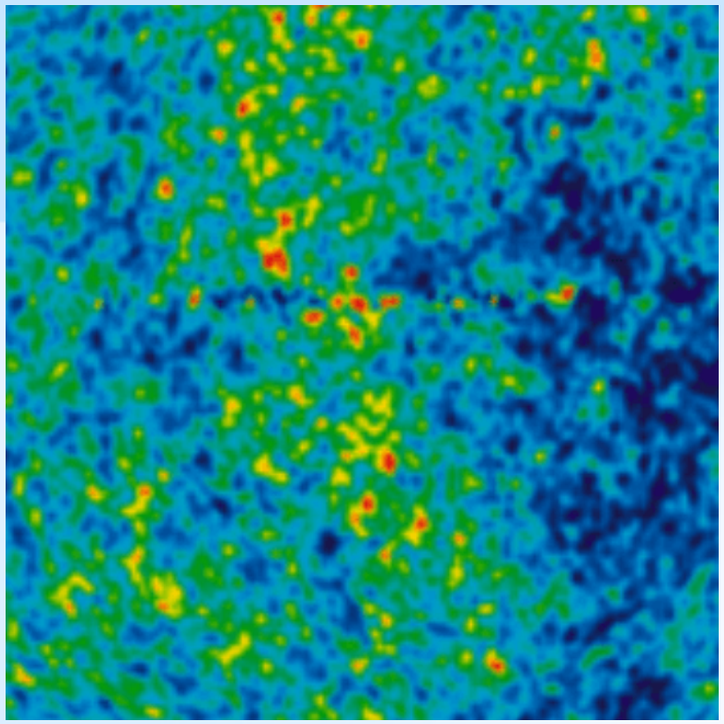
Cosmic Overdensities in the early Universe

Validating Protoclusters across Radio and Submm using Machine Learning

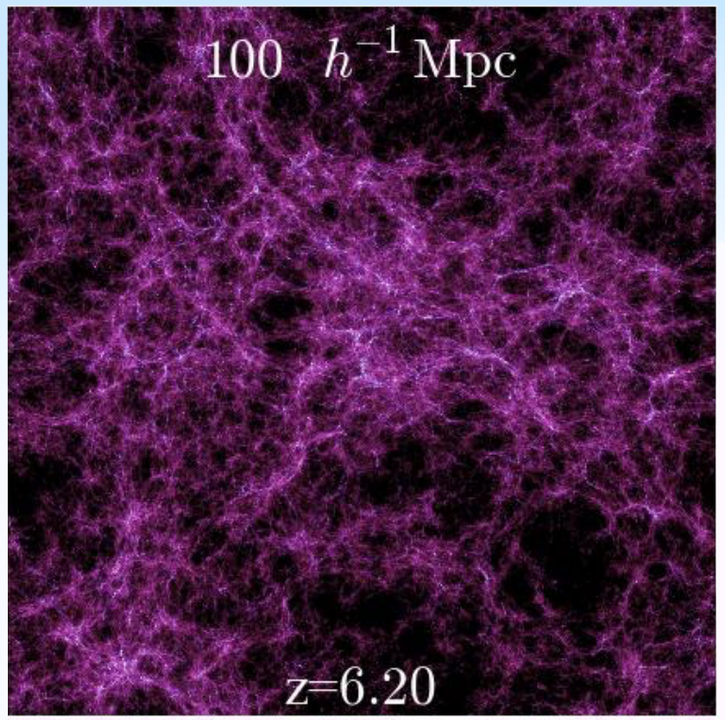
Tom Bakx **Chalmers UoT**

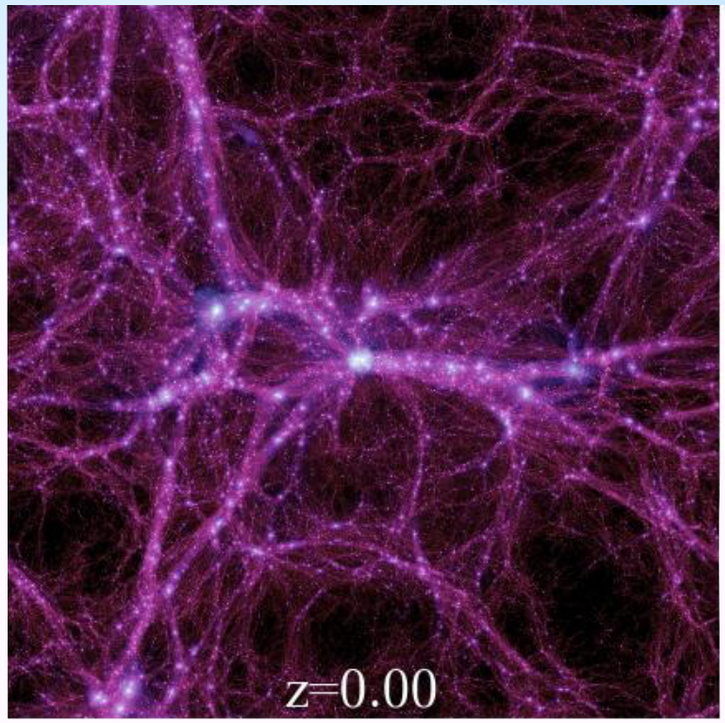
Galaxy clusters are solidified quantum-fluctuations left over from the big bang



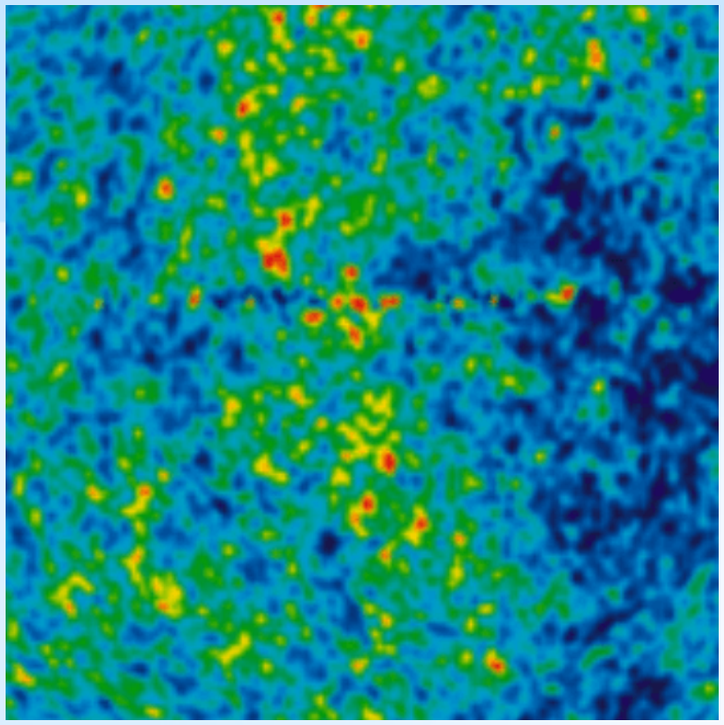
z = 1100

Millennium simulation



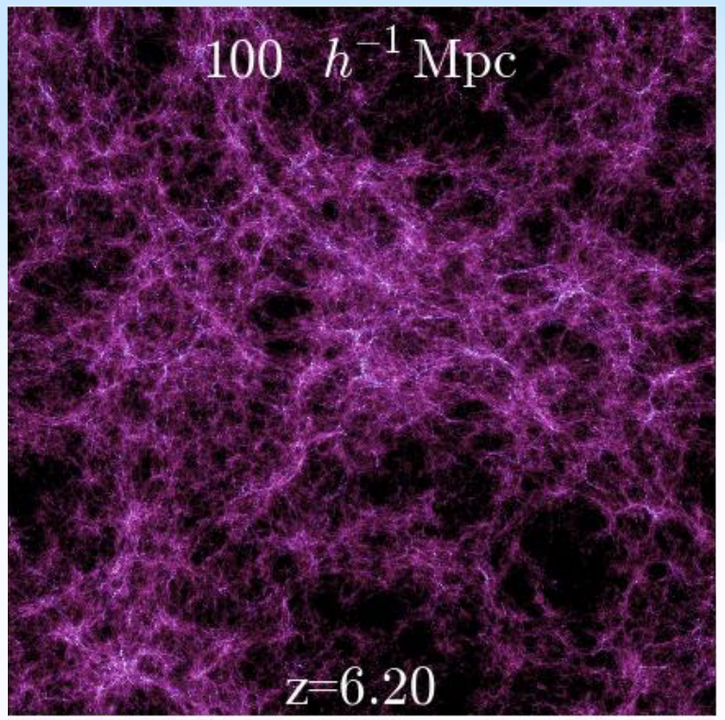


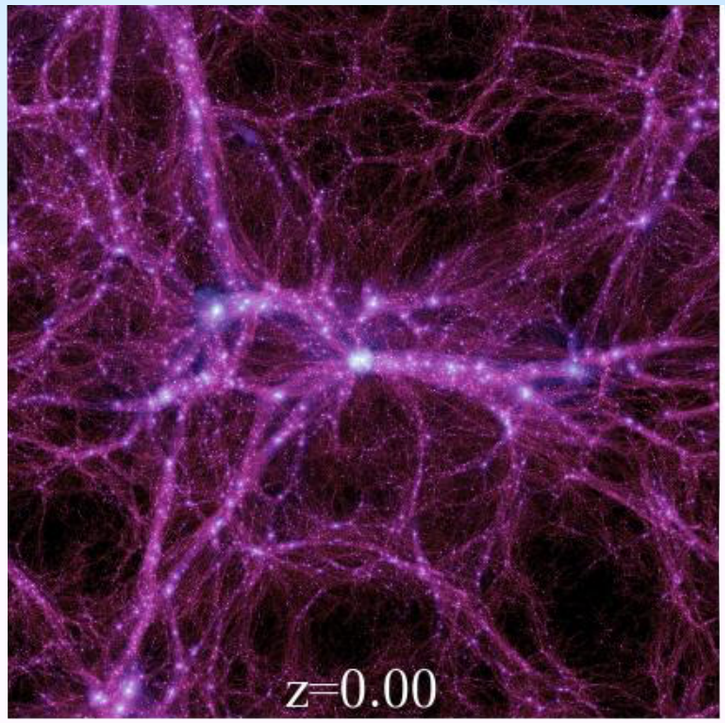
Galaxy clusters are solidified quantum-fluctuations left over from the big bang



