LEARNING THE RELATIONSHIP BETWEEN A GALAXIES SPECTRA AND ITS STAR FORMATION HISTORY

Christopher C. Lovell
Prof. Viviana Acquaviva

Kartheik Iyer, Prof. Eric Gawiser,
Prof. Peter Thomas, Dr. Stephen Wilkins

OUTLINE

Introduction

Spectral Energy Distribution Fitting

Star Formation Histories

Method

Convolutional Neural Networks

Hydrodynamic Simulations

Results

Error estimation

SDSS predictions, VESPA comparison

Conclusions & Questions

(please ask questions anytime)

GALAXY SPECTRAL ENERGY DISTRIBUTION

- **HAVE:**

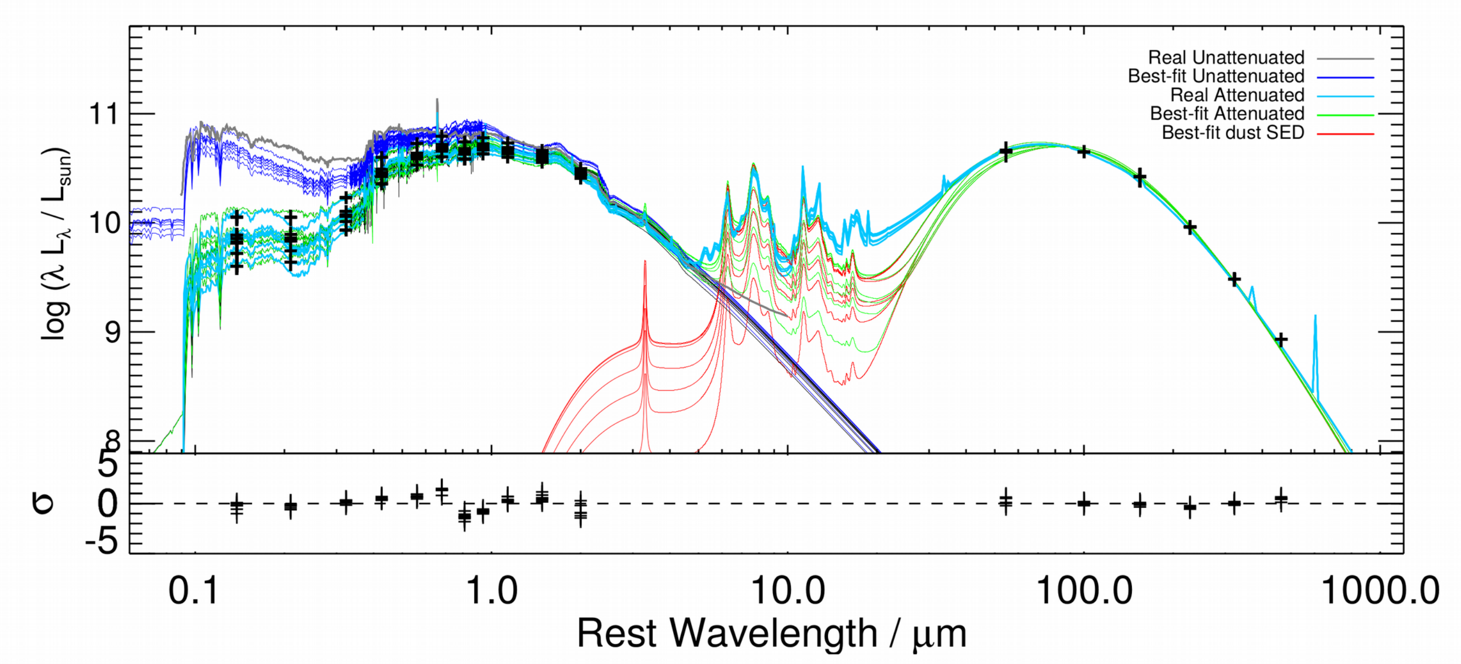
  Flux at different wavelengths / bands

  Spatially unresolved

- **WANT:**

  Physical properties

  Age, Mass, *Star Formation History*, Dust Content, Metallicity…



Hayward & Smith, 2014

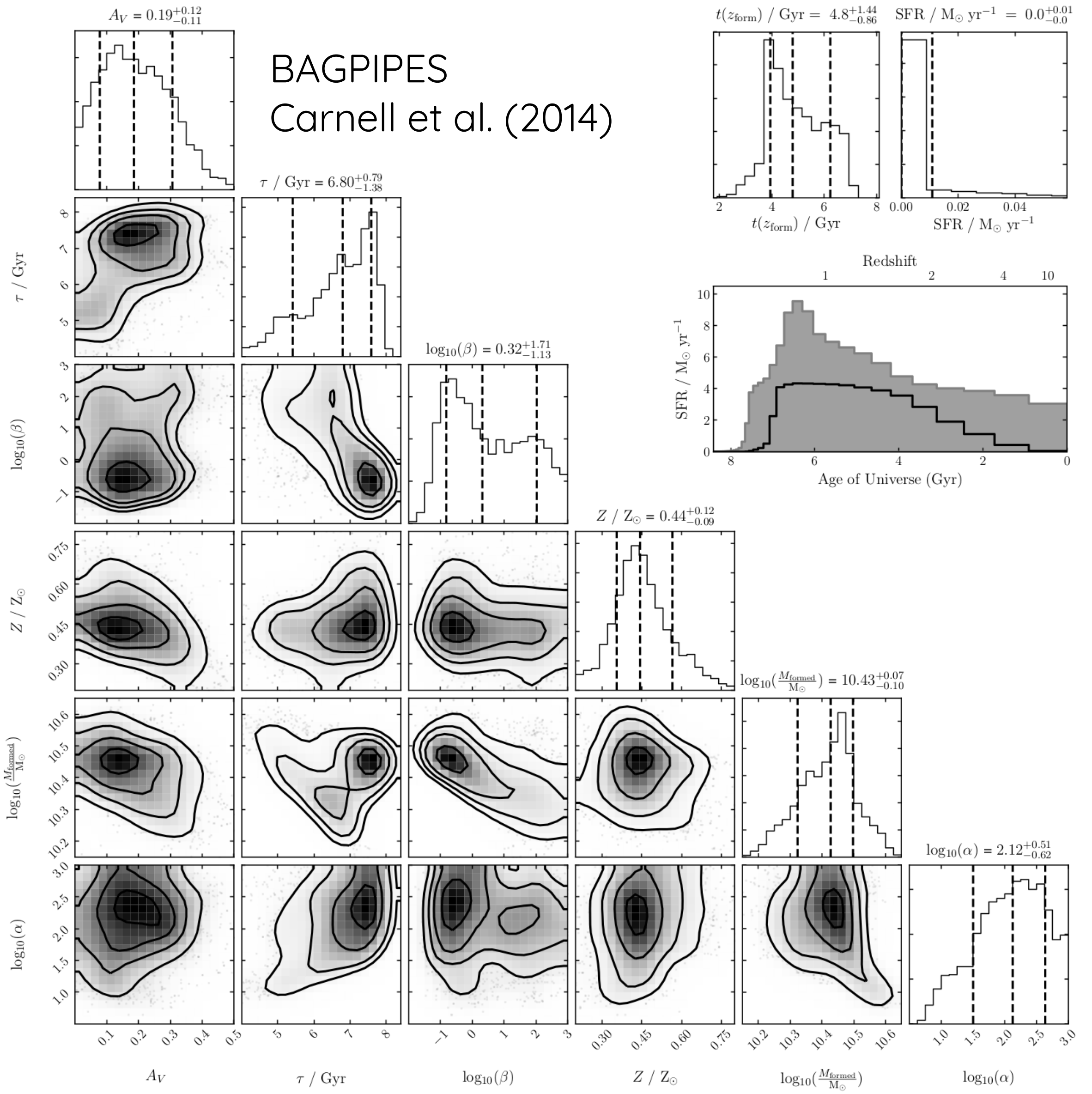
SED FITTING

- Use models with known properties, fit to observational data

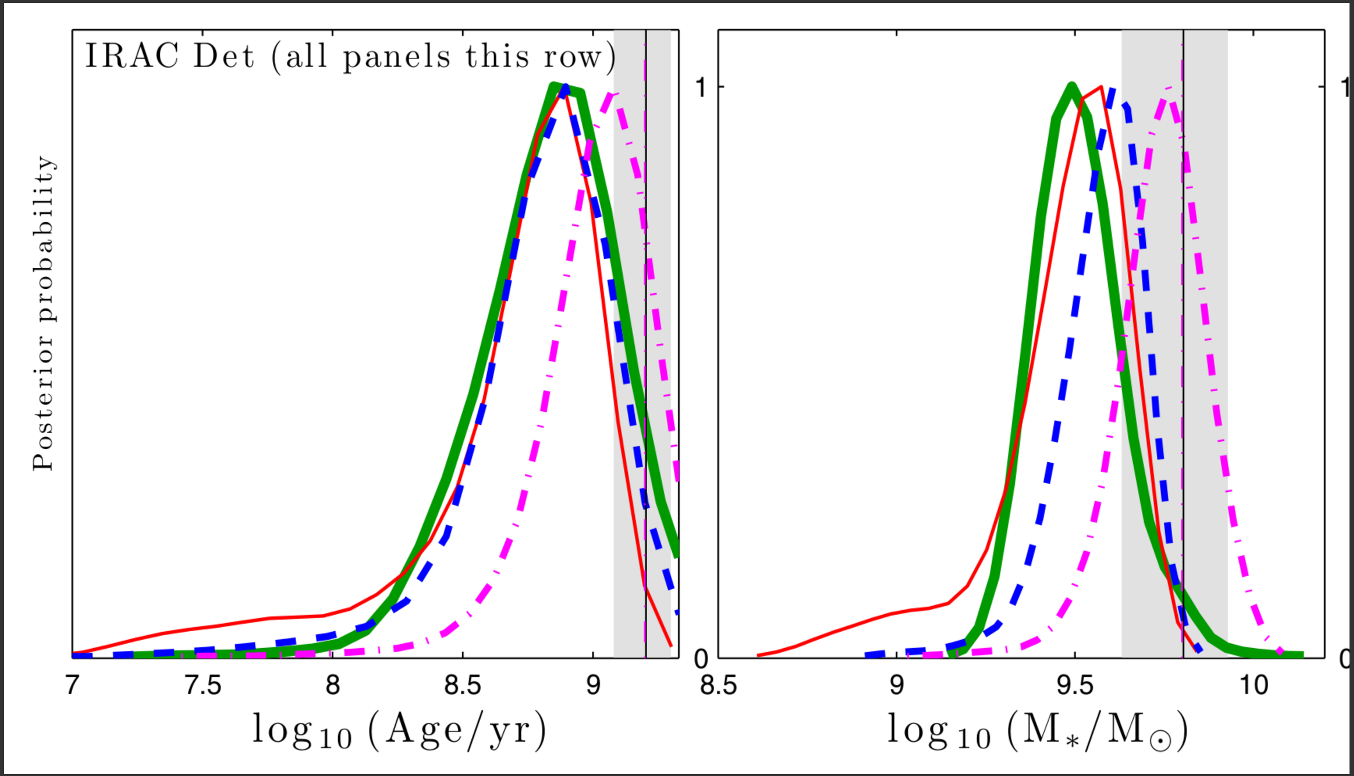
  → infer properties

- There are a **lot** of codes for doing this

  GalMC, Interrogator, BEAGLE, Prospector, VESPA, MAGPHYS, BayeSED, CIGALE, SEABASs, FAST, BAGPIPES…..



