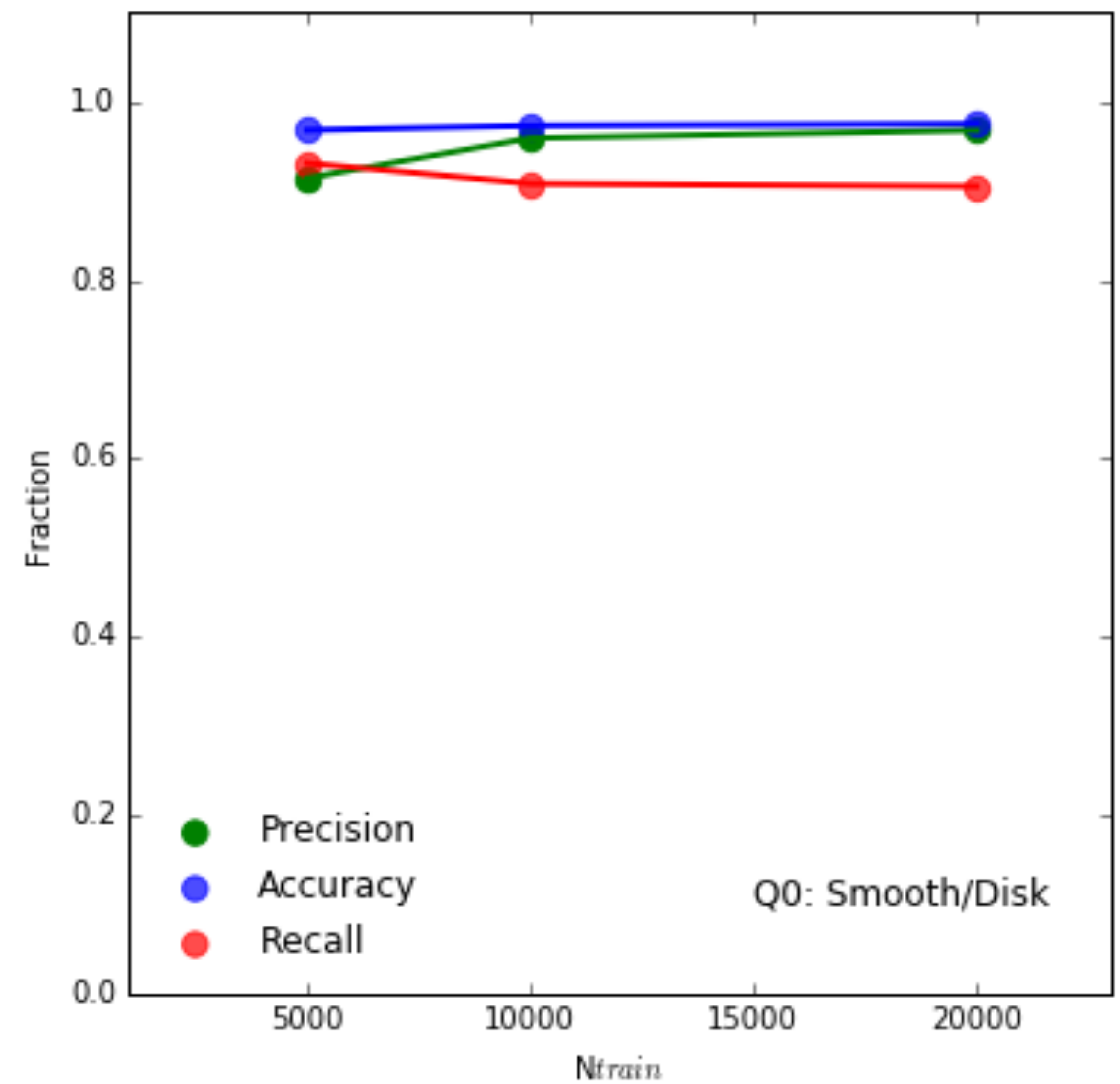

1.0
0.8
0.6
0.4
0.2
0.0
Fraction
5000
10000
15000
20000
N_{train}
Precision
Accuracy
Recall
Q0: Smooth/Disk



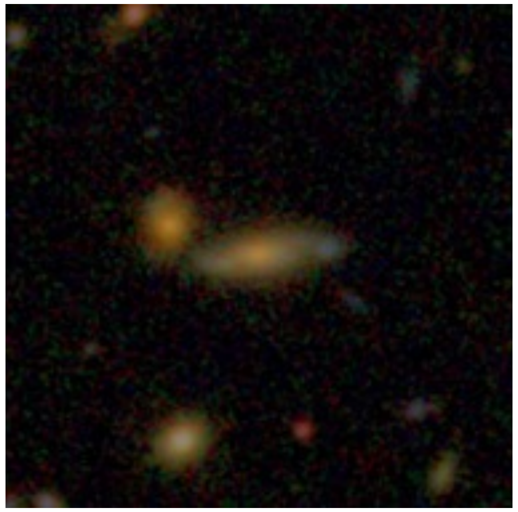
Accuracy
Agreement between classifiers
Nobj=2000
Nobj=5000
Nobj=10000
Nobj=20000
Nobj=30000
Nobj=40000
Nobj=50000
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0



Accuracy
N_{train}
1.00
0.95
0.90
0.85
0.80
0.75
0.70
10000
20000
30000
40000
50000

