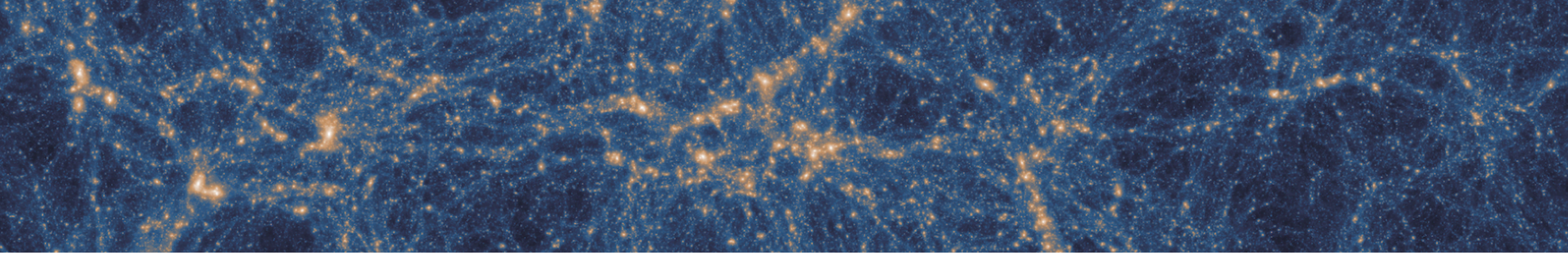
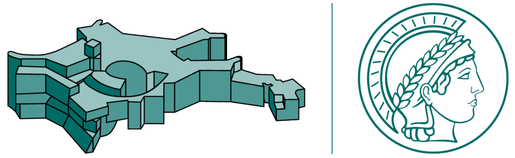


Explainable deep learning models for cosmological structure formation

Luisa Lucie-Smith
Postdoctoral Research Fellow @ Max-Planck-Institute for Astrophysics, Garching

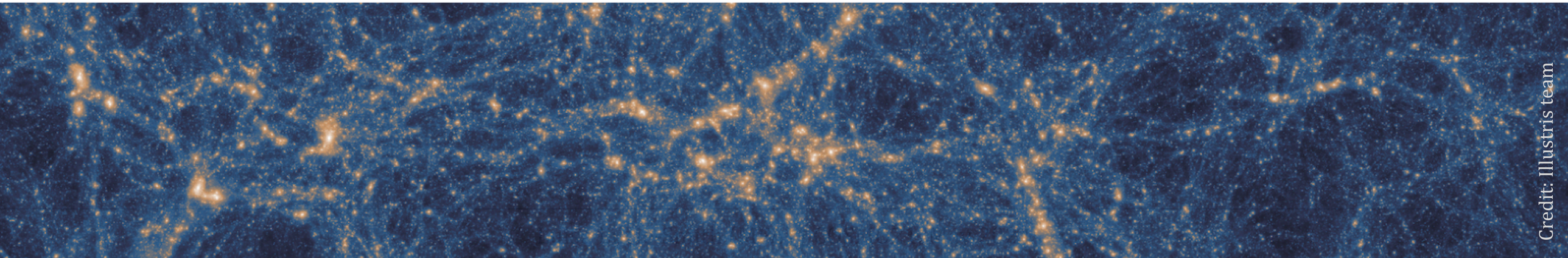
COSMO21 Statistical Challenges in 21st Century Cosmology
Chania, 23rd May 2024



MAX PLANCK INSTITUTE
FOR ASTROPHYSICS

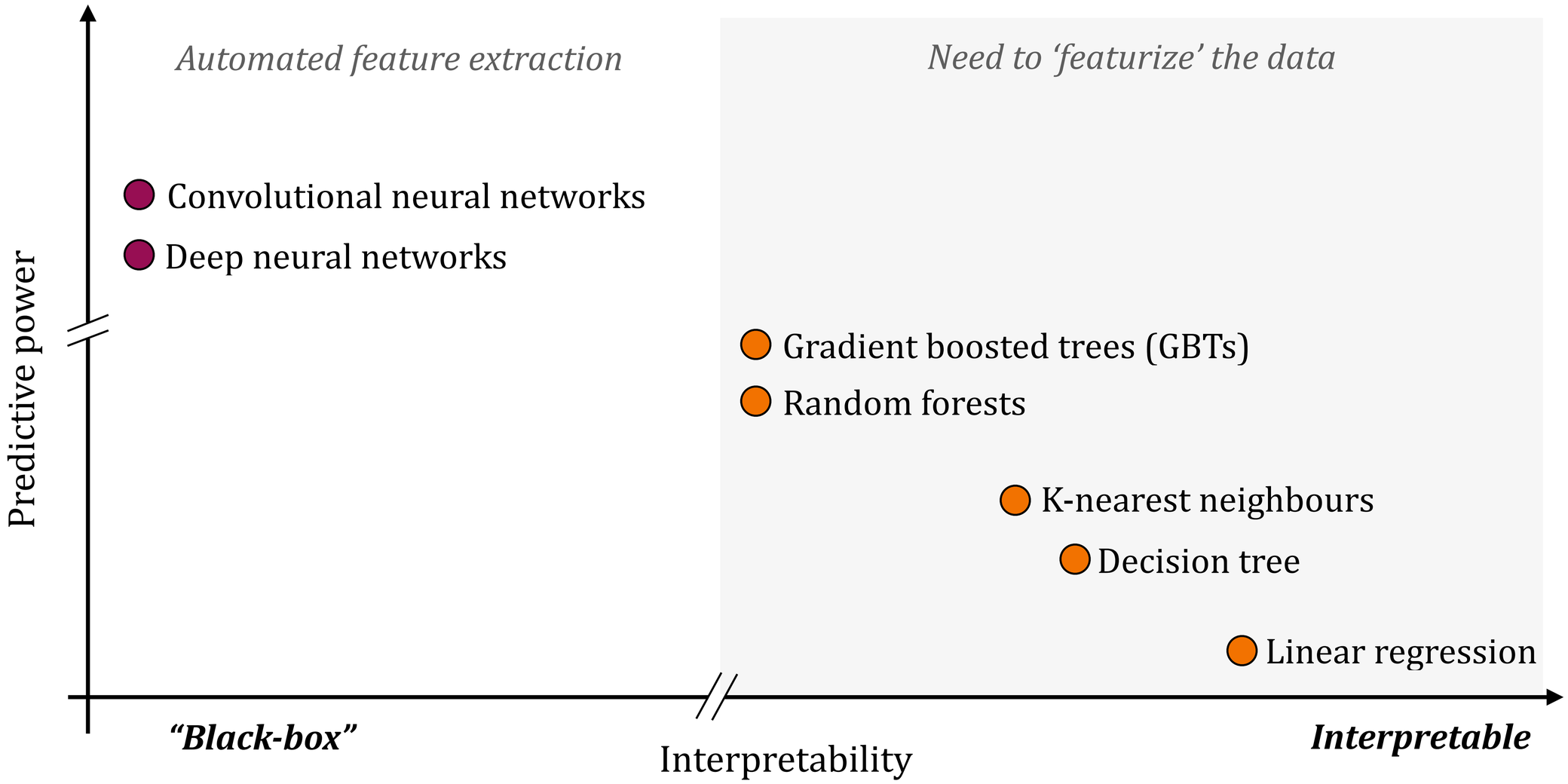
Machine learning in Astrophysics

- *Automating/accelerating known physical models (e.g. emulators)*
- *Enabling new parameter inference paradigm via simulation based inference*
- *Can we **extract new knowledge** about the underlying physics from deep learning models by interpreting their outputs? Requires **explainable AI***



Credit: Illustris team

Current landscape for explainable AI

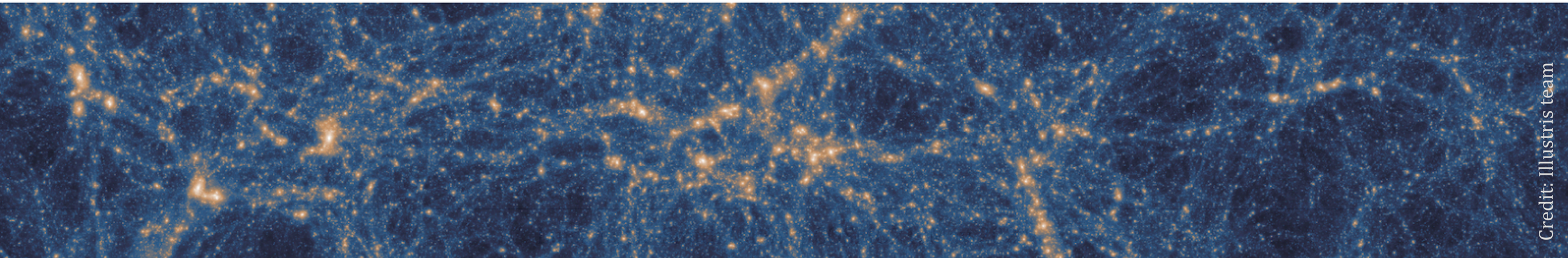


Requirements for explainable AI

1. ***Interpretability**:* account for why the ML model reaches its predictions

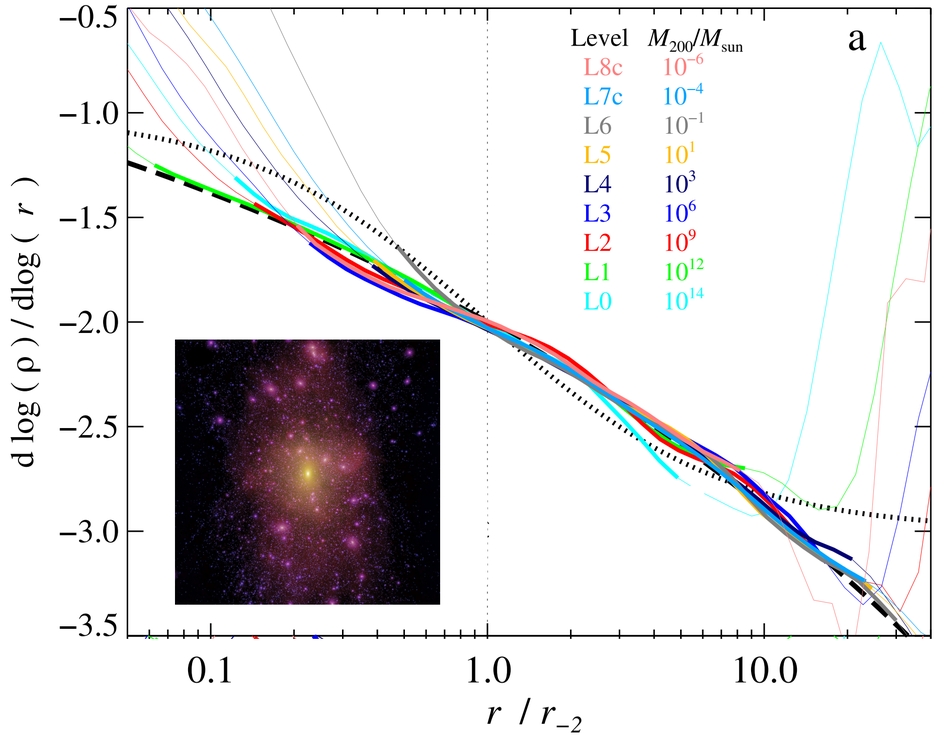
2. ***Explainability**:* map this account onto existing knowledge in the relevant science domain

N.B.: *many physical models in cosmology are also not explainable!*



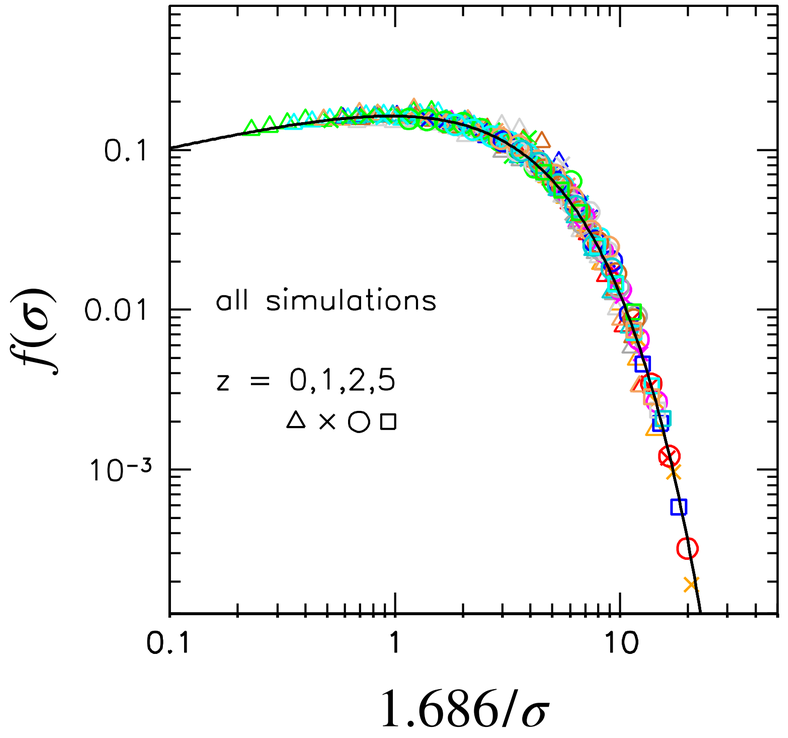
'Emergent' universal properties of the large-scale structure