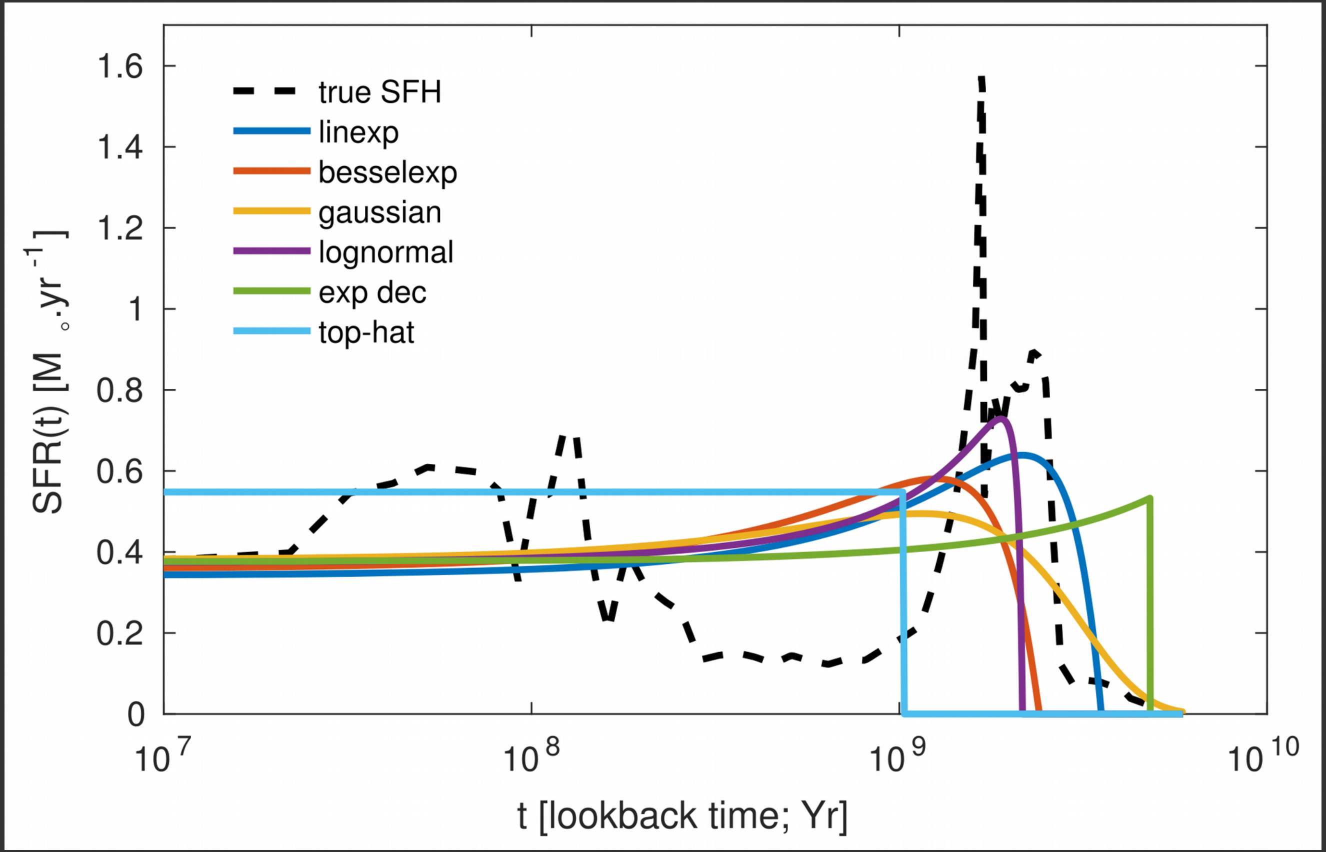
ASSUMPTIONS DOMINATE OVER ERRORS



- Choice of SPS model, extinction law, IMF…
- **Simplistic SFHs** lead to high bias in derived quantities
- All methods biased toward young stellar populations (outshining)

GalMC, Acquaviva et al. (2011)

ASSUMPTIONS DOMINATE OVER ERRORS



Iyer & Gawiser (2017)

- Choice of SPS model, extinction law, IMF…
- **Simplistic SFHs** lead to high bias in derived quantities
- All methods biased toward young stellar populations (outshining)

A different approach to estimating the SFH…

- Take SFHs from simulations (Illustris & EAGLE)
- Generate realistic synthetic SEDs
- **Teach a machine** the relationship between the spectra and the histories
- Test within and between simulations to evaluate generalisation properties

MACHINES OF LOVING GRACE

- Learn from single objects *and* the whole population

  *Analogous in Bayesian parameter estimation to learning the **likelihood** and the **priors***

- Highly non-linear model

  Able to discern higher level features

- Flexible SFH parametrisation

RAGE AGAINST THE MACHINE

- Less transparent generalisation properties
- Supervised machine learning methods limited by training data

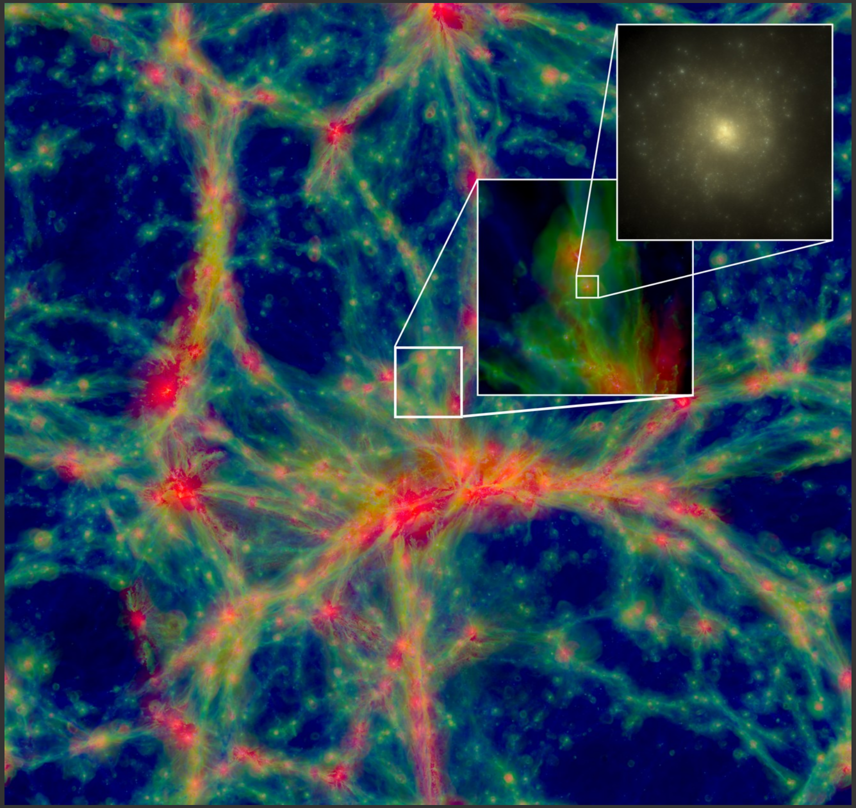
  Observational training data limited, must use simulations

  State of the art simulations volume limited

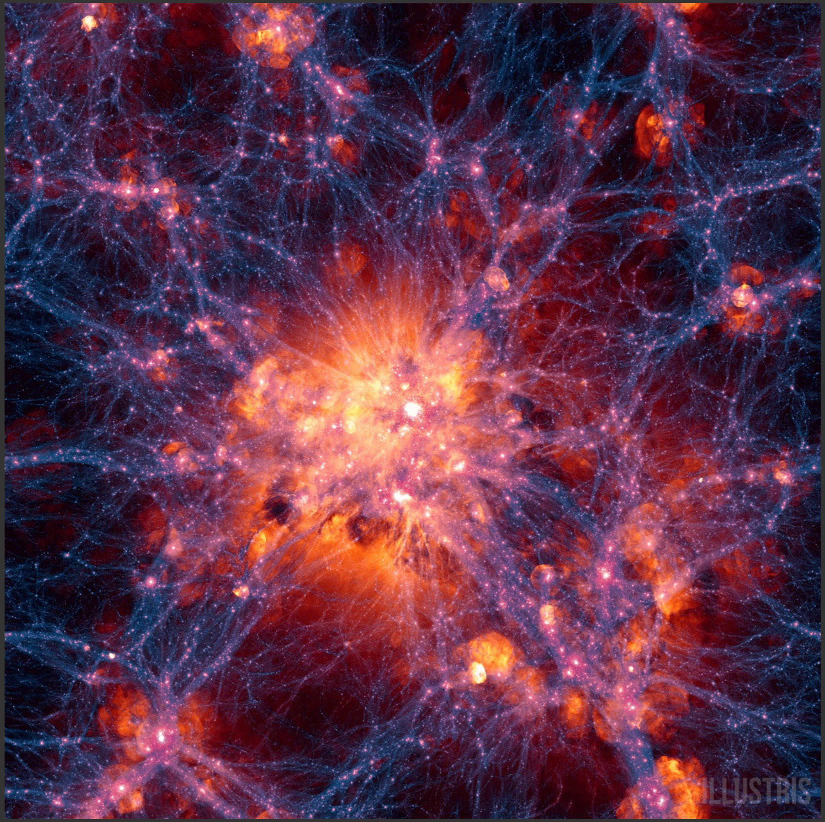
  Agreement between Hydrodynamic simulations still not great

COSMOLOGICAL HYDRODYNAMIC SIMULATIONS

EAGLE *Schaye+14*



Illustris *Vogelsberger+14*



MOTIVATION FOR MULTIPLE SIMULATIONS

- Get a much larger training sample of galaxies

  → helpful for the most massive objects with lower number densities
- Avoid overfitting to a single galaxy evolution model

  → use combined training set
- Can evaluate generalisation properties

  → train on a single simulation, test on another

  → **Assess whether we are learning the intrinsic relationship between galaxy SEDs and their SFHs, rather than overfitting to a particular simulation**