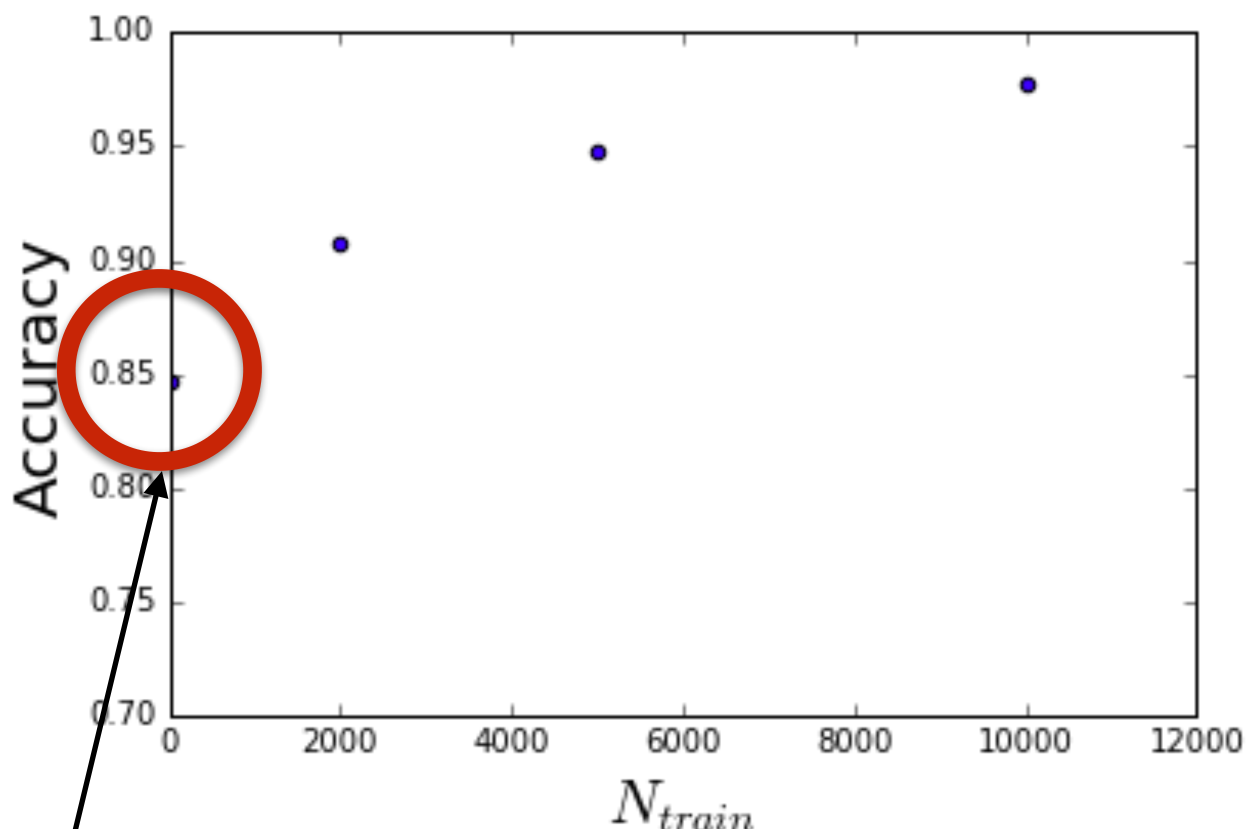


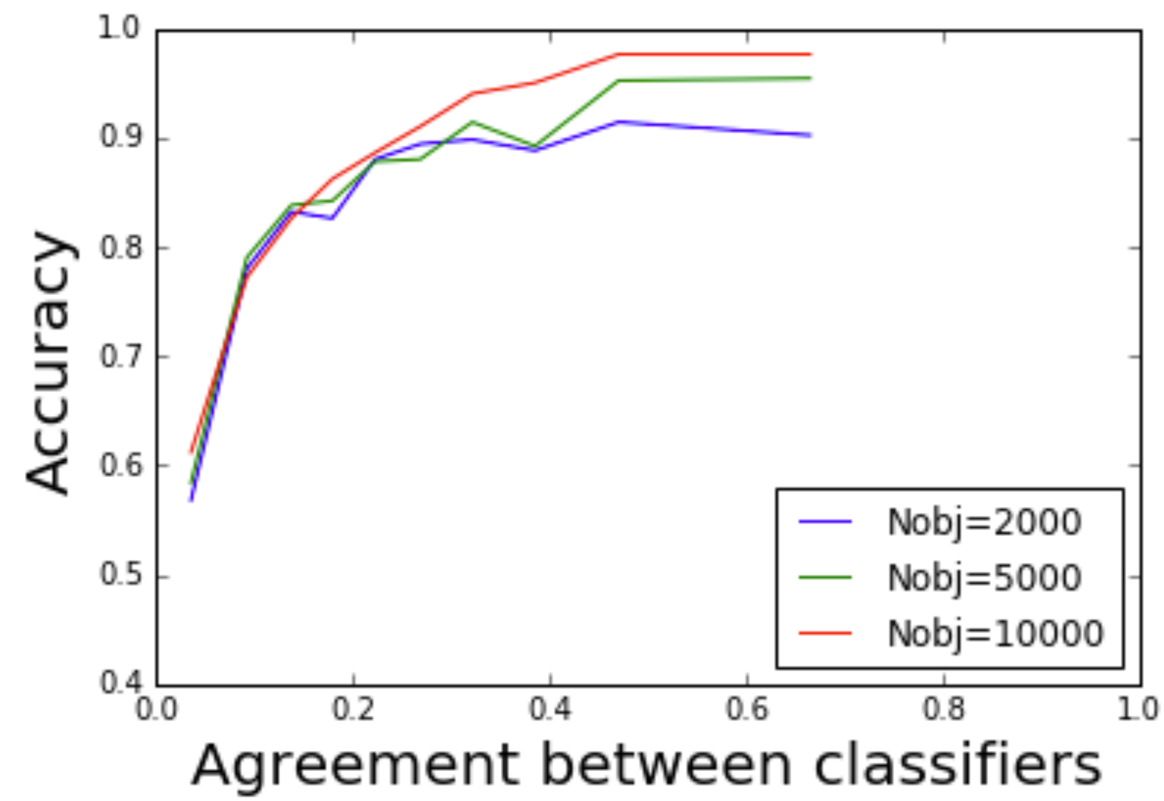
Fraction
N_{train}
Precision
Accuracy
Recall
Q0: Smooth/Disk
1.0
0.8
0.6
0.4
0.2
0.0
5000
10000
15000
20000