Halo density profiles



Wang et al. (2020)

Halo mass function

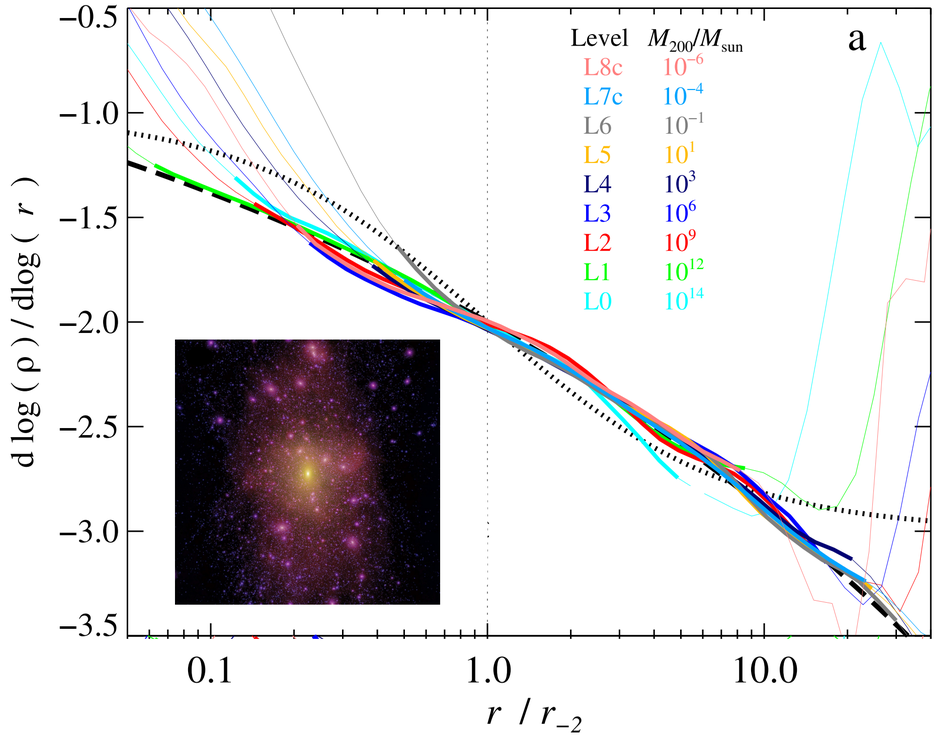


Despali et al. (2016)

*Challenging to **explain** how these emergent properties arise from simulations alone*

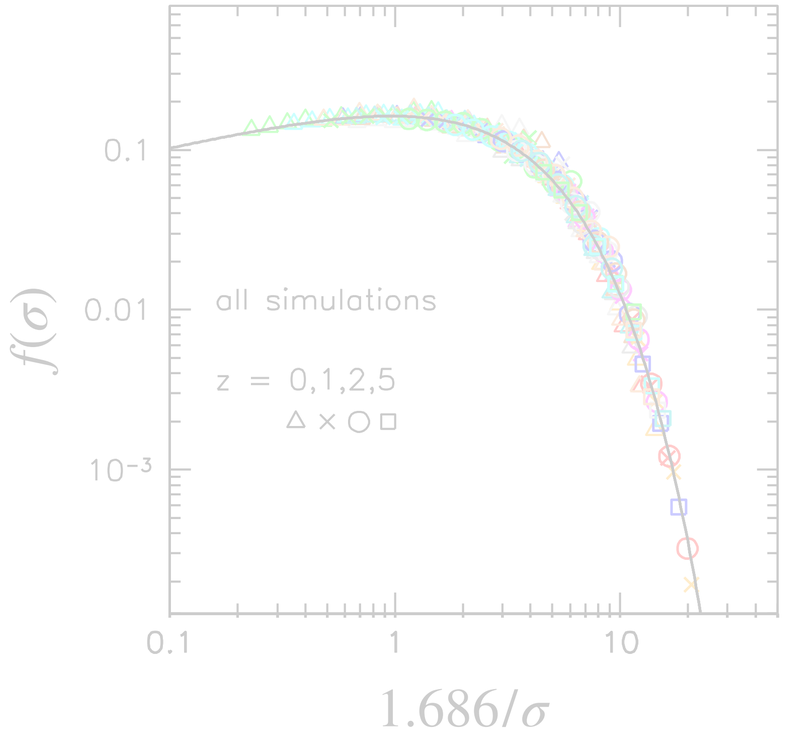
'Emergent' universal properties of the large-scale structure

Halo density profiles



Wang et al. (2020)

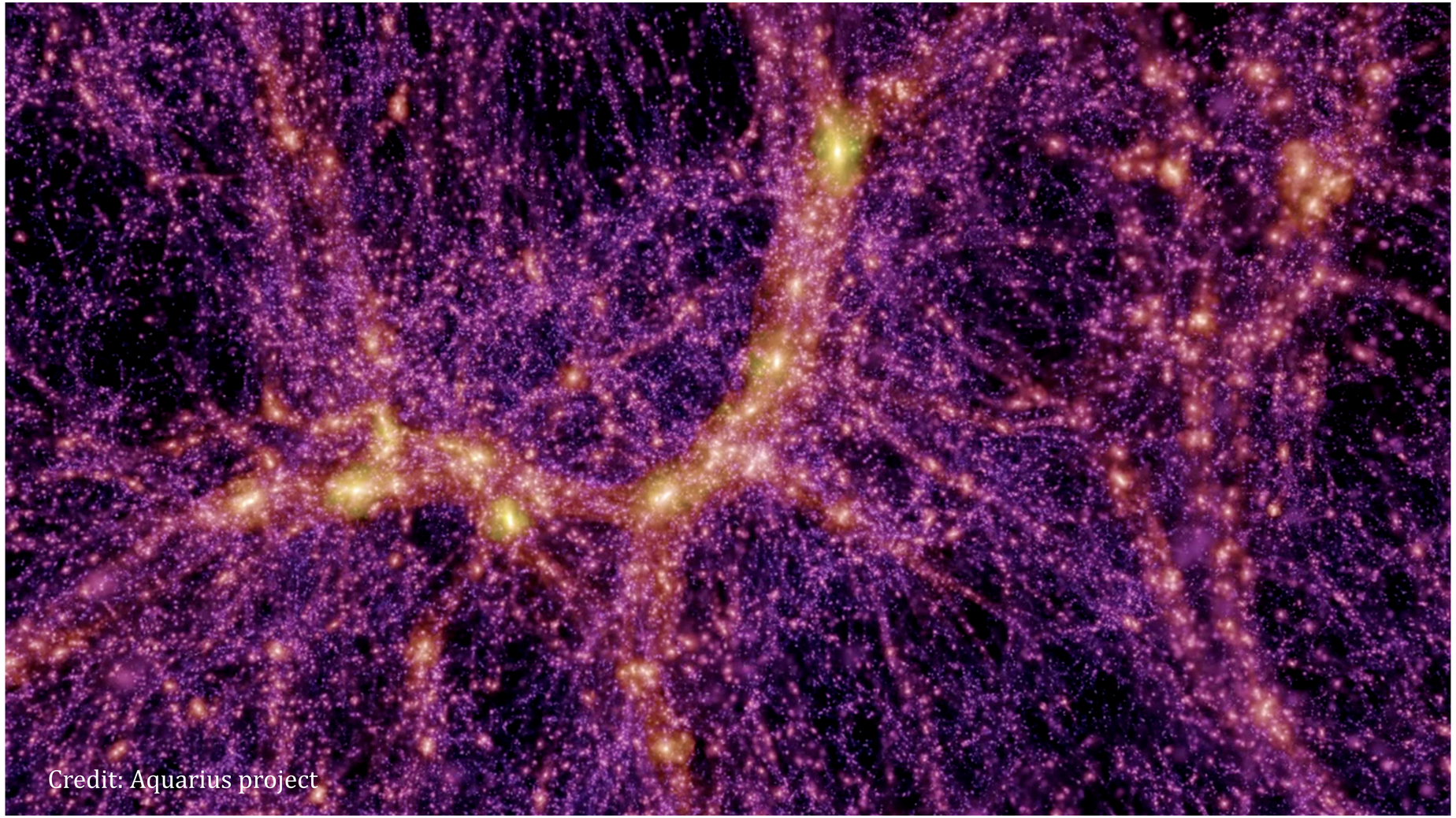
Halo mass function



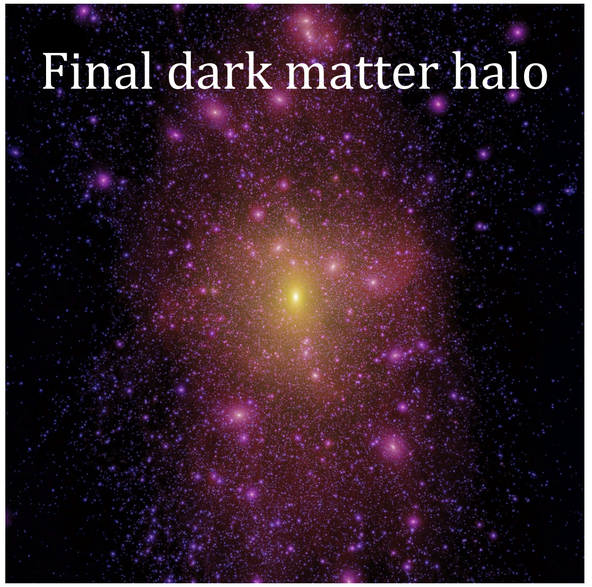
Despali et al. (2016)

*Challenging to **explain** how these emergent properties arise from simulations alone*

Building theoretical foundation: what determines the final structure of clusters?