SELECTION

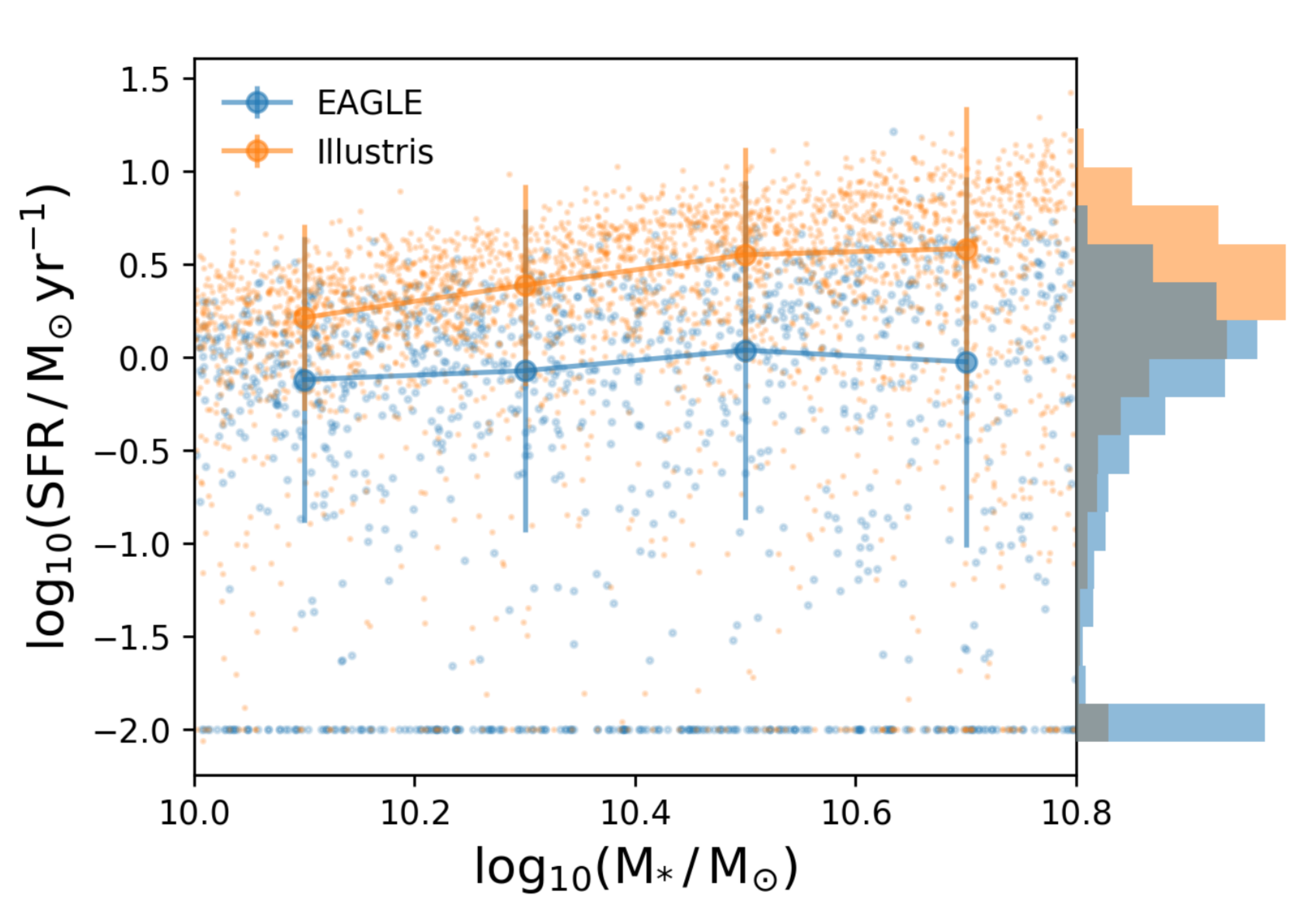
- $10^{10} < M^* / M_\odot < 10^{10.8}$ stratified sample in stellar mass

  → avoid overfitting to low mass galaxies that dominate the mass function

- Number of galaxies selected:

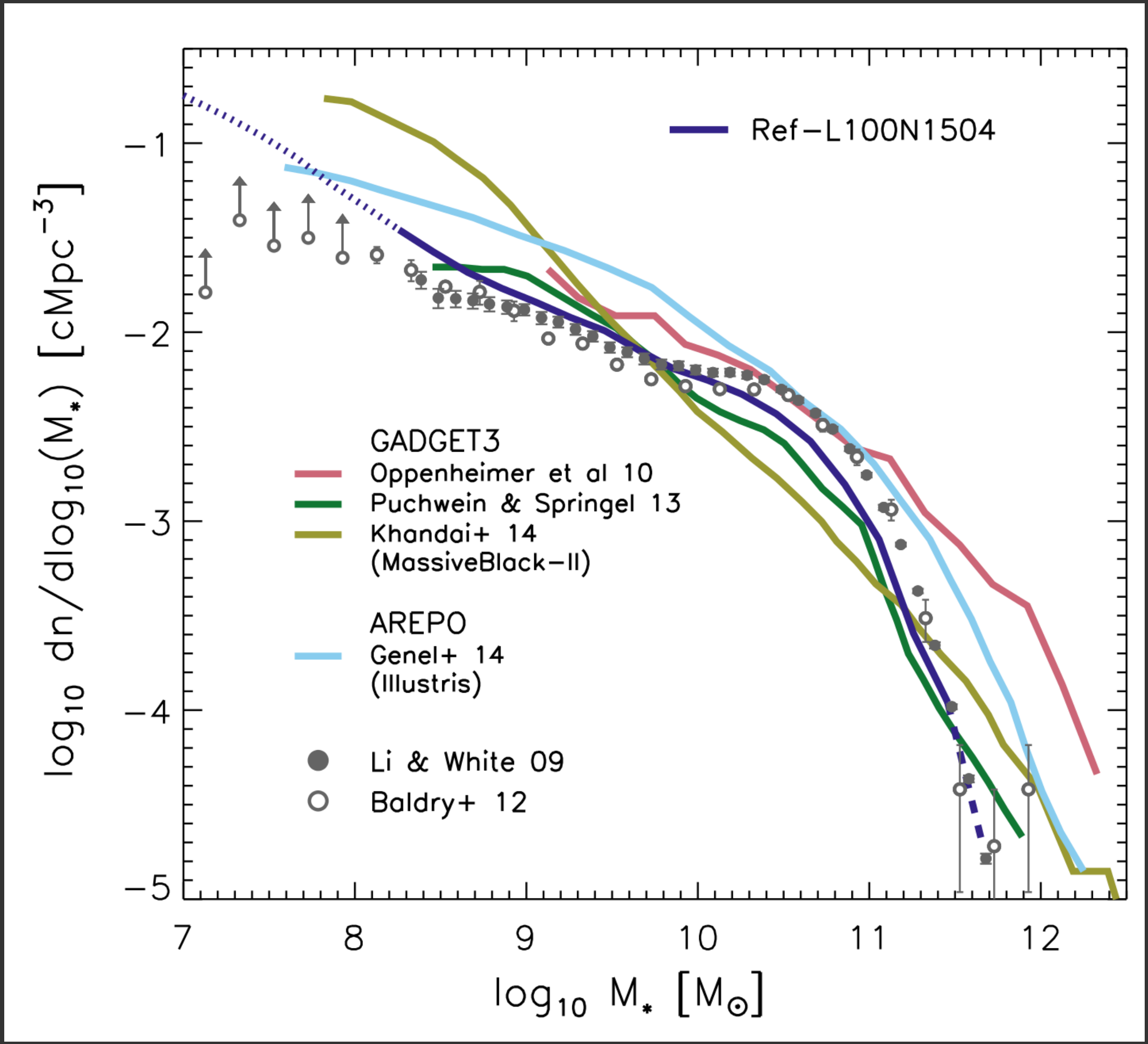
  ~2900 Illustris

  ~1000 EAGLE

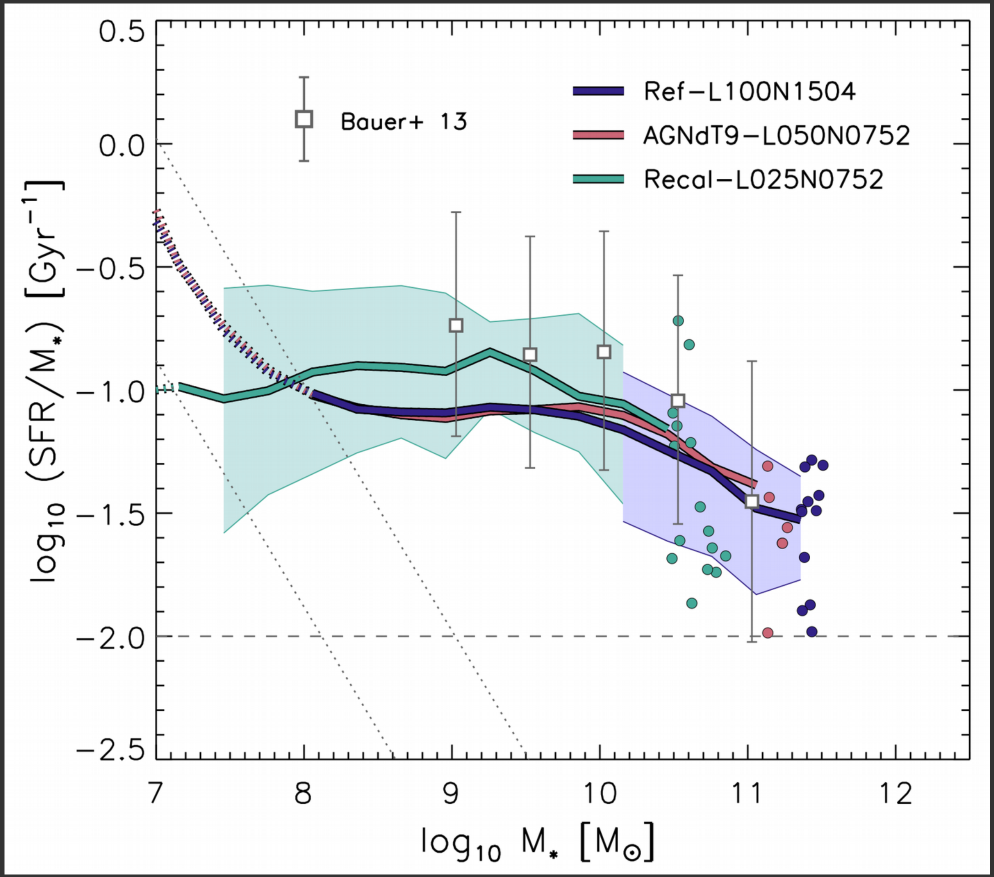


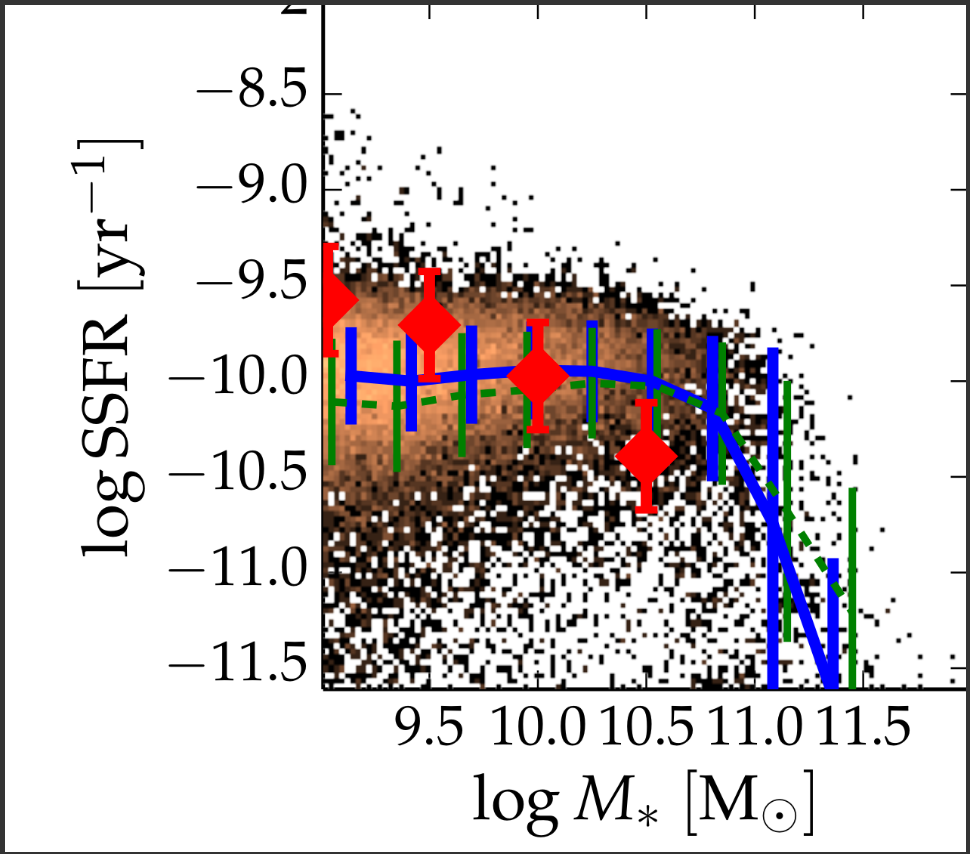
Galaxy Stellar Mass Function

Illustris GSMF has a higher normalisation at low and high masses, but fits the knee well → this is where most of the stellar mass is