z = 1100

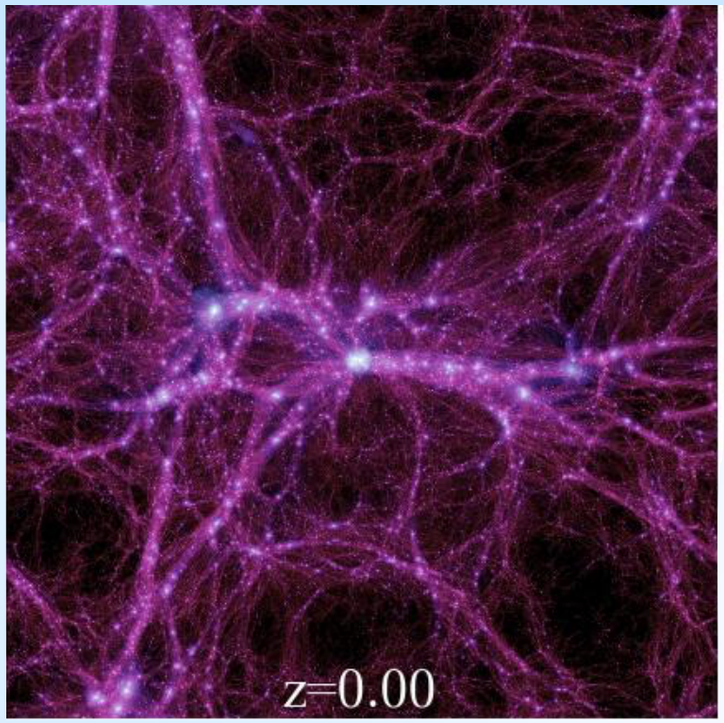
Millennium simulation

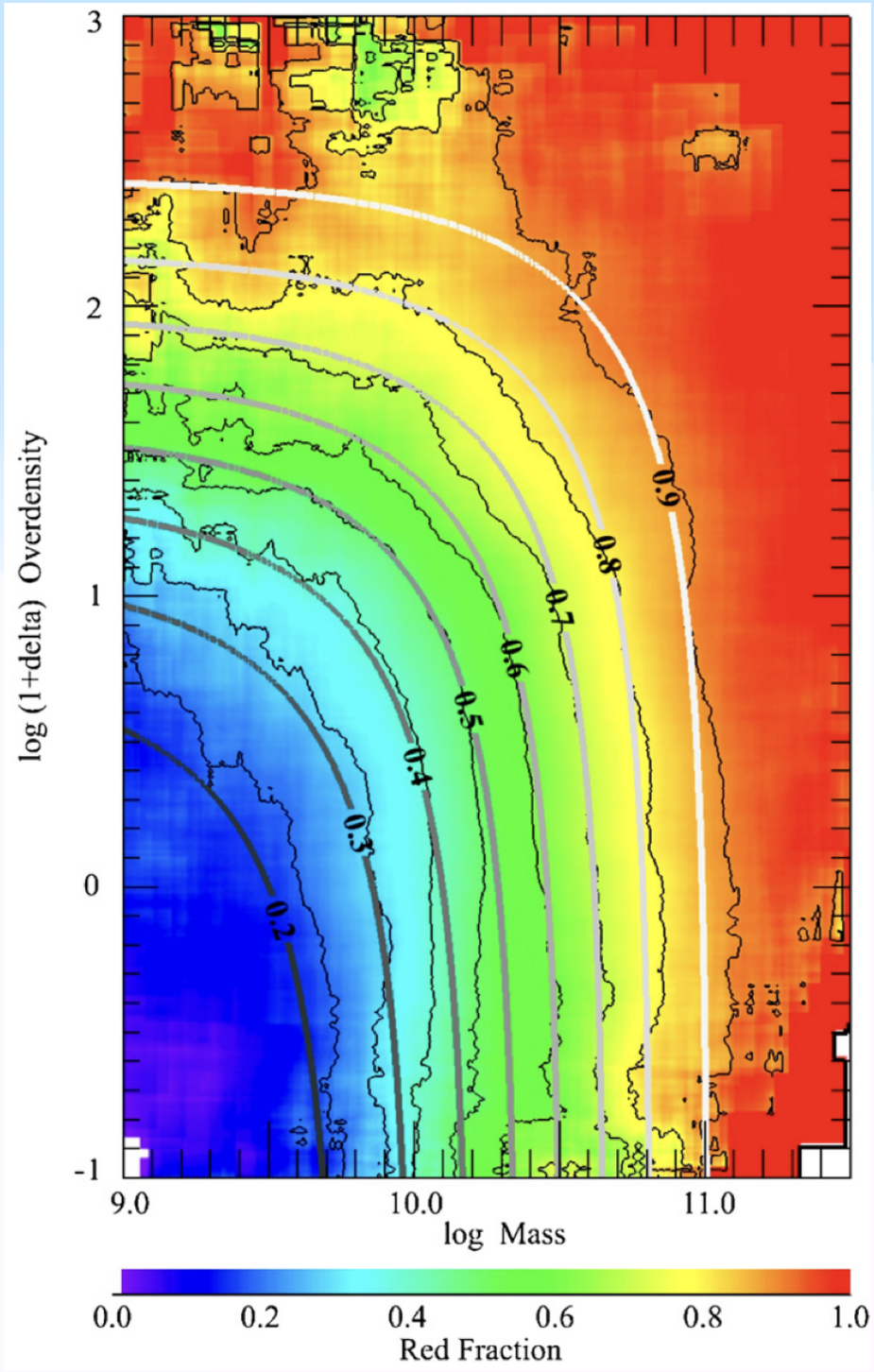




Large datasets (incl. SKA) are required to observationally link the populations and evolution, **necessitating ML techniques**

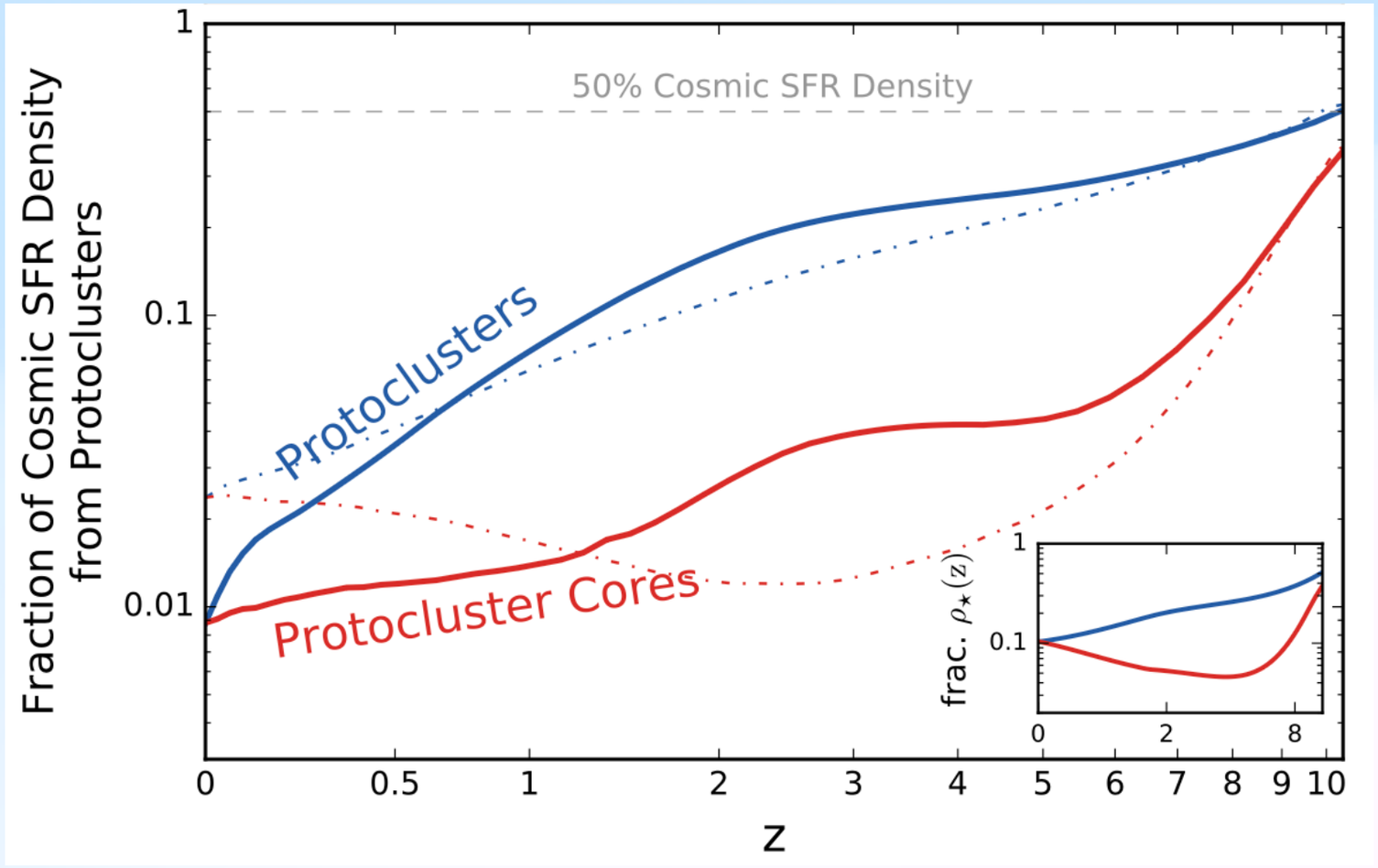
Galaxy evolution (appear to) happen at accelerated rates