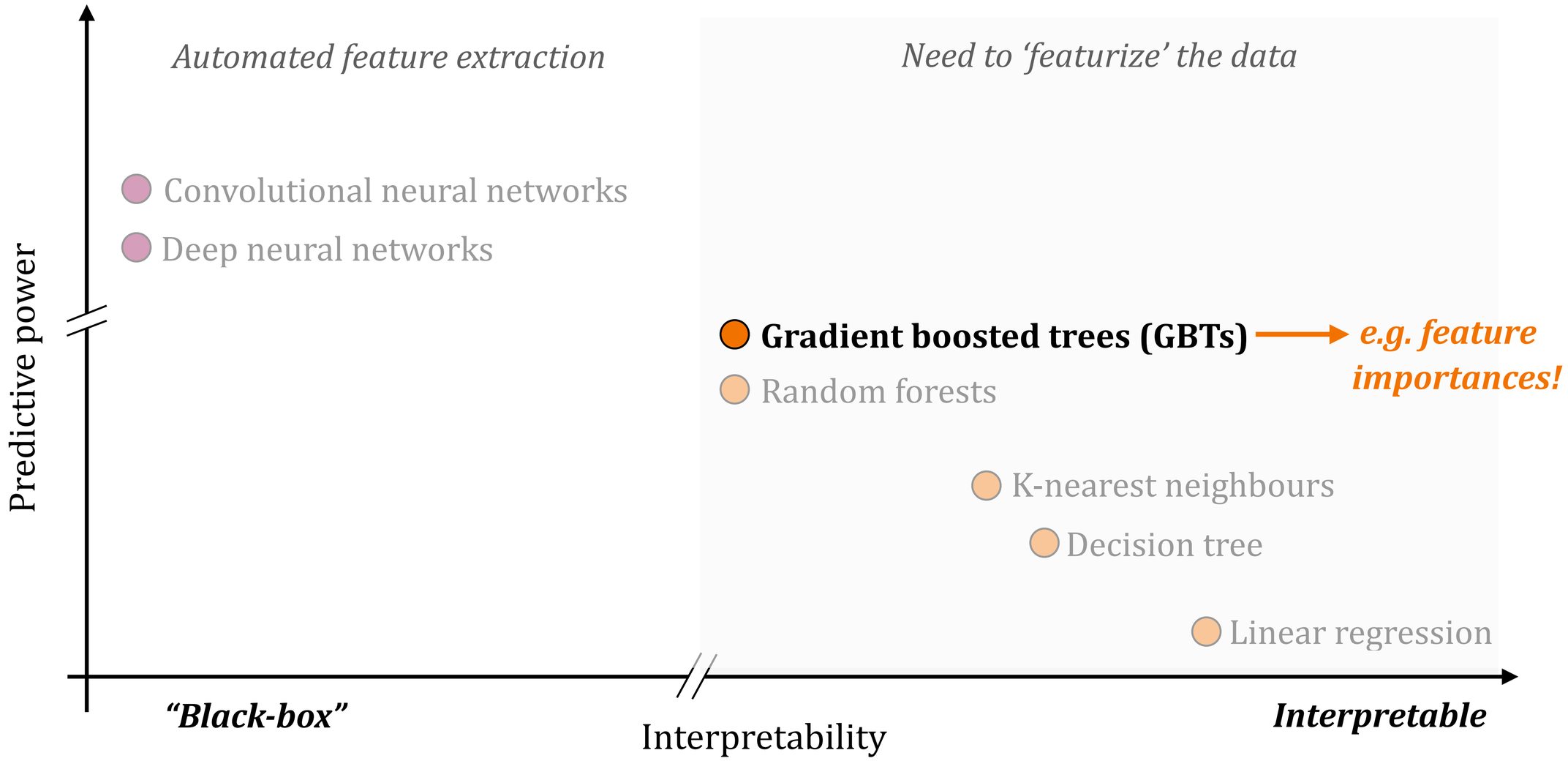


Credit: Aquarius project



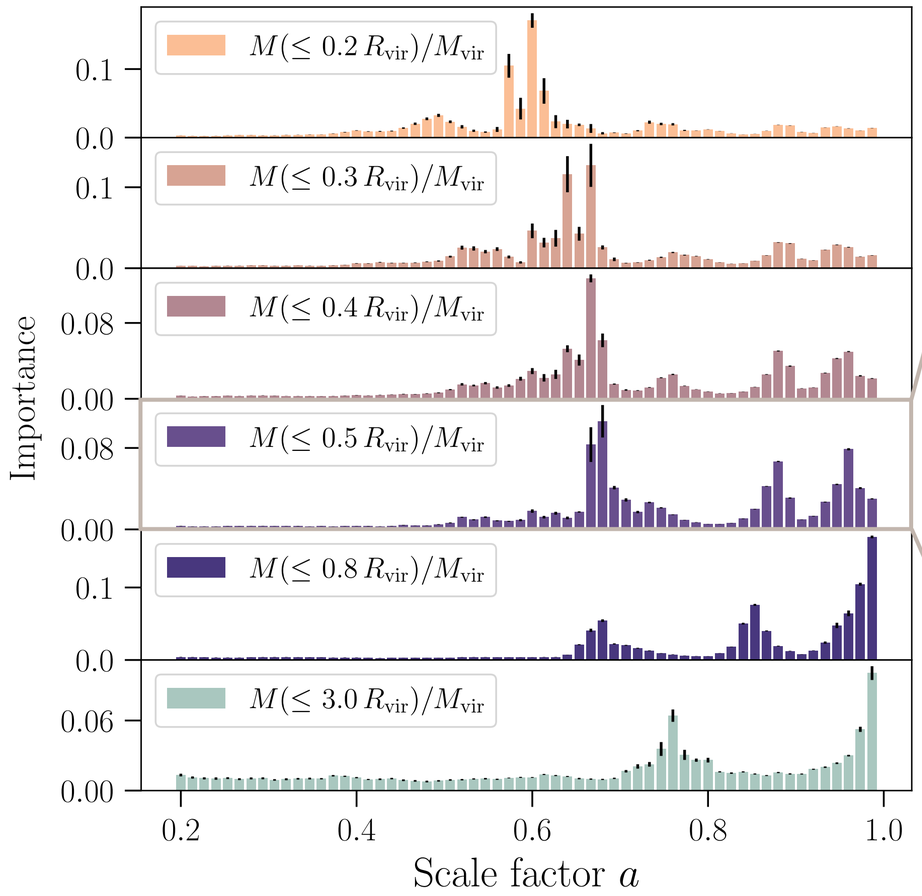
work with Susmita Adhikari (IISER) & Risa Wechsler (Stanford)

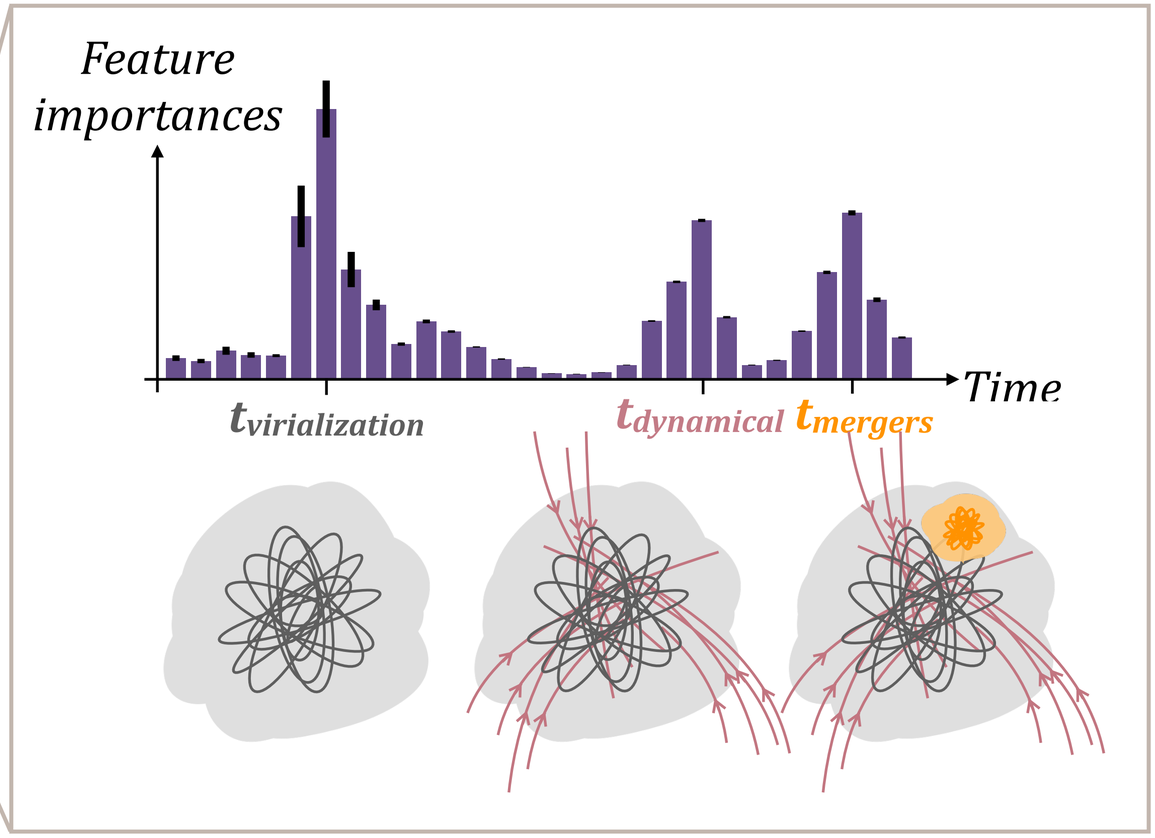
Gradient boosted trees allow for interpretability via feature importances



Three distinct timescales determine final halo structure

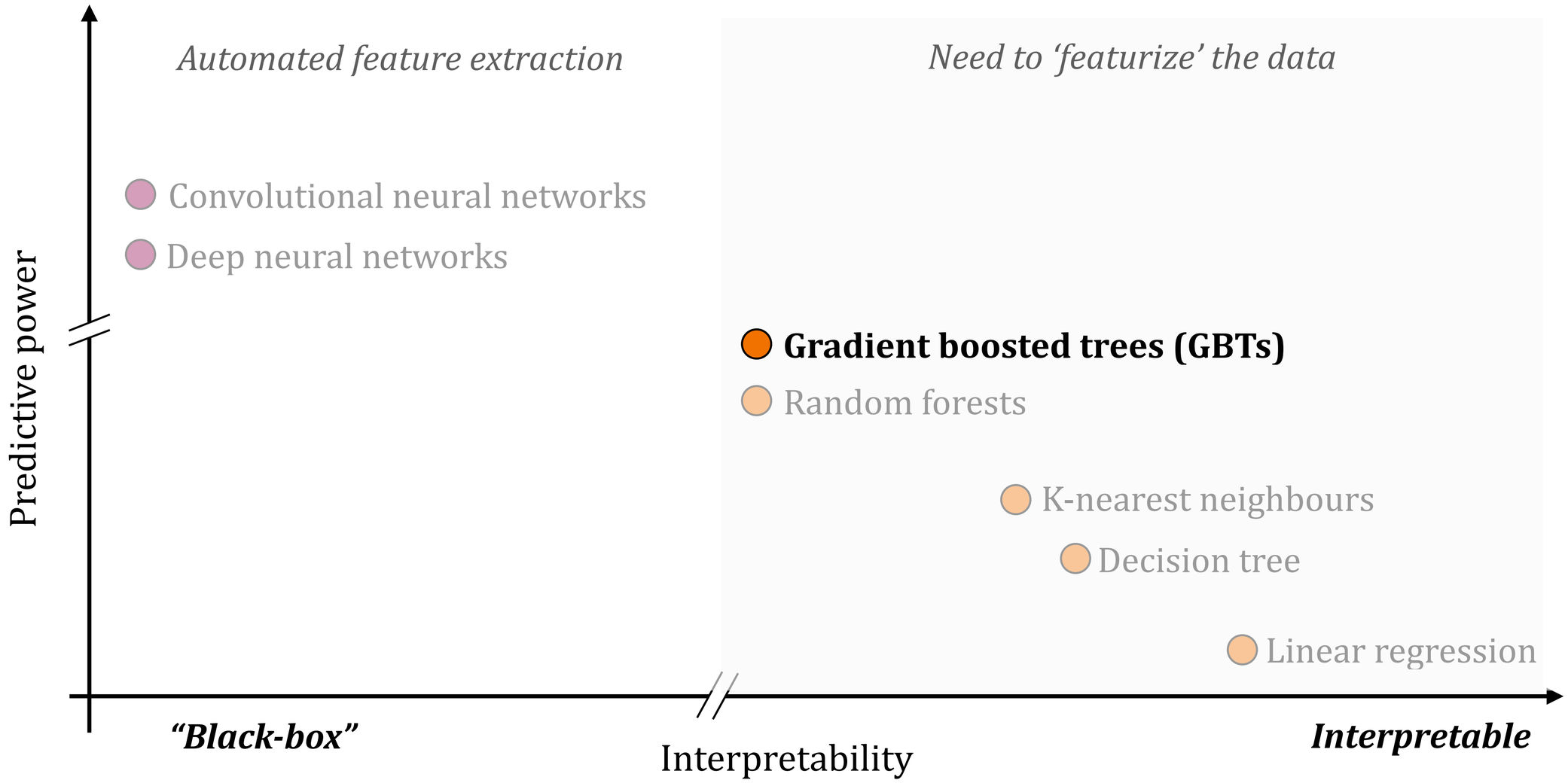
Input: Halo mass accretion history ⟶ GBT ⟶ *Output:* Final halo mass profile



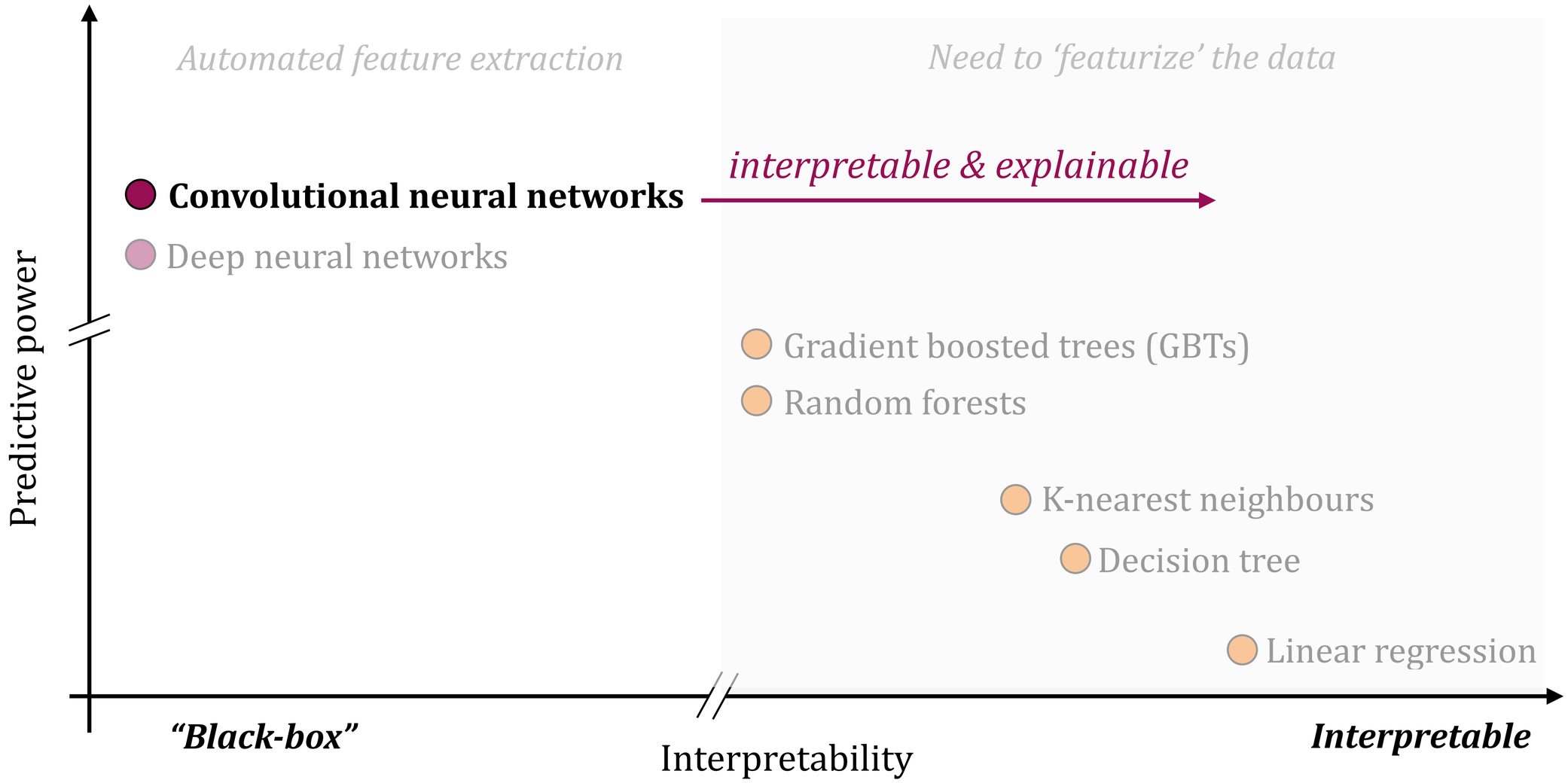


Lucie-Smith, Adhikari, Wechsler (MNRAS, 2022)

