Simulation-based Forward Modeling of Cross-Survey Cross-Correlations with Diffsky

Andrew Hearin
Gillian Beltz-Mohrmann

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

DiffStuff Team:


Andrew Hearin


Matt Becker


Alex Alarcon


Joseph Wick


Gillian Beltz-Mohrmann


Alan Pearl

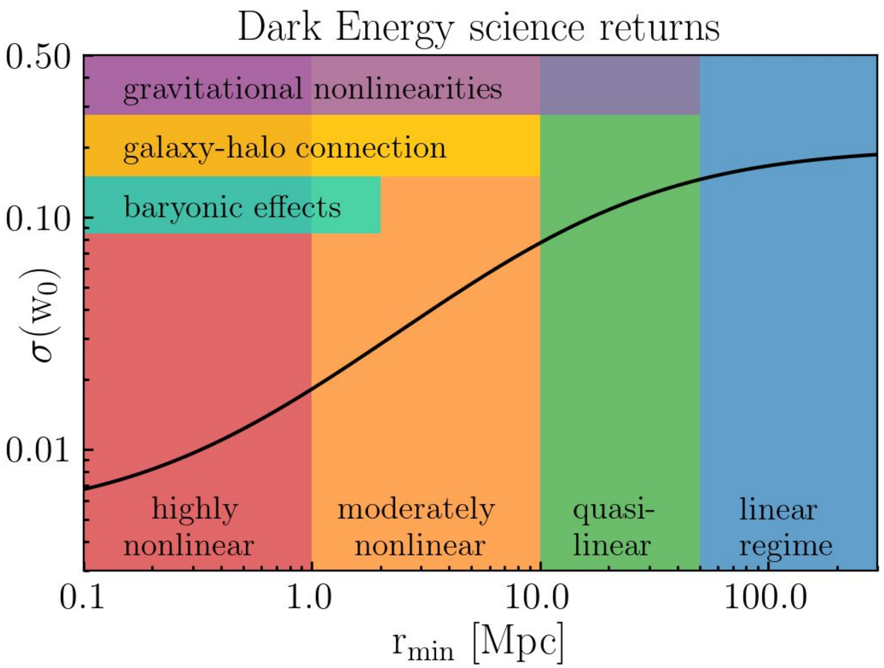

Georgios Zacharegkas

The next decade of cosmology

- The next generation of cosmological surveys will allow us to potentially explore the observational signatures of physics beyond the standard model
- Wealth of information contained in:
  - **Higher-order** clustering statistics
  - LSS measurements in the **nonlinear** regime
  - **Multi-redshift** constraints
  - **Cross-survey** analyses
- Many different approaches to extracting cosmology from "non-standard" observables
  - EFT extensions to higher-order sumstats
  - HOD-type models in highly nonlinear regime
  - Full-field emulators of hydro sims

Promise and challenges of the nonlinear regime

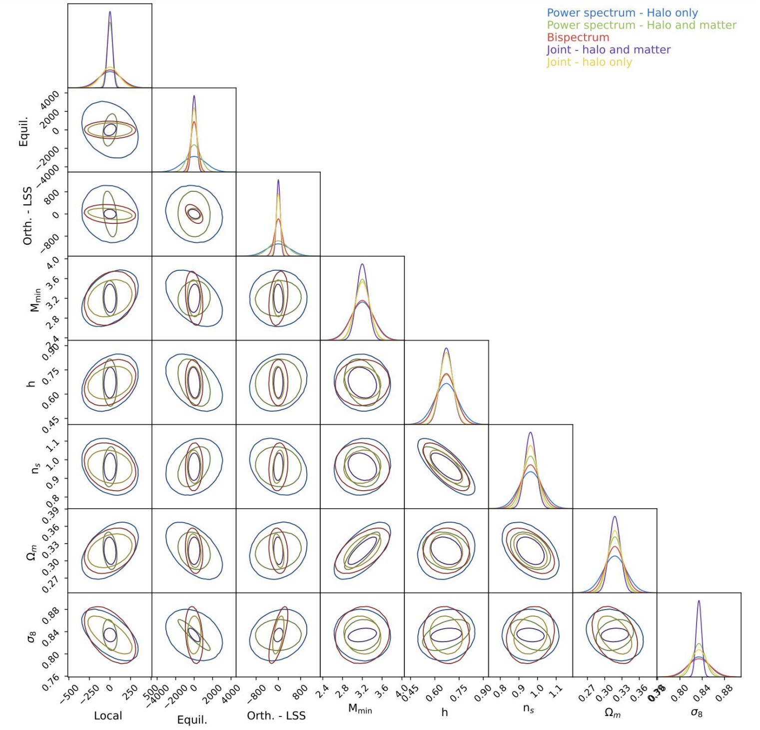
- Factor-of-many gains in constraining power on dark energy from nonlinear regime
- Nonlinear scales open up **entirely new probes of GR** inaccessible to quasi-linear regime, e.g., cluster RSD, splashback, etc.
- Modeling systematics dominate statistical uncertainty in the nonlinear regime
- Cross-x now widely adopted for systematic error control, overlapping surveys in 2020s enable joint cross-x analyses



(adapted from Zentner et al. 2013)

Promise and challenges of higher-order sumstats

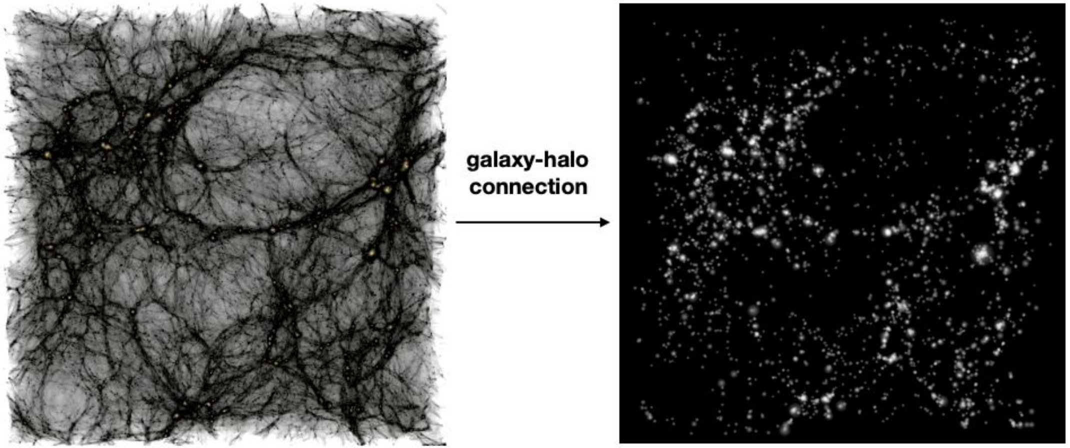
- Higher-order sumstats (bispectrum and beyond) can break degeneracies with cosmological and nuisance parameters
- For PNG, up to 4x increases in constraining power beyond P(k) made possible with bispectrum measurements in mildly nonlinear regime
- Modeling challenges are formidable!
  - Substantial expansion of param space required for even idealized theoretical predictions
  - Survey systematics (e.g., fiber collisions, window effects, etc) substantially more challenging
  - Computational demands can steeply increase



Coulton et al (2023)
(Quijote-PNG)

Can we use traditional models of the galaxy-halo connection to predict nonlinear & beyond-2pt clustering?

- E.g., HOD, SHAM
- How can these methods be extended for multi-z multi-tracer analyses?



Approaches to modeling the galaxy-halo connection

← physical models | empirical models →

Hydrodynamical Simulations	Semi-analytic Models	Empirical Forward Modeling	Subhalo Abundance Modeling	Halo Occupation Models
Simulate halos & gas; Star formation & feedback recipes	Evolution of density peaks plus recipes for gas cooling, star formation, feedback	Evolution of density peaks plus parameterized star formation rates	Density peaks (halos & subhalos) plus assumptions about galaxy—(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

Wechsler & Tinker 2018

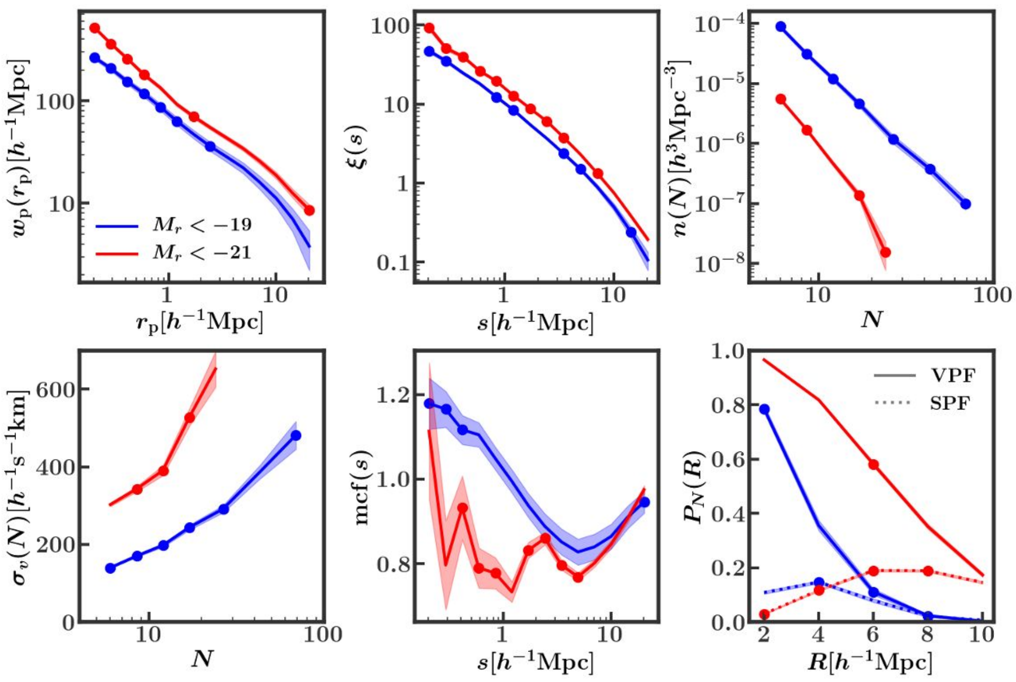
Small-scale clustering analyses with the standard HOD

Szewciw et al. (2022):

- SDSS: Mr < -19 & -21
- Standard HOD model, fixed cosmology
- Galaxy number density
- Projected correlation function
- Group multiplicity function

\+ Redshift-space Correlation Function
\+ Average group velocity dispersion function
\+ Mark Correlation Function
\+ Counts-in-cells statistics

* Selected combo of different scales of each statistic to optimize constraining power



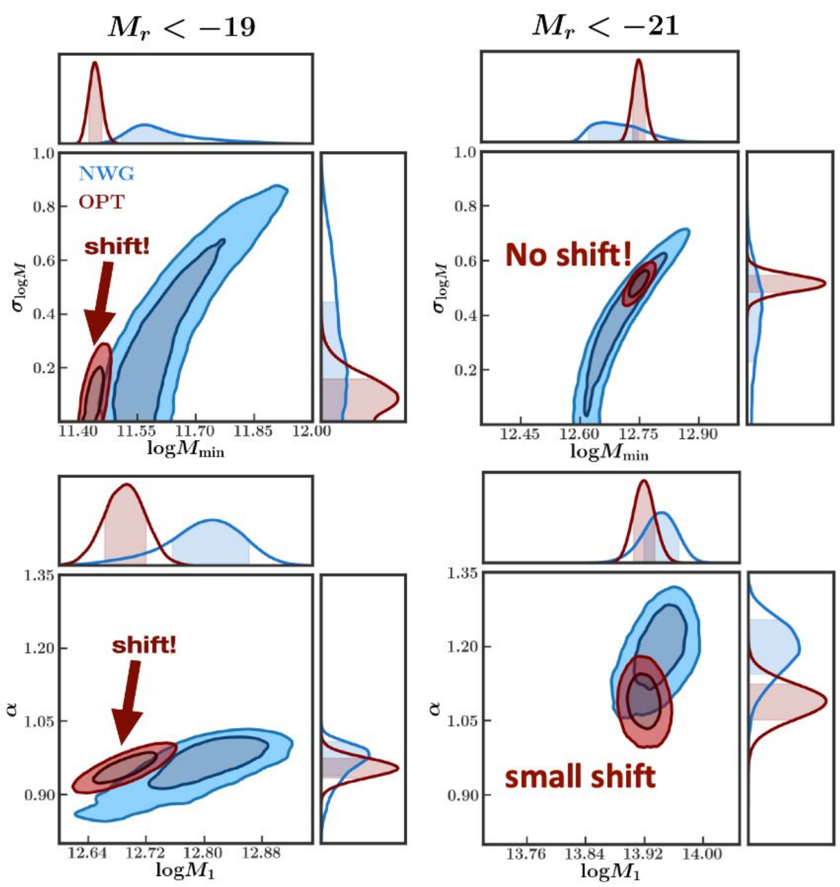
Szewciw et al. (2022)

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

Small-scale clustering analyses with the standard HOD

- Major shifts seen in best-fit parameter values compared to previous results
  - Shifts likely due to the inclusion of clustering statistics that are sensitive to non-standard effects (e.g. assembly bias)
- Major increase in constraining power
- >4σ tension for both samples