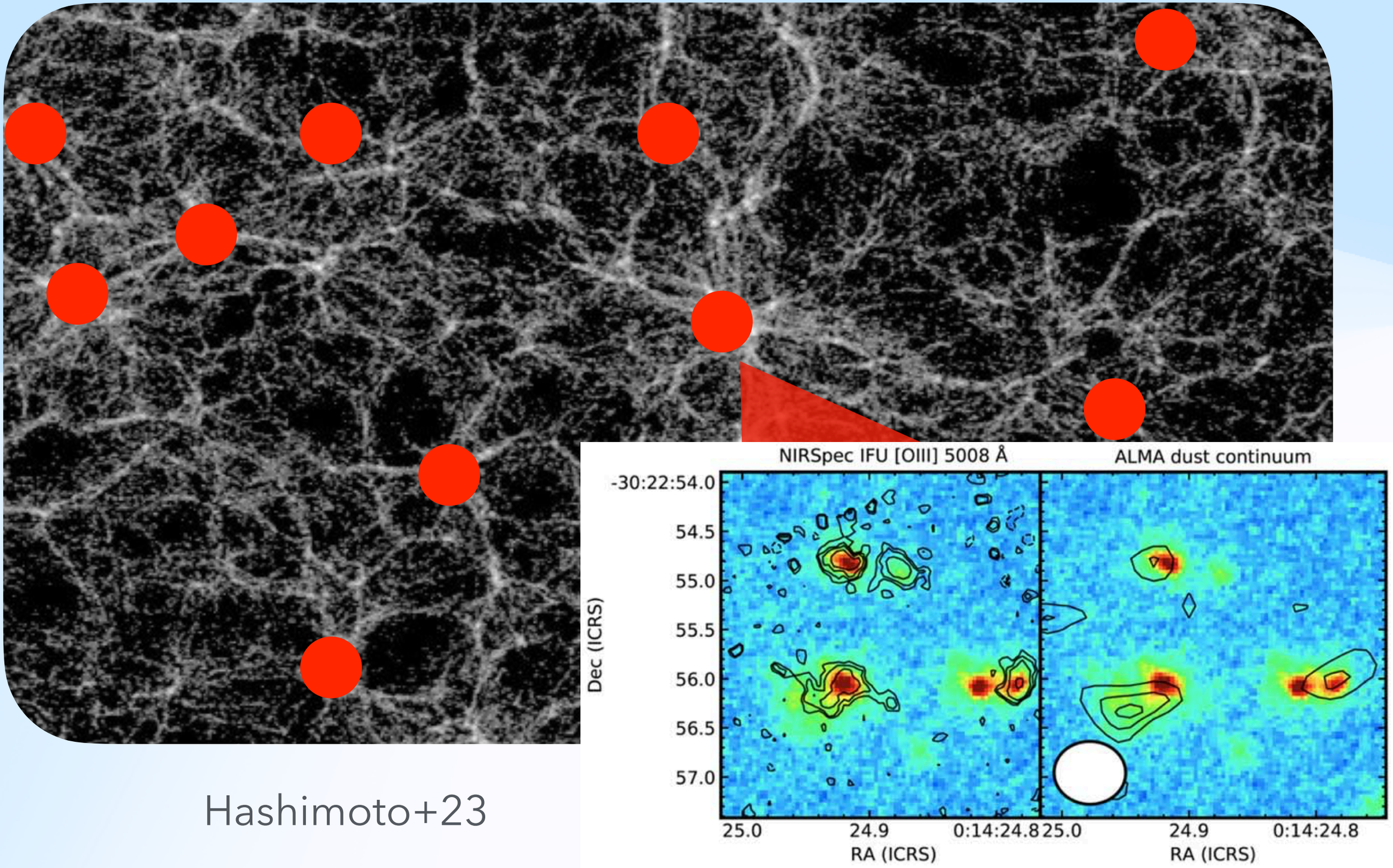
Peng et al. 2010

At z > 10, half of all SF occurs in clusters



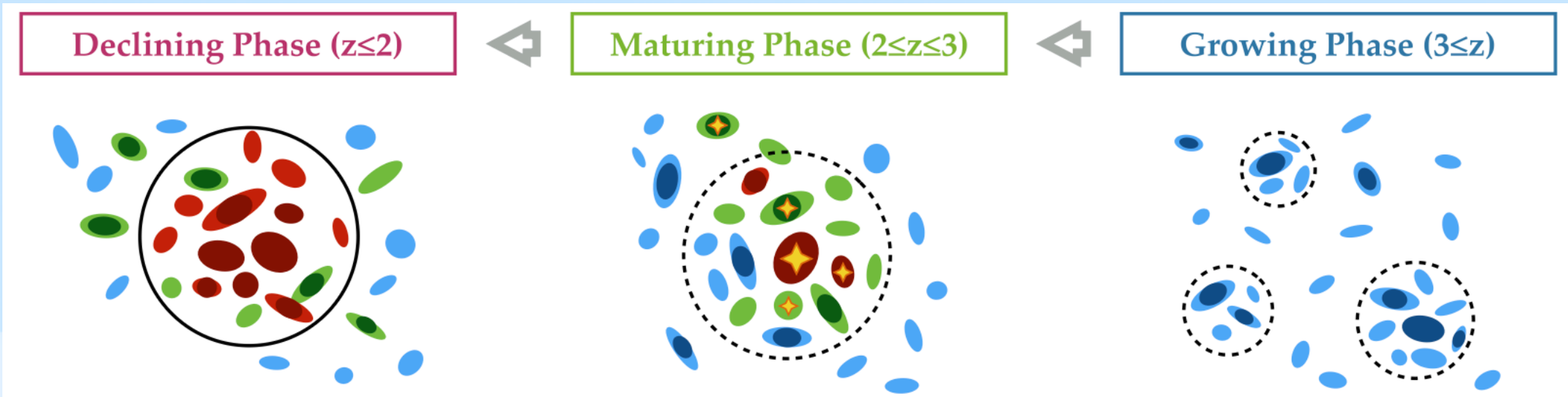
Chiang et al. 2013, 2017

Many high-z galaxies are found in clusters



Hashimoto+23

Overdensities evolve more rapidly, but remain poorly characterized

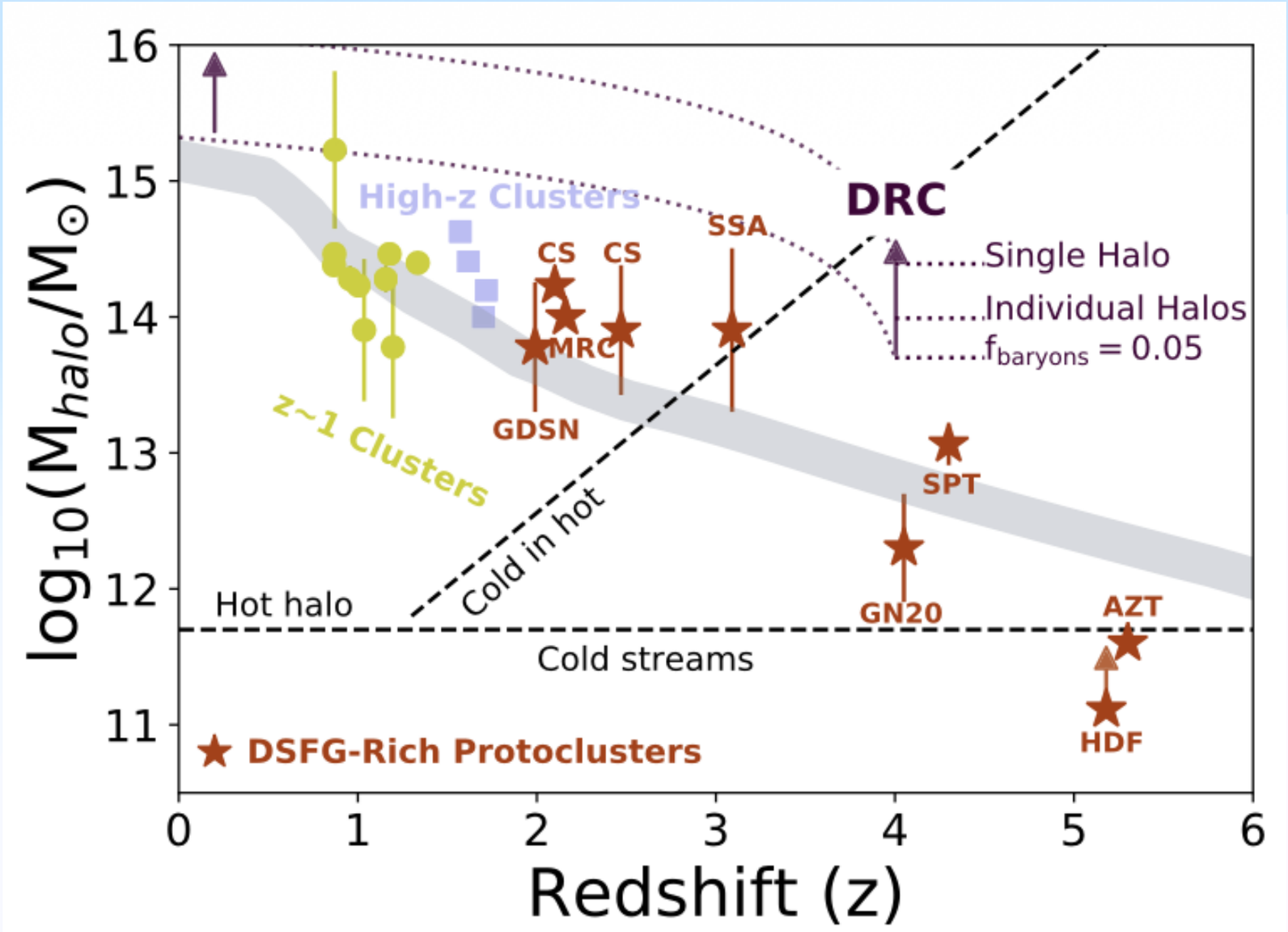


	Declining Phase	Maturing Phase	Growing Phase
Galaxies	Starvation / Environmental quenching	Energetic AGN feedback	Massive galaxy formation
Halo Gas	Cold gas depletion / Superheated plasma	Pre-heating / Core collapse	Cold gas stream in hot gas