Limitation: requires humans to come up with 'features' of data

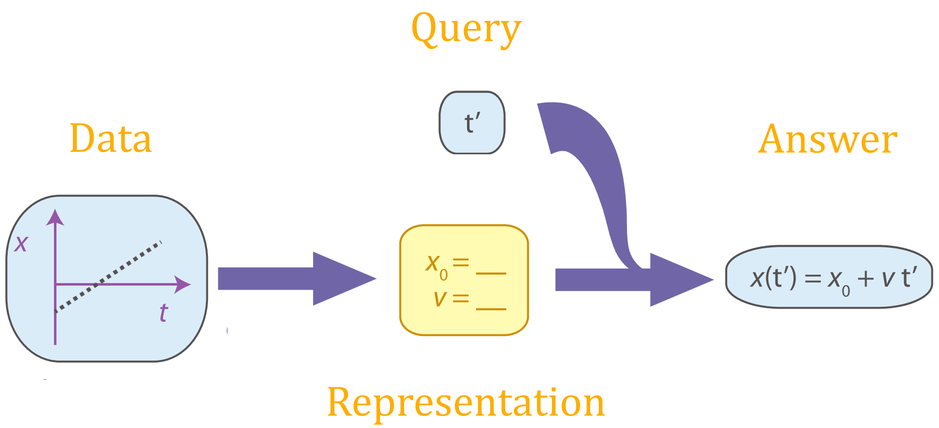


Knowledge extraction with deep learning

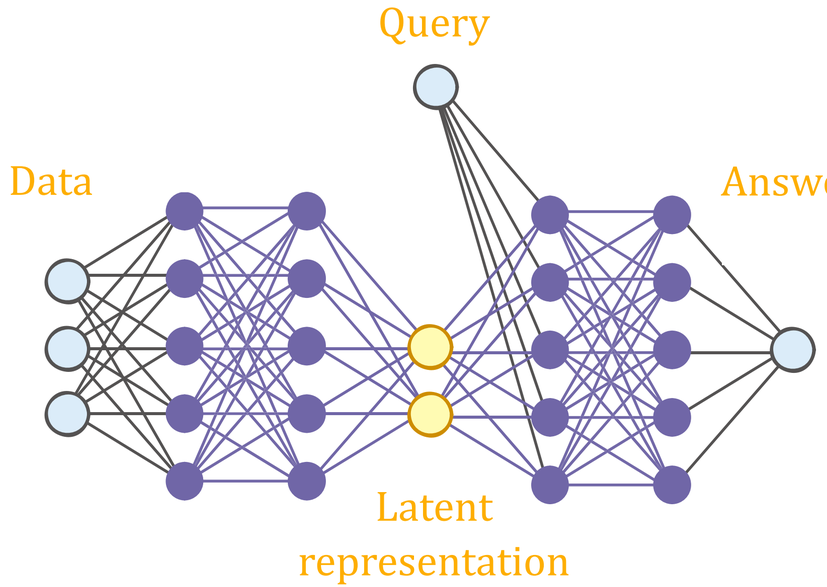


SciNet model

Human learning



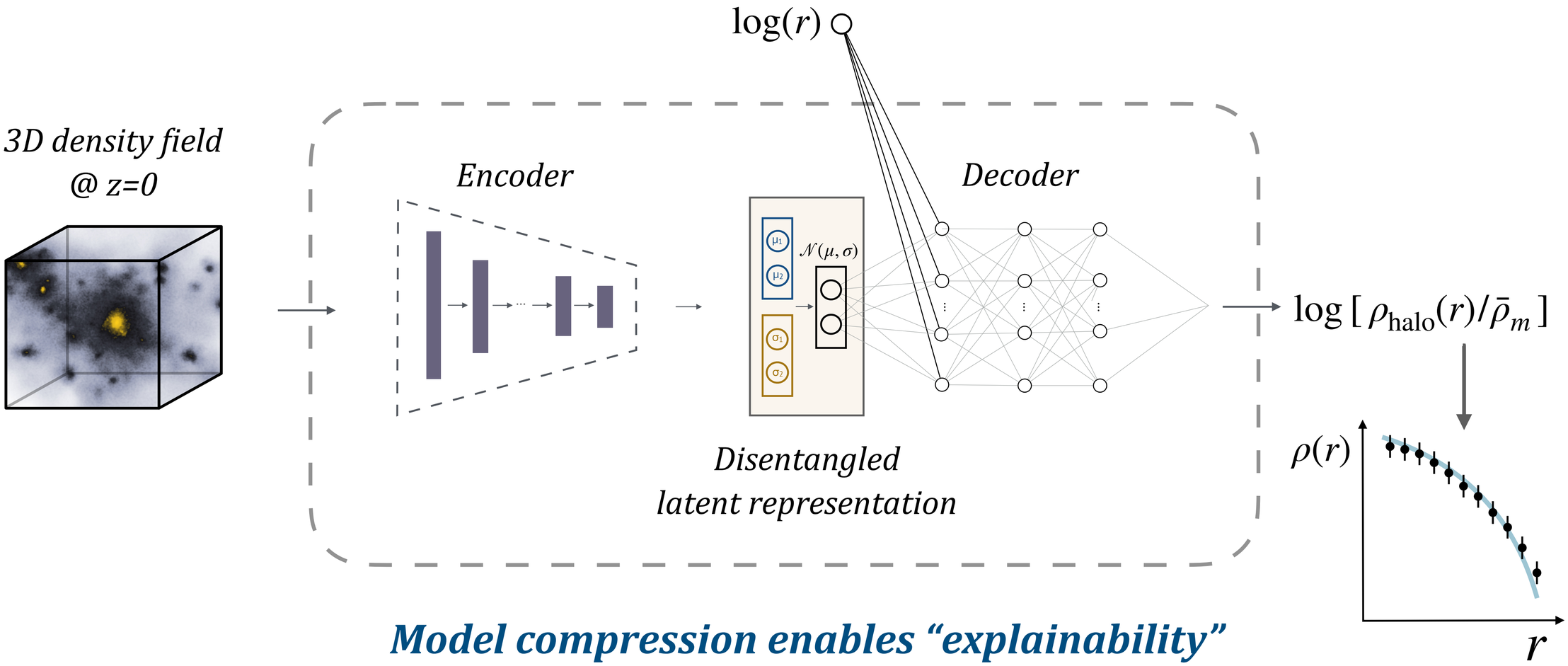
SciNet



- SciNet learns relevant physical parameters in toy 1D problems
- Relies on comparing latents with already-known physical parameters

Iten et al. (PRL, 2020)

An interpretable variational encoder (IVE) for halo density profiles out to their outskirts

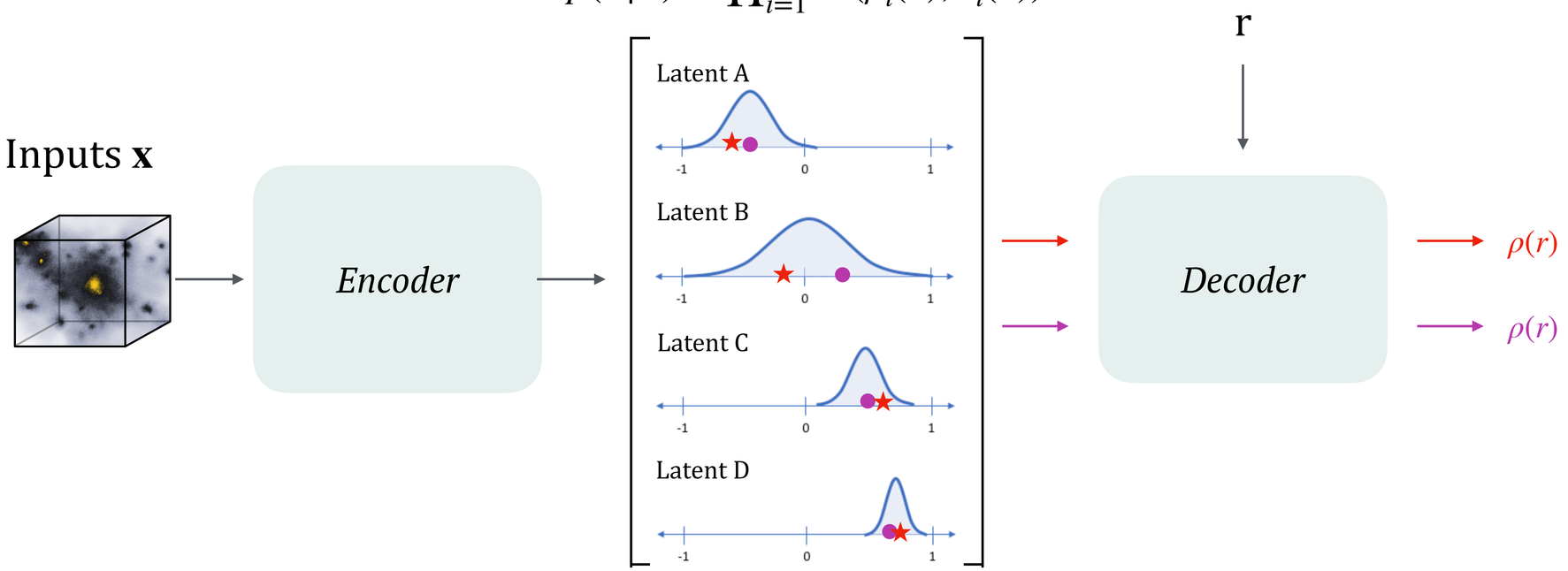


Model compression enables "explainability"

Lucie-Smith et al. (PRD, 2022); Lucie-Smith et al. (PRL, 2024)

Desired latent representation properties for interpretability

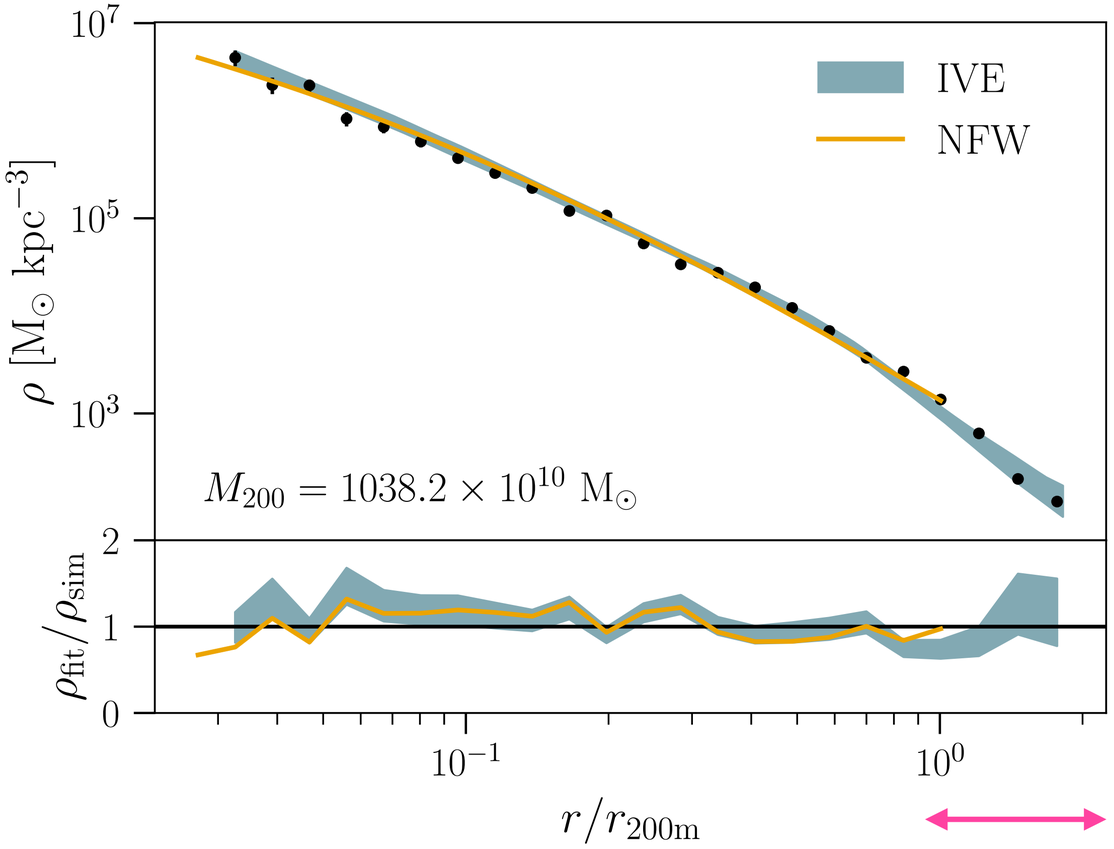
$$p(\mathbf{z}\,|\,\mathbf{x}) = \prod_{i=1}^{L} \mathcal{N}(\mu_i(\mathbf{x}), \sigma_i(\mathbf{x}))$$