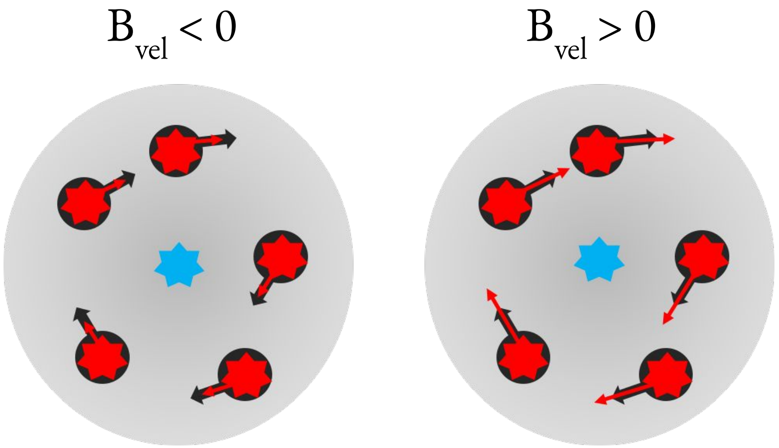
Szewciw et al. (2022)

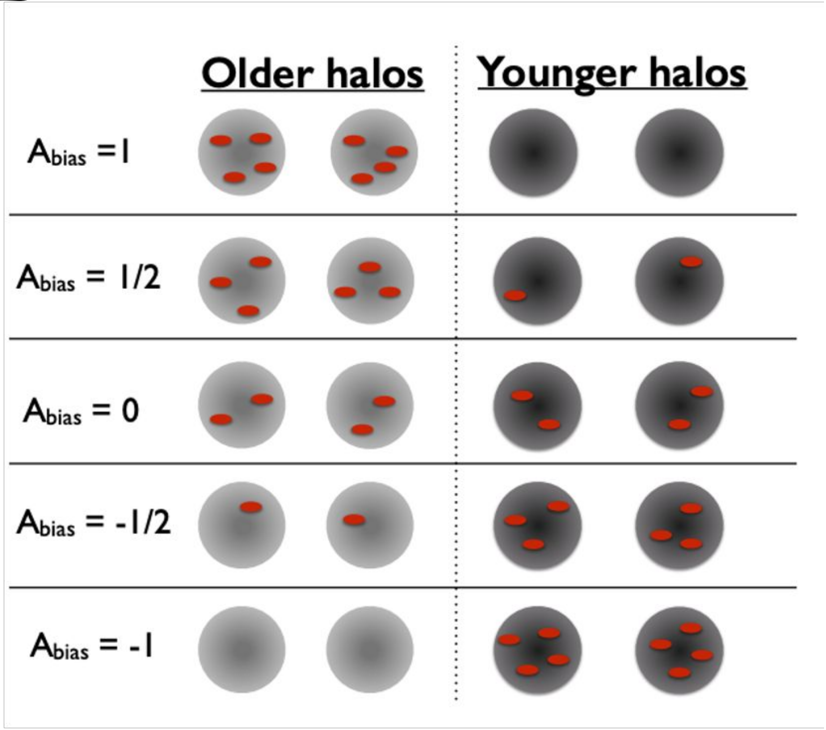
U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

Extensions to the standard HOD

- Comparisons with hydro simulations (e.g. Beltz-Mohrmann et al. 2020) indicate presence of assembly bias and velocity bias, particularly among low-luminosity galaxies
- We repeated our SDSS analysis with these extensions to the HOD





Hearin et al. (2016)

U.S. DEPARTMENT OF ENERGY
Argonne NATIONAL LABORATORY

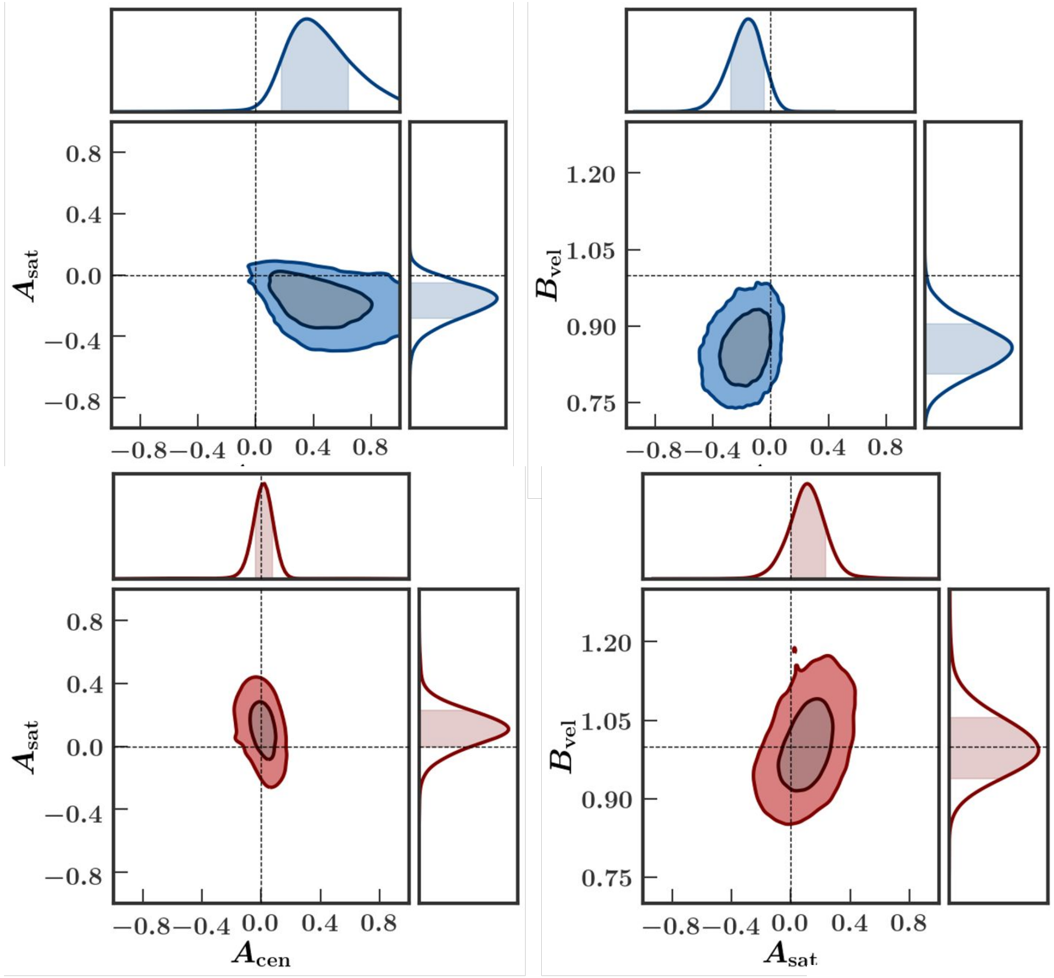
Results after HOD extensions

-19 sample:

- Tight constraints on HOD parameters
- Best model: environment dependent assembly bias + satellite velocity bias
- Significant detection of assembly bias and velocity bias
- No remaining tension with SDSS

-21 sample:

- No detection of assembly or velocity bias
- No relief of tension with SDSS (still 4.5σ)

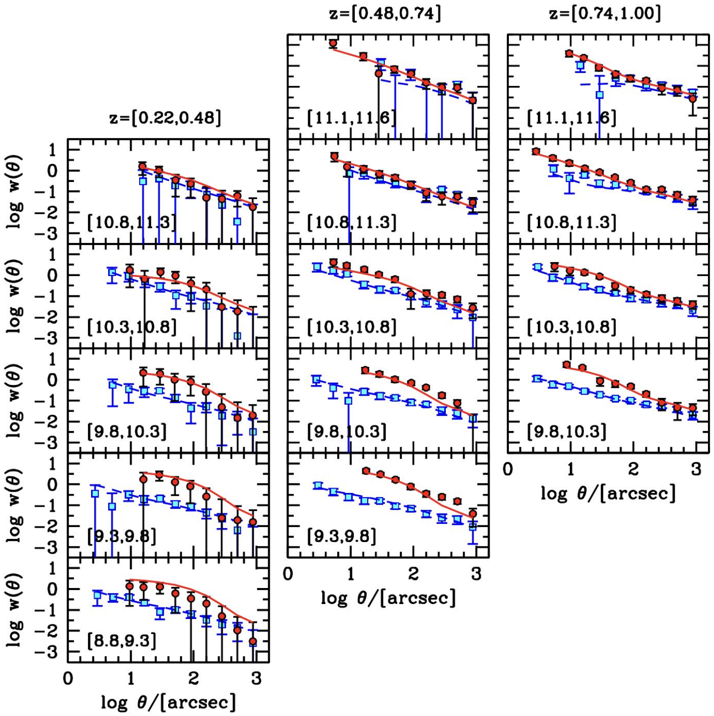


Beltz-Mohrmann et al. (2022)

U.S. DEPARTMENT OF ENERGY
Argonne NATIONAL LABORATORY

Can we continue to extend the HOD?

- Including additional freedom in the HOD allowed us to accurately model nonlinear clustering for one SDSS sample, but not another
- There is additional freedom we could have included (e.g. anisotropic satellite distributions) but we limited ourselves to freedom that was well-motivated based on hydro comparisons
- Each new degree of freedom adds to our parameter space
- **If we wanted to fit multiple galaxy samples at multiple redshifts *simultaneously* we would have a runaway parameter problem**
- The HOD is not the only model with this issue (e.g. EFT)
- Need a new model with physically motivated flexibility that is designed for multi-tracer, multi-z analyses



Tinker et al (2013)

Technological advancements in the last 20 years

- HOD was born in 2002, a time when:
    - Cosmological simulations could only reliably resolve host halos at a single redshift (no substructure, no merger trees)
    - SDSS and 2dF had freshly supplied single-tracer, single-redshift (z=0) galaxy samples
- *HOD limitations reflect the era in which it was born*

What has changed in the interim?

- N-body sims have improved dramatically in the last 22 years
    - Halo substructure (aka subhalos) and merger trees have become industry-standard tools
- GPUs and AI/ML techniques have transformed the computing landscape

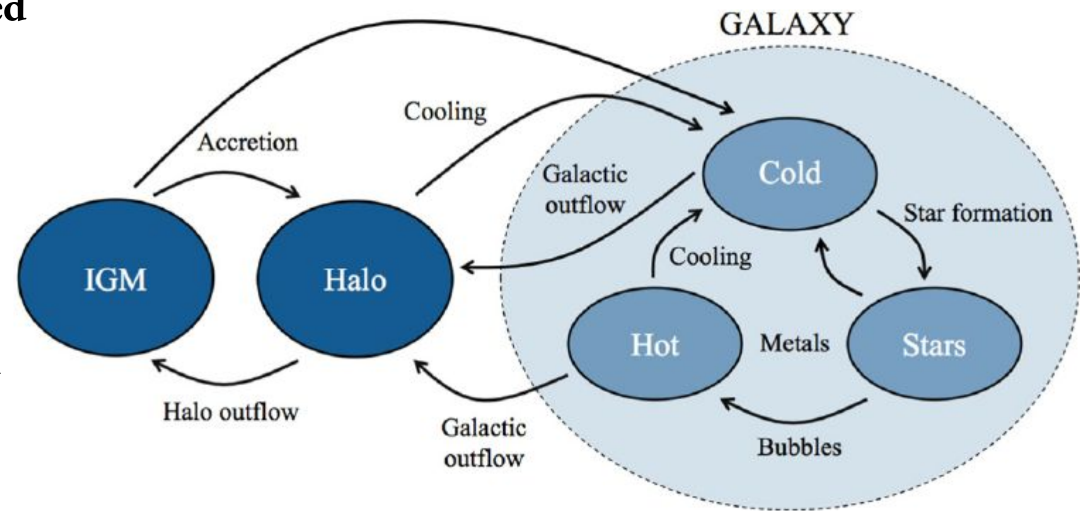
Let's create a new, physics-based model that leverages these advances!

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

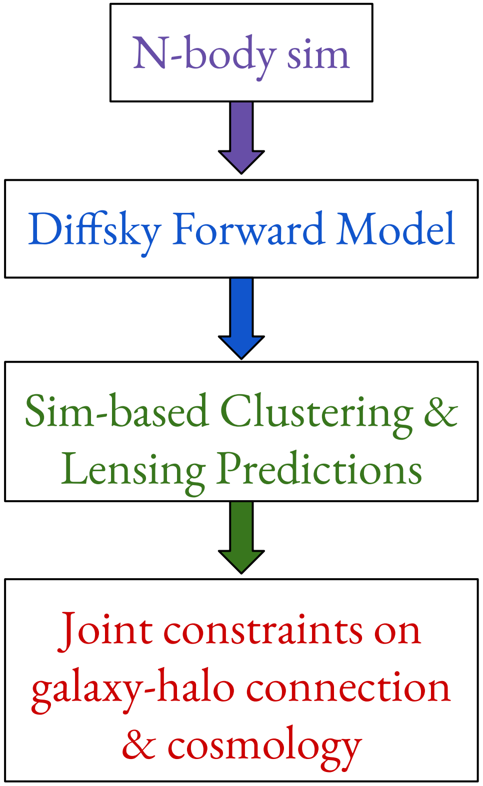
Traditional SAM approach to physical model of multi-λ predictions

- Root simulation data: high-res N-body sim with merger trees
- Physics assumptions formulated as **coupled ODE system** regulating exchange of mass/energy/momentum between collection of reservoirs
- **Fully deterministic**: merger tree + SAM ⇒ point-estimator for galaxy properties
- Predict LSS ⇒ solve ODE system for each individual simulated merger tree
- Cross-survey multi-λ predictions emerge naturally from simulated SEDs



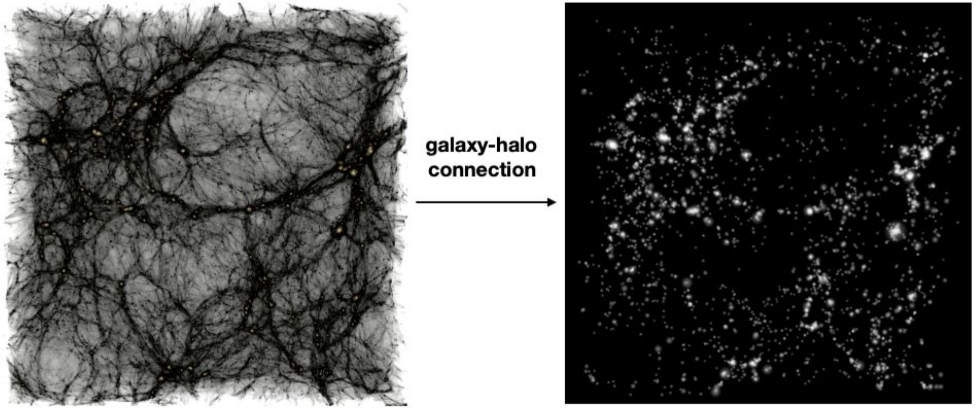
Diffsky: A New Forward Model of the Galaxy-Halo Connection

- **Goal**: develop new generation of galaxy–halo models
  - Suitable for multi-z, multi-λ predictions
  - Based on simple physical assumptions
- **Approach**:
  - Ground-up reformulation of predictions to be fully probabilistic & differentiable
  - Leverage GPU performance of modern autodiff
- **Long-term goals**:
  - full-scale, multi-z, multi-tracer, cross-survey cosmological analyses (including cross-x)
  - Informative priors for EFT analyses
  - Mocks for all!