CANDELS

DEEP-LEARNING
MACHINE
TUNED
FOR
SDSS





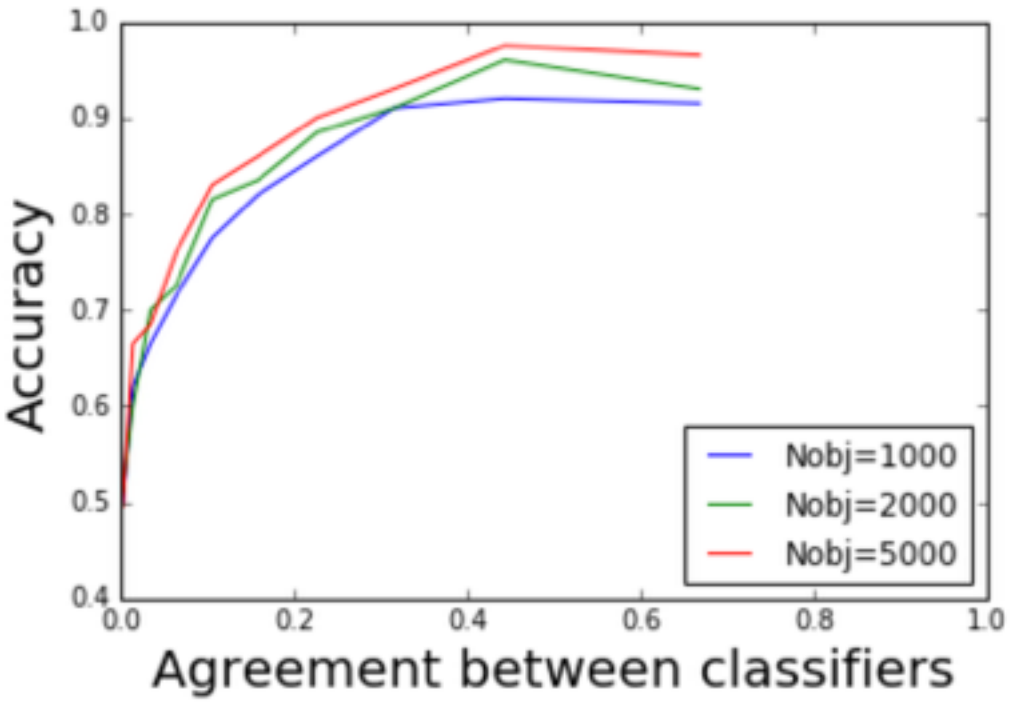
No training

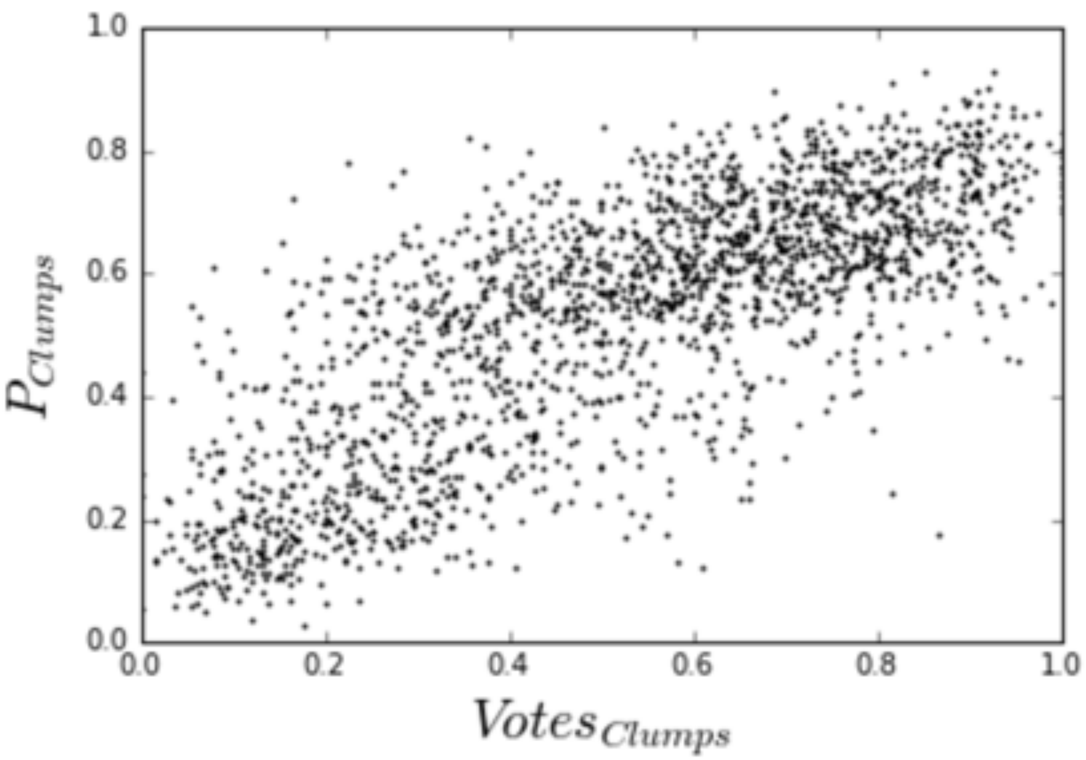
CLUMPS

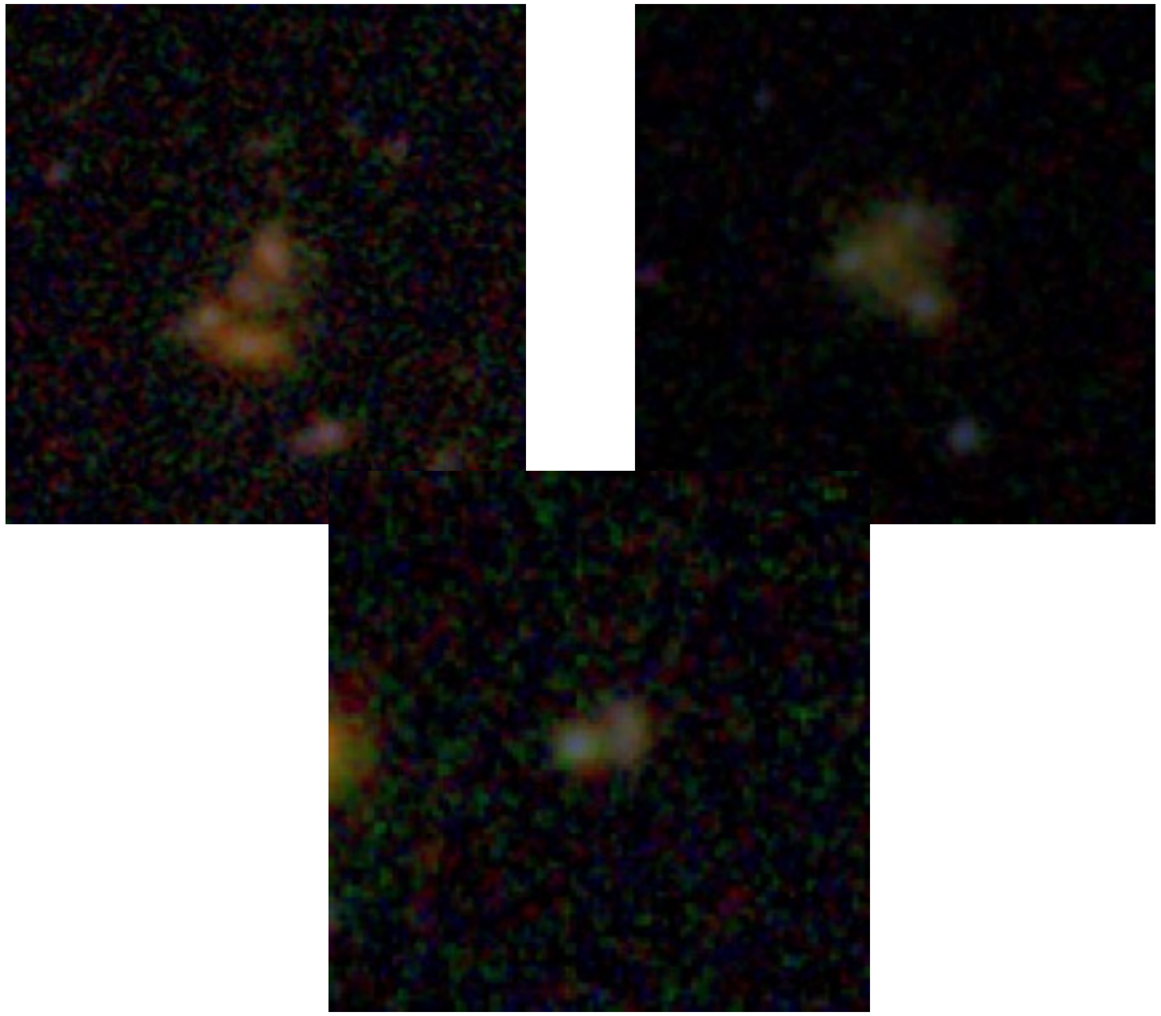
DL machine trained on CANDELS first question

update the weights

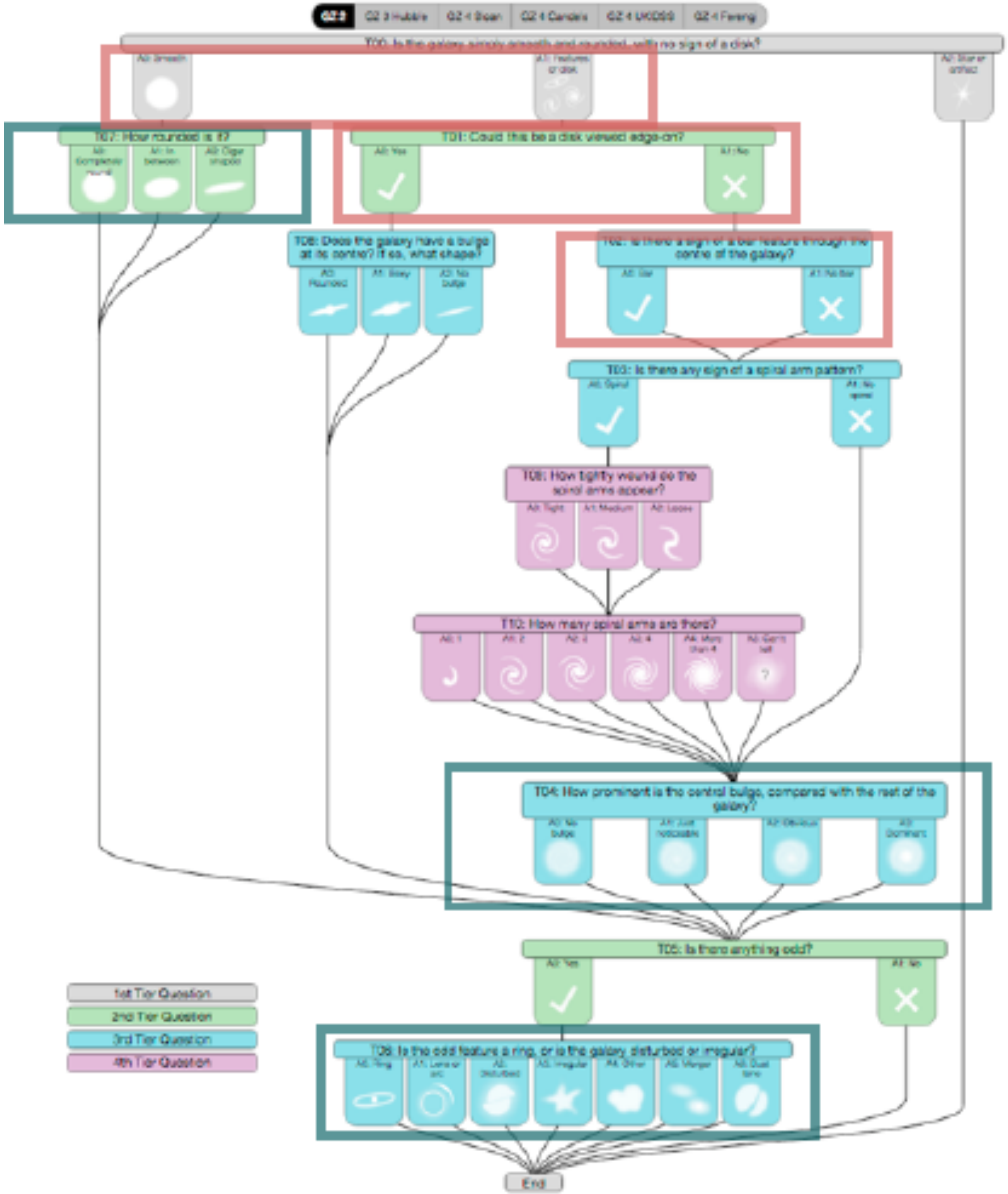
question: clumps (2nd level question)