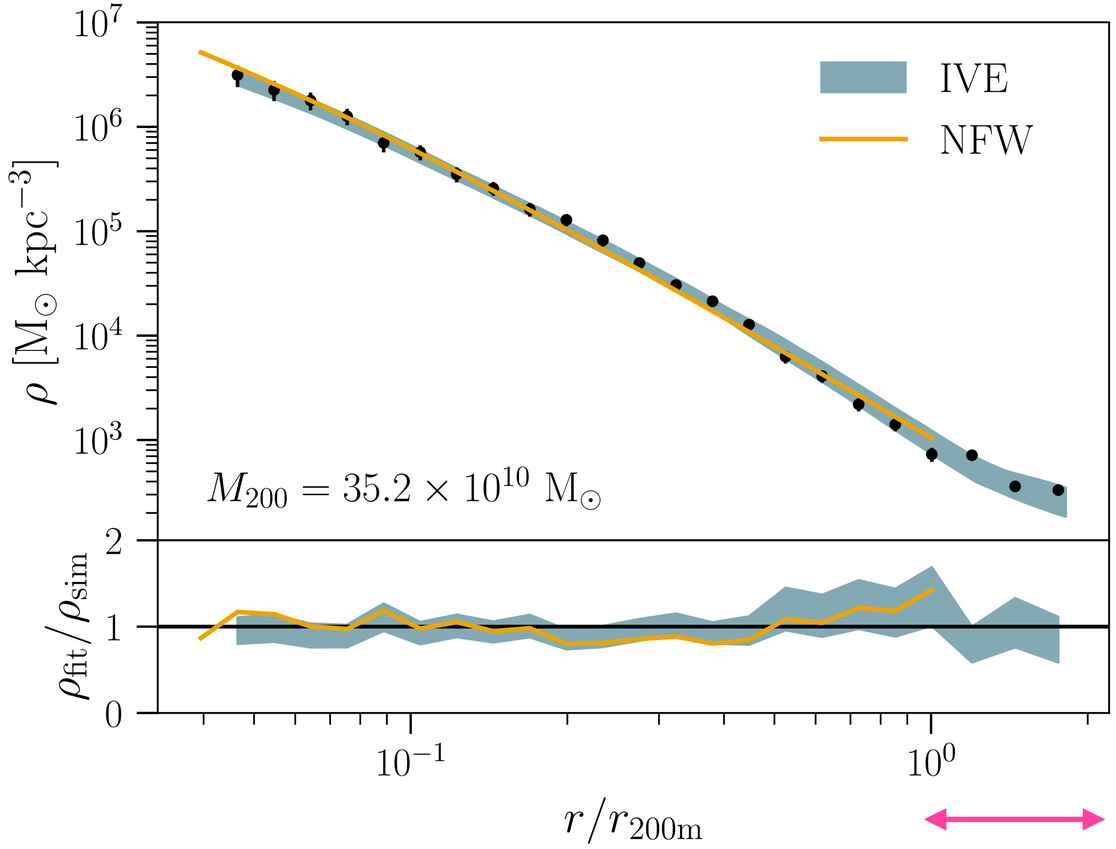


- **Interpretability** can be achieved if latent space is **disentangled:** independent factors of variation in profiles captured by different, independent latents
- Disentanglement encouraged via **loss function** optimised during training

Examples of fits produced by the interpretable variational encoder

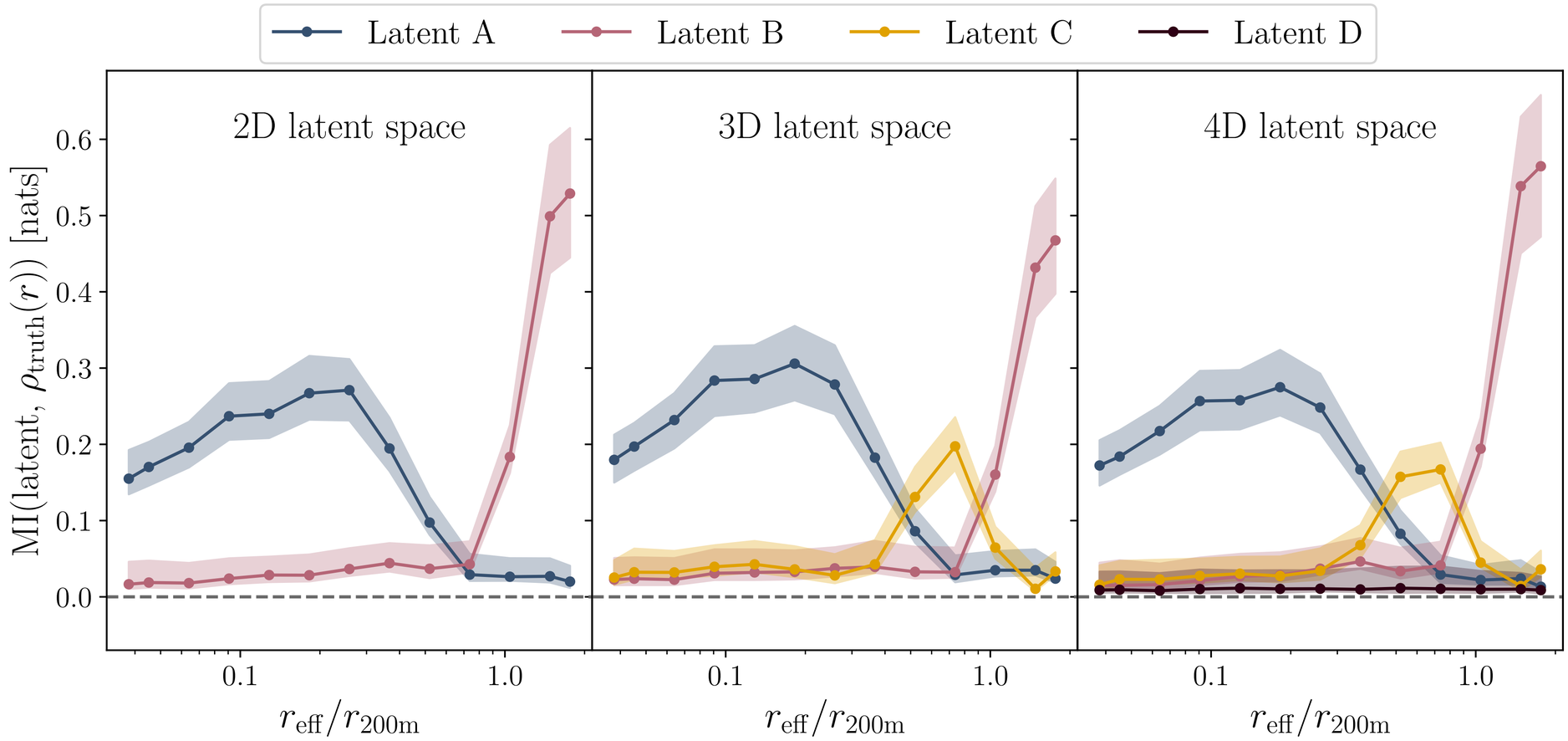
We fit halos of a broad range in mass (10^{11}— $10^{14}\,\mathrm{M}_\odot/h$) at z=0





Halo outskirts exhibit cosmologically interesting features related to the 'splashback radius'

Interpreting the latent representation using mutual information



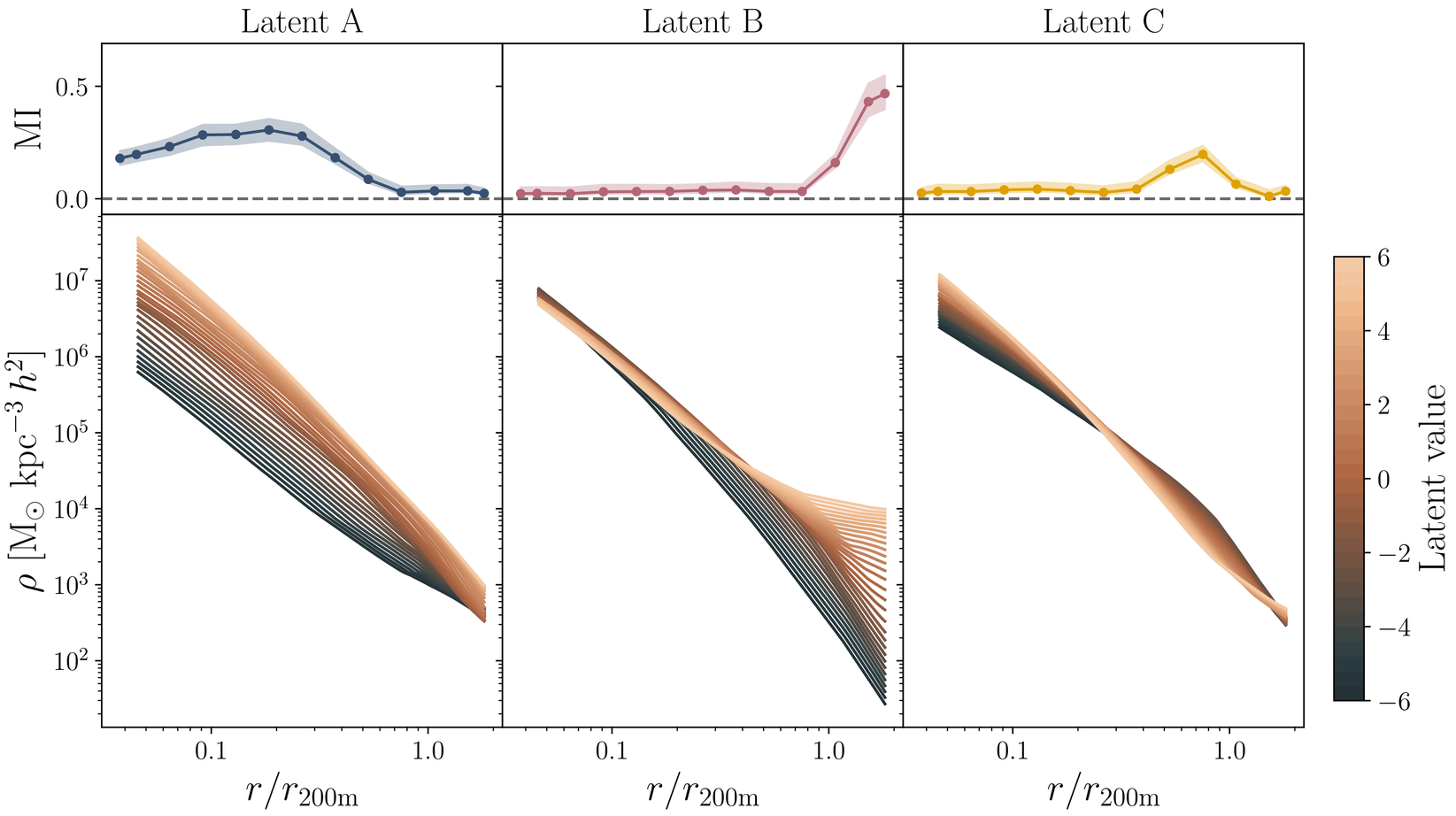
MI estimator:

HTTPS://GITHUB.COM/DPIRAS/GMM-MI

Piras, Peiris, Pontzen, Lucie-Smith et al. (MLST, 2023)

Lucie-Smith, Peiris, Pontzen, Nord et al. (Phys. Rev. D, 2022)

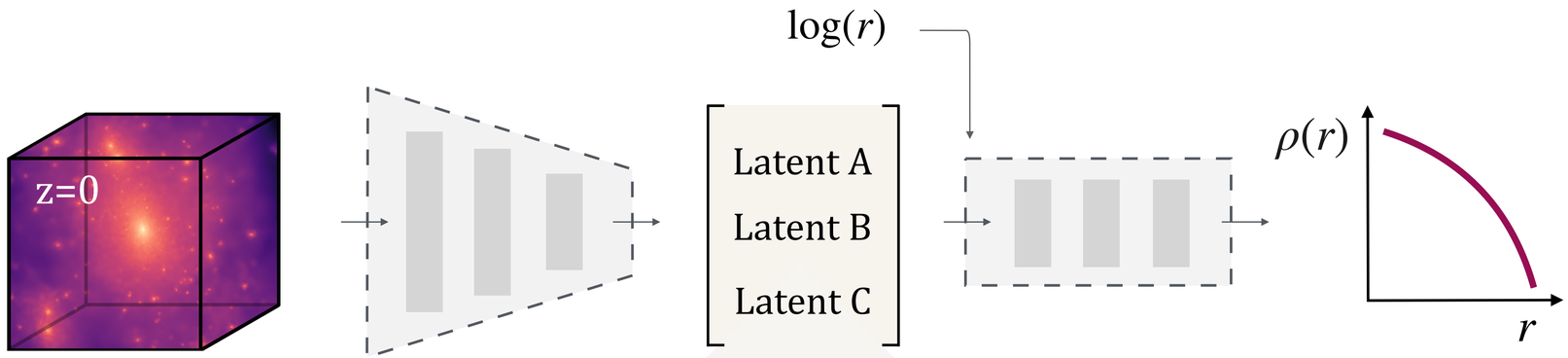
Systematically varying one latent at a time