What makes Diffsky different?

- Empirical forward model of SEDs
- Flexibility and multi-λ predictivity of a SAM (without directly solving ODEs)
- Orders-of-magnitude faster due to AI/ML techniques on GPUs
- Model parameters have direct, simple physical interpretation
- Methodically validate using hydro sims & SAMs
    - Only introducing freedom warranted by the data



Approaches to modeling the galaxy-halo connection

← physical models | empirical models →

Hydrodynamical Simulations	Semi-analytic Models	Empirical Forward Modeling	Subhalo Abundance Modeling	Halo Occupation Models
Simulate halos & gas; Star formation & feedback recipes	Evolution of density peaks plus recipes for gas cooling, star formation, feedback	Evolution of density peaks plus parameterized star formation rates	Density peaks (halos & subhalos) plus assumptions about galaxy—(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

Wechsler & Tinker 2018

Multi-λ predictions

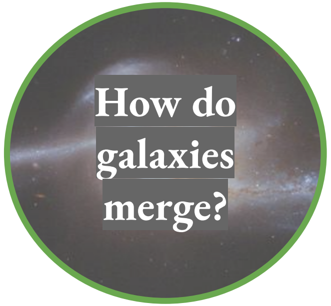
Differentiable **sky** predictions

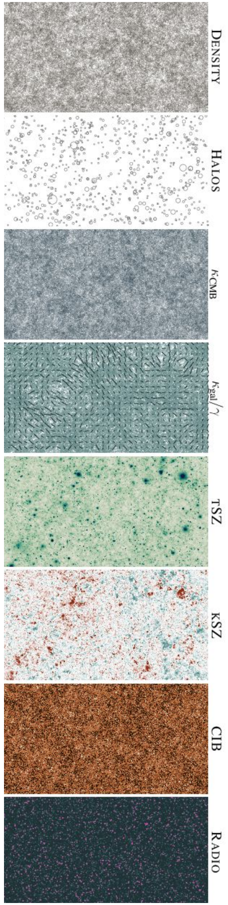
Diffmah
(Hearin et al. 2021)

Diffstar
(Alarcon et al. 2023)

DSPS
(Hearin et al. 2023)

Diffmerge
(Beltz-Mohrmann et al. in prep.)



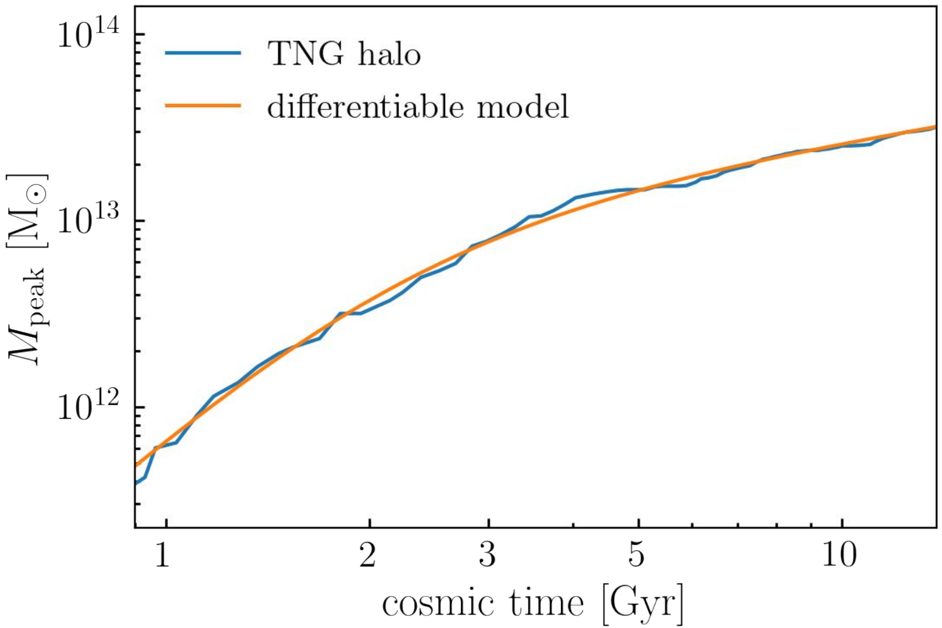


*All model parameters have physical interpretations. We seek the minimum interpretable parametric flexibility required to accurately capture the data.

Image credit: Millennium XXL simulation, NASA, ESA, Yuuki Omori/Agora simulation
Slide credit: Alex Alarcon

Differentiable Halo Mass Evolution

- Root simdata = high-res N-body with merger tree
- **Diffmah** approximates $M_{halo}(t)$ with Θ_{MAH}
- *Preprocessing step:* replace main progenitor of every simulated merger tree with a differentiable approximation



Diffmah: Hearin et al. 2021

Differentiable Approach to Galaxy Evolution

- Root simdata: analytic Θ_{halo} for every halo

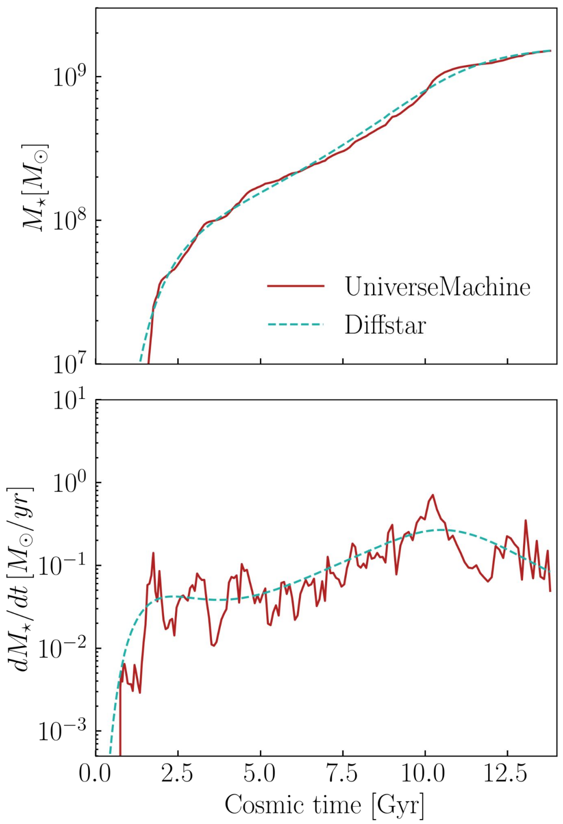
Key idea

Seek parametric family of solutions to galaxy formation ODEs as function of Θ_{halo}

Application to SFH

- Diffstar: SFH approximation based parametric model of SFR efficiency
- More info in Alarcon+22
- Upshot: SFH(t) parametrization Θ_{SFH} based on physical ingredients:
  - main sequence efficiency
  - gas consumption timescale
  - quenching (and possible rejuvenation)

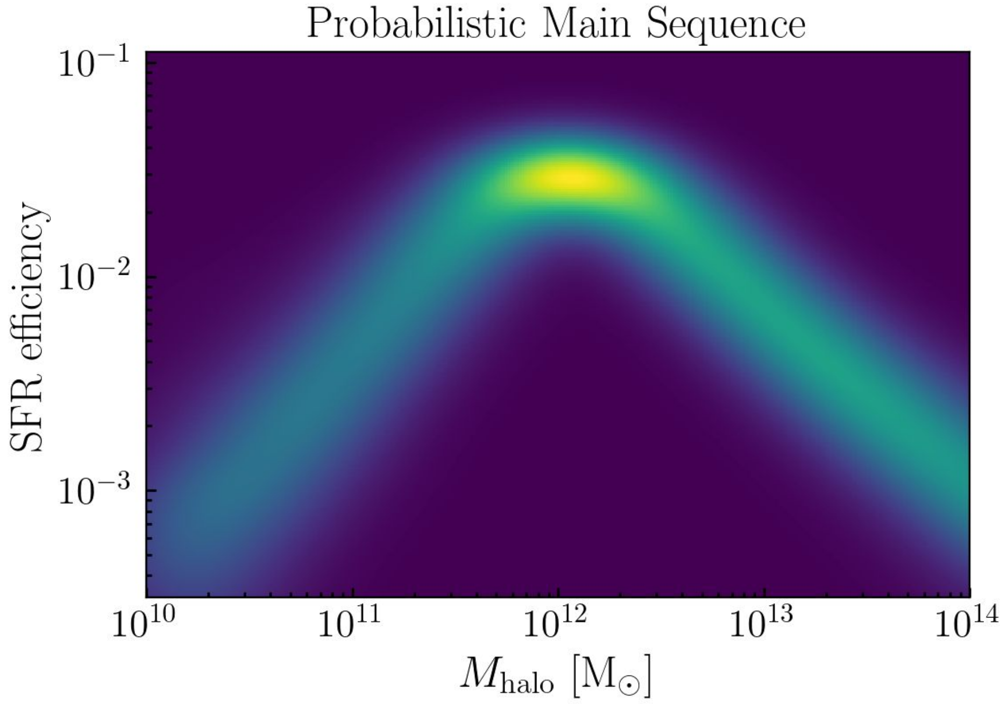
Diffstar approximation to SFH(t)



U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

Fully Probabilistic Formulation



- Traditional SAMs make a *deterministic* prediction for the galaxy evolving in a halo
- But an N-body halo does not contain sufficient info for such a prediction!
  - Quite different galaxies could live in a DM halo with same assembly history
  - Predictions should have variance from physics missing in the underlying sim

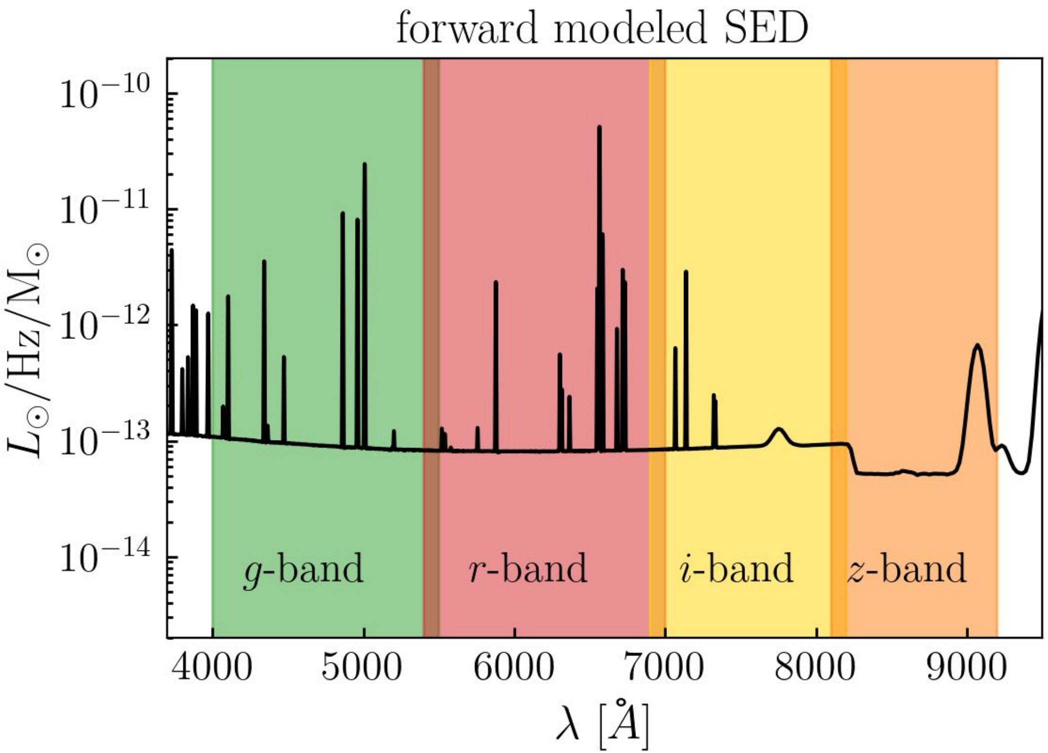
Key idea

Parameterize a *probabilistic* galaxy that lives in each simulated dark matter halo

- Technical detail: requires propagation of parametrized PDF of individual galaxies through to population-level sumstats