SPECIFIC STAR FORMATION RATE





GENERATING SYNTHETIC SPECTRA

- Star particles represent ~10^6 solar masses
- Combination of the initial mass, age and metallicity, coupled with assumed IMF, determines intrinsic SED
- Dust in the ISM leads to attenuation. Amount of dust linked to mass and metallicity of star forming gas
- Young star particles (Age < 100 Myr) are still enshrouded in their birth clouds

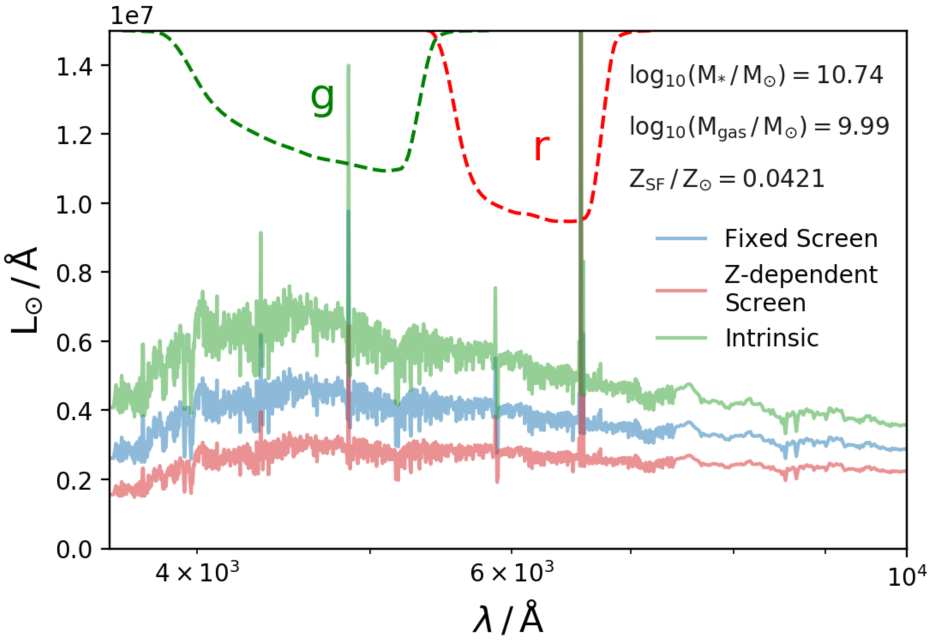
  → leads to nebular attenuation + further dust attenuation
- Ignore the contribution of AGN

SPS MODELLING

- Treat each particle as a Simple Stellar Population (SSP)
- Resample recent star formation, as Poisson noise can significantly affect colours
- Flexible Stellar Population Synthesis (**FSPS**; Conroy+09, Foreman-Mackay+14)
- Includes nebular attenuation contribution for young populations (< 100 Myr); function of incident ionising radiation, computed using **CLOUDY** (Byler+17)

pythonFSPS

Cloudy



DUST MODELLING

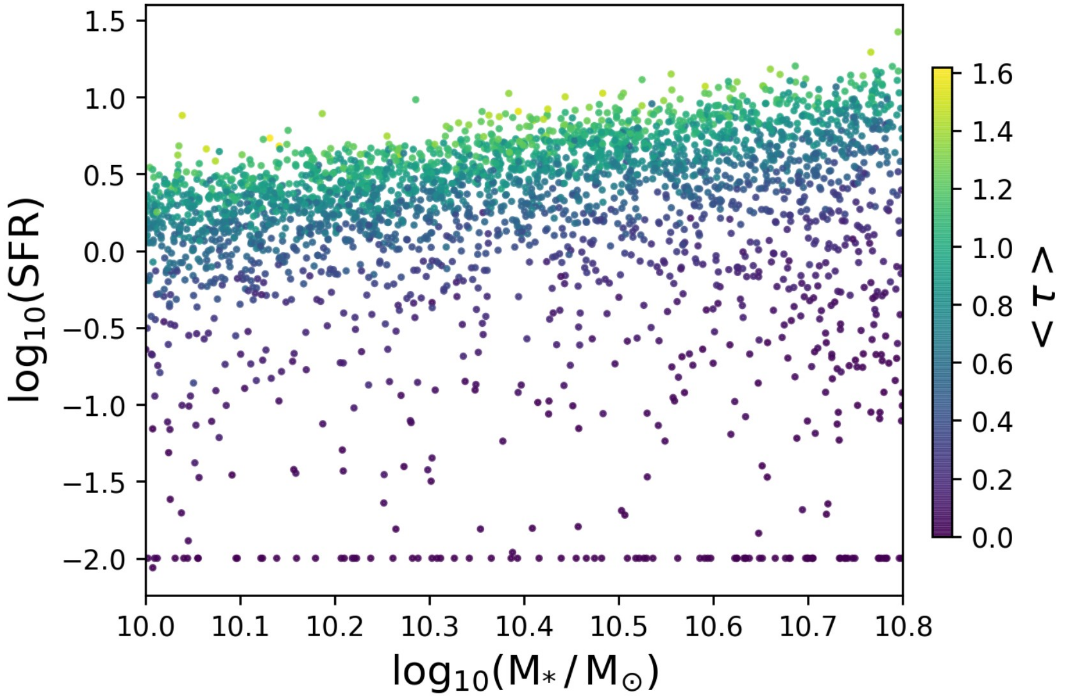
- Two component Charlot & Fall screen model as in Trayford+15

  → Orientation independent, can be applied to EAGLE and Illustris equally

- Attenuation coefficient dependent on **total** mass and metallicity of star forming gas

$$\gamma = \frac{Z_{\mathrm{SF}}}{Z_{\mathrm{Z14}}} \left(\frac{M_{\mathrm{SF}}}{M_*} \frac{1}{\beta} \right) \qquad T(\lambda, t) = \exp\left[-\tau(t) \left(\frac{\lambda}{\lambda_V} \right)^{\alpha(t)} \right]$$

$$\begin{aligned} t \leqslant t_{\mathrm{disp}} &: \ \tau = \gamma \tau_{\mathrm{cloud}} + \gamma \tau_{\mathrm{ISM}} \,; \quad \alpha = -0.7 \\ t \geq t_{\mathrm{disp}} &: \ \tau = \gamma \tau_{\mathrm{ISM}} \,; \quad \alpha = -1.3 \end{aligned}$$



CNN ARCHITECTURE

- 2 x Convolutional layers
  - First applied direct to standardised (mean zero, unit variance) 1D spectra
  - Second applied to output of first, to learn higher order features
- 1 x max-pooling layer
  - Reduces dimensionality → reduced training time
- Traditional fully connected network
  - ‘shallow and wide’
- Hyperparameter optimisation with HYPERAS *github:maxpumperla/hyperas*

Talk to me after for further details

EXTRA DETAILS

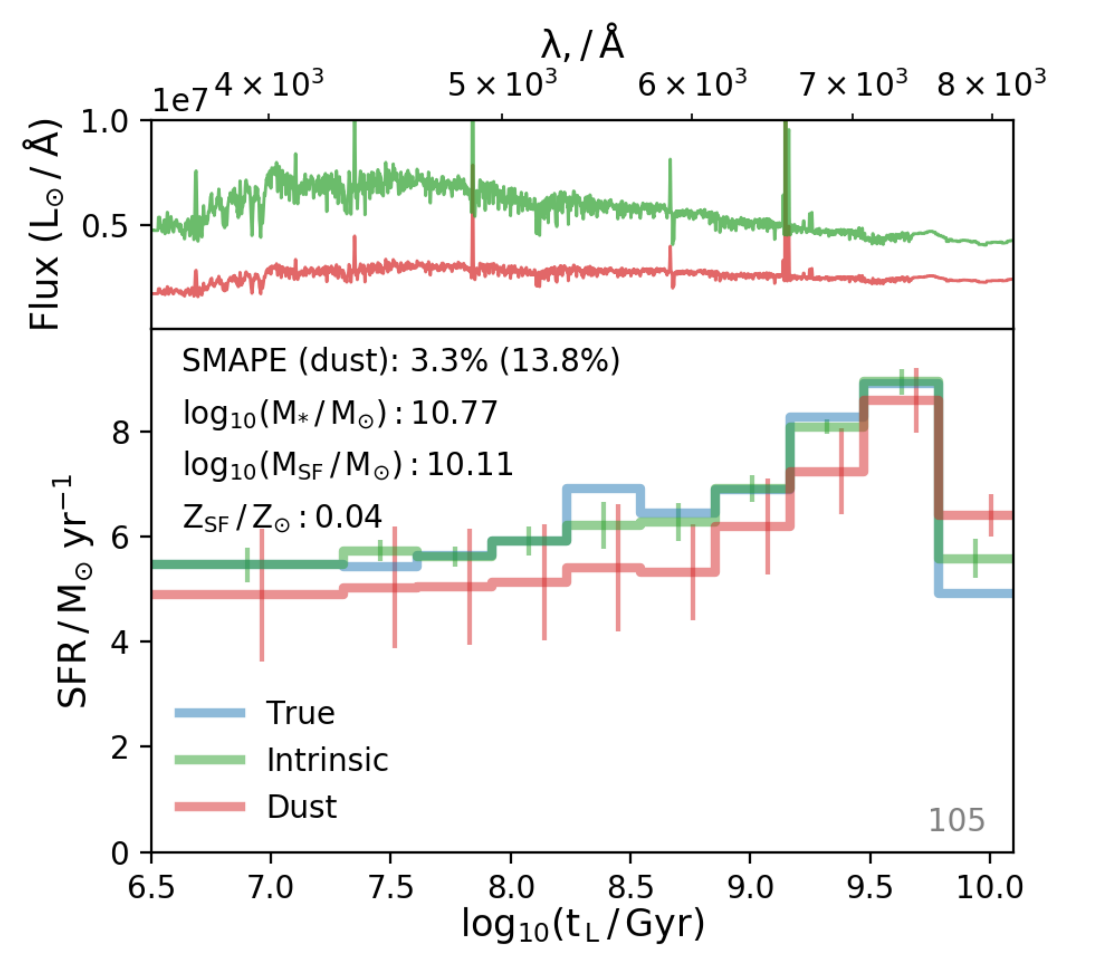
- 10 uniform bins in log lookback time
- → encoded bias towards more recent bins where greater constraints possible
- Spectral coverage matched to SDSS DR7