Shimakawa+18

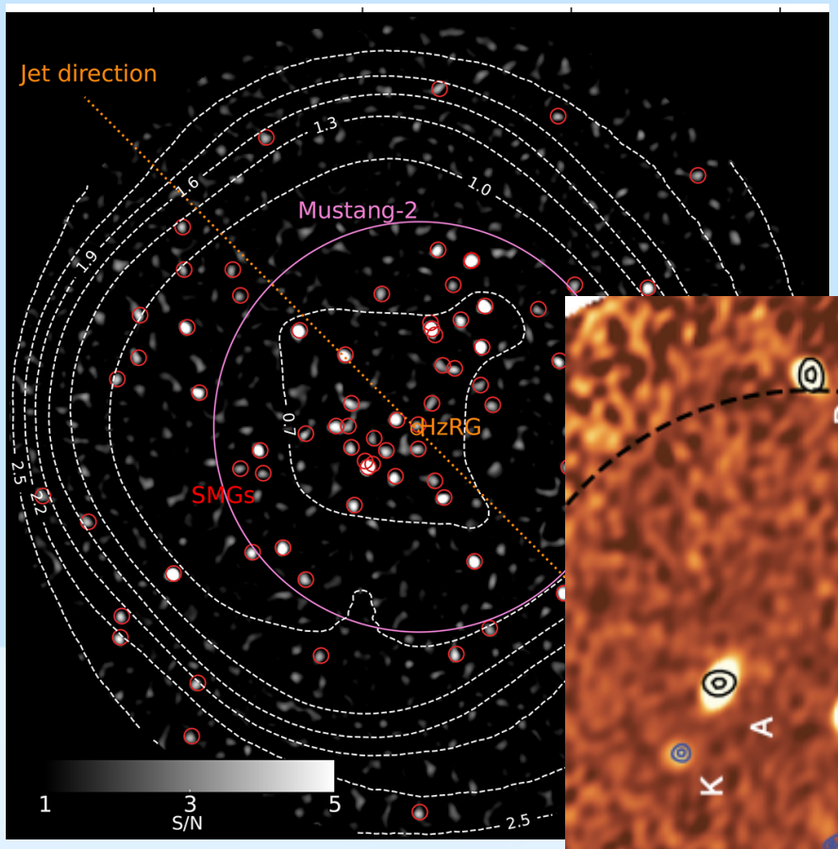
A single protocluster can destroy Λ-CDM

Λ-CDM limit
from models

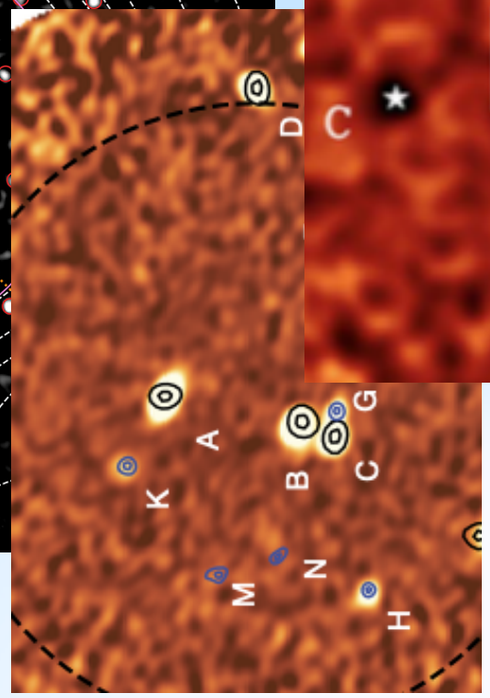


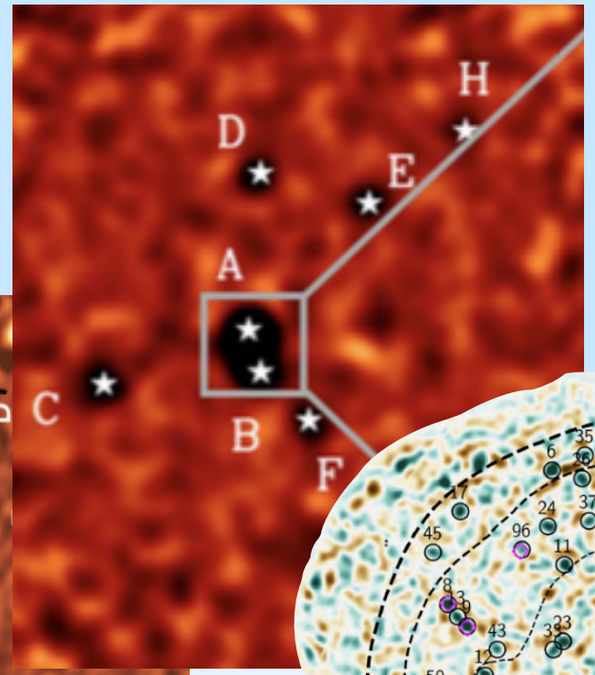
Long et al. 2020

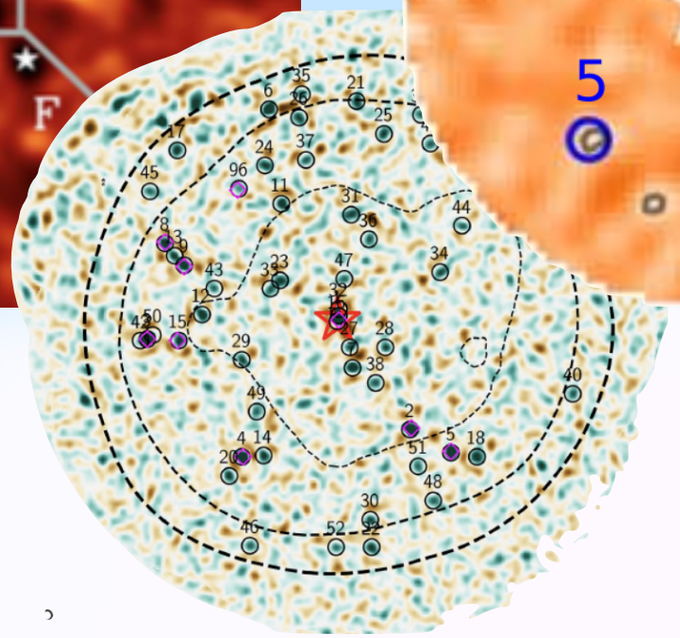
The full wavelength regime is used to study these overdensities