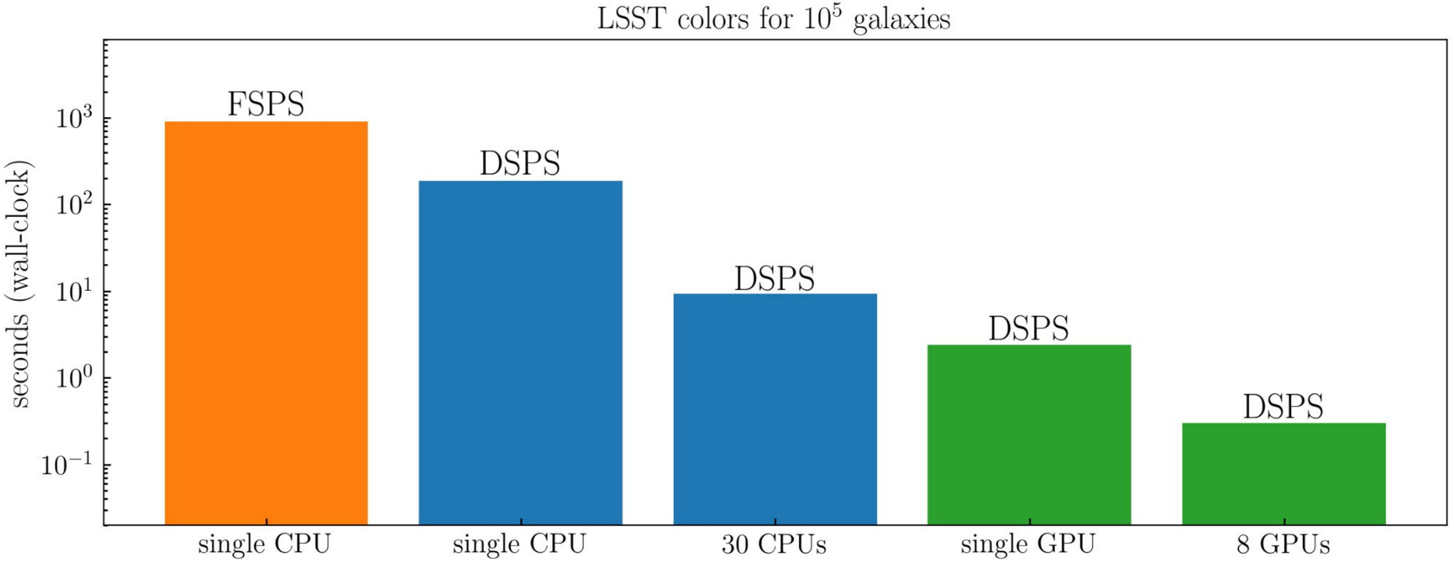
Differentiable Approach to SEDs/Photometry

- Use Stellar Population Synthesis to predict SED from SFH
- SPS models include ingredients Θ_{SPS} for dust, bursty star formation, metallicity, etc.
    - *Diffsky includes new probabilistic ingredients for each of these*



Differentiable Approach to SEDs/Photometry

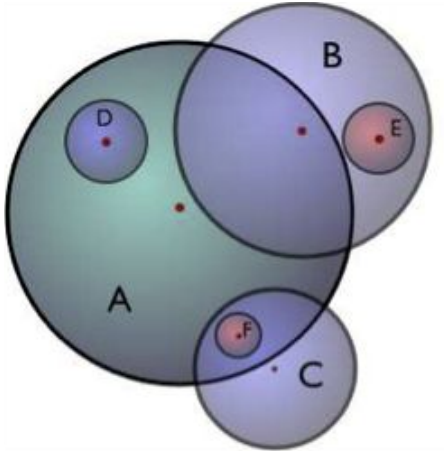
- Use Stellar Population Synthesis to predict SED from star formation history
- SPS models include ingredients Θ_{SPS} for dust, bursty star formation, metallicity, etc.
  - Diffsky includes new probabilistic ingredients for each of these
- **Enormous performance gains from DSPS: a JAX-based implementation of SPS**



U.S. DEPARTMENT OF ENERGY
Argonne NATIONAL LABORATORY

Differentiable Merging

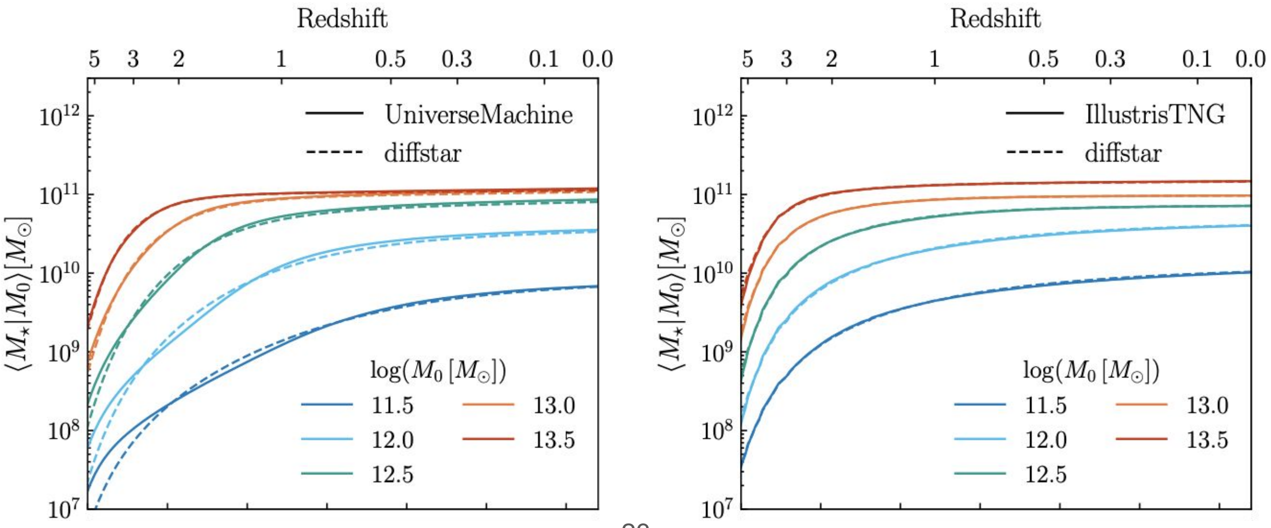
- Probabilistic model for when a satellite galaxy deposits some/all of its stellar content onto the central galaxy
- Depends on:
    - t_{infall}
    - $M_{\text{host, infall}}$
    - $M_{\text{sub, infall}}$
- Includes two rounds of merging to account for satellite preprocessing prior to final infall
- Validated with a version of UniverseMachine in which merging was turned off and then reintroduced with our model (i.e. sats retain their stellar mass until z=0)
- Designed for future use on Argonne sims with *cores* (50 most bound subhalo particles which are tracked to z=0 to account for artificial disruption)



Fitting the model - a programmatic approach

Key principle: Seek the minimum interpretable parametric flexibility required to accurately capture the data

1. Build & validate each piece of the model using existing SAMs & hydro sims (e.g. UniverseMachine, TNG)
2. Fit to increasingly complex target data to validate and stress-test flexibility of the model
3. Incorporate each new ingredient into unified forward modeling pipeline for observational predictions



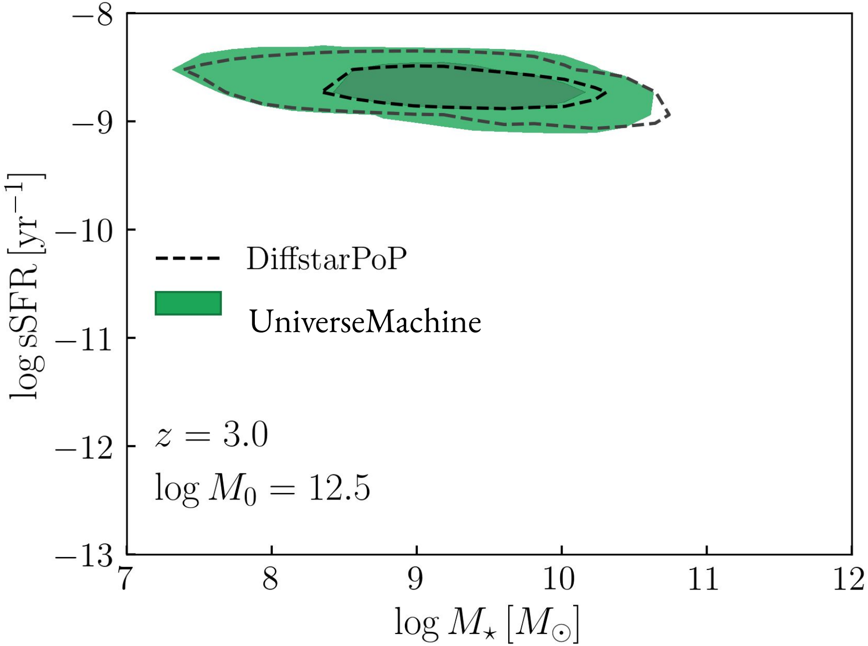
U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

DiffstarPop: Mstar vs sSFR – Redshift evolution

Simultaneously fit the 2D Mstar-sSFR distributions as a function of redshift and present-day halo mass M0.

Plot on the right shows the galaxy population evolution as a function of redshift at fixed log M0=12.5.



Gif credit: Alex Alarcon

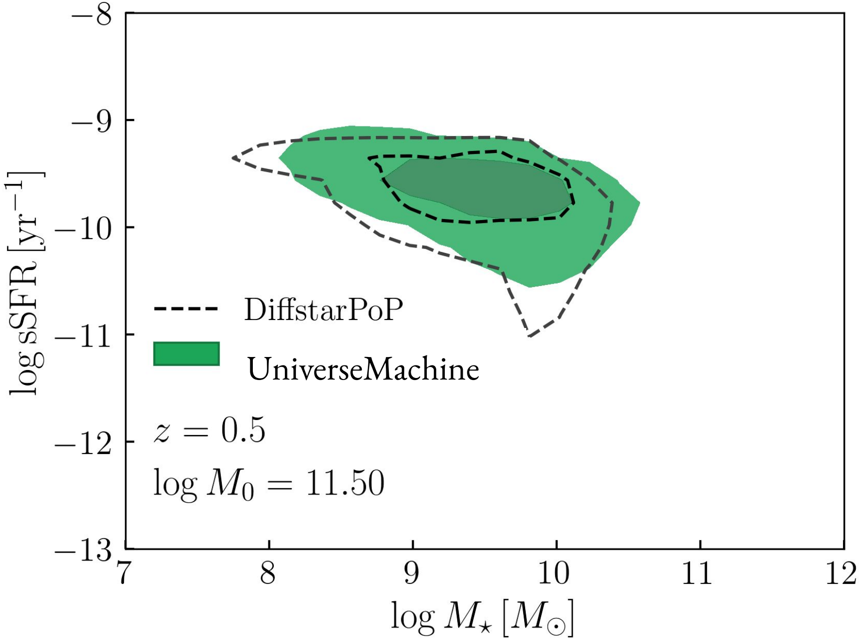
U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

DiffstarPop: Mstar vs sSFR – M0 evolution

Simultaneously fit the 2D Mstar-sSFR distributions as a function of redshift and present-day halo mass M0.

Plot on the right shows the galaxy population evolution as a function of M0 at fixed *z*=0.5.

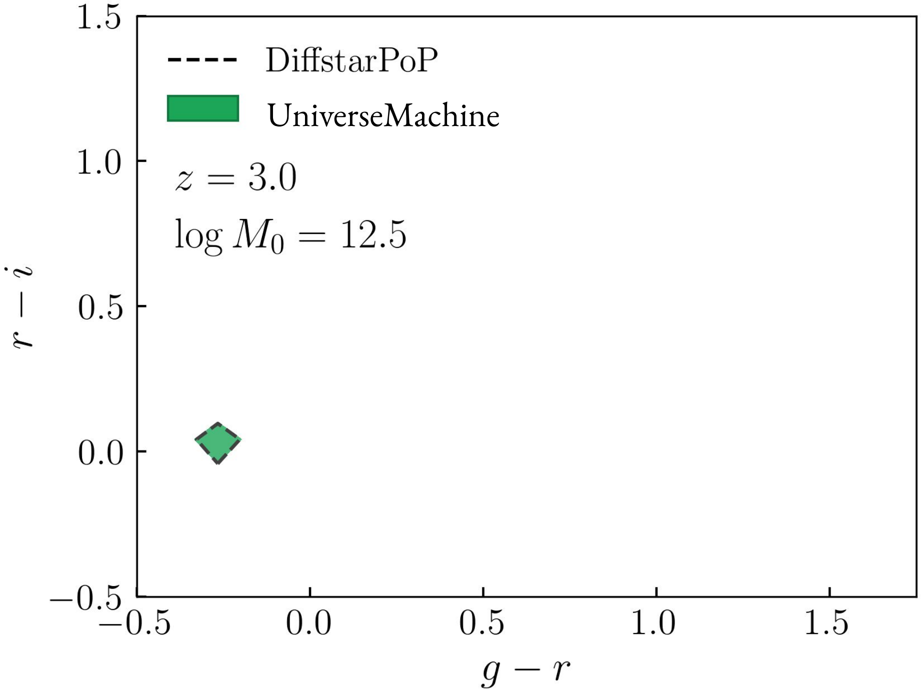


Gif credit: Alex Alarcon

DiffstarPop + DSPS colors: g-r vs r-i – Redshift evolution

Simultaneously fit the 2D Mstar-sSFR distributions as a function of redshift and present-day halo mass M0.