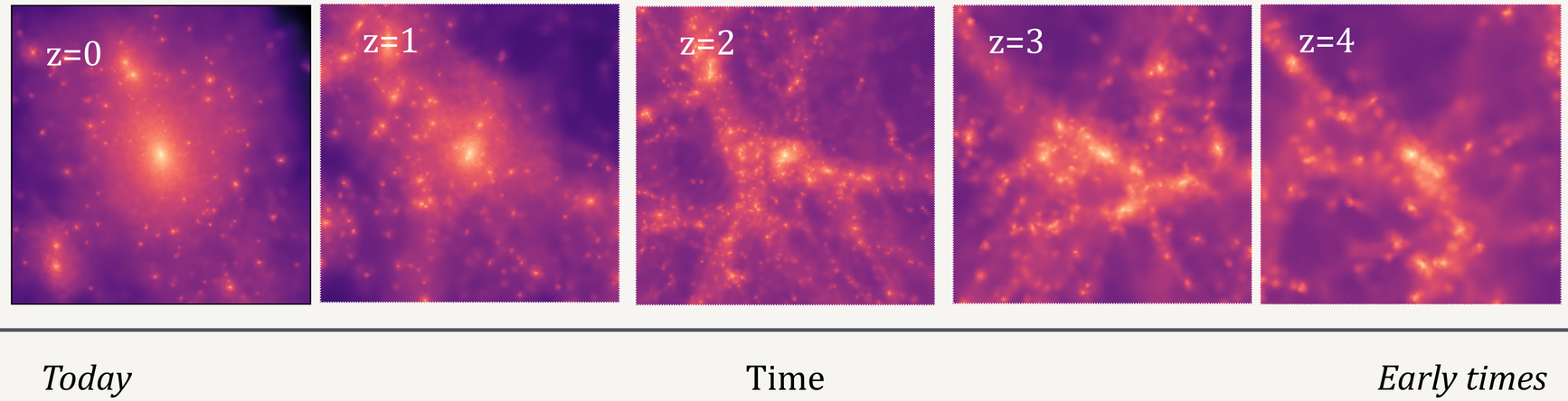
Latent A = ***normalisation***; Latent B = ***outer slope;*** Latent C = ***inner slope***

Lucie-Smith, Peiris, Pontzen, Nord et al. (Phys. Rev. D, 2022)

Exploiting the latent representation beyond its original training task

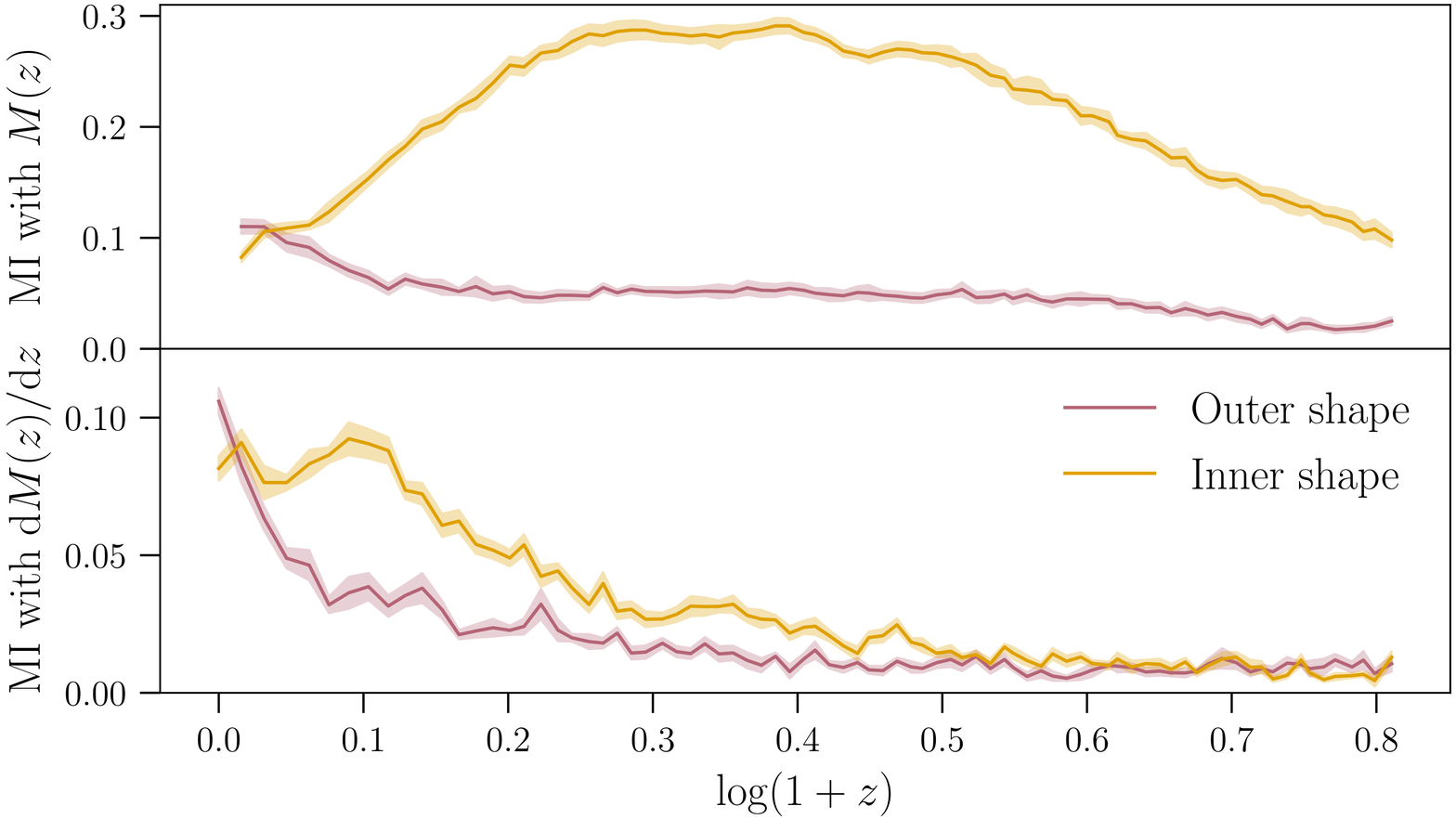


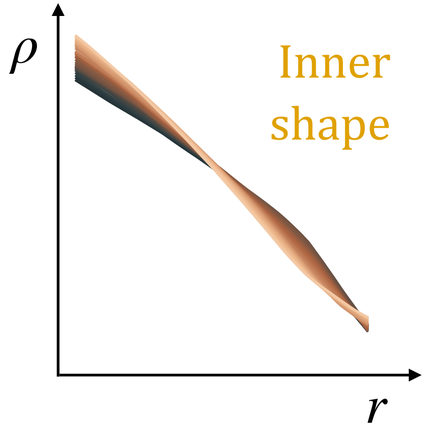
Does the latent space retain information about the origin of the halo density structure?

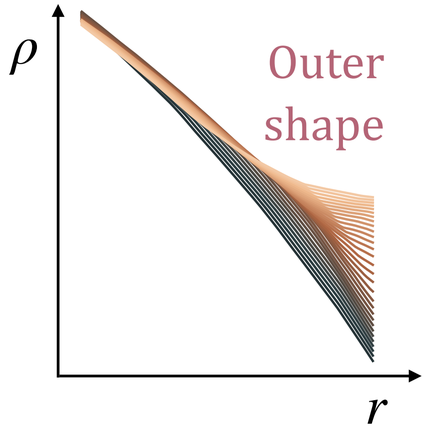


Lucie-Smith, Peiris, Pontzen (Phys. Rev. Lett., 2024)

Relation between the latents and the halo evolution history



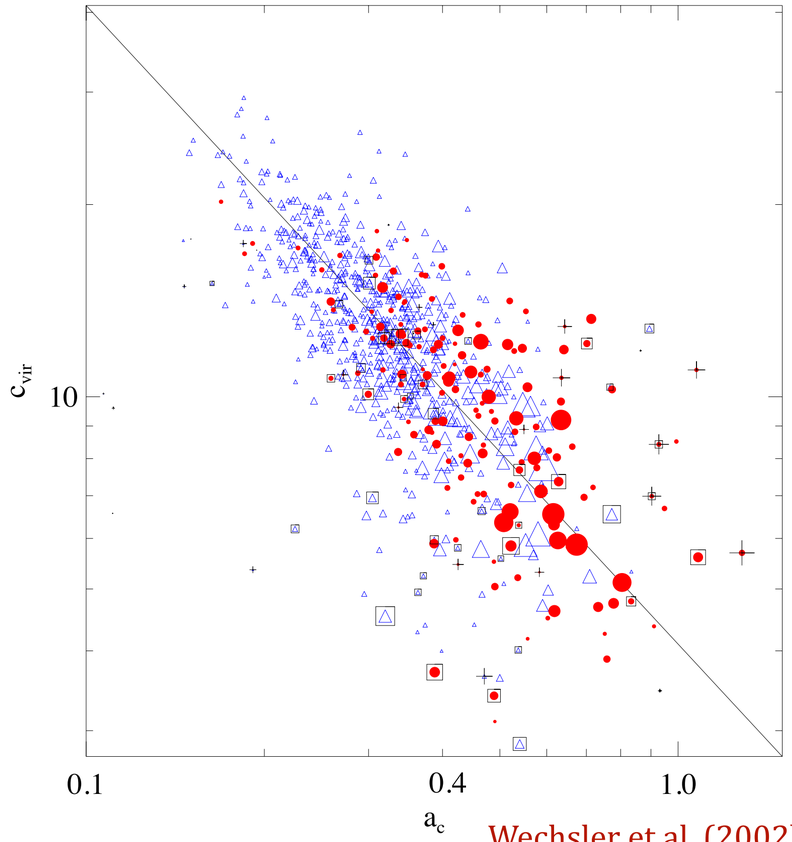




Lucie-Smith, Peiris, Pontzen (Phys. Rev. Lett., 2024)

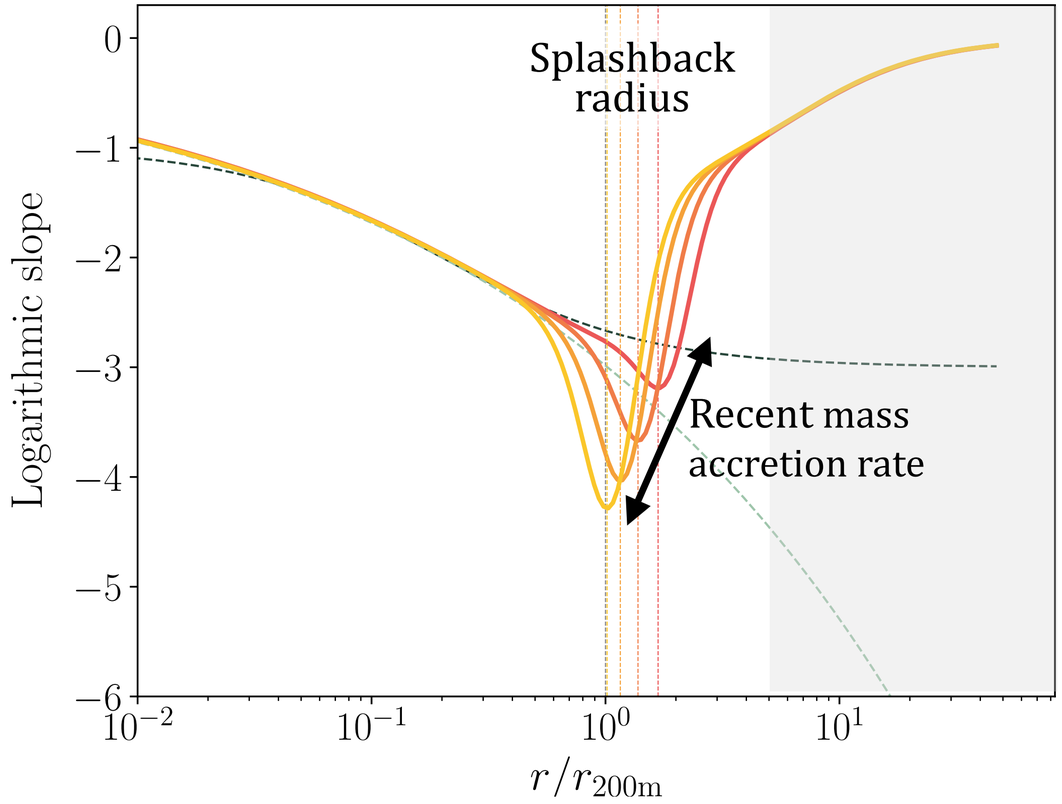
Connection to current explanation of density profiles