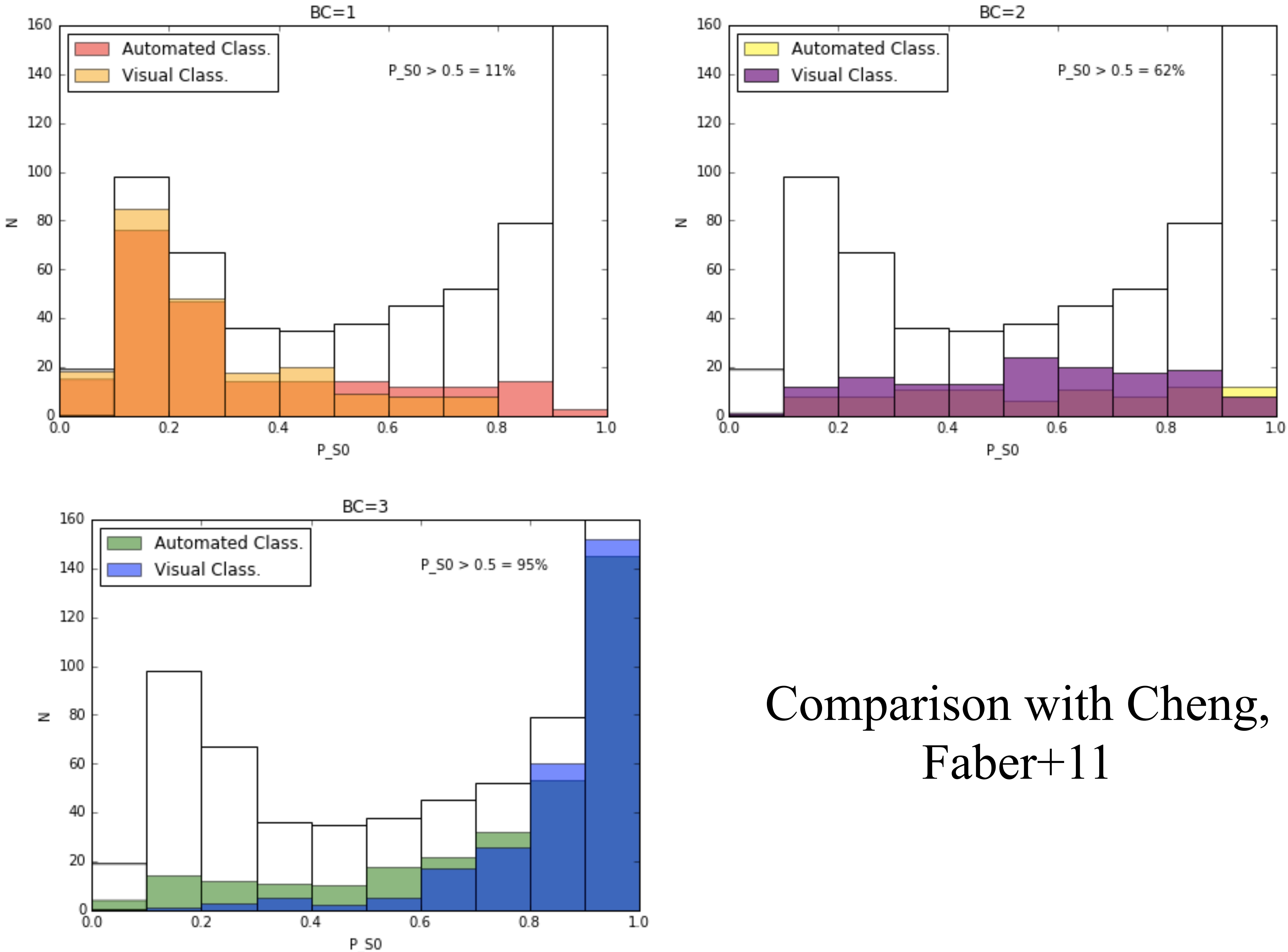






Galaxy Zoo Decision Trees





Comparison with Cheng, Faber+11

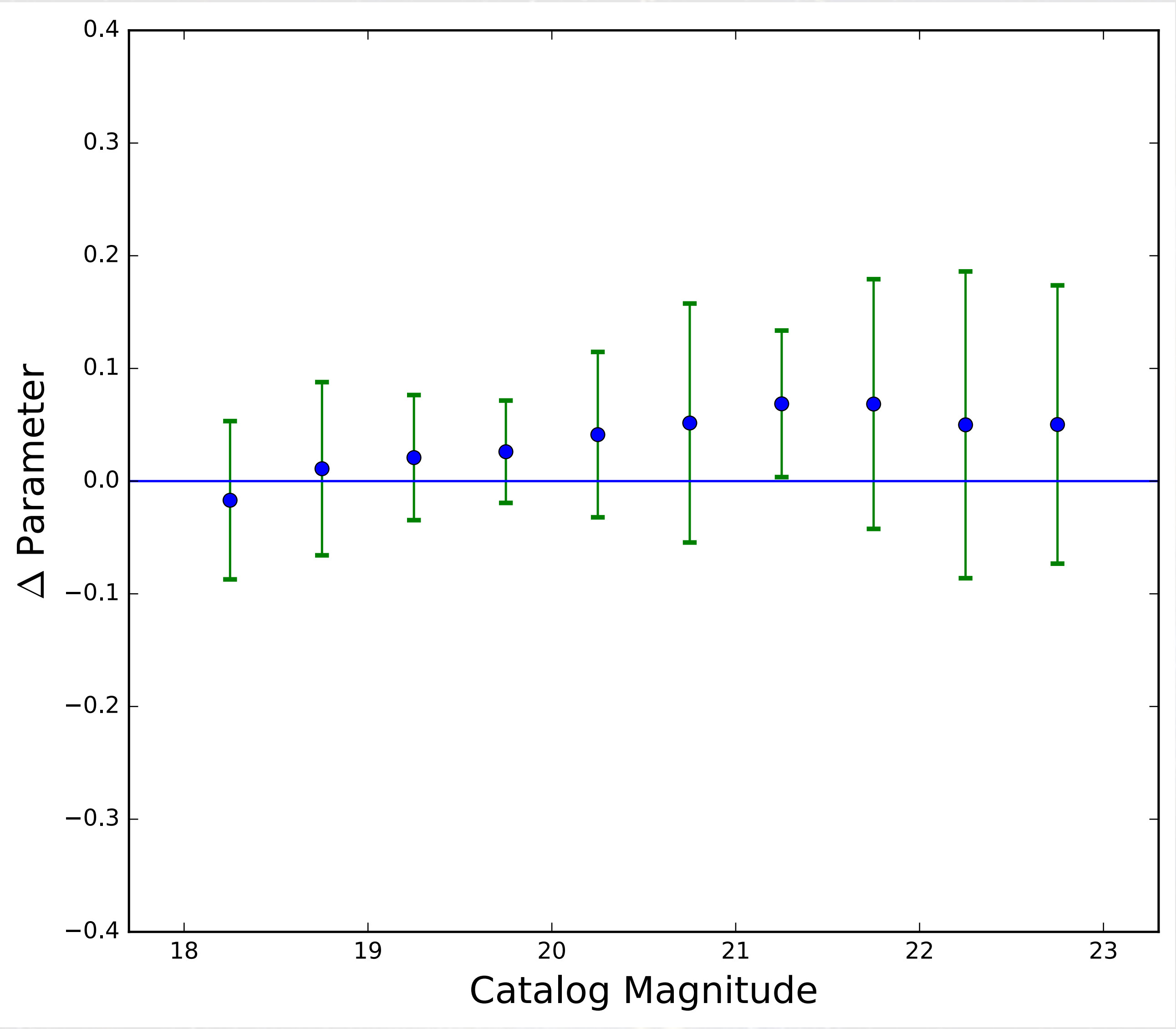
Dominguez-Sanchez, MHC+17

Group #2: Quantitative measurements

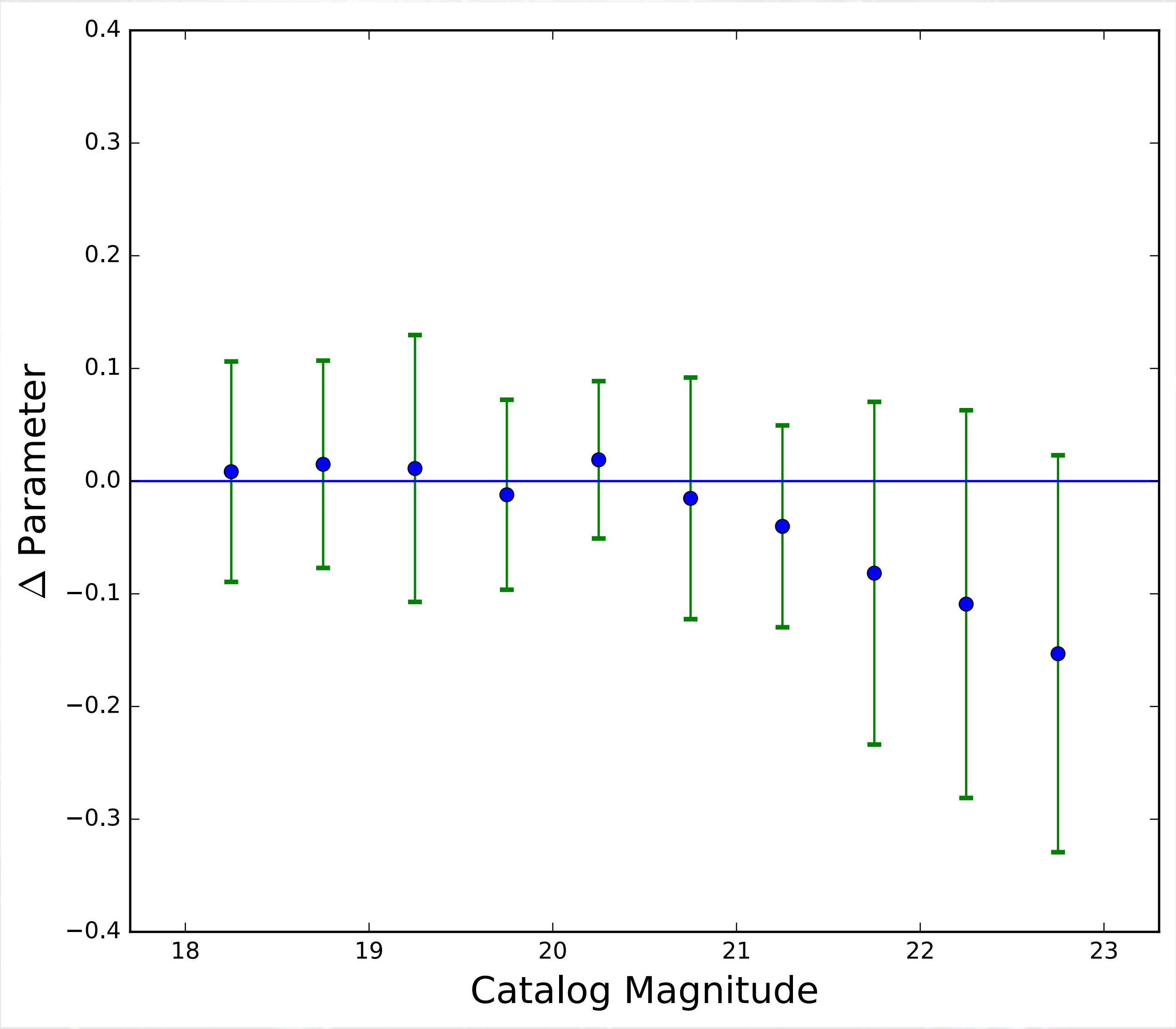
Predictions on Simulated Data

5000 stamps

Magnitude



DNN

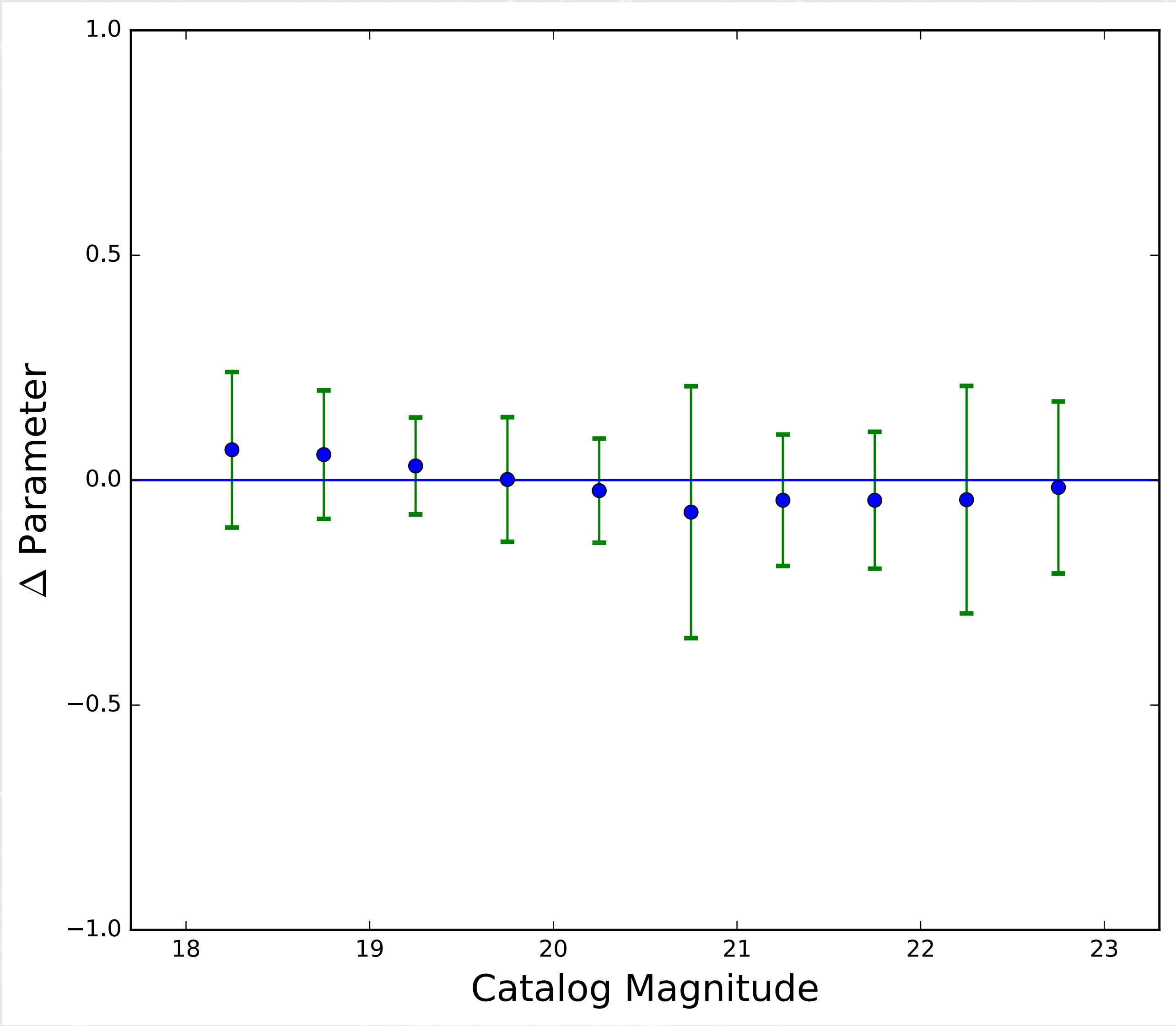


GALFIT-SExtractor

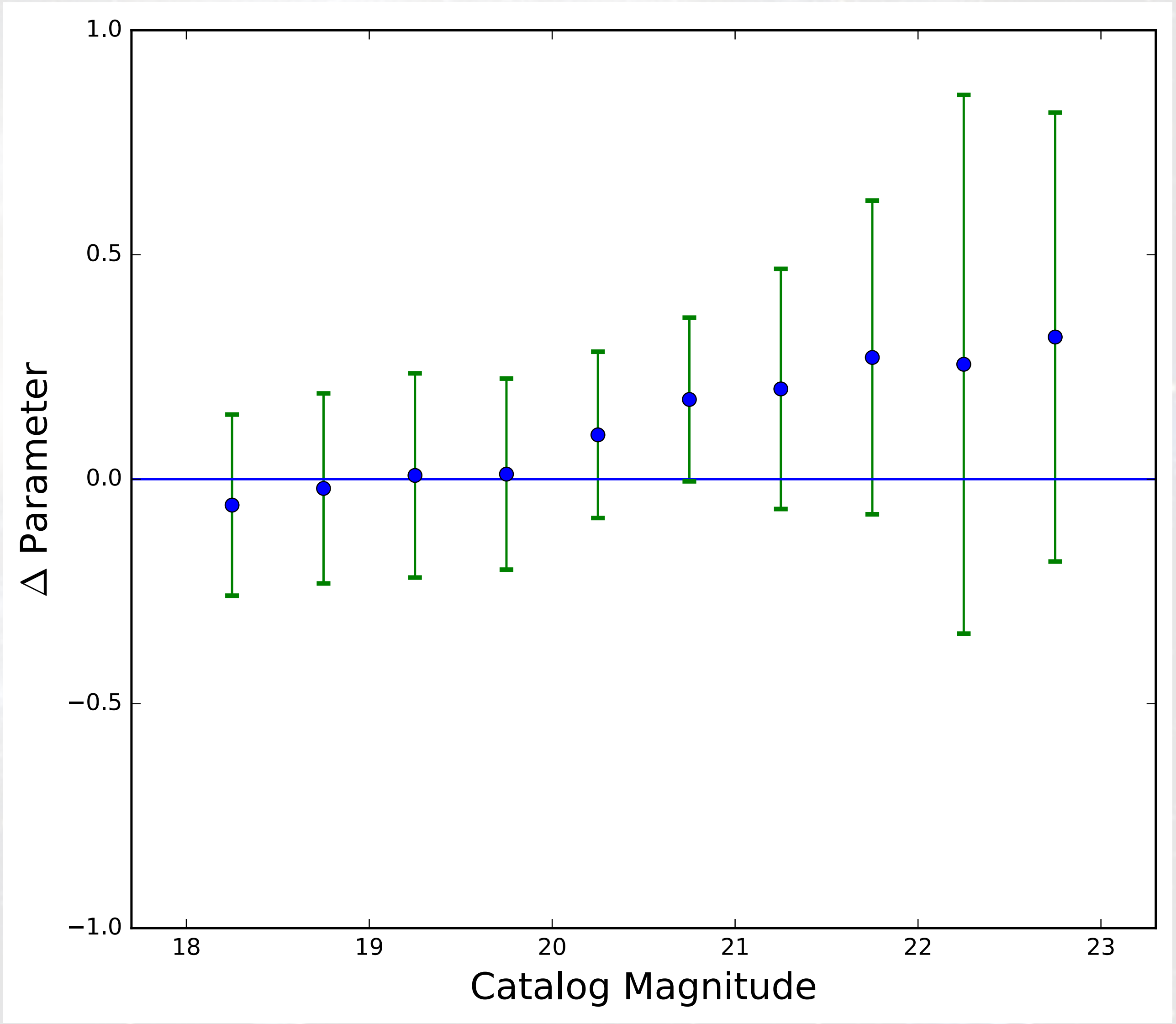
Predictions on Simulated Data

5000 stamps

Sersic index



DNN



GALFIT-SExtractor

Summary of predictions on simulation

R^2 simulated data			
Parameter	Architecture 1	Architecture 2	GALFIT
Magnitude	0.947	0.995	0.986
Radius	0.892	0.955	0.738
Sérsic index	0.887	0.348	0.292
Ellipticity	0.755	0.603	0.896
Position Angle	0.941	nc	0.825

coefficient of determination

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

~5000 sources fitted

GALFIT/GALAPAGOS

Time: ~ 4 Hours

CNN (ONCE TRAINED)

Time: < 3 seconds

Summary of Predictions on Real Data (after domain adaptation)

R^2 Real data			
Parameter	Before TL	After TL	2 GALFIT
Magnitude	0.788	0.982	0.985
Radius	-1.639	0.856	0.860
Sérsic index	-0.768	0.718	0.735
Ellipticity	0.256	0.897	0.904
Position Angle	0.132	0.893	0.863