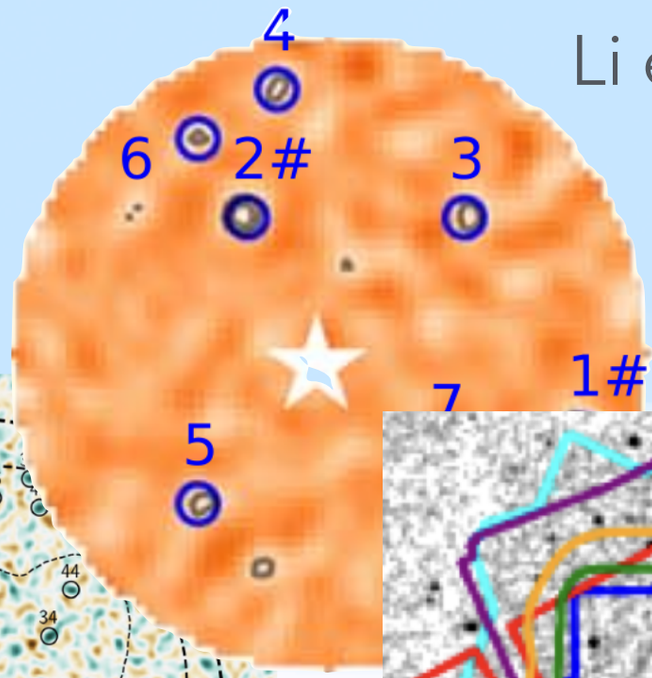


Zhou et al. 2024
Miller et al. 2018
Oteo et al. 2017

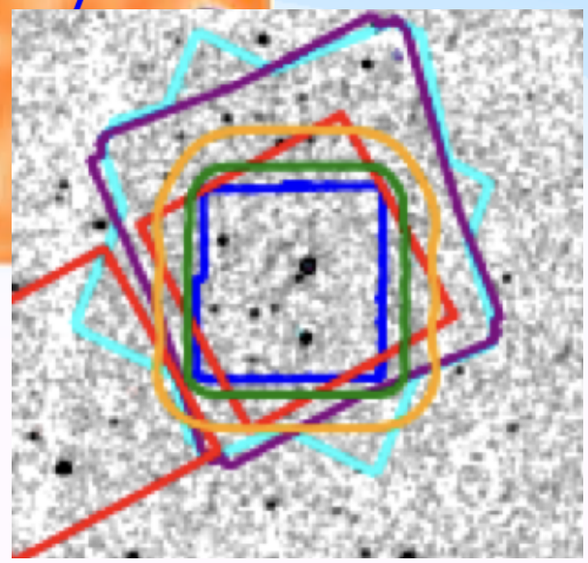






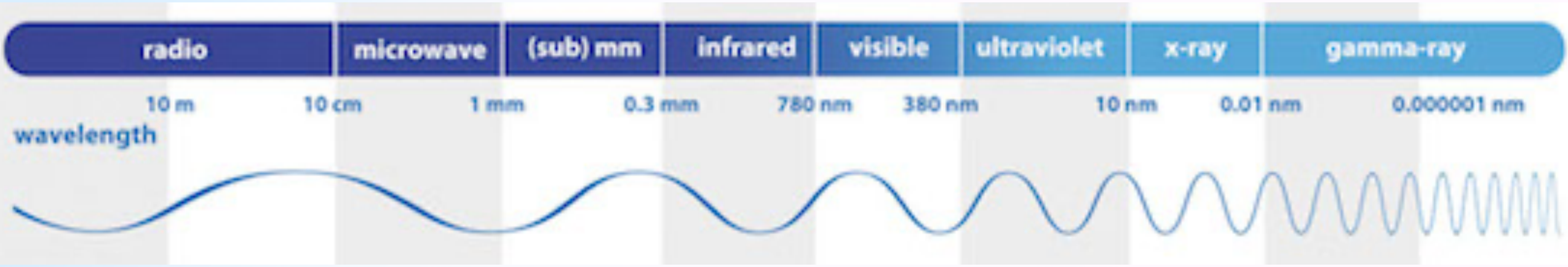


Li et al. 2023

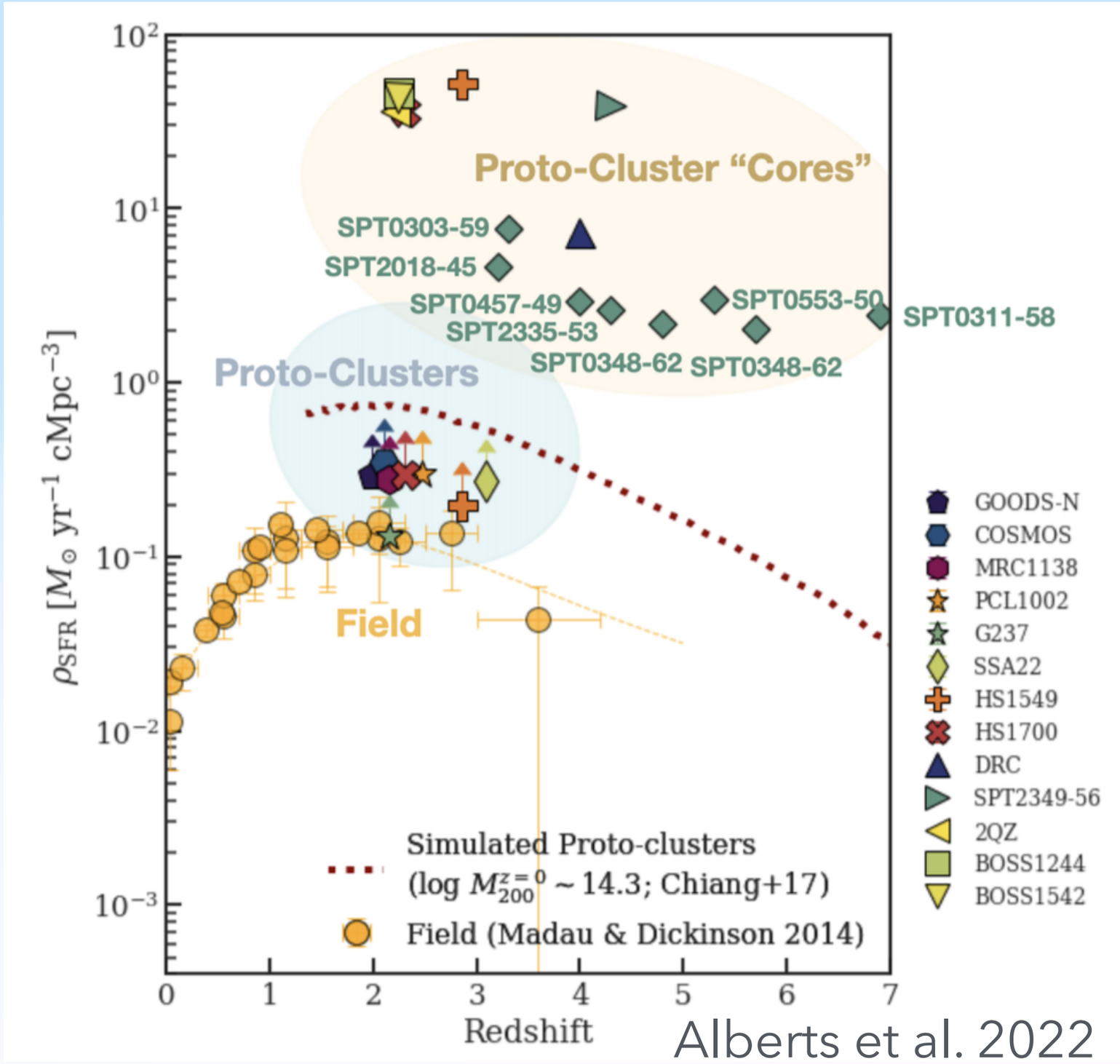


Arrigoni-Battaia et al. 2023

Travascio et al. 2025

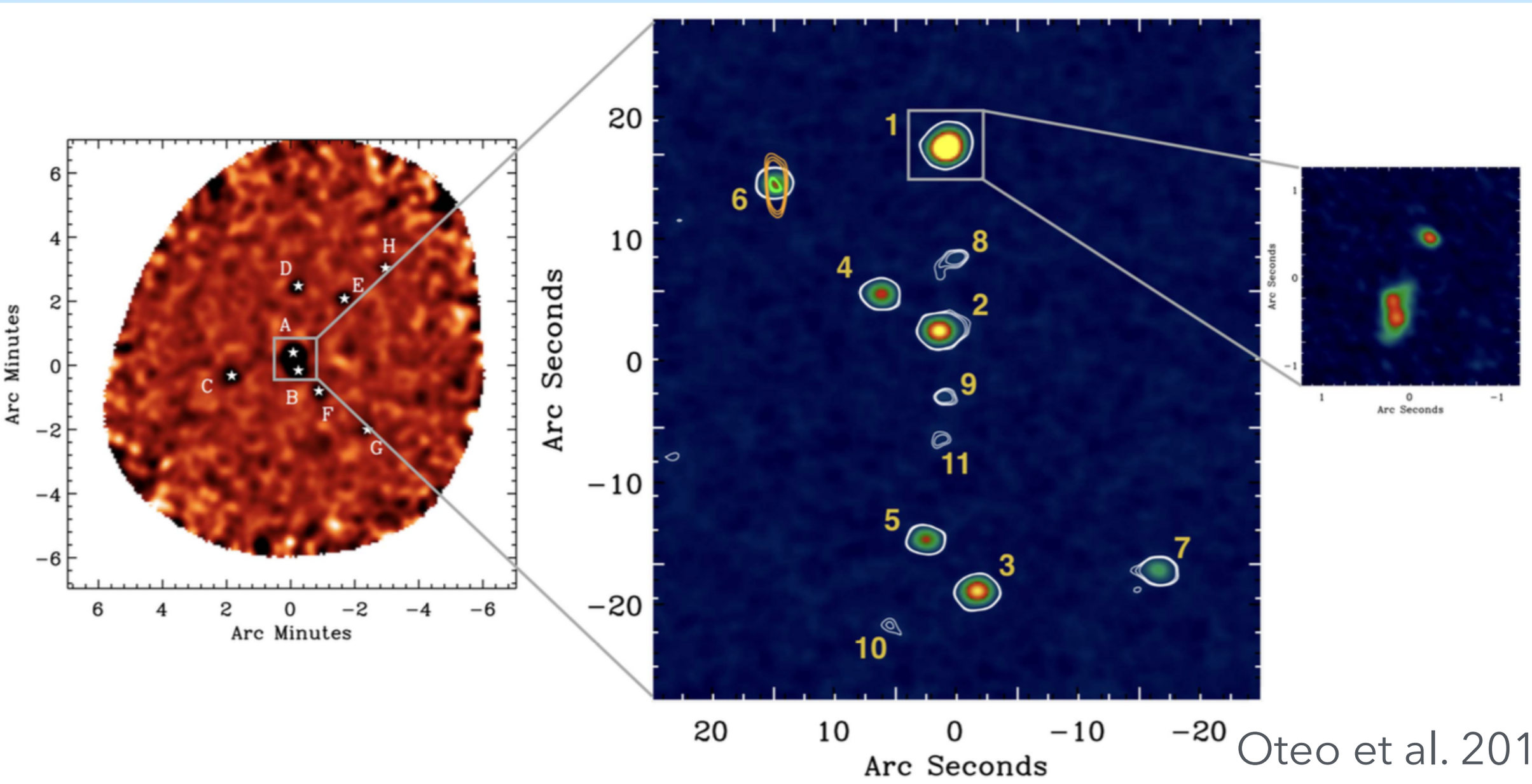


The majority of its ‘evolution’ is hidden behind a cosmic veil of dust



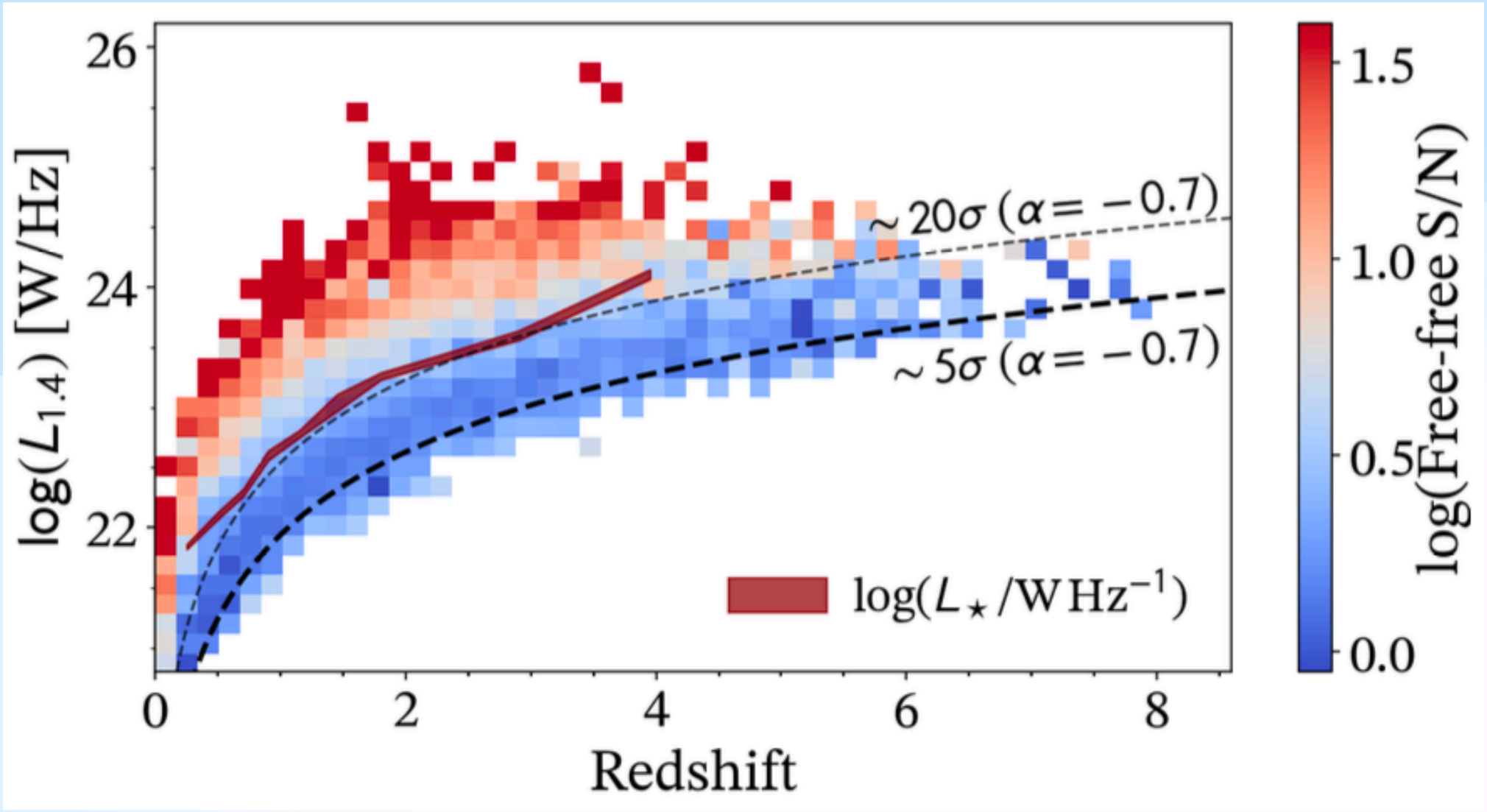
Alberts et al. 2022

Submm or radio observations are needed to characterize protoclusters