IVE recovers known relation between ***inner profile*** and ***early assembly history***

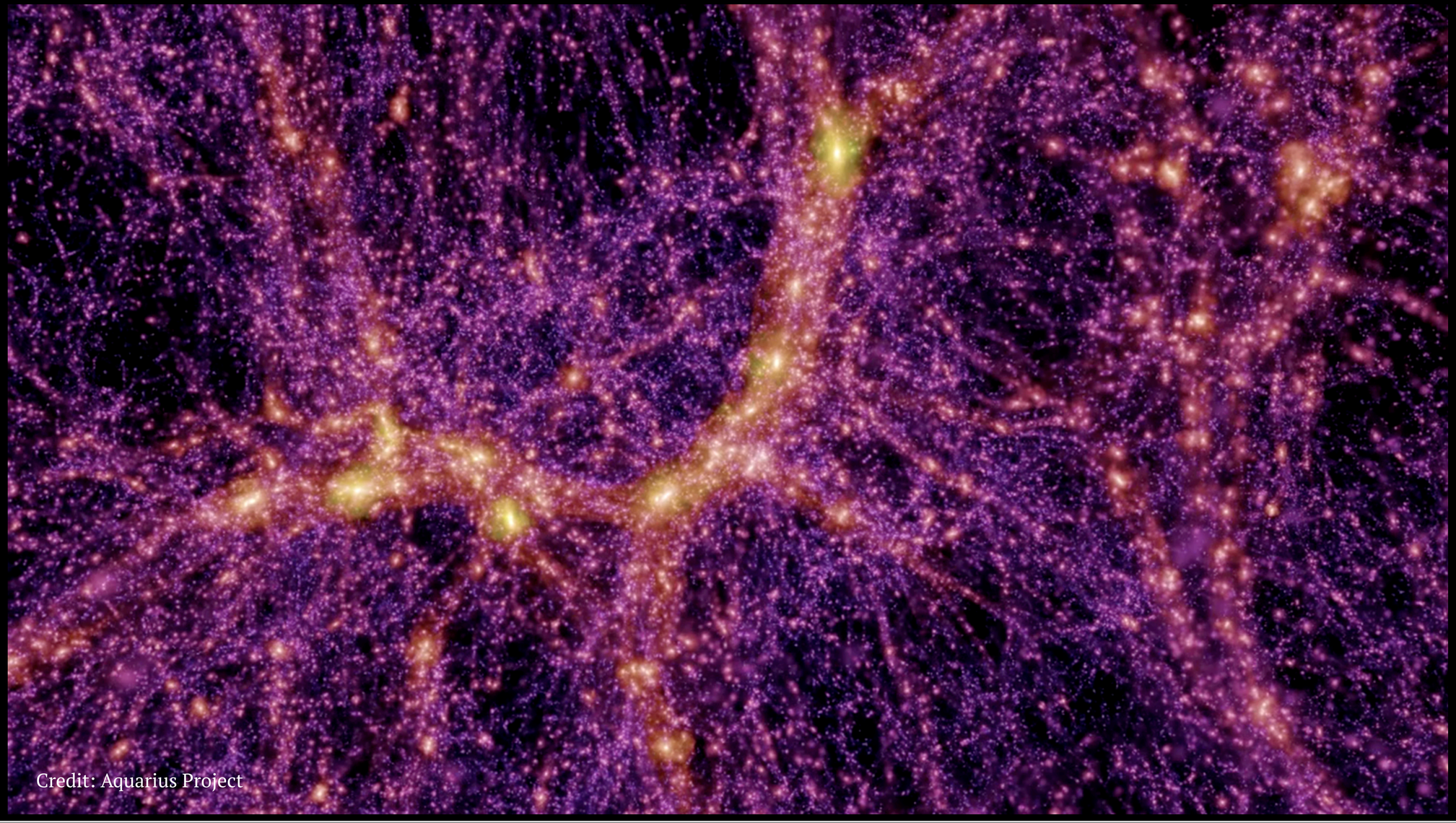


Wechsler et al. (2002)

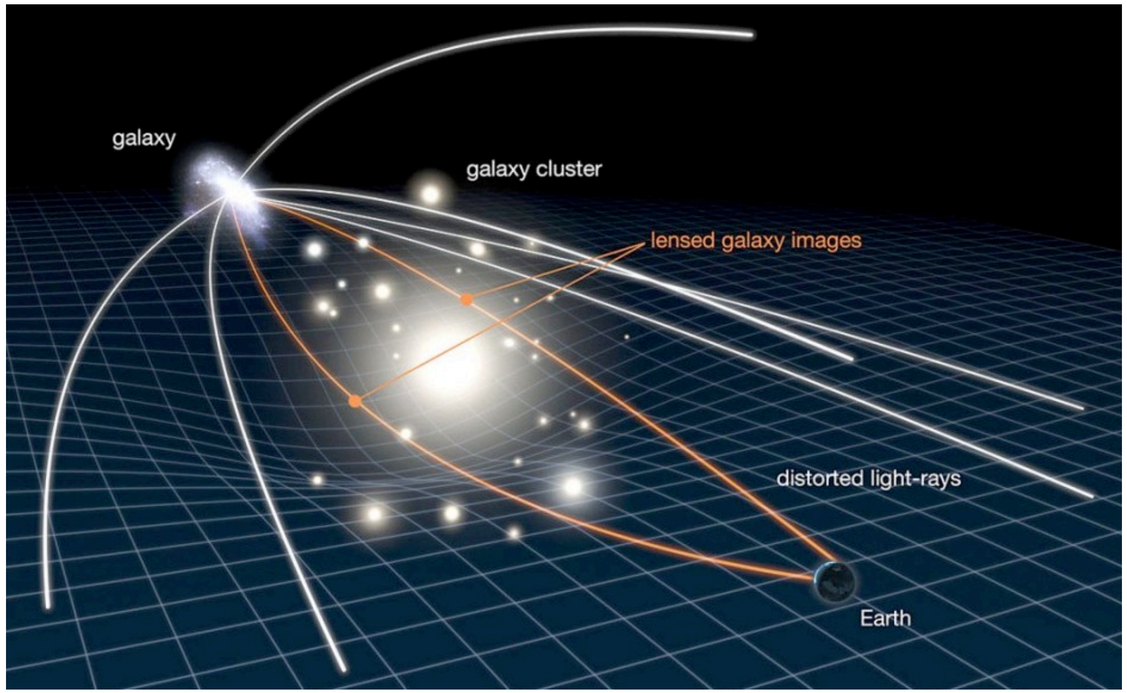
IVE discovers that ***outer profile*** depends on ***only one component*** related to ***most recent accretion rate***

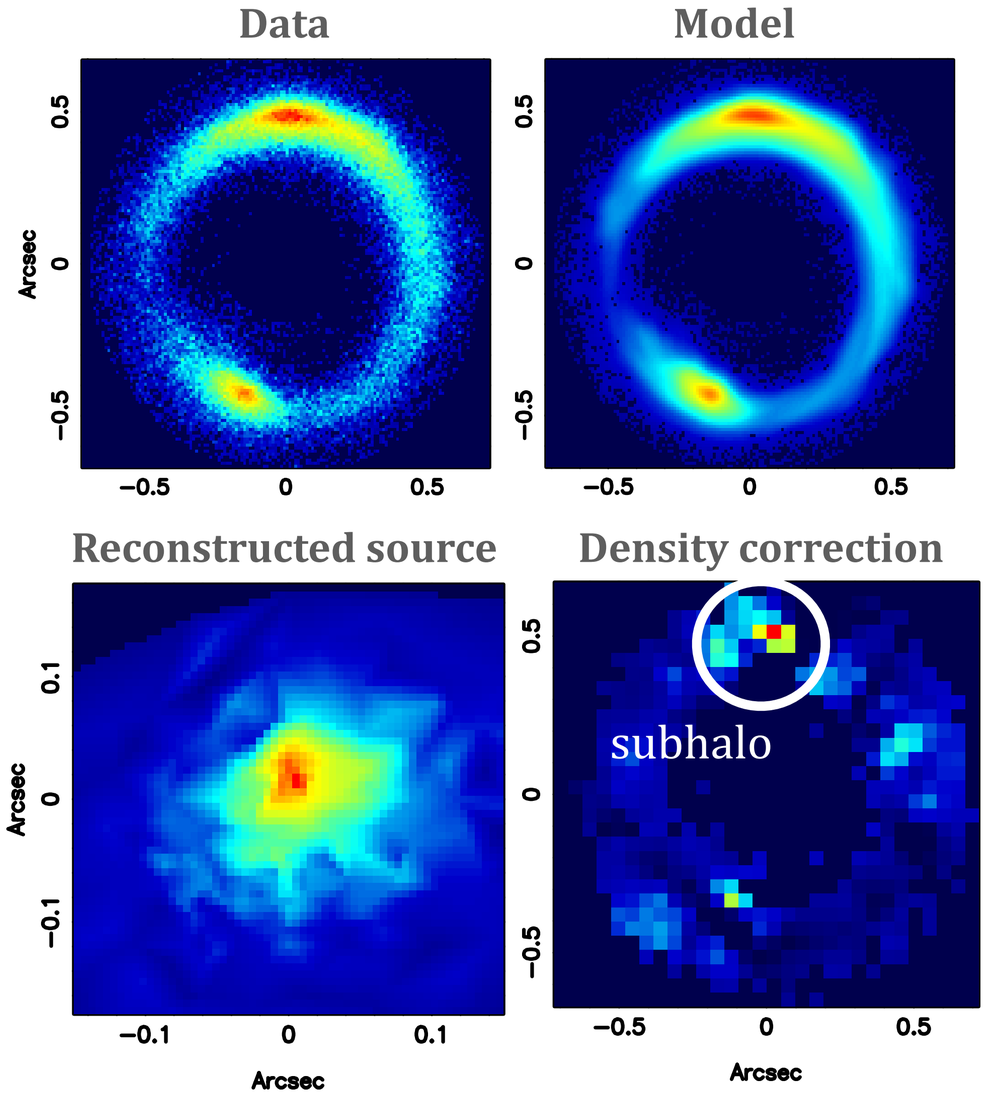


What about substructures within host halos?



Detecting small-scale structures via strong gravitational lensing



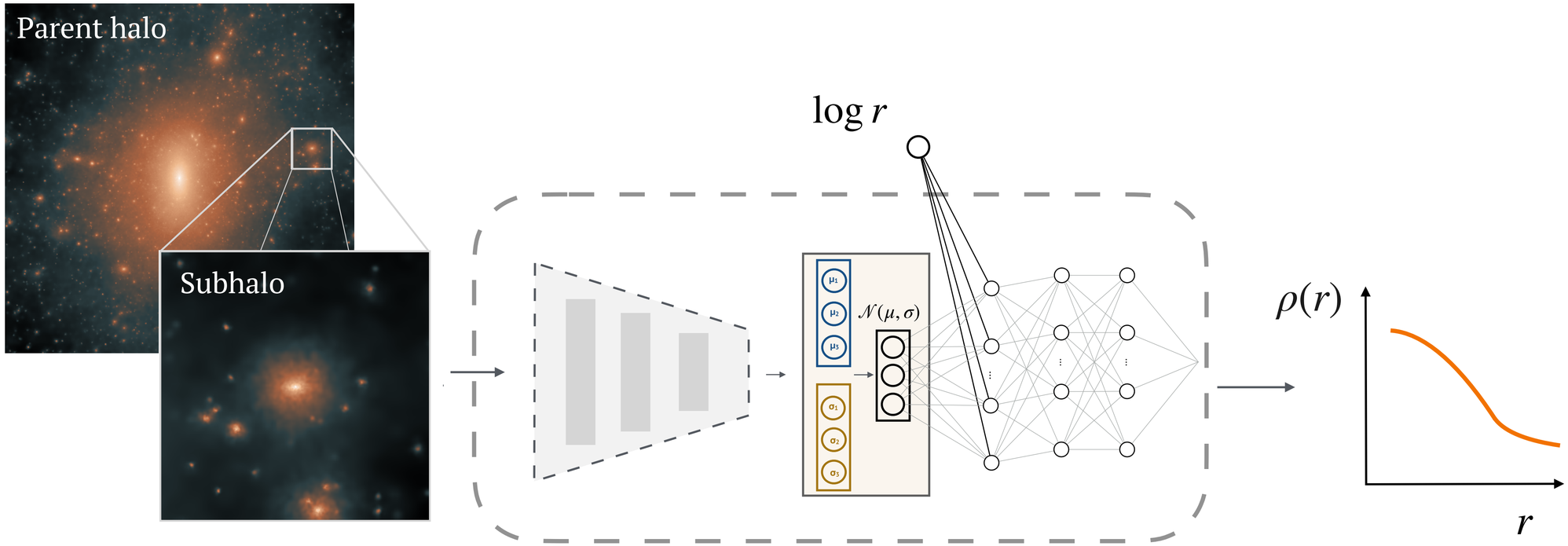


JVAS B1938+666; Vegetti et al. (2012)