coefficient of determination

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

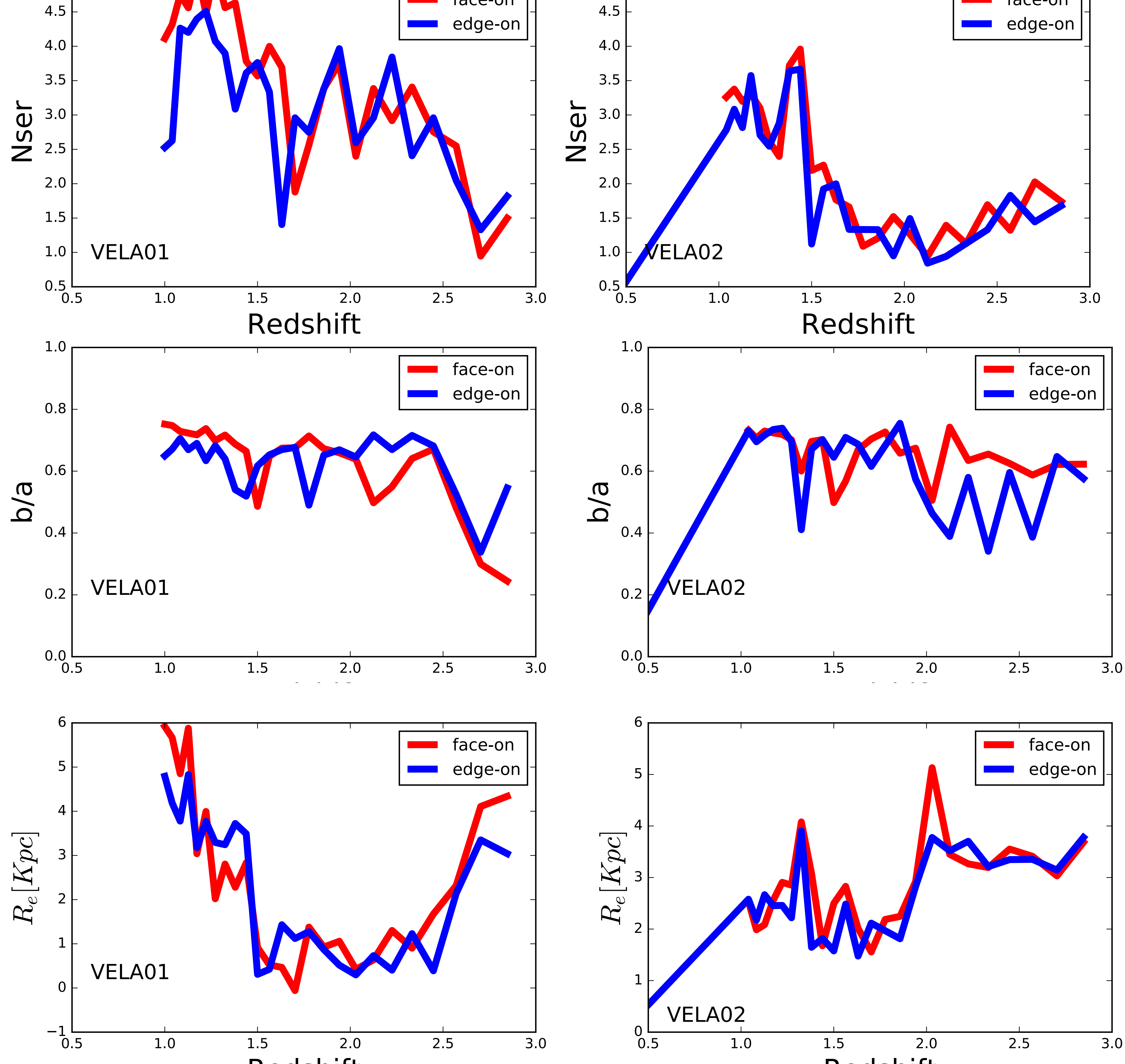
~3000 sources fitted

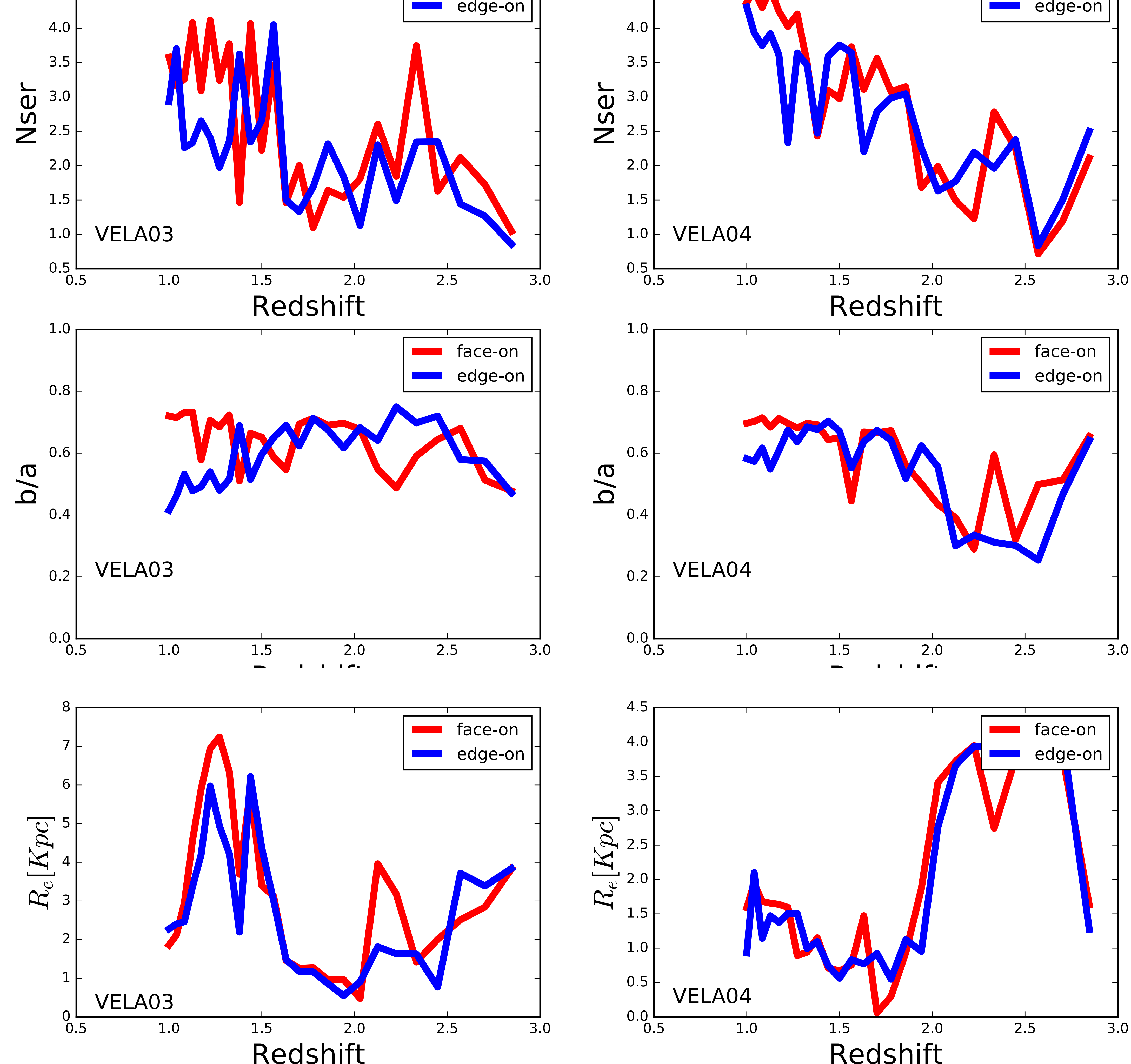
GALFIT/GALAPAGOS

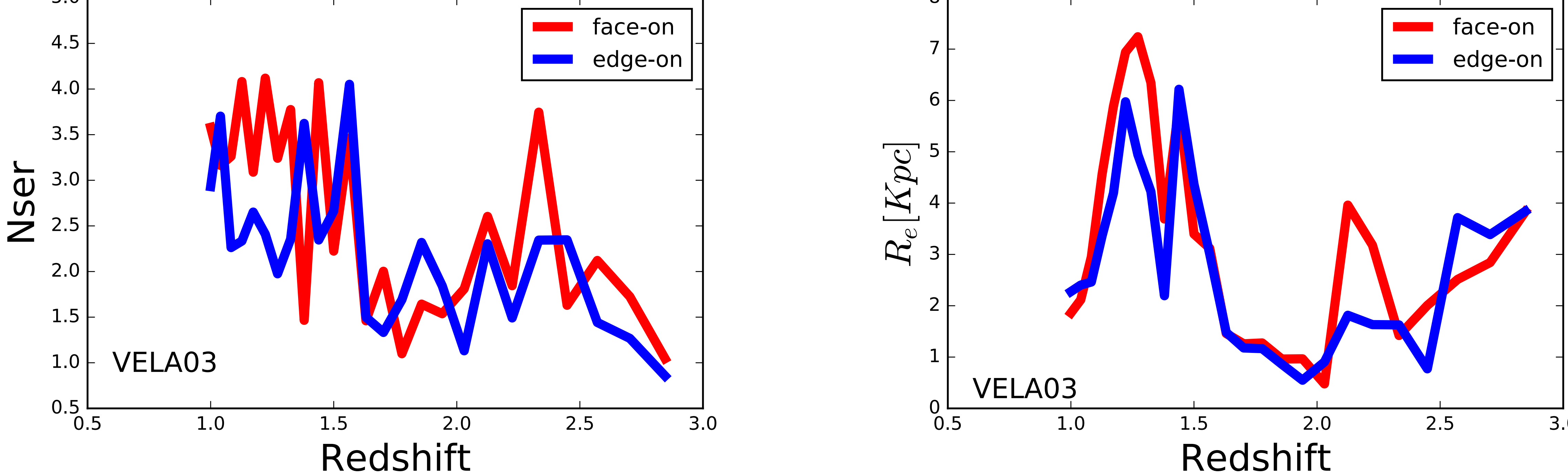
Time: ~ 2h30

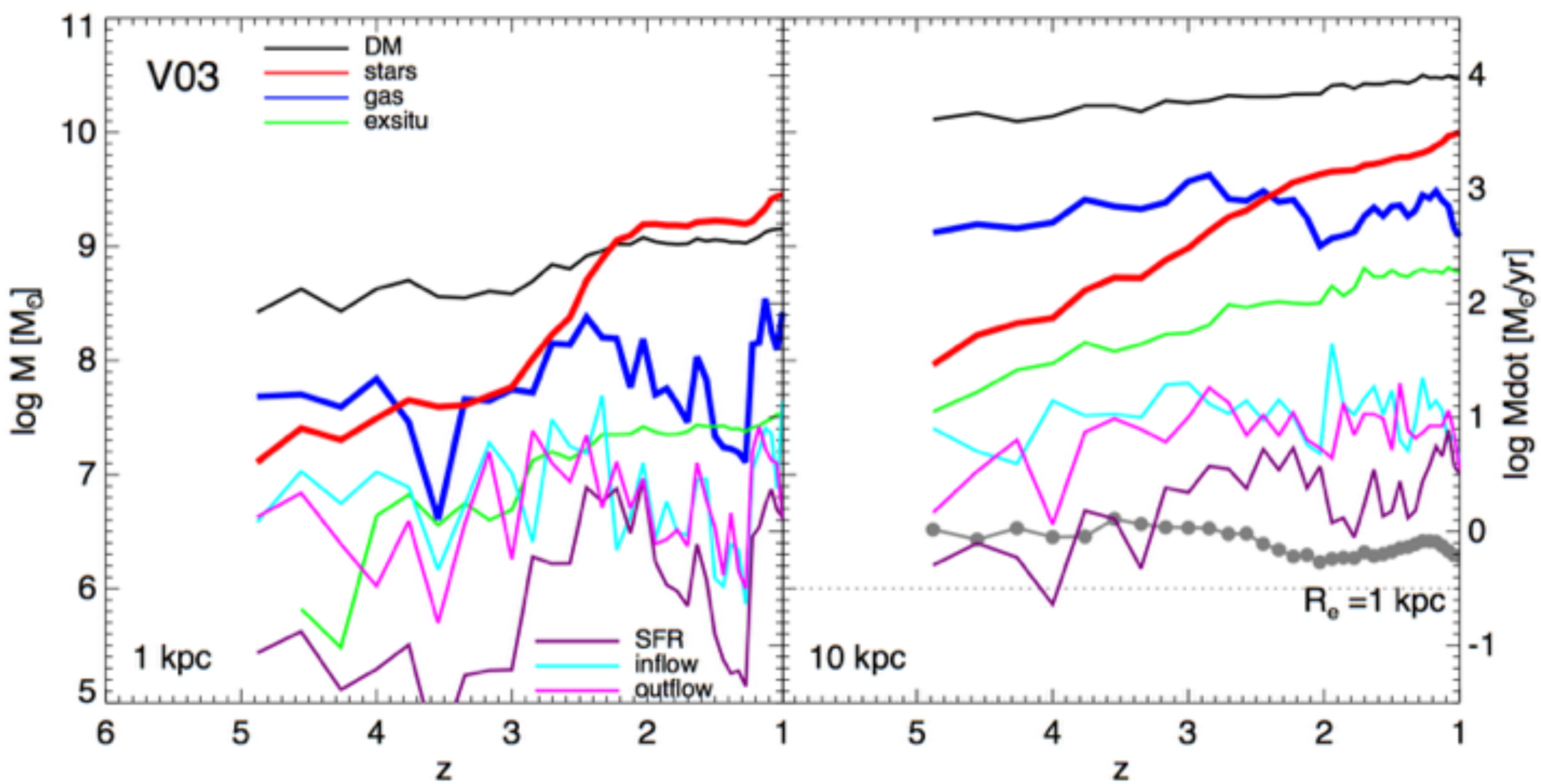
CNN

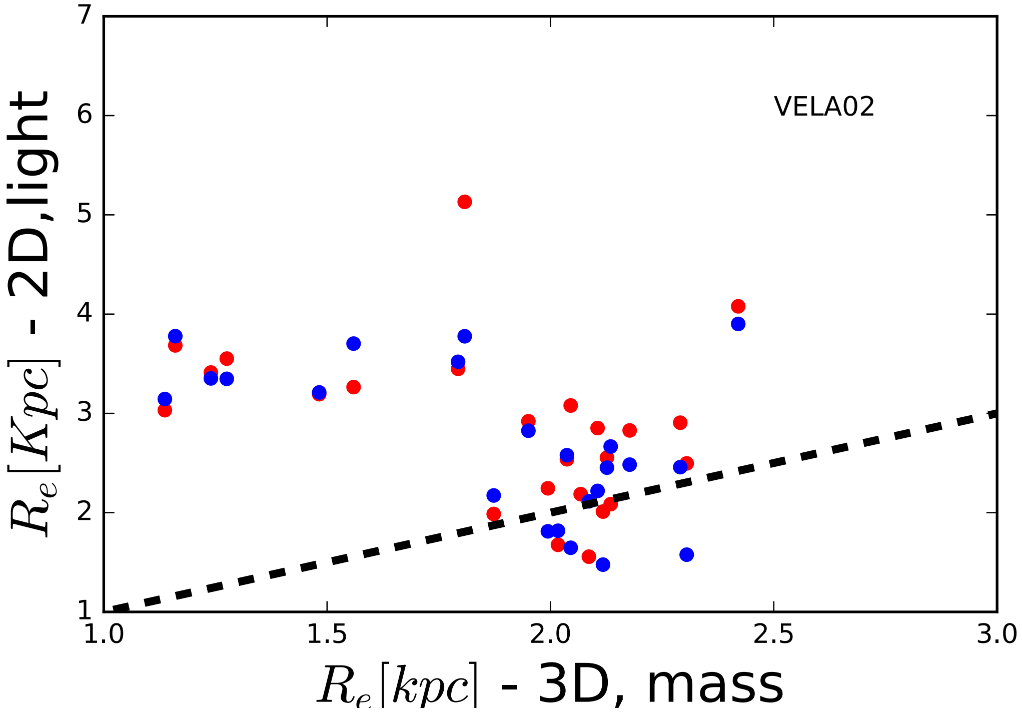
Time: 10 min domain adaptation + 2s