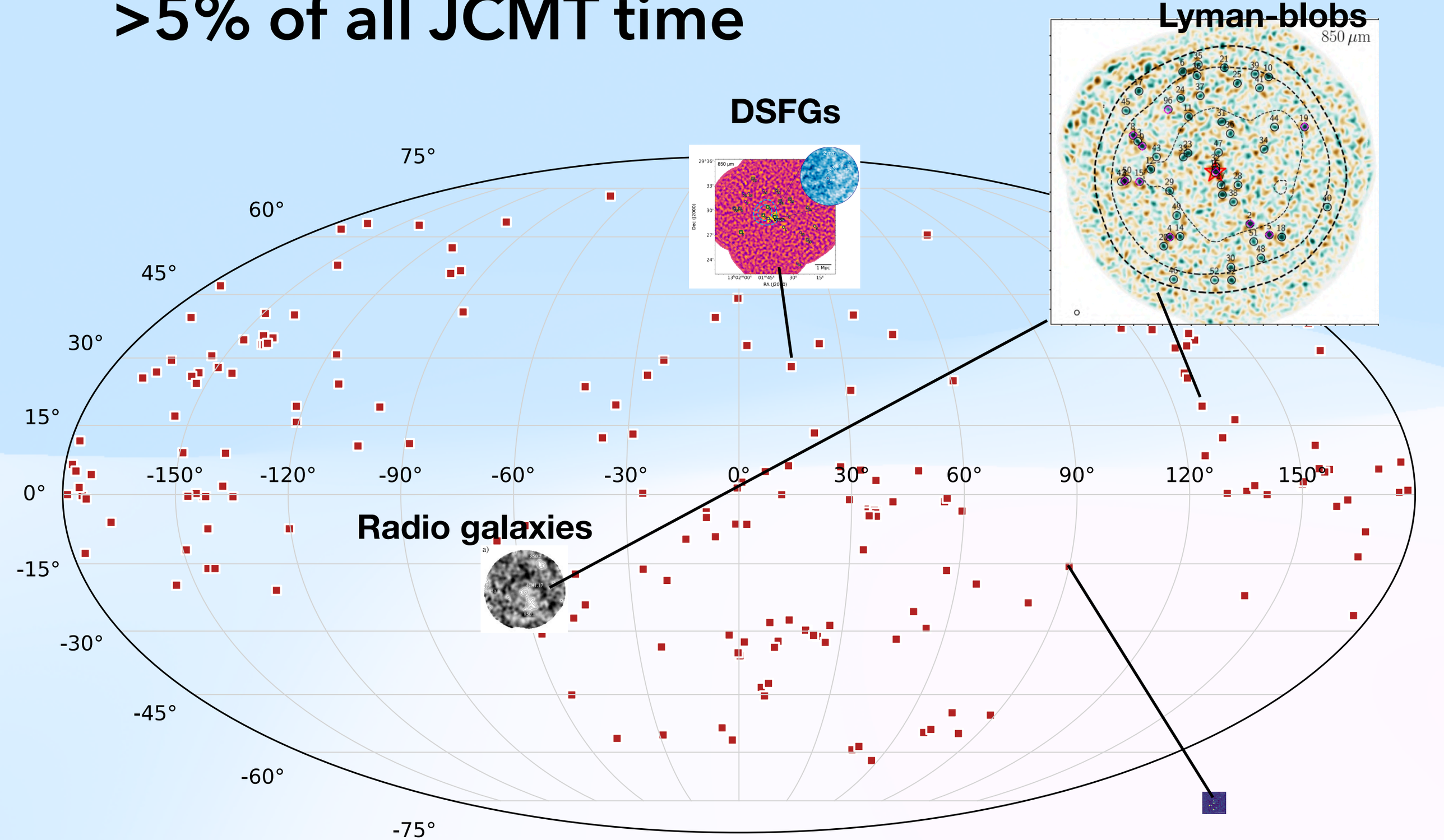
Oteo et al. 2018

SKA-mid will be able to measure very accurate star-formation rates

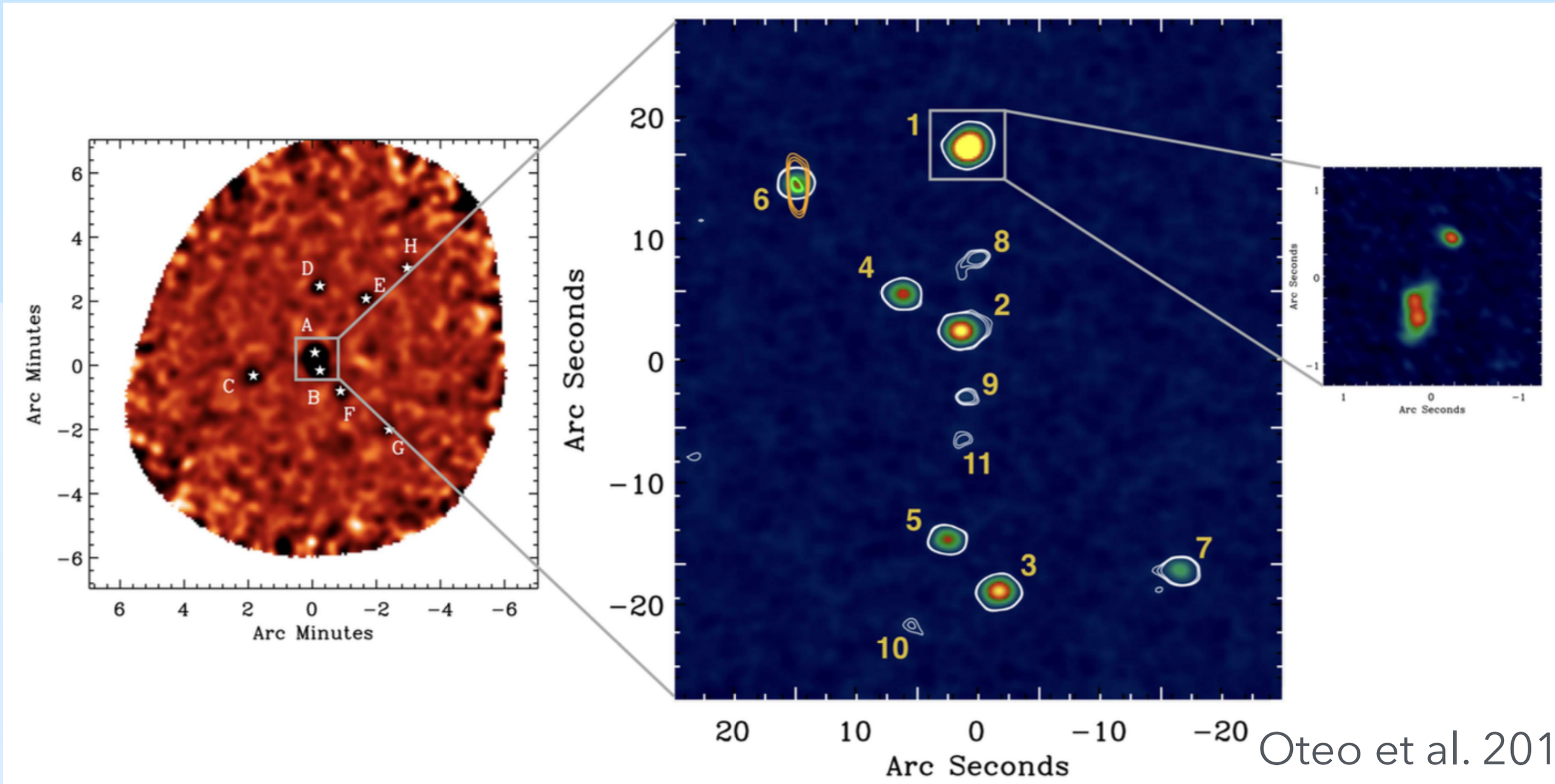


Algera et al. 2025

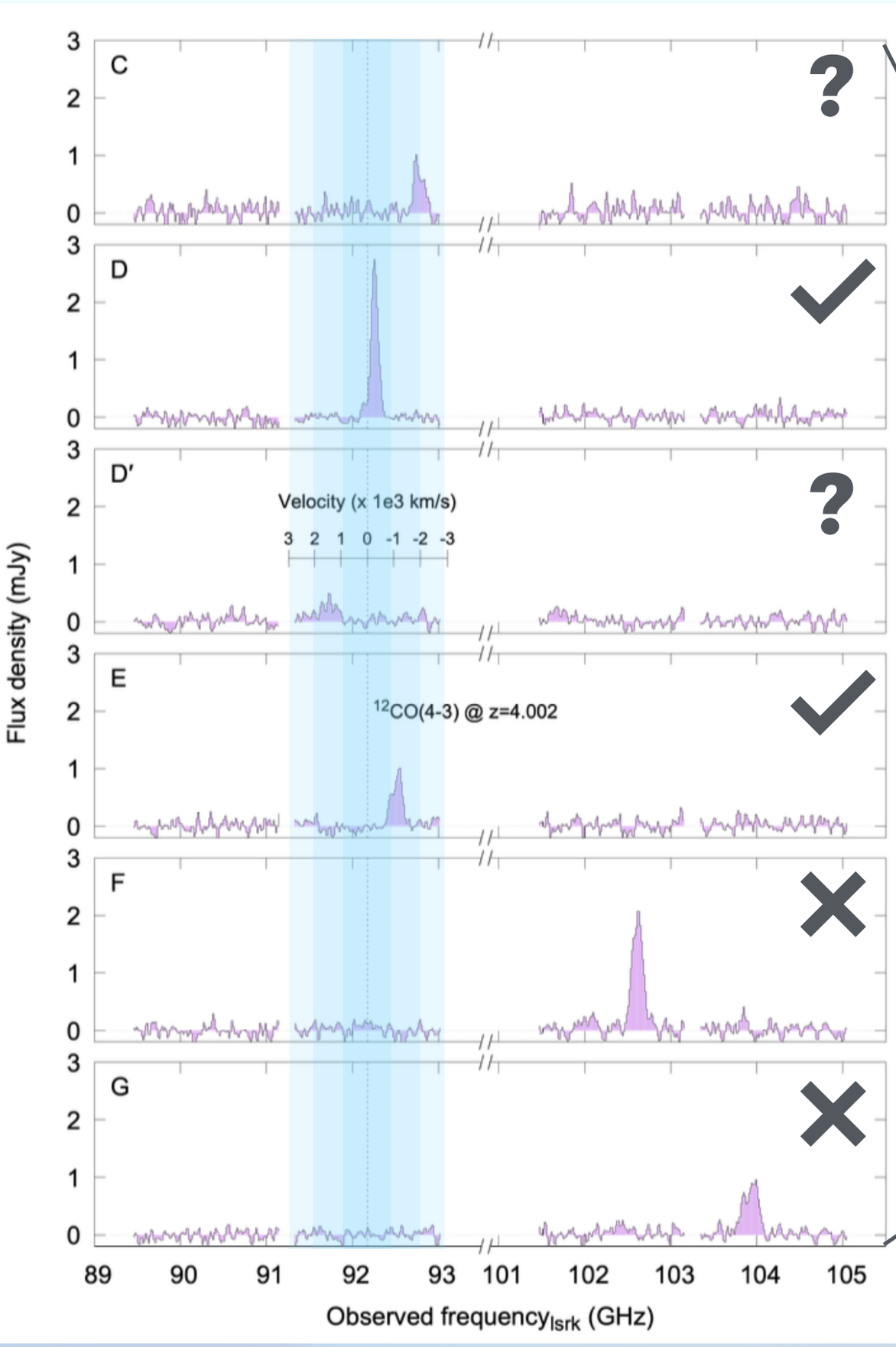
I collated all submm protocluster science, >5% of all JCMT time

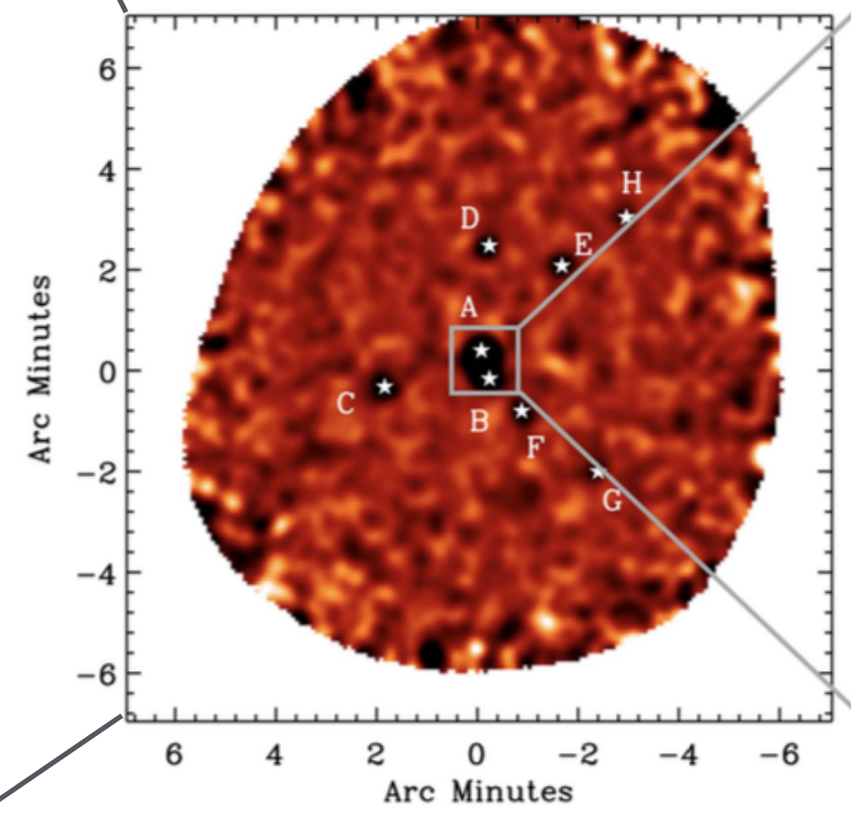


But not is all what it seems



Oteo et al. 2018





Ivison et al. 2020

Chen+23

Linking datasets to conclusions is hard…

Non-linear relationships

Overdensity signal depends on complex interplay of area × depth × wavelength × beam size

Missing data

Not all fields have all wavelengths
Ancillary data varies

Heterogeneous observations

SCUBA-2: 13" beam, 850 μm
LABOCA: 18" beam, 870 μm
AzTEC: 36" beam, 1.1 mm
Different sensitivities, areas, thresholds

Selection biases

DSFG-selected ≠ Radio-selected ≠ LBG-selected
Each method has different contamination

???

Does ~20-50% of star formation at z>4 occur in protoclusters?

What fraction of published protocluster candidates are false positives?

How do halo masses and spatial distributions test ΛCDM?