  ~ 3000 – 8000 Å
- 30 pkpc aperture to match SDSS Petrosian aperture at z = 0.1
- Evaluate with Symmetric Mean Absolute Percentage Error (**SMAPE**)

$$\mathrm{SMAPE} = \frac{\Sigma_i \, |\mathrm{Y}_i^{\mathrm{true}} - \mathrm{Y}_i^{\mathrm{pred}}|}{\Sigma_i \, (\mathrm{Y}_i^{\mathrm{true}} + \mathrm{Y}_i^{\mathrm{pred}})}$$

RESULTS

Example Fit



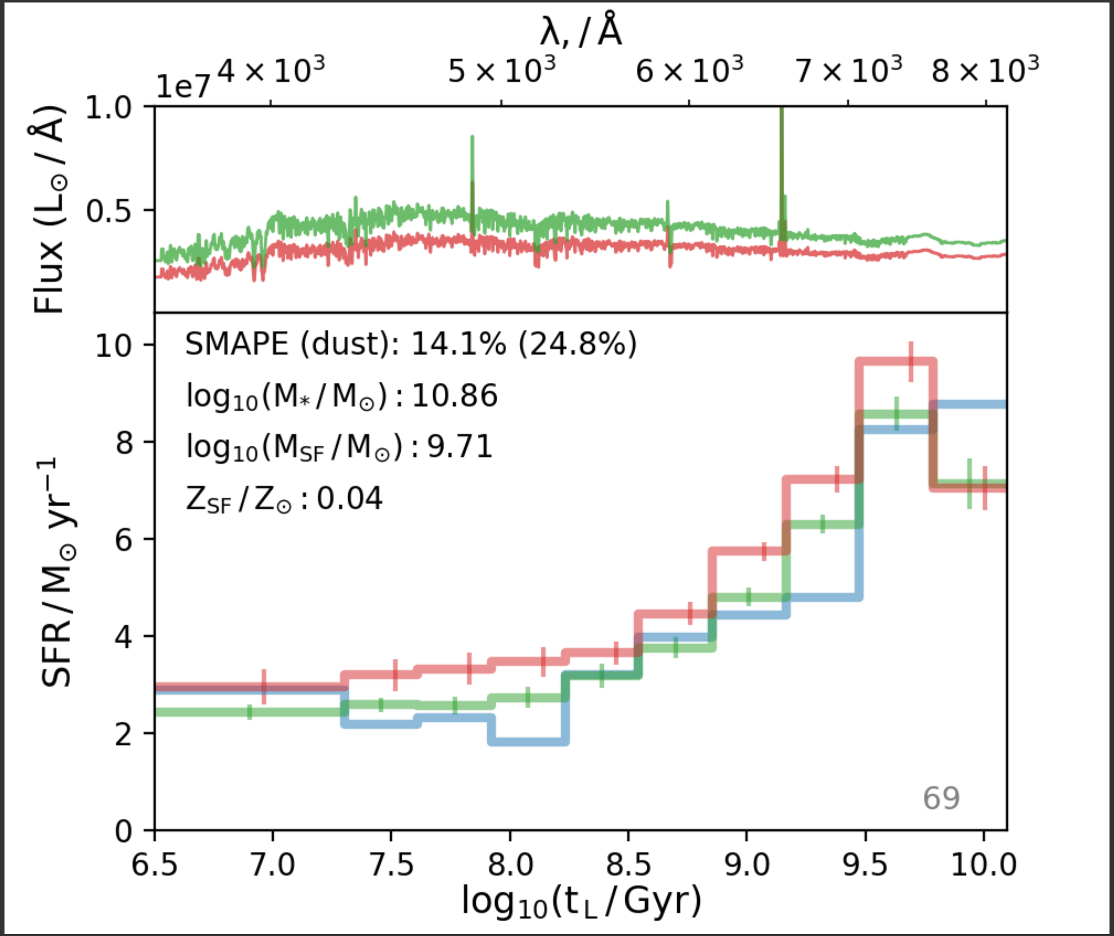
- Illustris galaxy

  In top quartile of SMAPE distribution

- Intrinsic + Dust attenuated SEDs

  → SMAPE for dust attenuated spectra higher than intrinsic

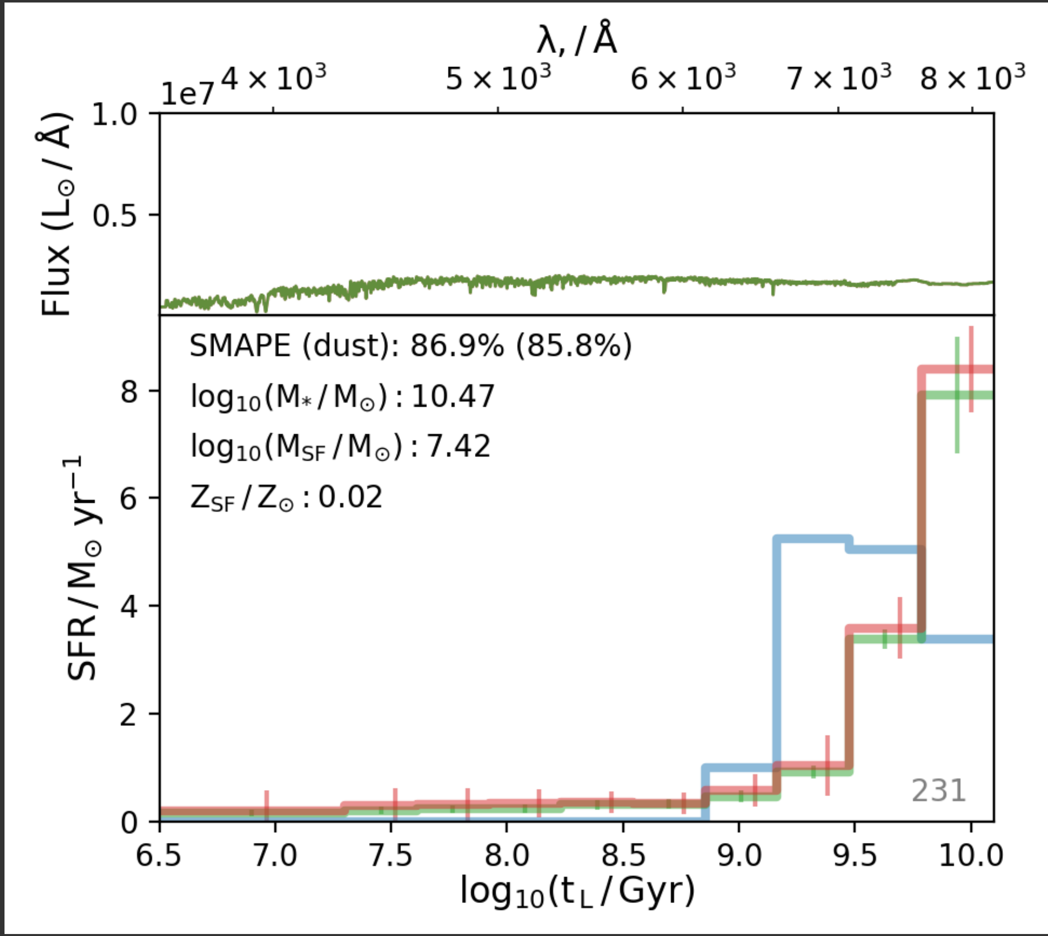
EXAMPLE FIT



Intrinsic (Green)
Dust (Red)

Median of SMAPE distribution

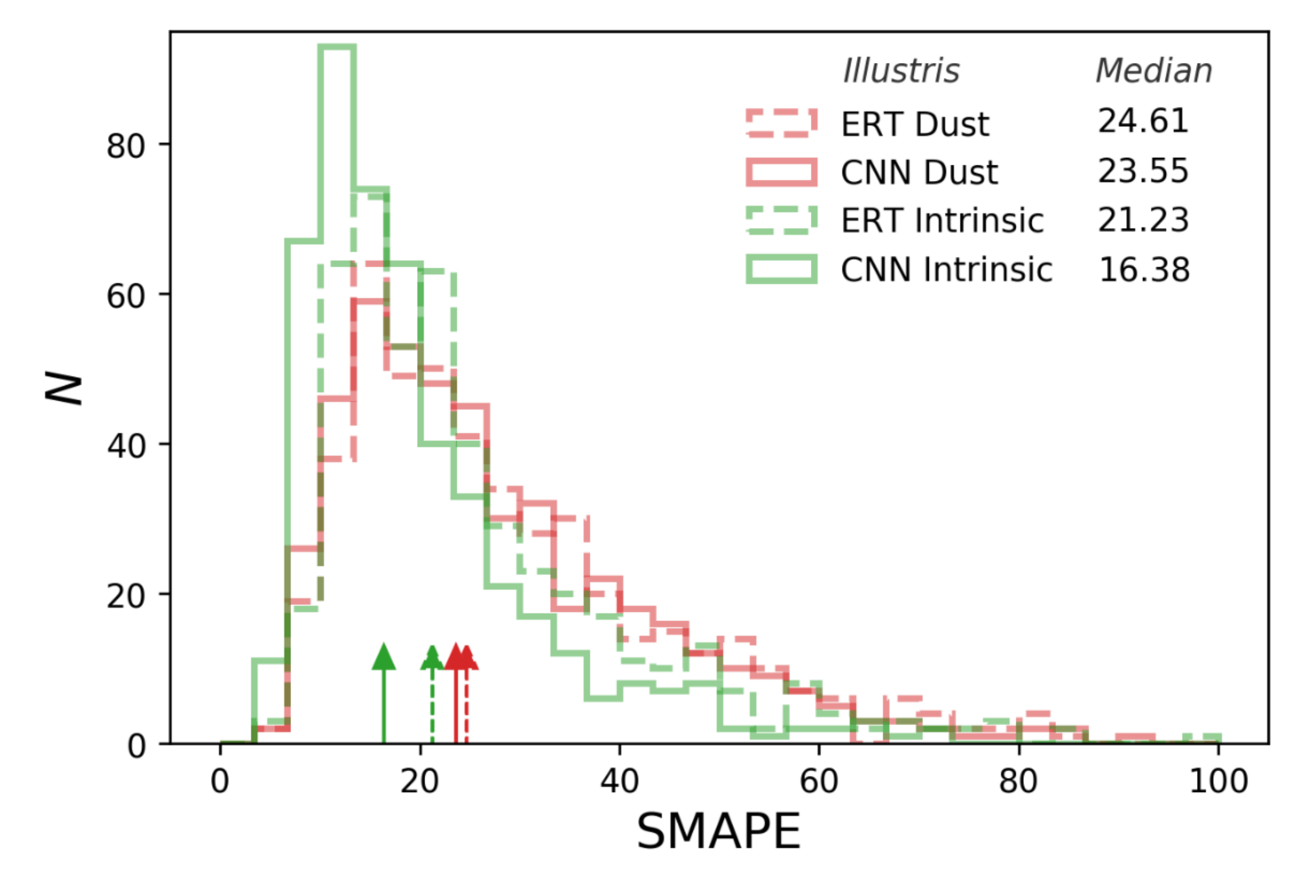
EXAMPLE FIT



Intrinsic (Green)
Dust (Red)