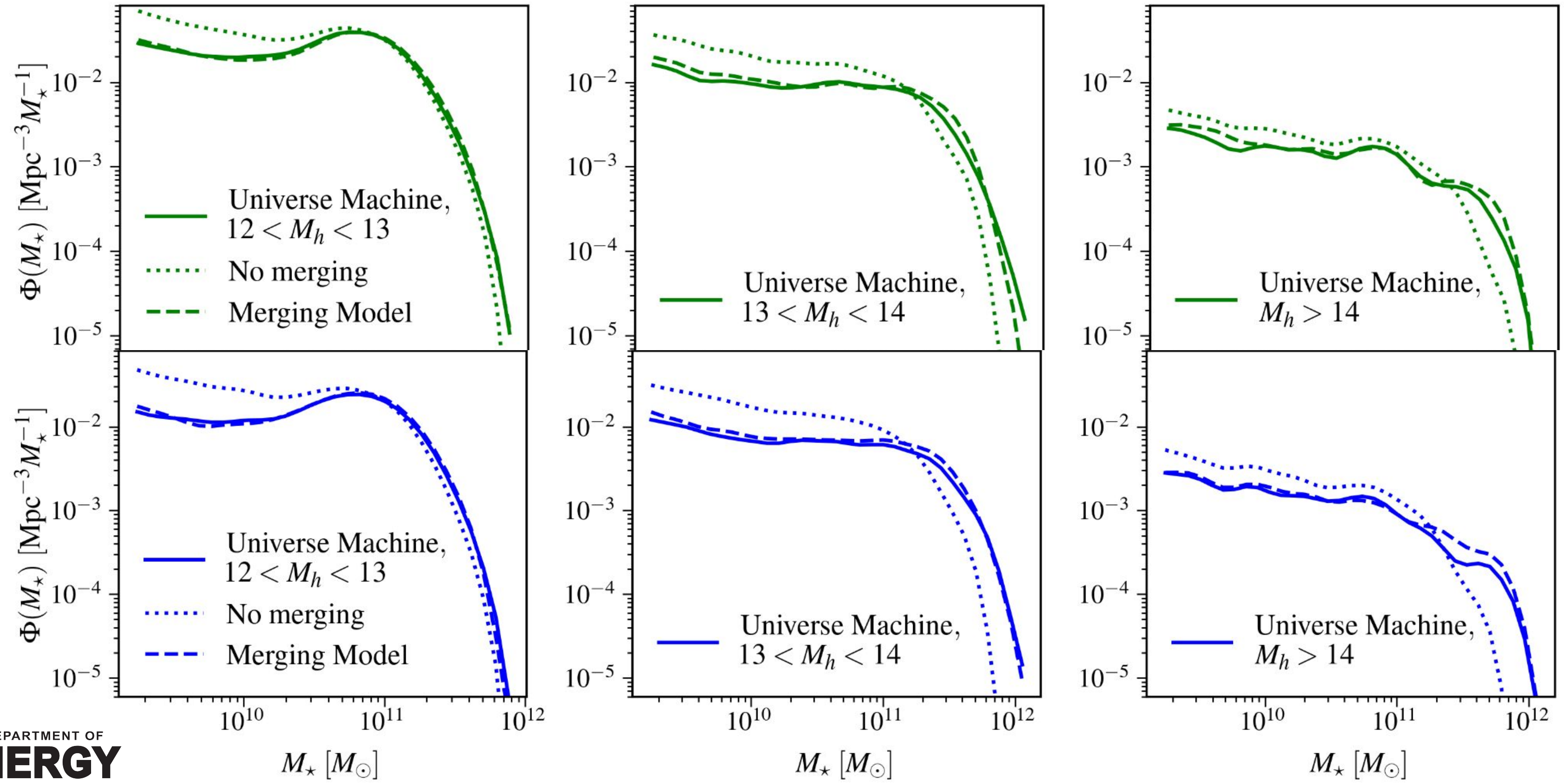
Color-color predictions using preliminary Diffburst + Diffdust models calibrated to COSMOS griz data by Gillian.



Gif credit: Alex Alarcon

Differentiable Merging

Simultaneously reproduces multi-z Conditional Stellar Mass function!



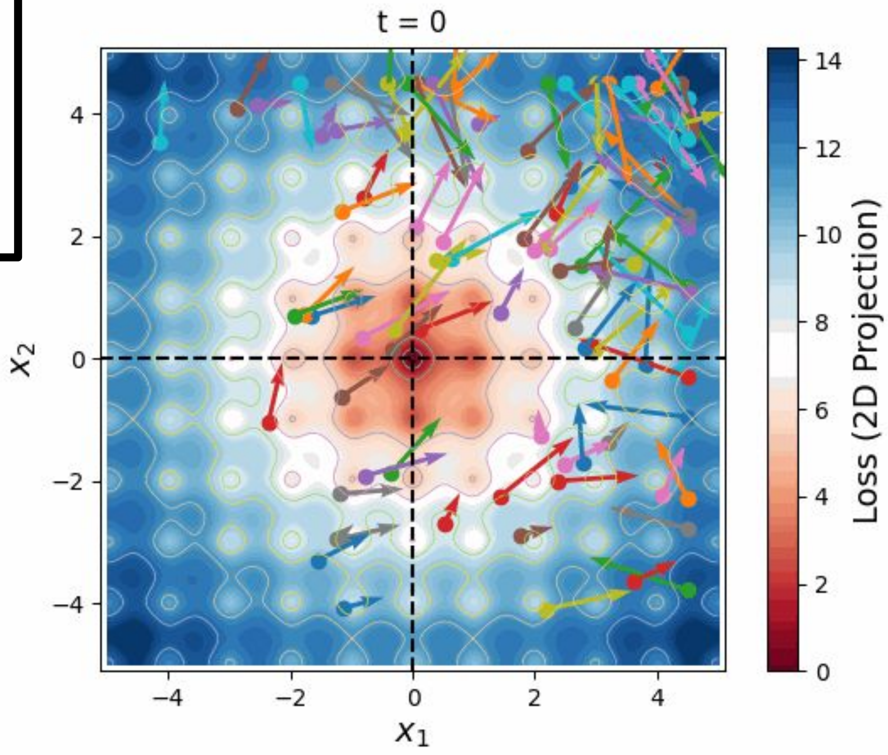
U.S. DEPARTMENT OF ENERGY

High-dimensional Optimization Techniques

Key idea

Use same techniques used in AI/ML optimization, but apply to differentiable physical models

- Particle Swarm Optimization to scan param space in parallel for global minima
- Stochastic mini-batch gradient descent to optimize predictions for multi-dim summary statistics
- Kernel density estimation for fine-grained PDF fitting



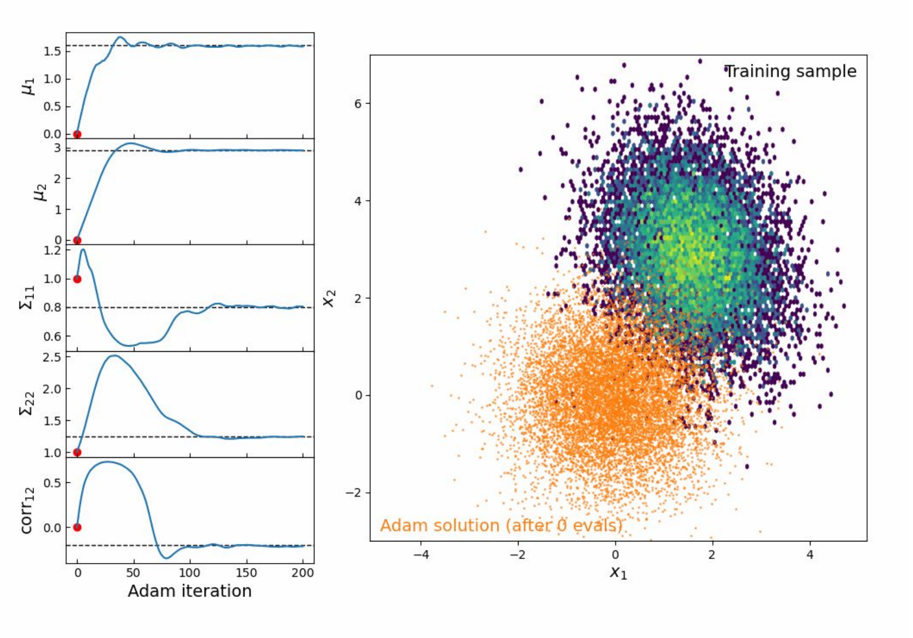
Gif credit: Alan Pearl

High-dimensional Optimization Techniques

Key idea

Use same techniques used in AI/ML optimization, but apply to differentiable physical models

- Particle Swarm Optimization to scan param space in parallel for global minima
- Stochastic mini-batch gradient descent to optimize predictions for multi-dim summary statistics
- Kernel density estimation for fine-grained PDF fitting



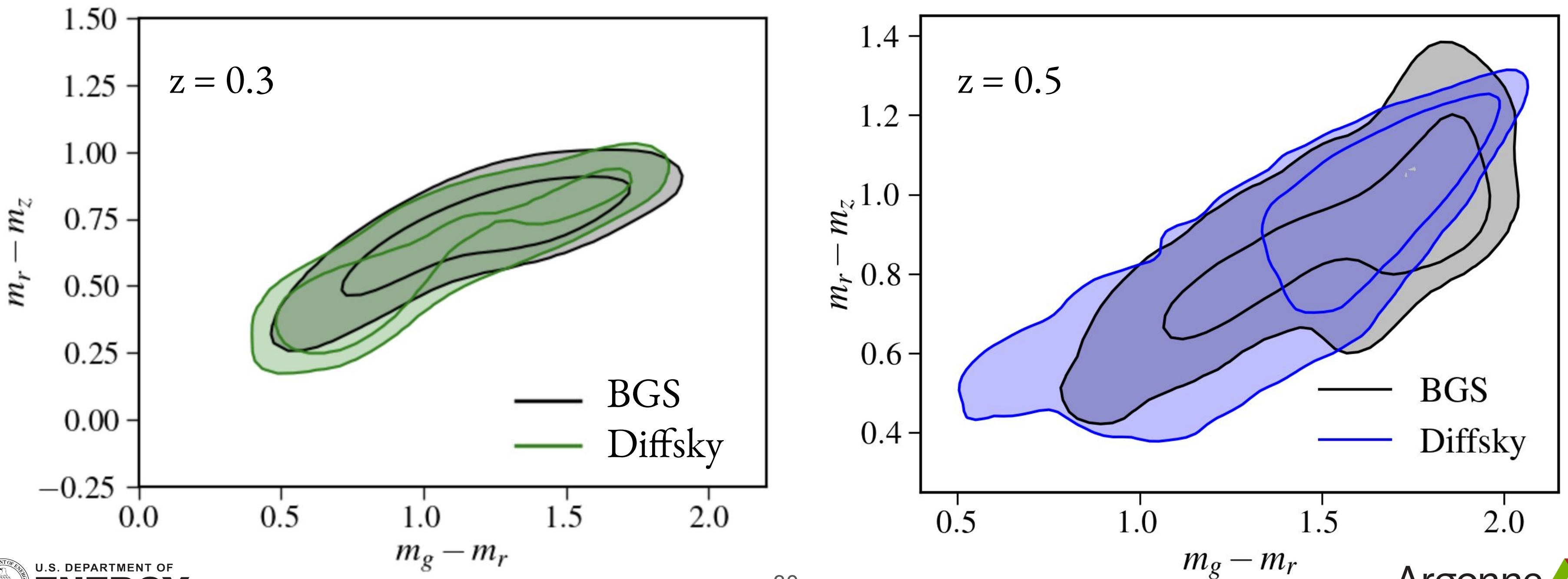
Gif credit: Alan Pearl

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

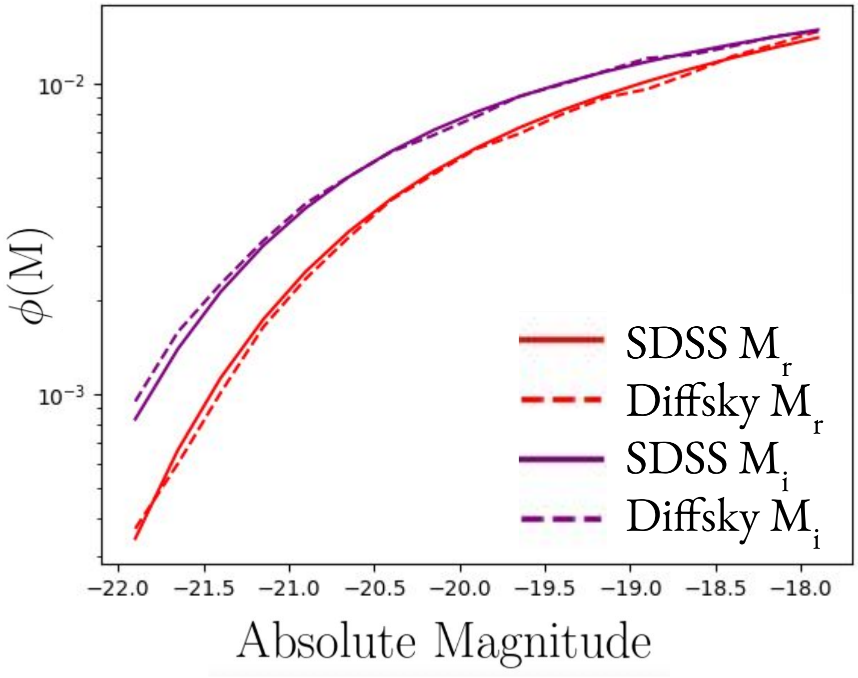
Fitting the model to DESI data

- Good agreement with BGS colors, number densities and satellite fractions at z=0.3 & z=0.5
- Also good agreement with LRG number densities and satellite fractions at z=0.5 & z=0.8