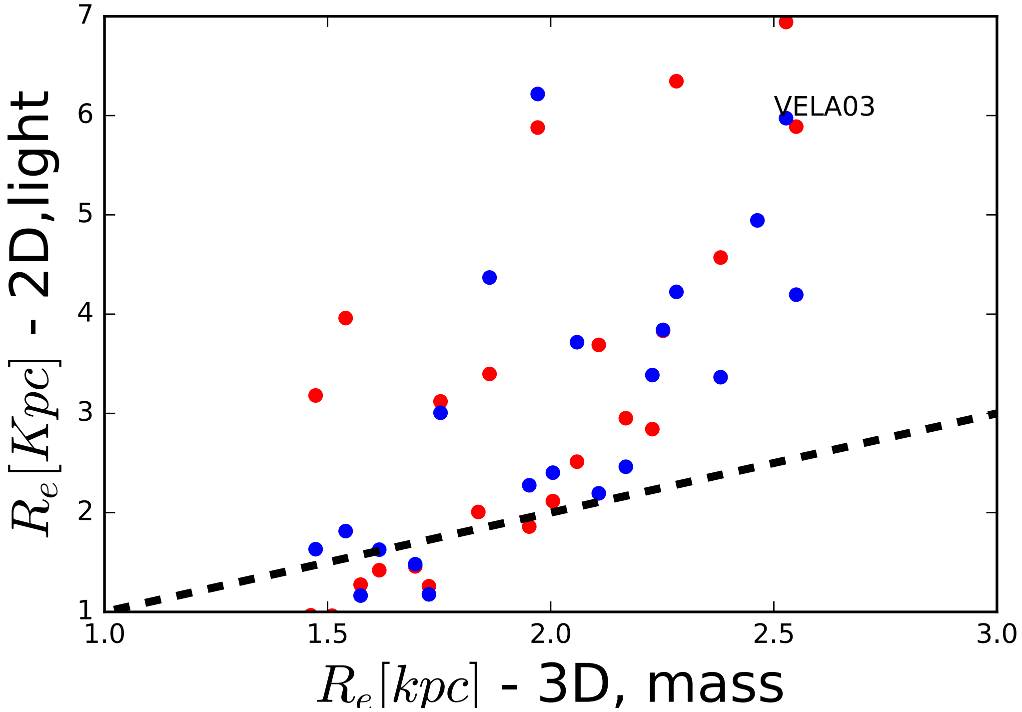

face-on
edge-on
Nser
Redshift
VELA01
VELA02
b/a
$R_e\,[Kpc]$

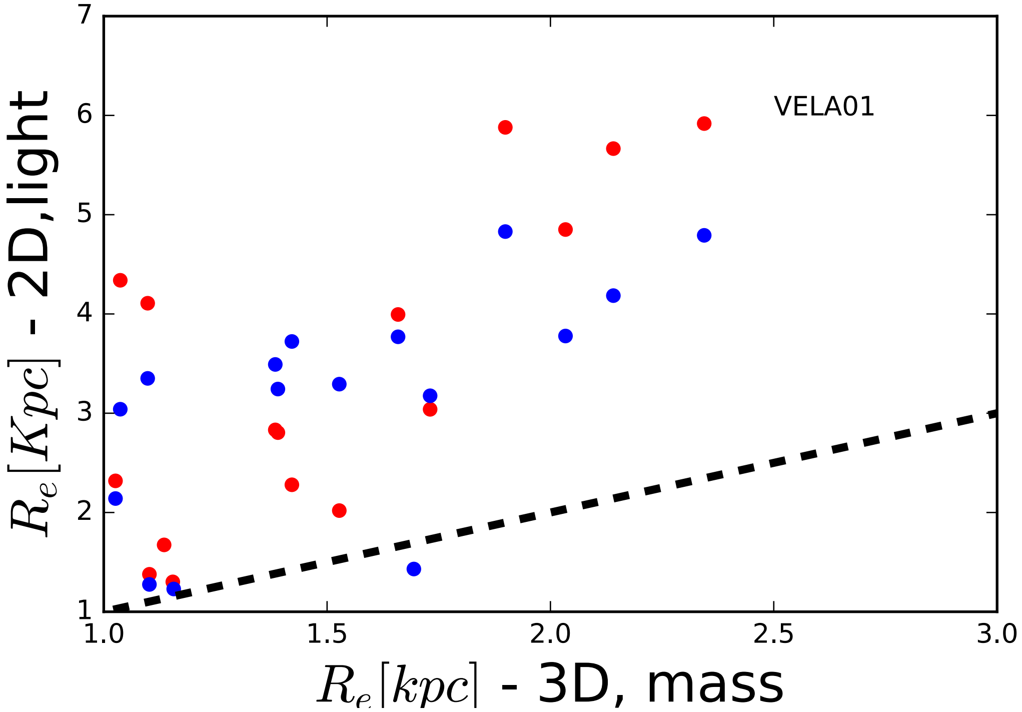

edge-on
Nser
VELA03
Redshift
edge-on
Nser
VELA04
Redshift
face-on
edge-on
b/a
VELA03
face-on
edge-on
b/a
VELA04
face-on
edge-on
$R_e [Kpc]$
VELA03
Redshift
face-on
edge-on
$R_e [Kpc]$
VELA04
Redshift


face-on
edge-on
Nser
VELA03
Redshift
$R_e\,[Kpc]$
VELA03
Redshift


V03
DM
stars
gas
exsitu
log M [M⊙]
1 kpc
SFR
inflow
outflow
z
10 kpc
R_e =1 kpc
log Mdot [M⊙/yr]
z


VELA02
$R_e[Kpc]$ - 2D,light
$R_e[kpc]$ - 3D, mass


VELA03
$R_e[Kpc]$ - 2D,light
$R_e[kpc]$ - 3D, mass


VELA01
$R_e[Kpc]$ - 2D,light
$R_e[kpc]$ - 3D, mass

Group #3: Hidden observables / correlations