Bottom quartile of SMAPE distribution

SMAPE DISTRIBUTION



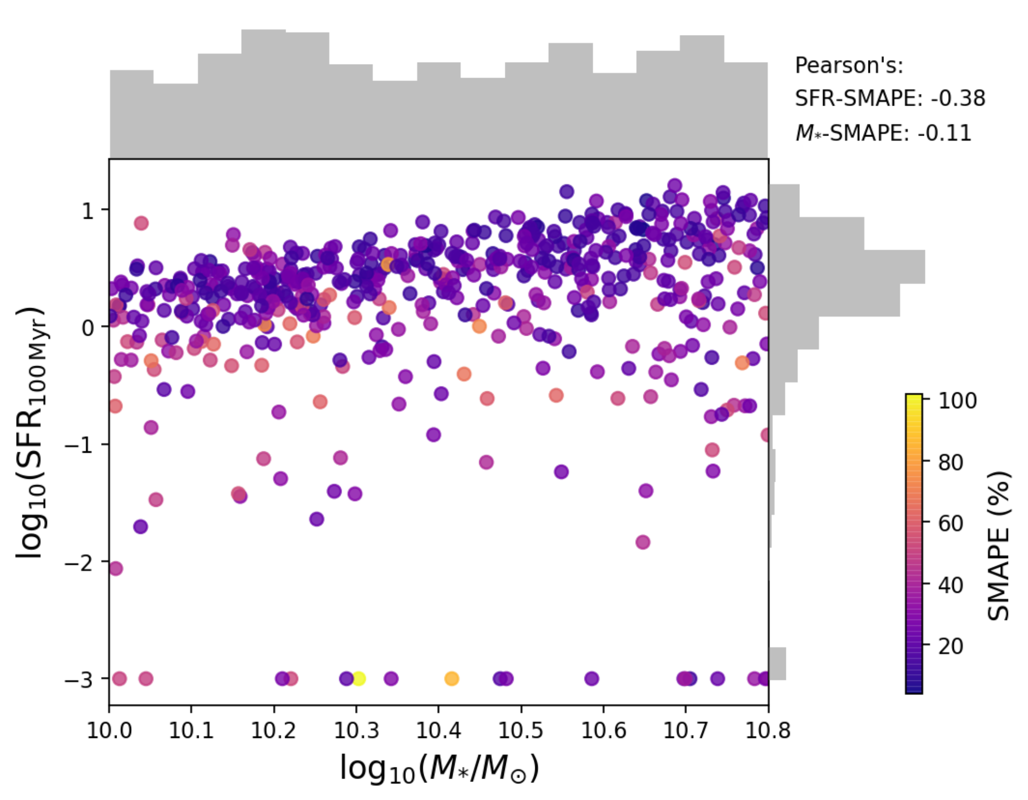
- Median shown by arrows at bottom
- CNN outperforms **Extremely Randomised Trees**, an ensemble decision tree method

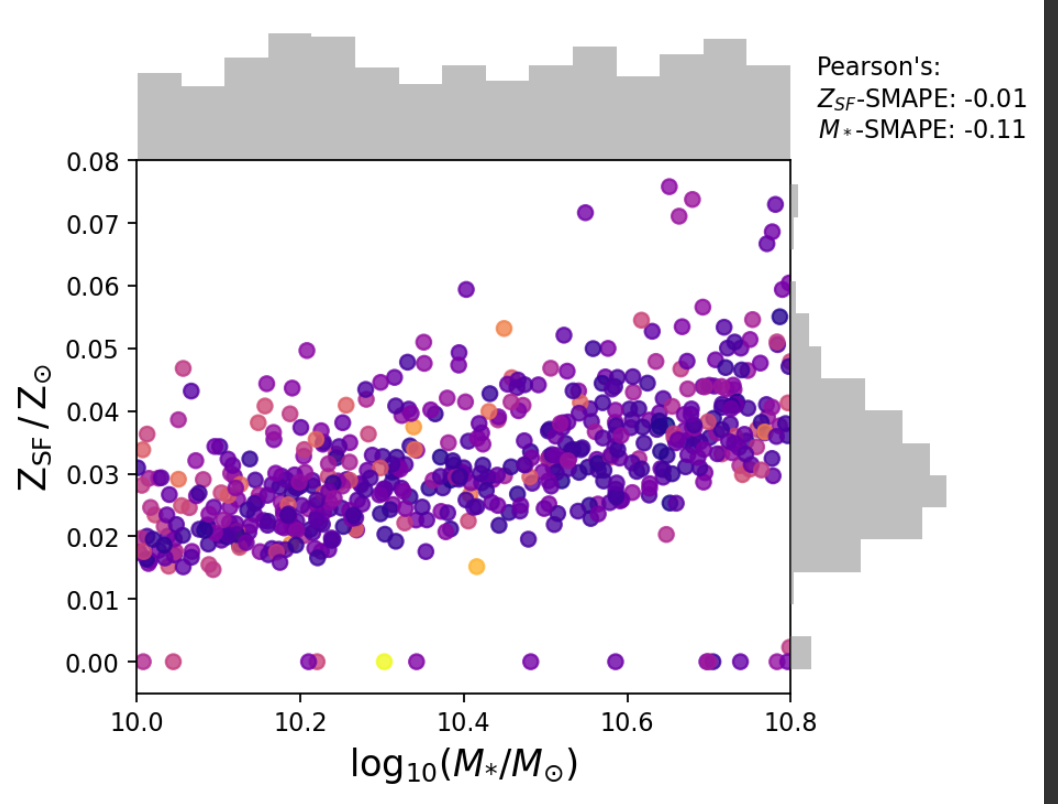
PHYSICAL CORRELATIONS

- SMAPE negatively correlated with recent SFR

  → Opposite to expectation from outshining bias

- Small negative correlation with stellar mass





ESTIMATING ERRORS

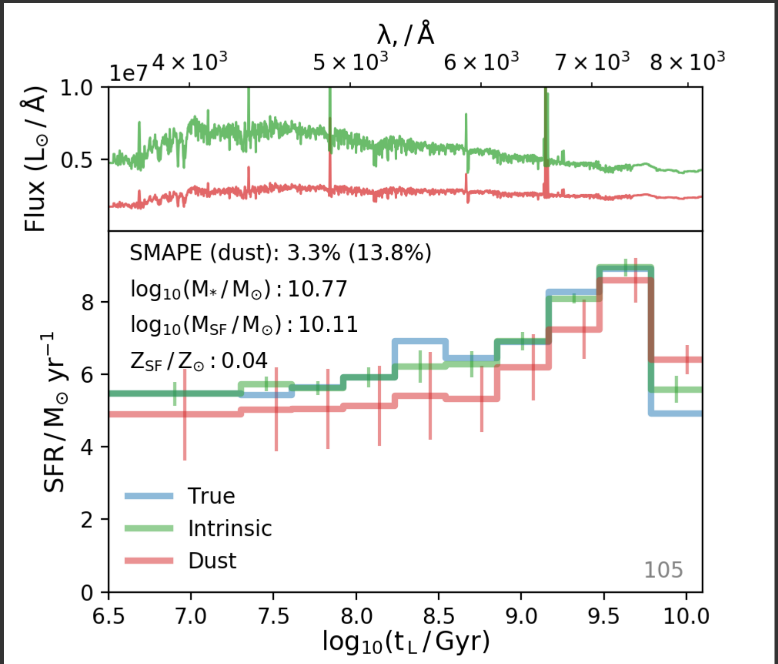
- We identify two main sources of error:

  → Spectral errors

  → Model errors

- For **spectral** errors, use average SDSS DR7 error spectrum from sample (details later)

- Create N_{err} realisations of each spectra + sampled noise, propagate through model

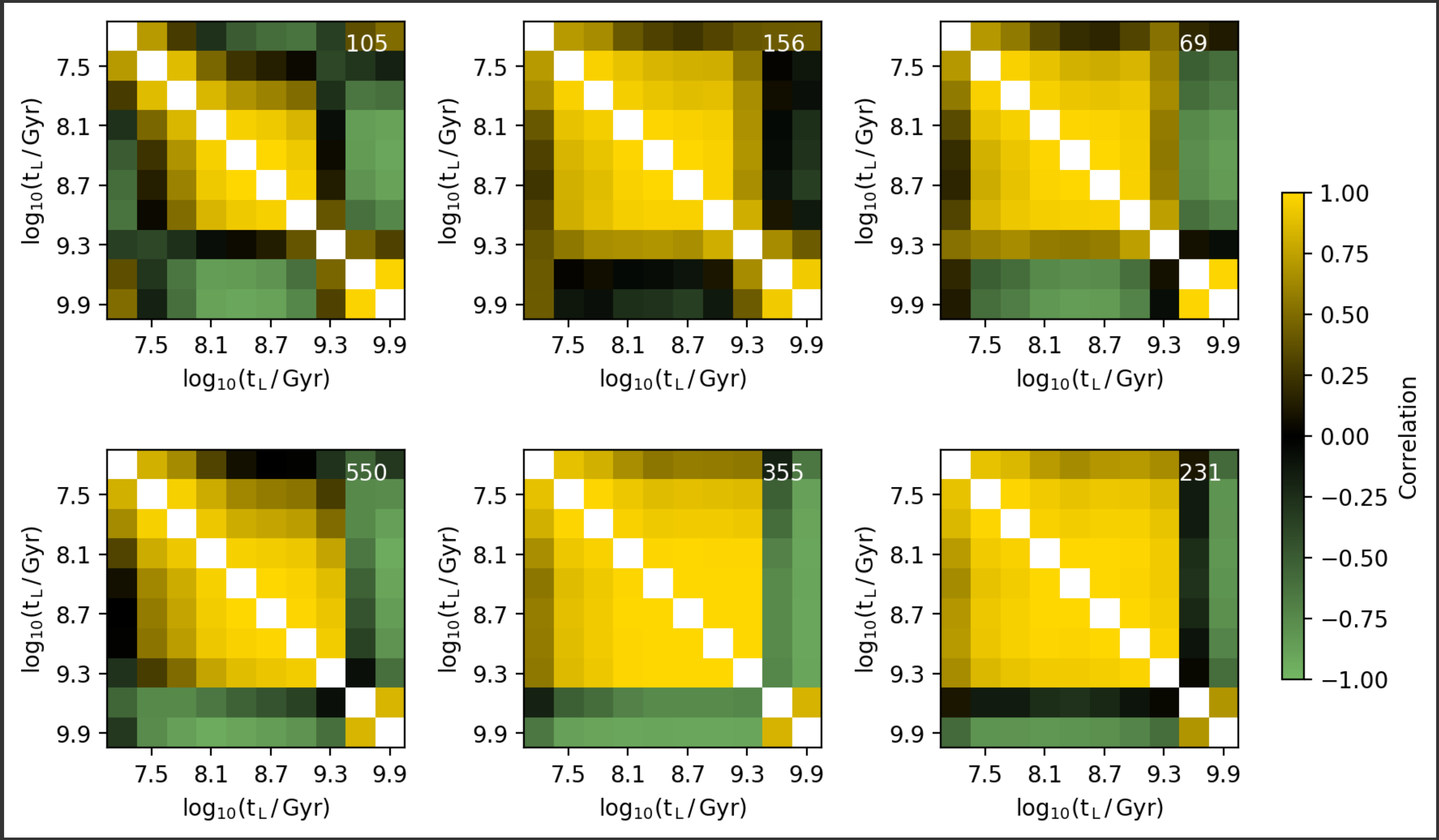


$$C_{ij} = \langle (x_i - \hat{x}_i)(x_j - \hat{x}_j) \rangle$$

$$\sigma_i = \sqrt{C_{ij}}$$

$$r_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j} \qquad r_{ij} \in [-1,1]$$