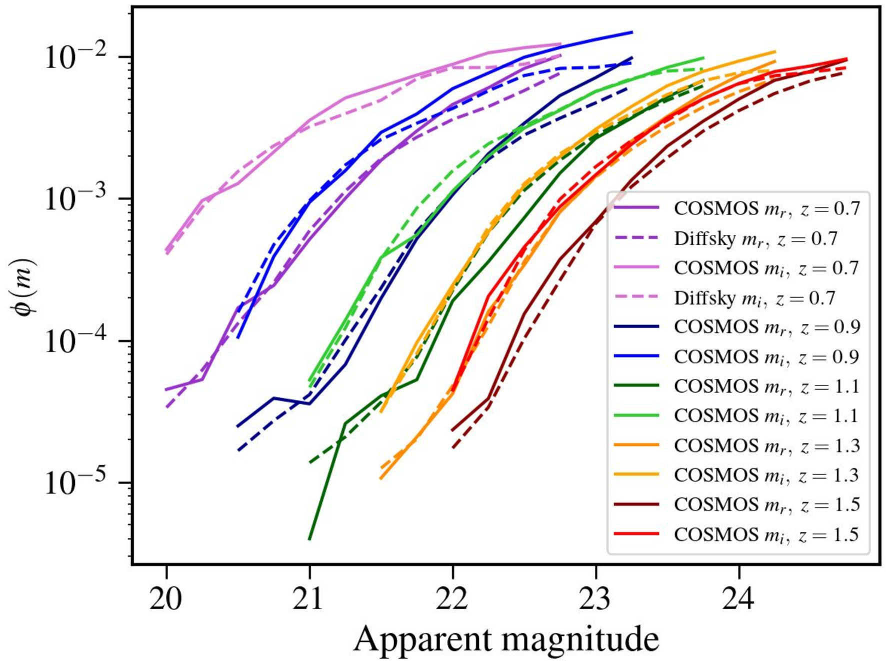


Fitting to SDSS & COSMOS Luminosity Functions

SDSS Main Galaxy Sample



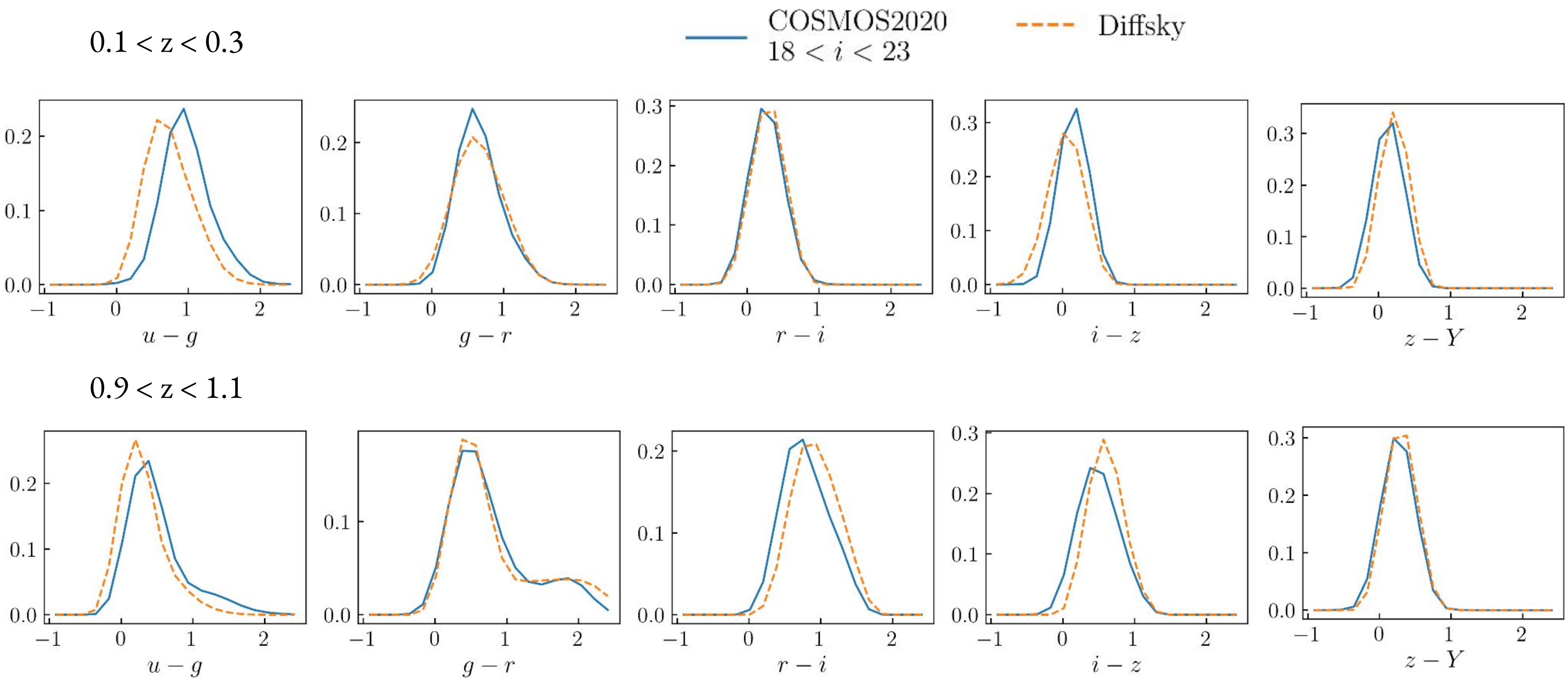
COSMOS 0.7 < z < 1.5



U.S. DEPARTMENT OF ENERGY

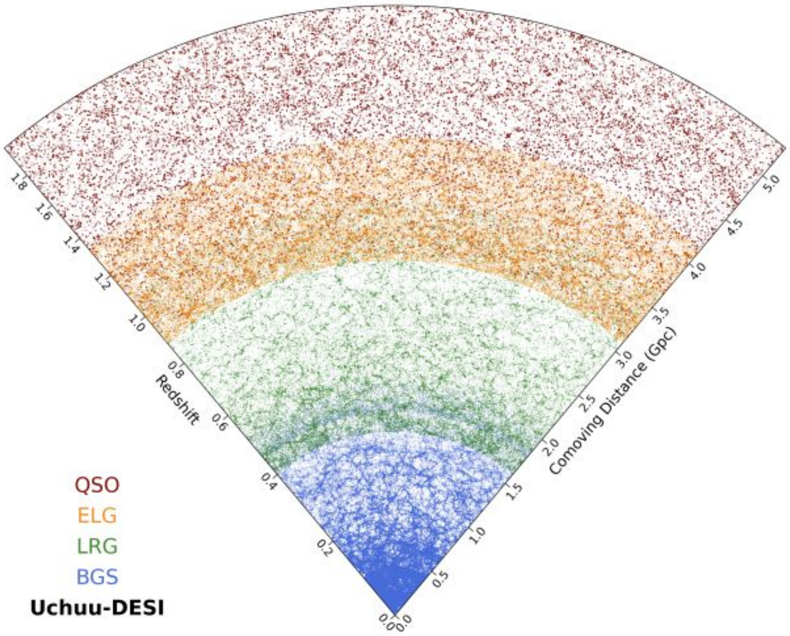
Argonne NATIONAL LABORATORY

Fitting to COSMOS colors



Model capability

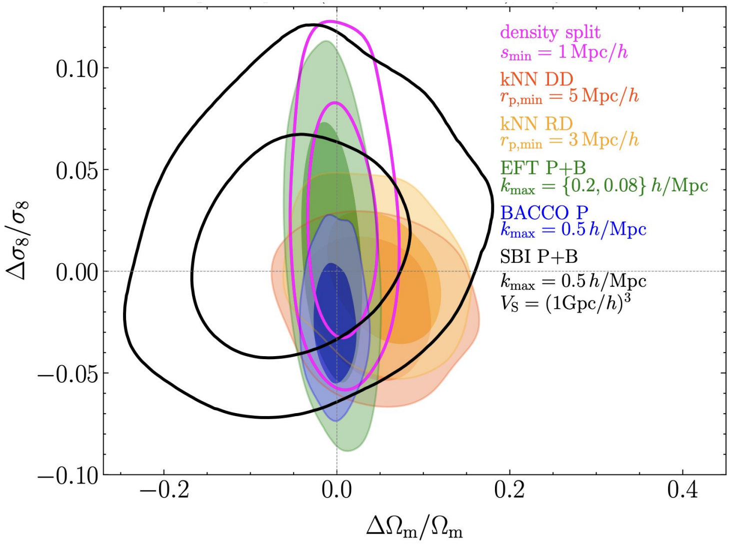
- New capability to fit data:
  - Multi-redshift, multi-wavelength, multi-tracer predictions
- Ideal for cross-survey analyses
- Allows for modeling systematics in a physically meaningful and sufficiently complex way
- We can provide validation data for other pipelines to test robustness (i.e. through mock challenges)
- We can populate simulations with different cosmologies (e.g. Abacus) to make mock galaxy catalogs



Prada et al. 2023

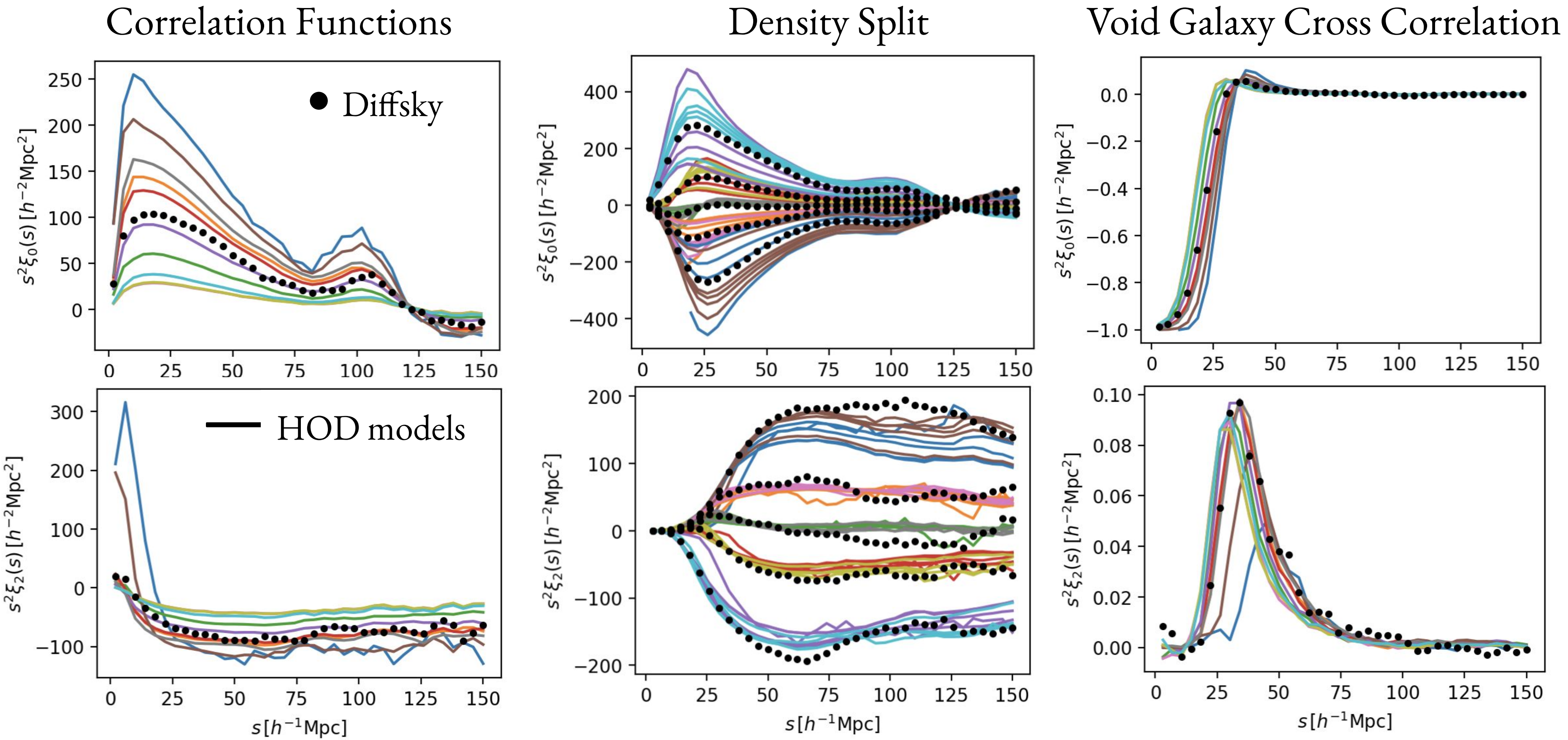
Critical role of mock validation tests

- Mock galaxy catalogs created are ideal for robust validation tests of LSS cosmology pipelines
- Mock challenges are a ubiquitous trend to validate cosmological analyses, test systematics, etc
  - Figure shows recent work from Beyond-2pt Collaboration on parameter-Masked Mock Challenge
  - Similar effort using Diffsky on the DESI Emulator Mock Challenge (discussed later in this talk)
- **Key features needed for compelling validation:**
  - Close agreement between mocks and target data
  - Mock-generating model should rely upon different assumptions from the analysis being validated
  - Ideally have *suite of mocks* spanning physically plausible range of systematic uncertainty



Krause et al (2024)
arXiv:2405.02252

DESI Emulator Mock Challenge: Alternative Clustering Methods

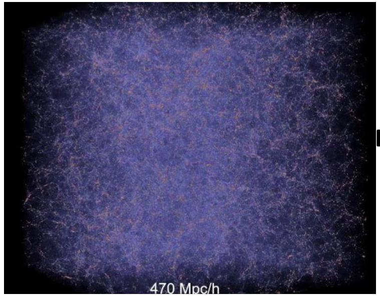


DESI Alternative Clustering Measurements Topical Group
(Enrique Paillas, Carolina Cuesta, Tristan Fraser, et al.)