… but ML techniques can link them:

1 Handles heterogeneity

Survey parameters as input features (not nuisances)

Learns instrument-specific biases

2 Discovers non-linear patterns

No assumptions about feature distributions

Finds unexpected correlations

3 Graceful degradation

Works with incomplete data

Uncertainty quantification through ensemble

4 Probabilistic predictions

Not binary yes/no

Confidence intervals: 0-100% probability

ML feature engineering can homogeneize data interpretation and validation

200 fields with photometric data

14 fields with spectroscopic data

Voting ensemble reduces model-specific biases

Surface density features (δ_{bias})
Accounts for different survey depths and areas

Spatial clustering (2-point-cor)
Clumpy vs scattered distribution of sources

Flux distribution (compactness)
Mergers have 1 bright source; protoclusters have many

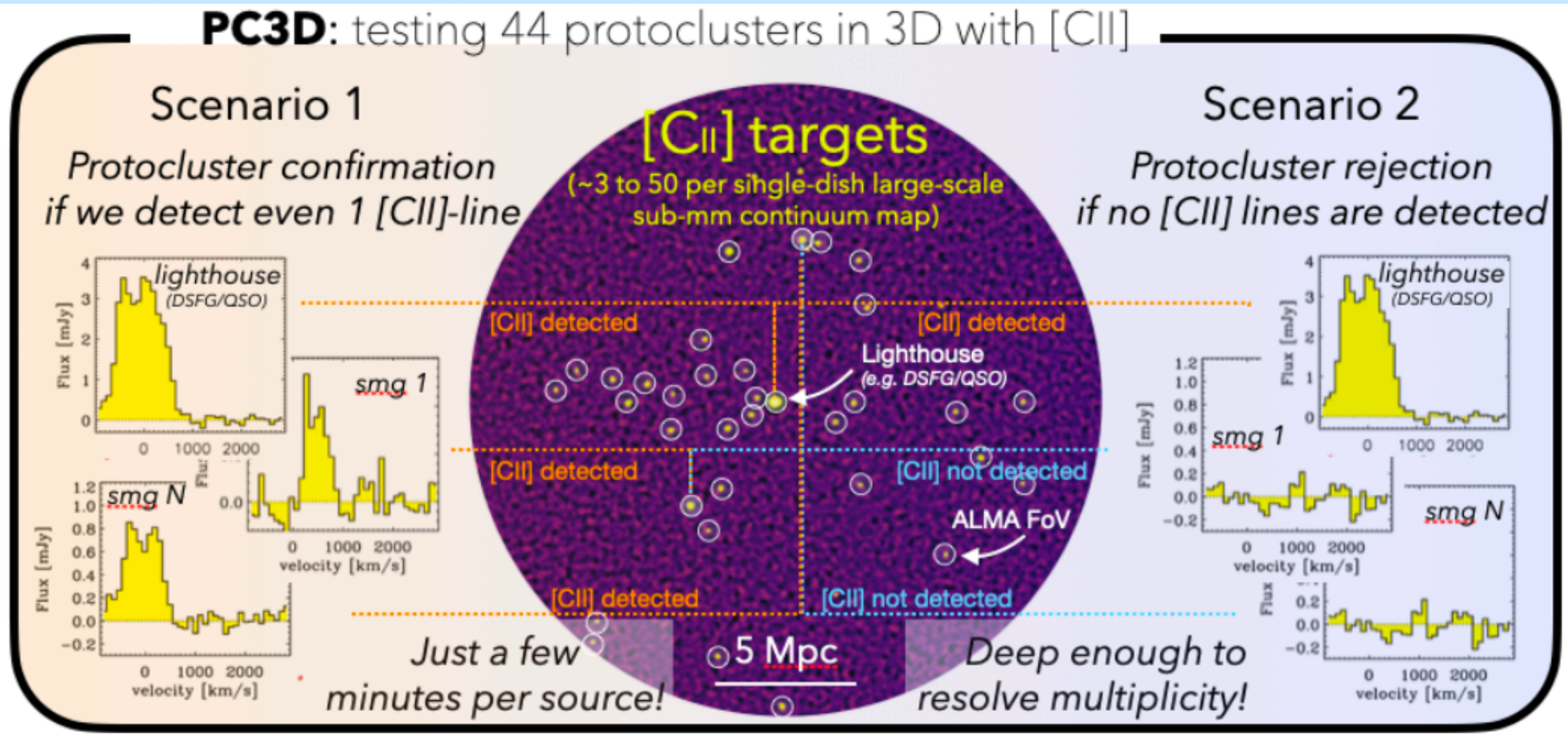
Random Forest
500 trees, small sample robust

Gradient Boosting
sequential error correction

Logistic Regression
interpretable baseline

Build framework for my 40hr ALMA project **PC3D**

Some systems are already in tension with Λ-CDM, but spectroscopic sampling is needed



ML feature engineering can homogeneize data interpretation and validation

200 fields with photometric data

14 fields with spectroscopic data

Surface density features (δ_{bias})
Accounts for different survey depths and areas

Spatial clustering (2-point-cor)
Clumpy vs scattered distribution of sources

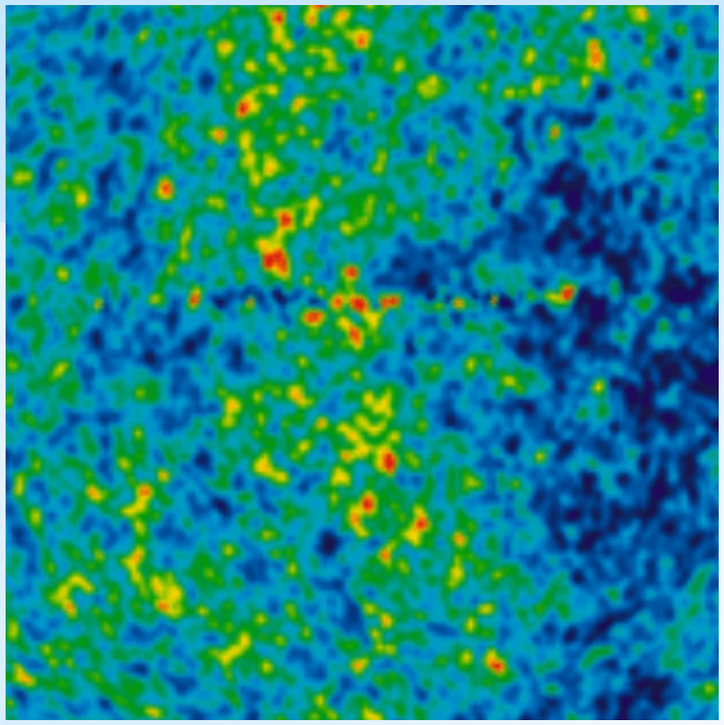
Flux distribution (compactness)
Mergers have 1 bright source; protoclusters have many

Voting ensemble reduces model-specific biases

Random Forest
500 trees, small sample robust
Gradient Boosting
sequential error correction
Logistic Regression
interpretable baseline

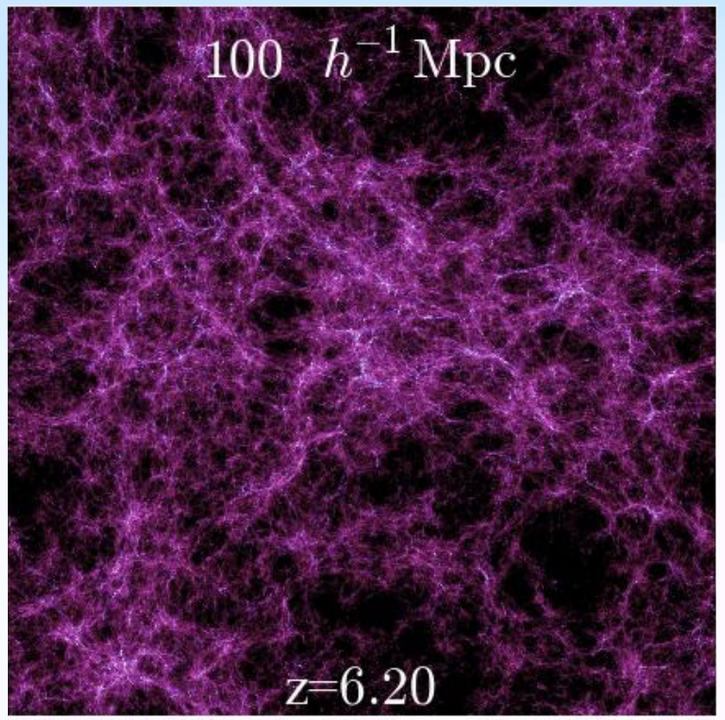
Build framework for my 40hr ALMA project **PC3D**

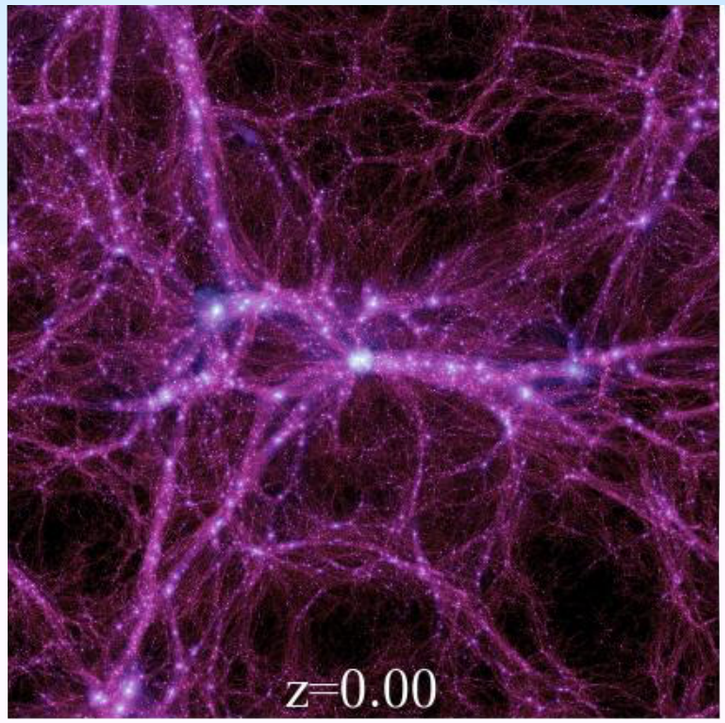
Galaxy clusters are solidified quantum-fluctuations left over from the big bang



z = 1100

Millennium simulation





Large datasets (incl. SKA) are required to observationally link the populations and evolution, **necessitating ML techniques**