CORRELATION MATRICES



Neighbouring bins correlated

Recent star formation negatively correlated with early star formation

Modelling errors

~ 10000 parameters in CNN

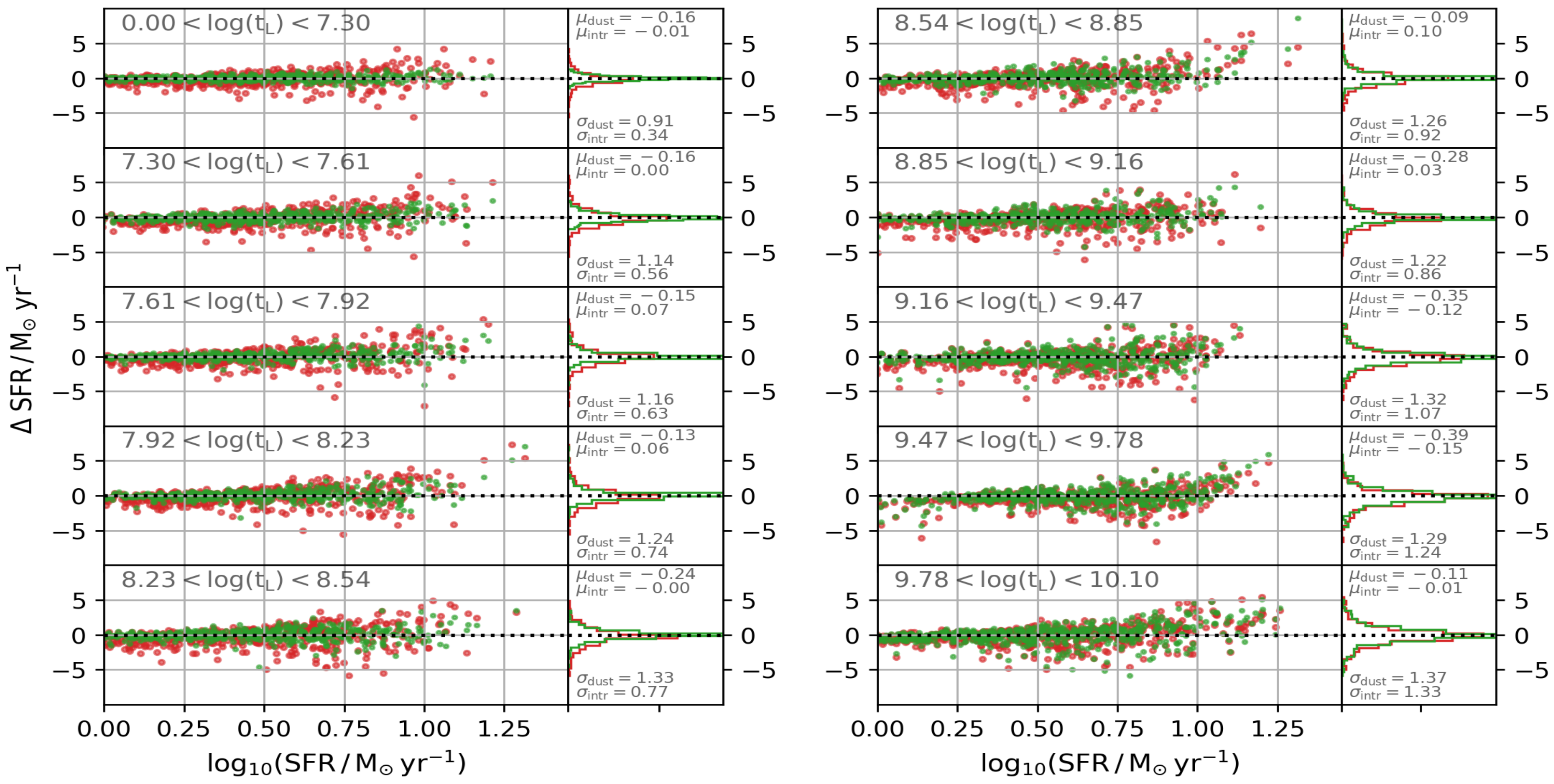
Impossible to estimate errors on all

Empirical approach:

Use residuals in test set

Estimate of total error from quadrature sum of spectra & model errors

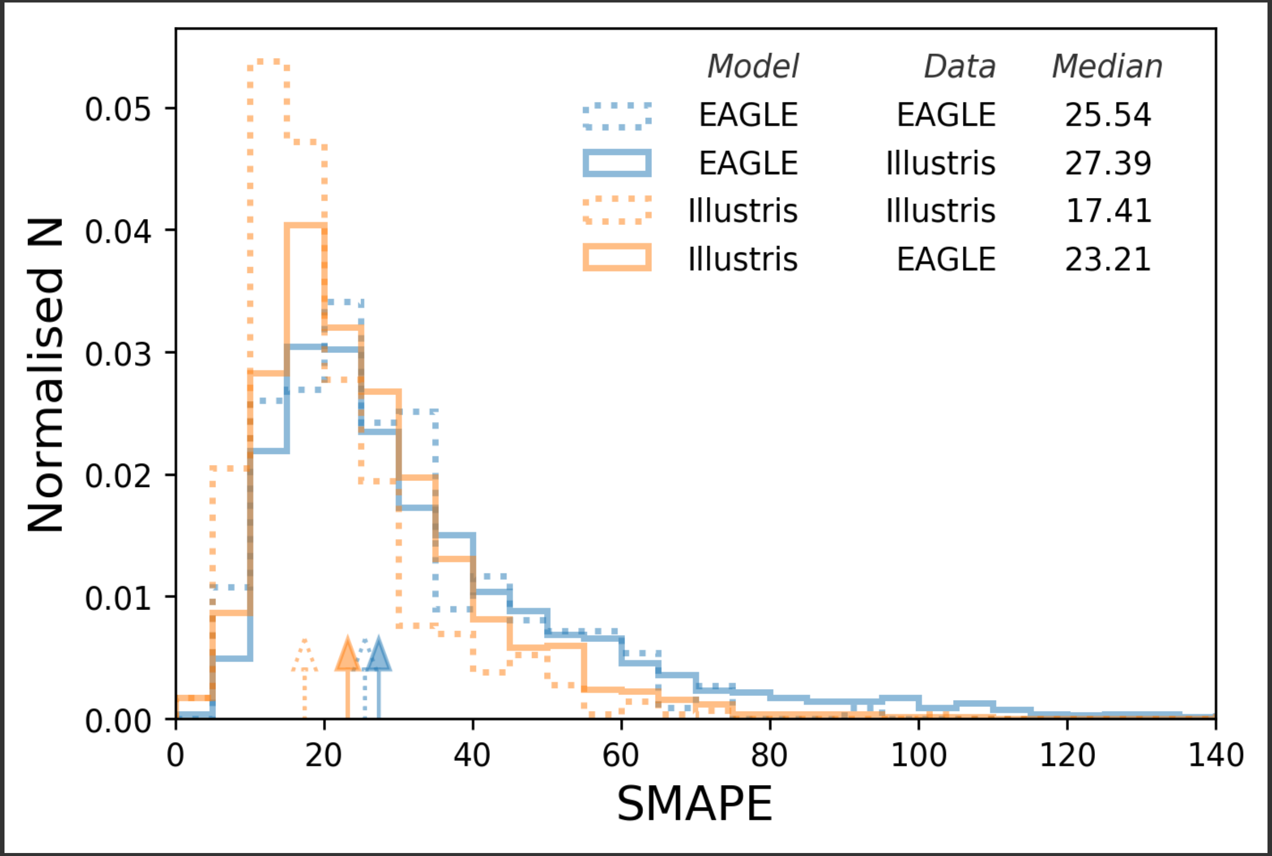
TEST RESIDUALS



Intra-model performance

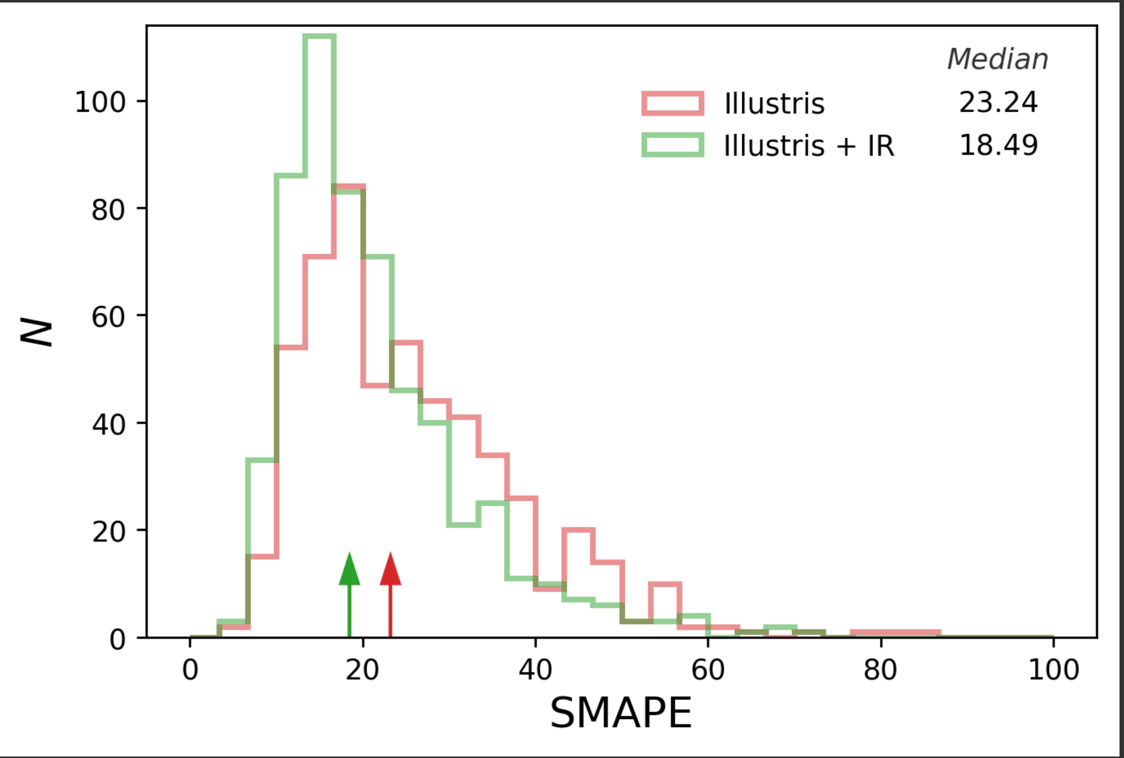
Model trained on one simulation then used to predict SFHs from another