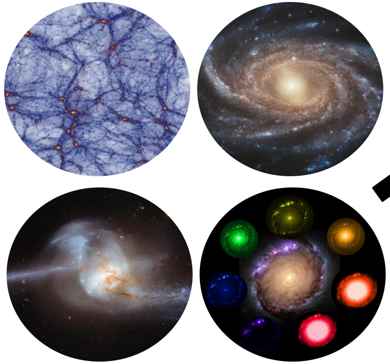
Argonne NATIONAL LABORATORY

Future cosmology analysis

We plan to perform our own **full-scale, multi-redshift, multi-tracer, cross-survey cosmological analysis** (including cross-correlations) with the diffsky pipeline.

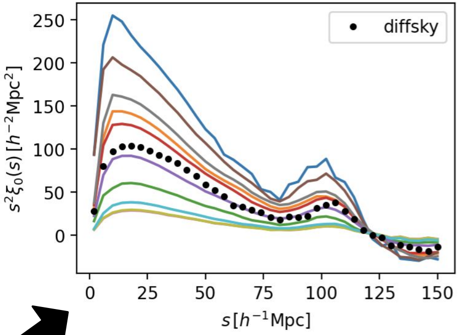


N-body simulation

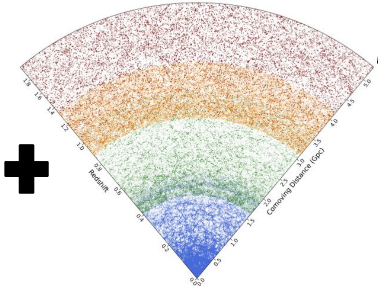


Diffsky Pipeline

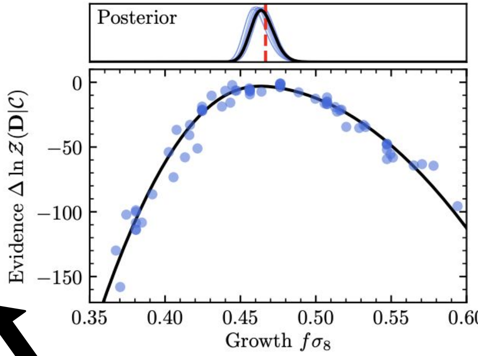
Measurements + Data



+



Constraints!



Cosmological Evidence Modeling (Lange et al. 2019)

Deploying Diffsky on Exascale Machines

- Model built to scale to very-large-volume high-res N-body sims with merger trees
- Model is targeting new HACC sims:
  - Farpoint: 1 Gpc, m_p ~3e7
  - Q-Continuum: 1 Gpc, m_p ~2e8
  - Last Journey: 5 Gpc, m_p ~3e9
- New HACC sims beginning to run on Frontier exascale machine at Oak Ridge
- Aurora exascale machine now at Argonne
  - 50,000+ GPUs in unified memory
- Aurora Early Science Projects:
  - New gen of extreme-scale HACC sims (N-body & hydro)
  - Expansive calibration of Diffsky

Thank you!
Questions?

aphearin@anl.gov
gbeltzmohrmann@anl.gov

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

Differentiable SED fits

Diffsky can also be used in individual galaxy SED-fitting!

Key technical advance

Deploy the gradient-based techniques to derive Bayesian posteriors on physical properties of individual galaxies

Novel feature

Fit photometry/SED of individual galaxy with a physical model of a co-evolving galaxy/halo (e.g., SFR efficiency, gas consumption timescale, etc)

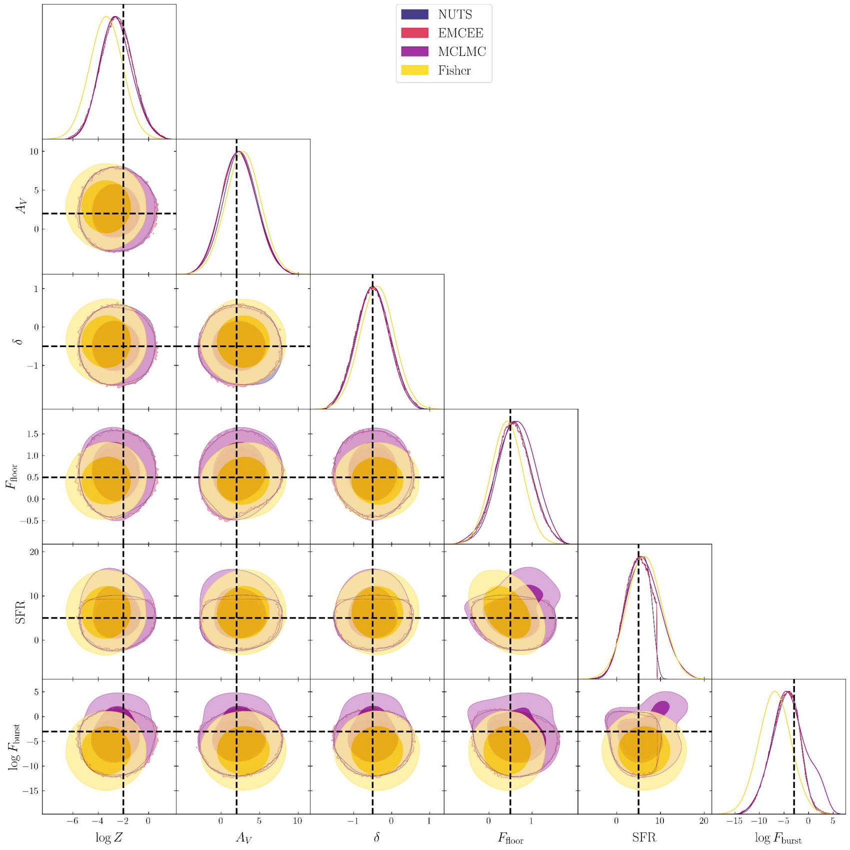
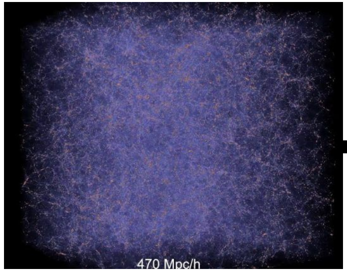


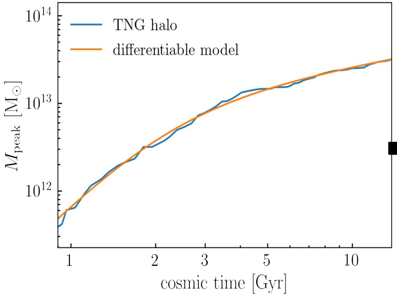
Image credit: Georgios Zacharegkas

Diffsky Pipeline

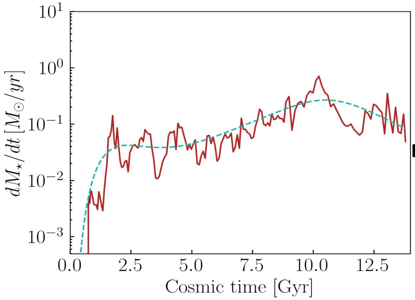
N-body simulation



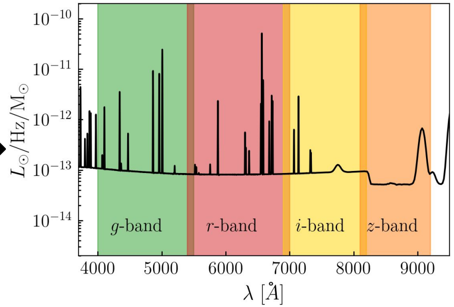
Diffmah:
Hearin et al. 2021

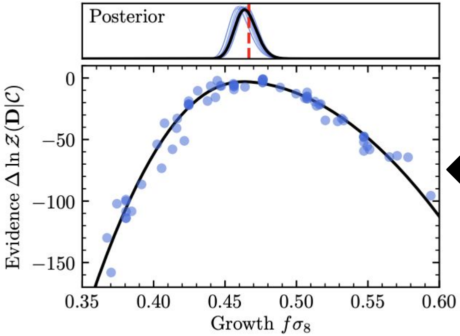


Diffstar:
Alarcon et al. 2023

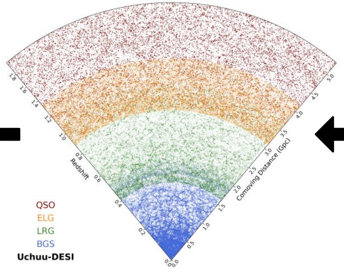


DSPS:
Hearin et al. 2023

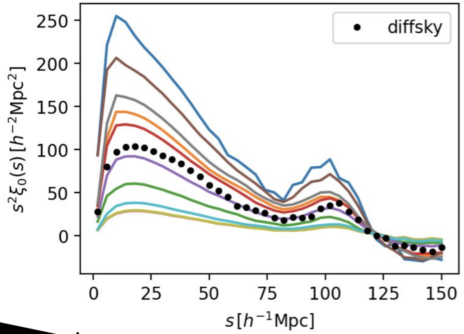




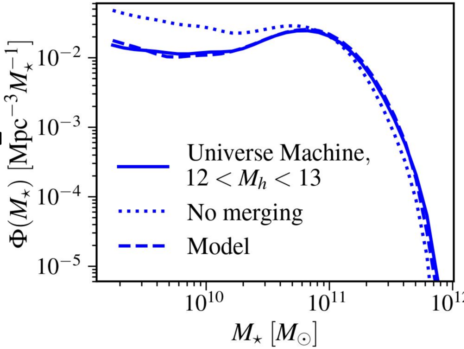
Cosmological Evidence
Modeling (Lange et al. 2019)



Data



Measurements

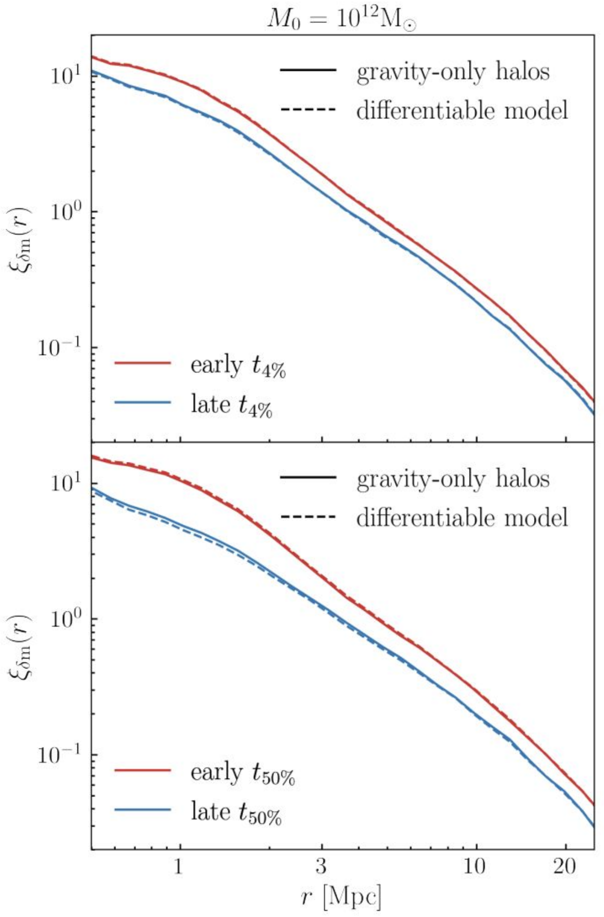


Diffmerge:
BM et al. in prep.

HMC

Differentiable Halo Mass Evolution

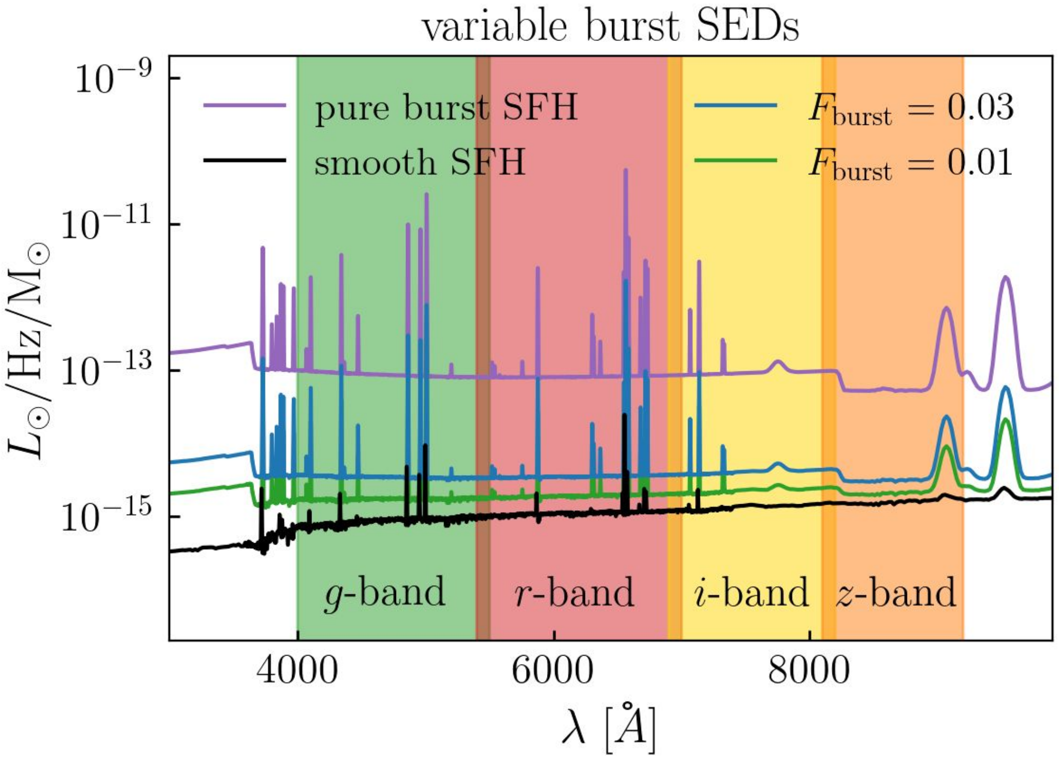
- Using a sample of host halos in BPL, we divide the sample in half according to the median value of halo formation time for the sample
- For each subsample, we compute the cross-correlation between halos and dark matter particles
- This demonstrates that the **correlation between halo formation time and the density field is retained when simulated merger trees are approximated with Diffmah**