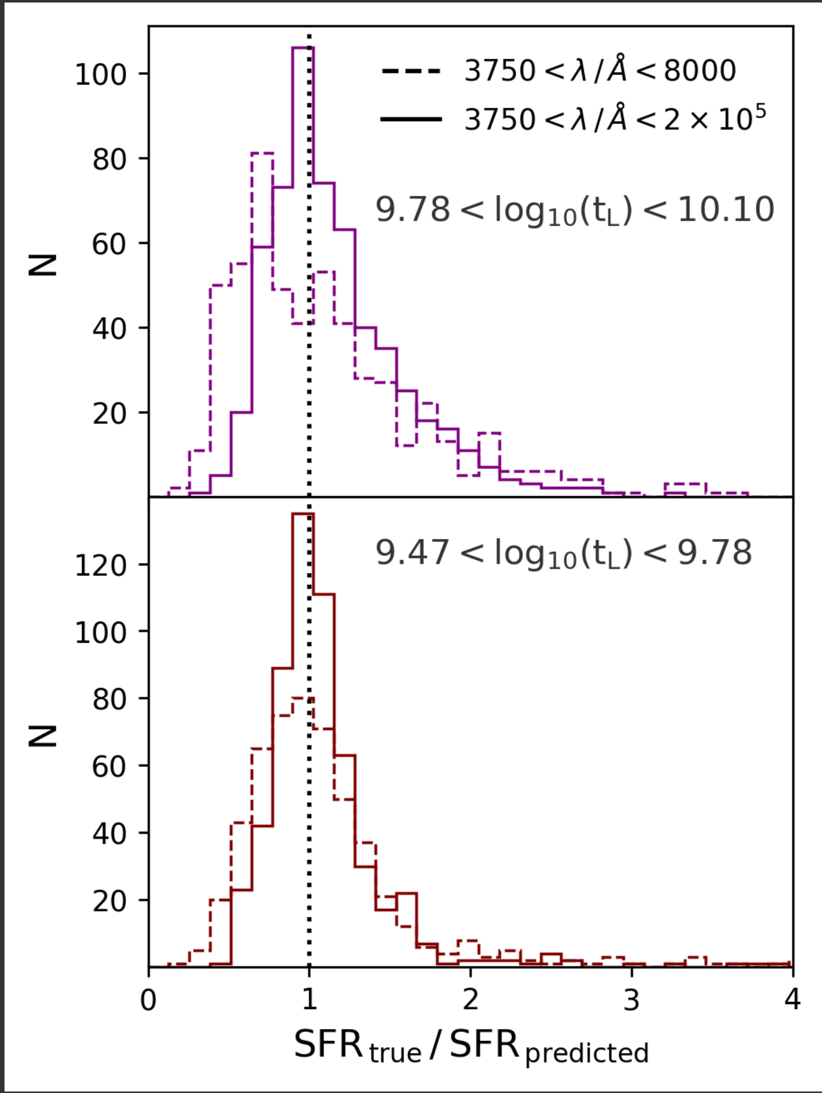
→ suggests we are learning the general relationship, and not overfitting to a single simulation



- Expanded wavelength range to NIR

  → leads to much improved fit,
  particularly for older stellar populations



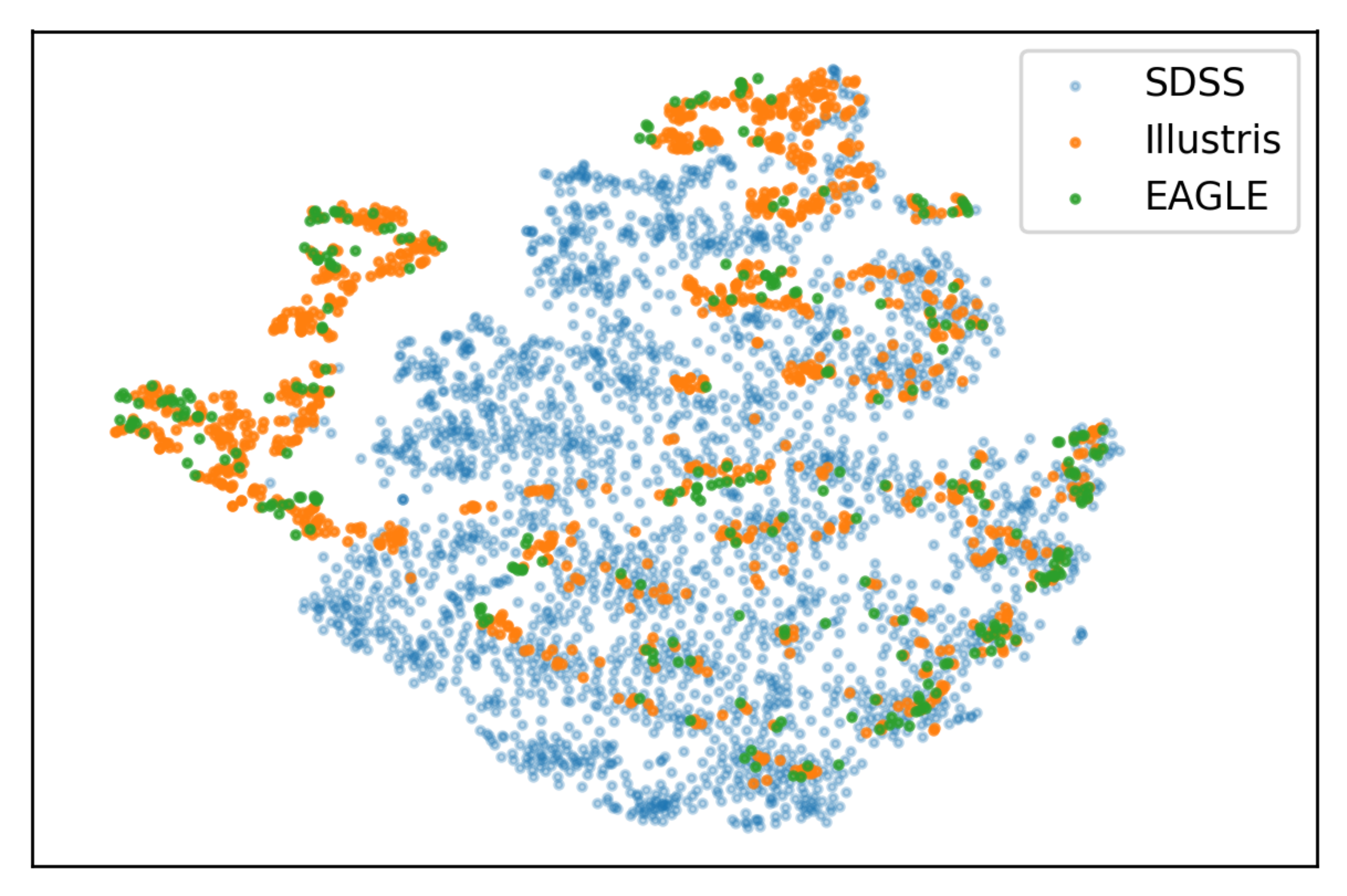


SDSS DR7

- Select sample based on g & r absolute magnitudes (colour + magnitude selection)
- 2400 galaxies
- t-distributed Stochastic Neighbour Embedding

t-SNE

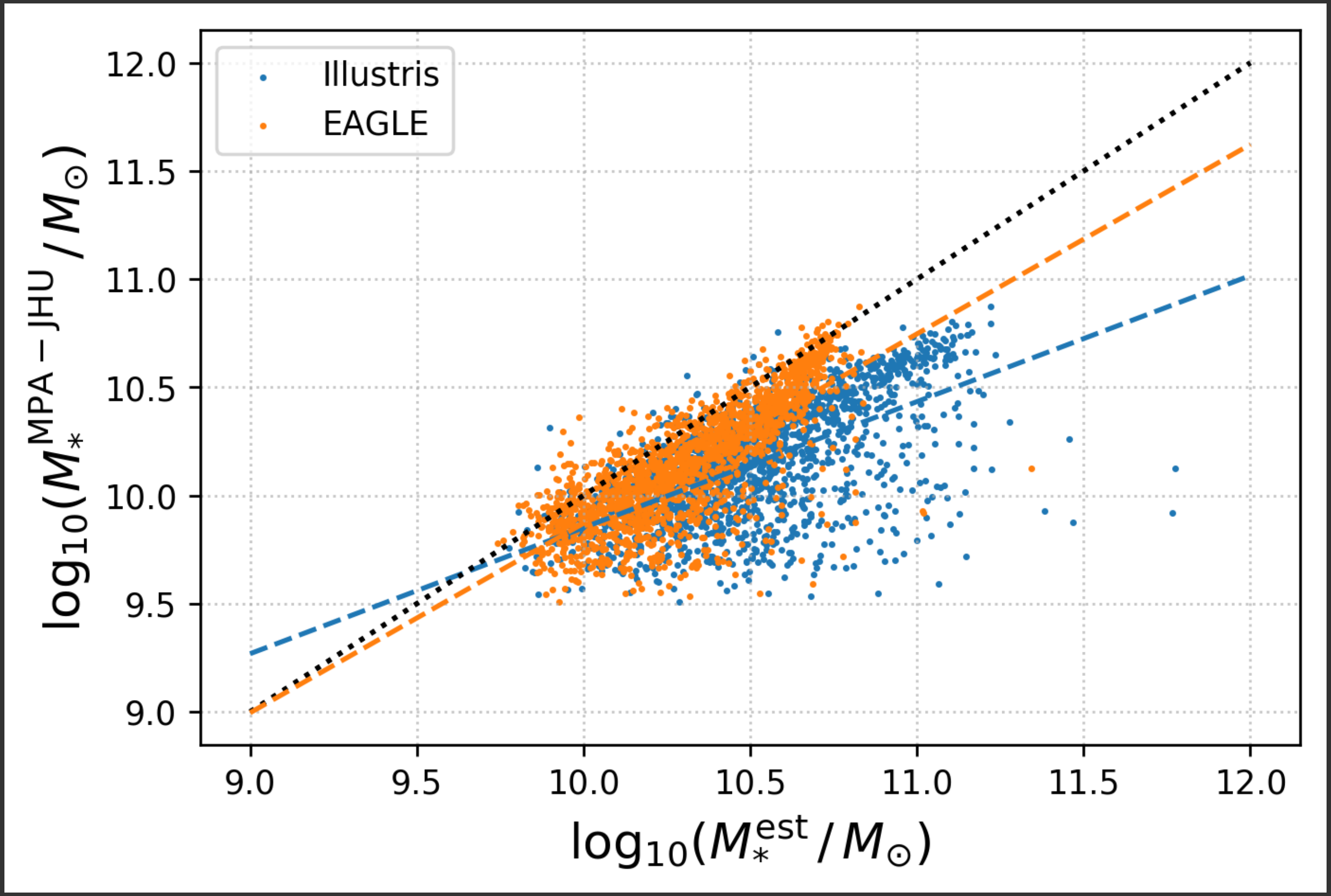
- Non-linear dimensionality reduction for visualisation



MPA-JHU MASS COMPARISON

Apply recycling fraction correction to SFH

Bias at higher masses



NEXT STEPS...

- Photometry using ERT
- More sophisticated dust modeling

  → Line of sight, e.g. LOSER (Davé+18)

  → full radiative transfer e.g. SKIRT (Camps+17, Trayford+17)
- Feature Importance
- More simulations (MUFASA, SAMs...)

Conclusions

- We have used supervised machine learning + cosmological simulations to estimate star formation histories
- We generated realistic spectra for EAGLE and Illustris simulations, including the effects of dust + nebular attenuation
- We achieved high accuracy in intra-simulation tests, suggesting good generalisation properties
- We estimate the error contribution from both the spectra and the model
- We applied the model to SDSS DR7 data and compared to the VESPA catalogue

Thanks for listening!

Cosmological Hydrodynamic Simulations

EAGLE *Schaye+14*

- Smoothed Particle Hydrodynamics (GADGET-3)
- Pressure-dependent star formation recipe
- 100 Mpc^3

Illustris *Genel+14*

- Adaptive Mesh Refinement (AREPO)
- Fixed density-dependent star formation recipe
- 106.5 Mpc^3

Typical gas element masses ~ 10^6 solar masses
Subgrid models for stellar and AGN feedback