Diffmah: Hearin et al. 2021

Bursty star formation

- SED of a burst is much brighter than for a smooth SFH (due to brightness of O and B stars)
- Burst SED is also bluer and has stronger emission lines
- Even tiny values of F_{burst} have a huge impact on the SED
- Burstiness depends on:
  - Stellar mass
  - sSFR



Dust attenuation

- A fraction of the starlight in a galaxy is obscured by dust (Salim et al. 2018):

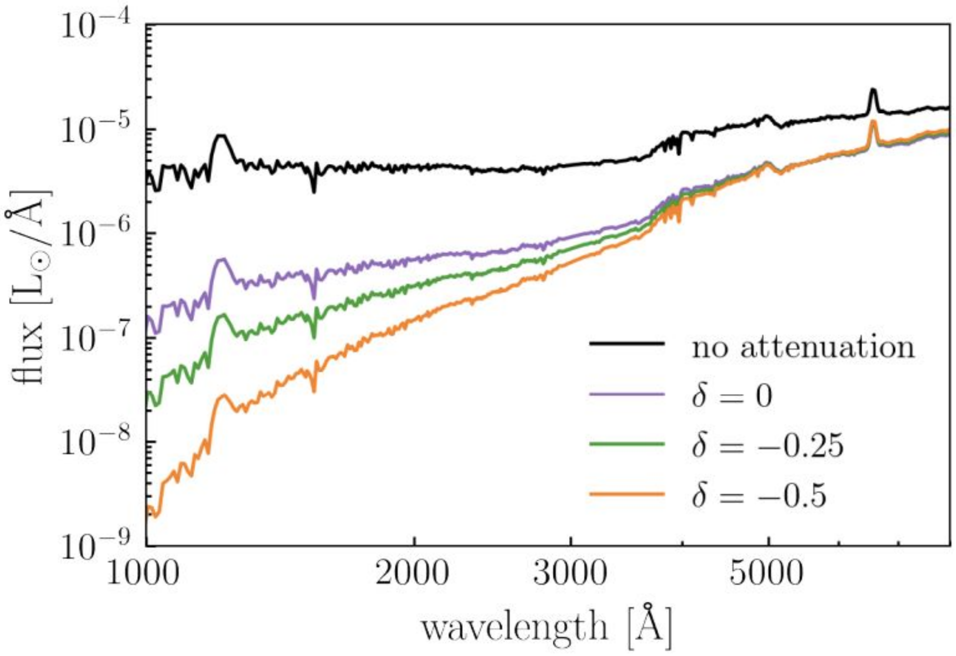
$$F_{att}(\lambda) = 10^{-0.4A_\lambda}$$

Attenuation curve →

$$A_\lambda = \frac{A_V}{4.05} \cdot k_\lambda$$

$$k_\lambda = k_0(\lambda) \cdot \left(\frac{\lambda}{\lambda_V}\right)^{\delta} + D_\lambda$$

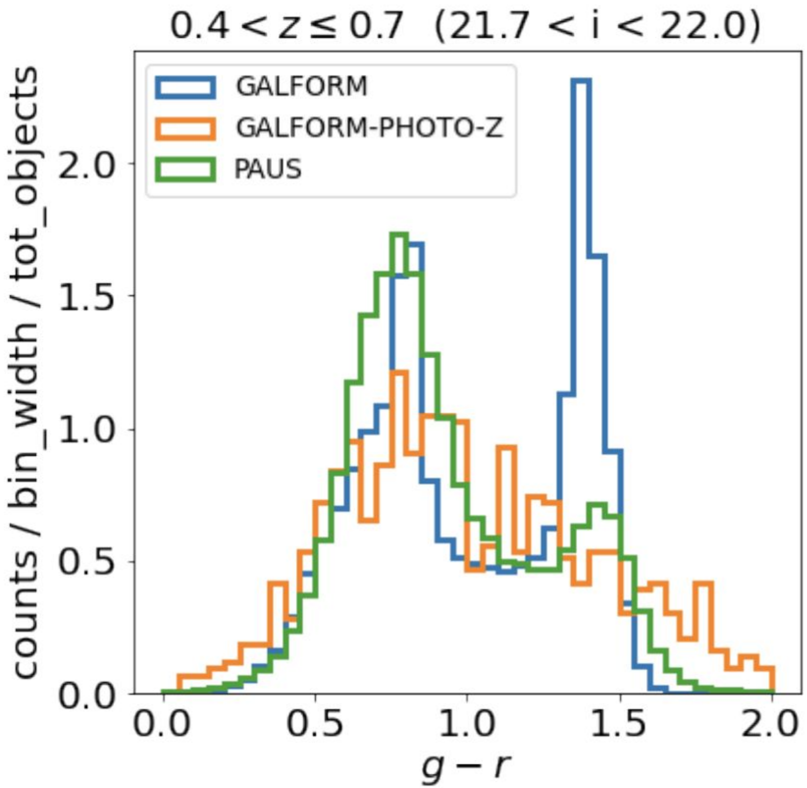
- A_V and δ depend on stellar mass & sSFR
- Additionally, some fraction of the light from a galaxy may be *un*obscured by dust (Lower et al. 2022)



DSPS: Hearin et al. (2023)

Forward modeled photometry from recent SAMs

- SAMs originally developed in 1990s
- Typically predict highly processed observational data (catalogs of stellar mass and SFR)
- Recent trend to predict directly observed quantities: apparent magnitudes & line flux
  - Eliminates source of uncharacterized systematic uncertainty on SPS
- Powerful idea, but quite difficult!
  - GALFORM and SHARK present sharp bimodality not seen in data
  - Only remedied with large empirical correction to predictions
- **Diffsky constraints take the same approach**

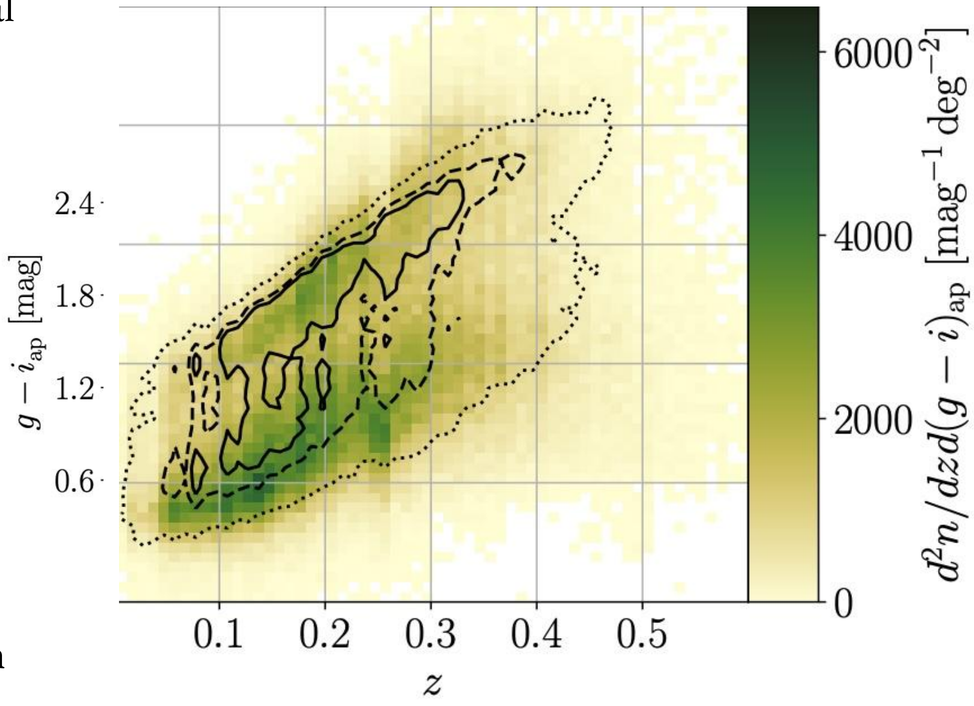


Manzoni et al 2023

U.S. DEPARTMENT OF ENERGY
Argonne NATIONAL LABORATORY

Forward modeled photometry from recent SAMs

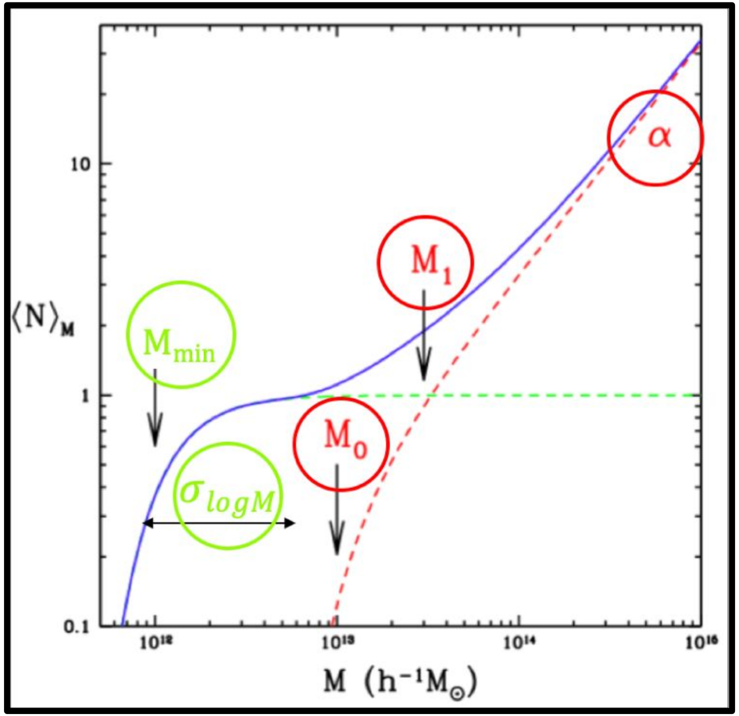
- SAMs originally developed in 1990s
- Typically predict highly processed observational data (catalogs of stellar mass and SFR)
- Recent trend to predict directly observed quantities: apparent magnitudes & line flux
  - Eliminates source of uncharacterized systematic uncertainty on SPS
- Powerful idea, but quite difficult!
  - GALFORM and SHARK present sharp bimodality not seen in data
  - Only remedied with large empirical correction to predictions
- **Diffsky constraints take the same approach**



Bravo et al 2020

Standard HOD Model

- Designed for a single volume-limited sample at a single redshift
- Assign a number of central and satellite galaxies to a halo of mass M using 5 parameters
- Central parameters: M_{min} and σ_logM
- Satellite parameters: M_0, M_1, and α
- Number of galaxies assigned to halo is based only on halo mass
- Central galaxy is placed at the center of the halo and is at rest with respect to the halo
- Satellite galaxies trace dark matter particles within the halo

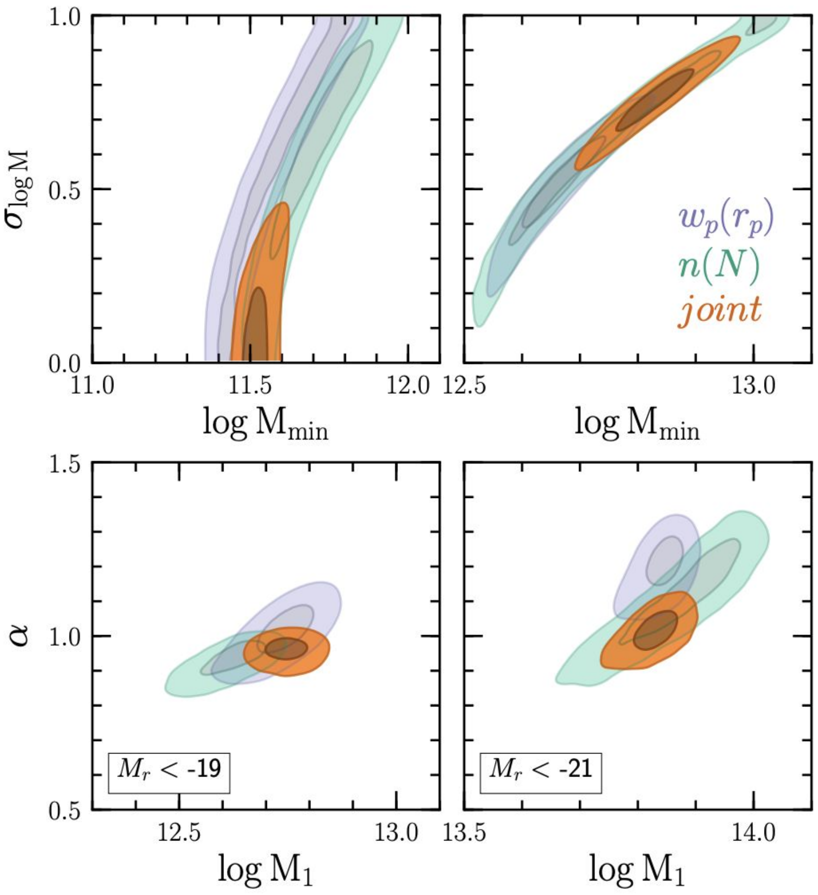


Berlind & Weinberg (2002), Kravtsov et al. (2004), Zheng et al. (2005), Zheng et al. (2007)

Small-scale clustering analyses with the standard HOD

Sinha et al. (2018):

- 2 volume limited samples in SDSS: -19 and -21
- Standard HOD model
- Galaxy number density
- Projected Correlation Function
- Group Multiplicity Function
- Mock covariance matrix



Sinha et al. (2018)