Inferred subhalo properties *depends strongly on* ***assumed density profiles***

work with Giulia Despali (Bologna) & Volker Springel (MPA)

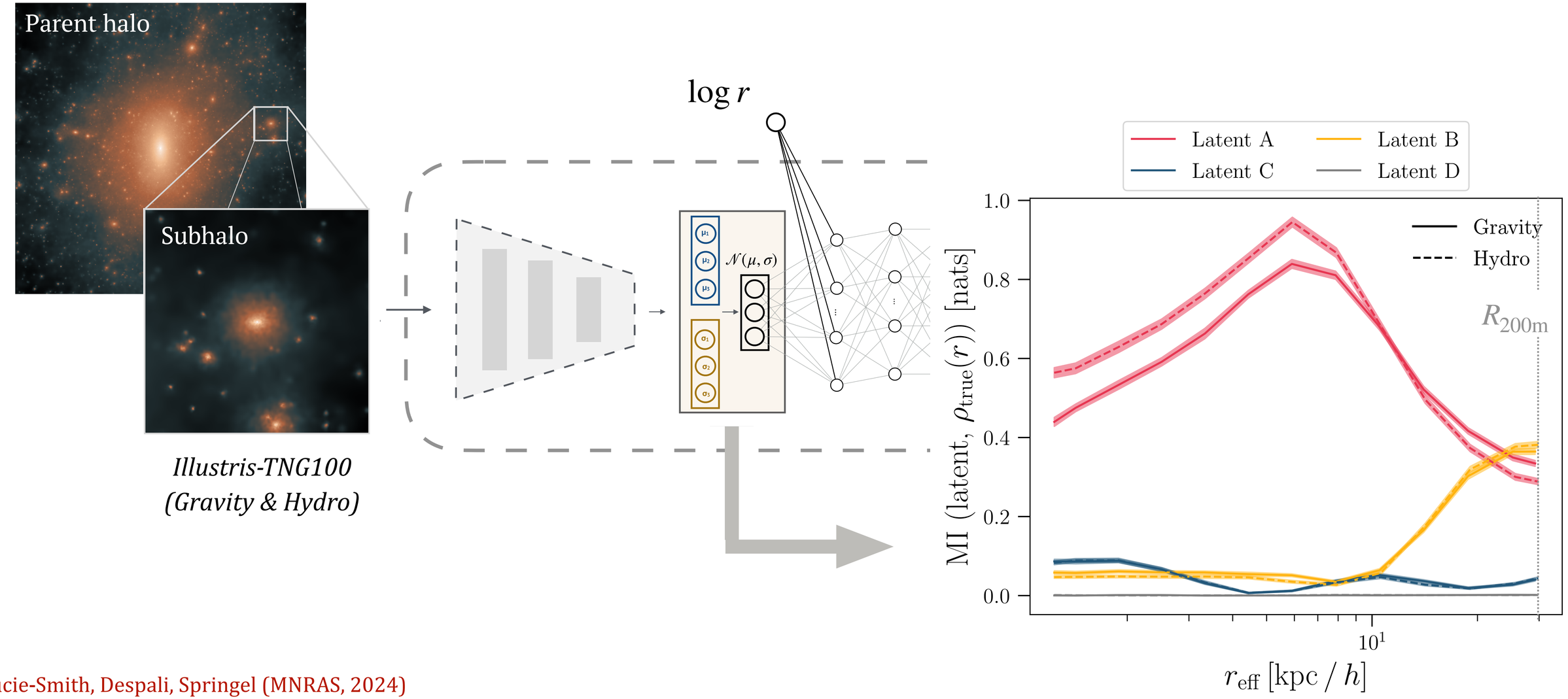
The subhalo density profile at $r < R_{200\mathrm{m}}$



Illustris-TNG100
(Gravity & Hydro)

Lucie-Smith, Despali, Springel (MNRAS, 2024)

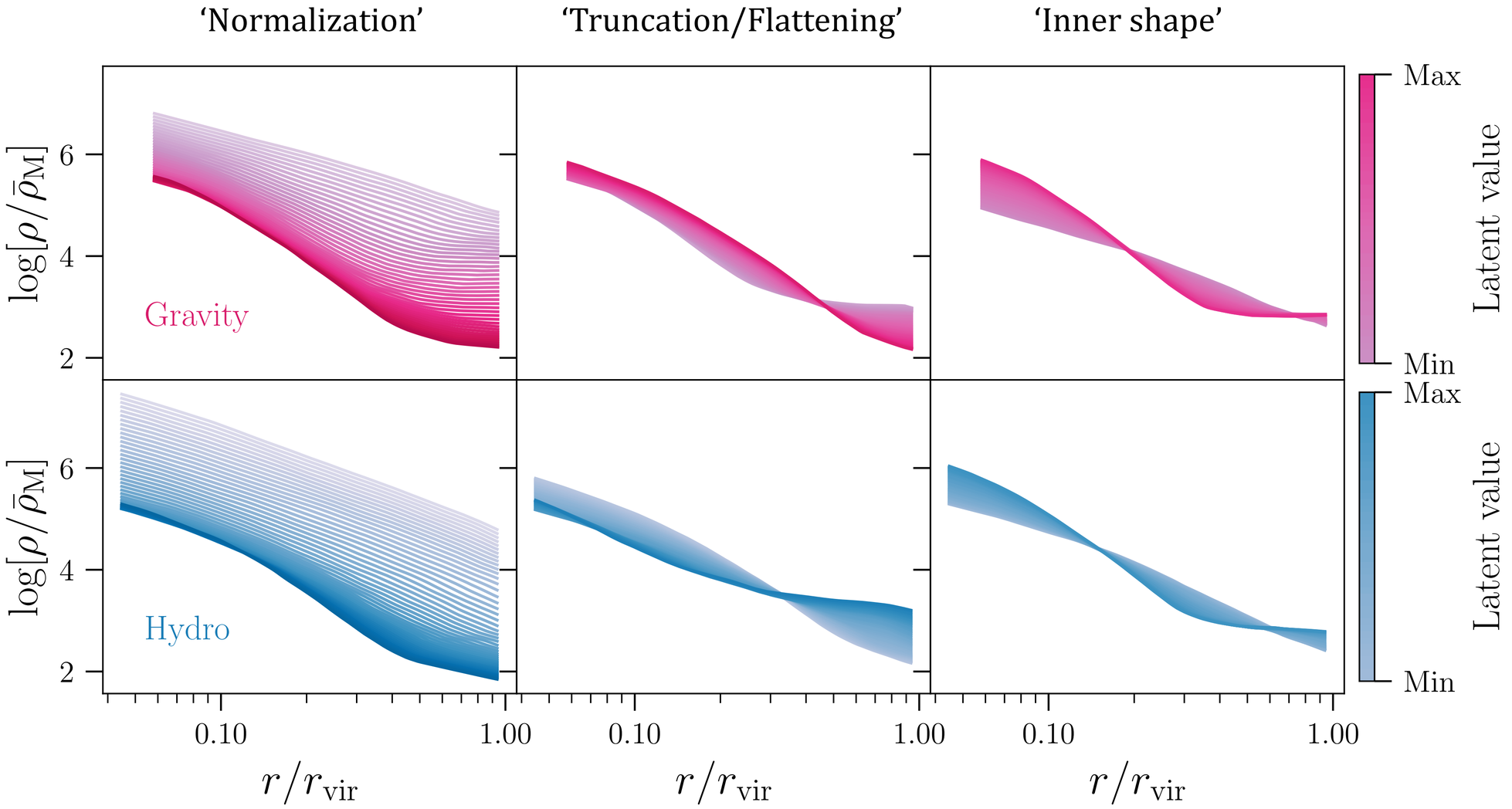
The subhalo density profile at $r < R_{200m}$



Illustris-TNG100
(Gravity & Hydro)

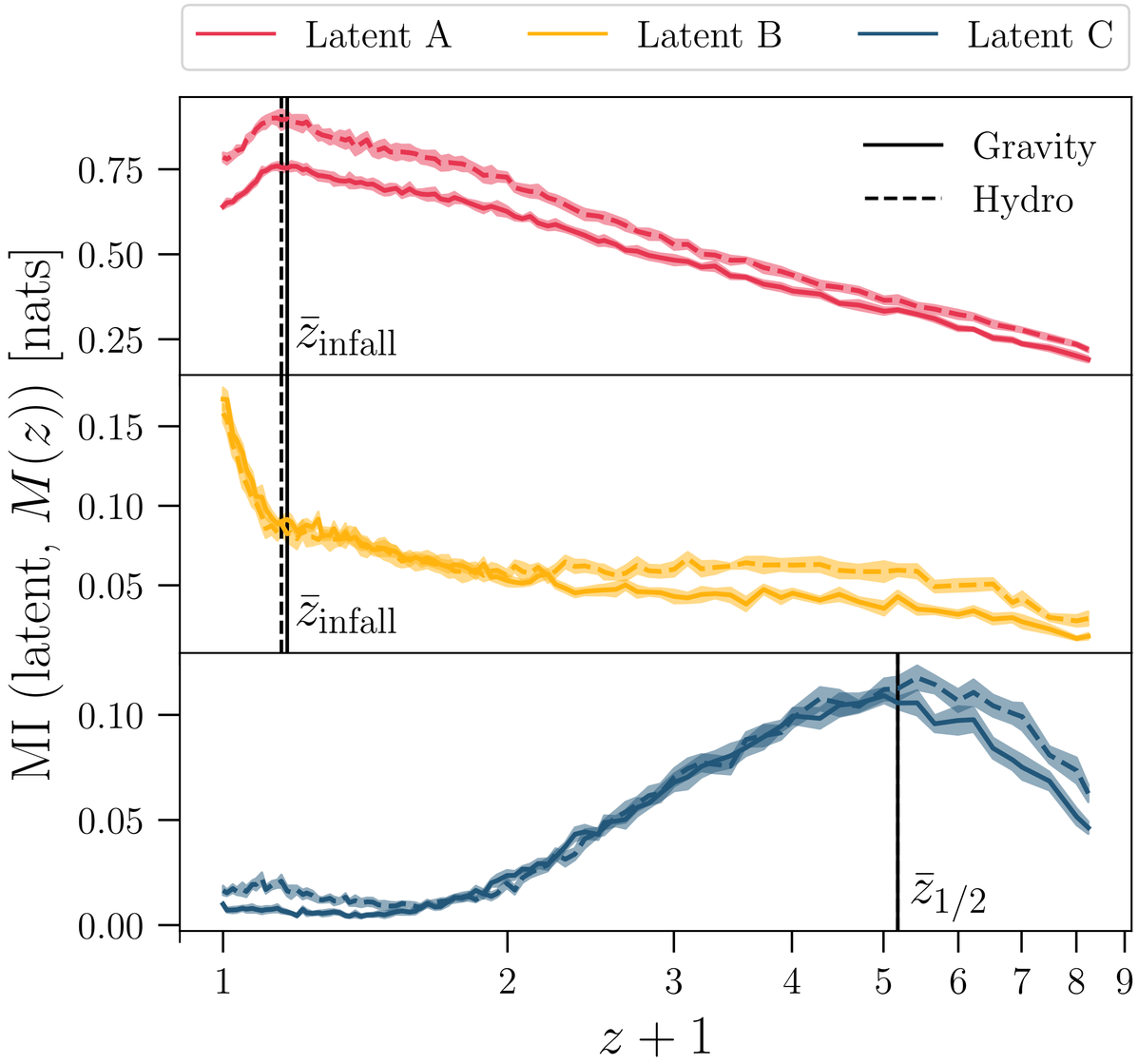
Lucie-Smith, Despali, Springel (MNRAS, 2024)

Subhalos require additional latent capturing tidal truncation



Lucie-Smith, Despali, Springel (MNRAS, 2024)

Mutual information between latents and M(z)



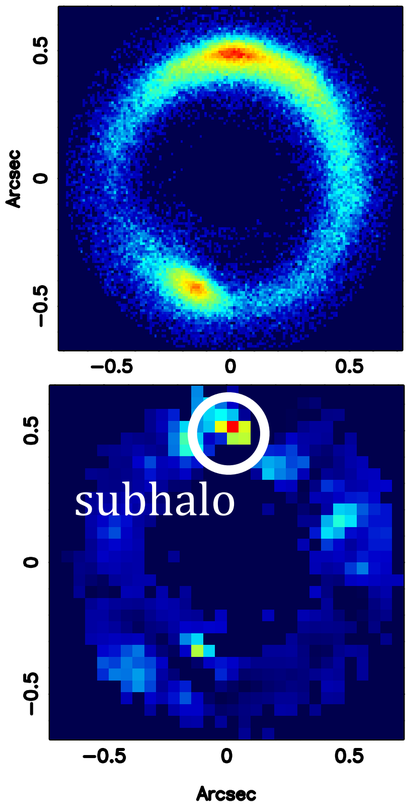
'Normalization' *latent sensitive to formation history* ***before infalling*** *into the main host halo*

'Truncation' *latent sensitive to formation history* ***after infalling*** *into the main host halo*

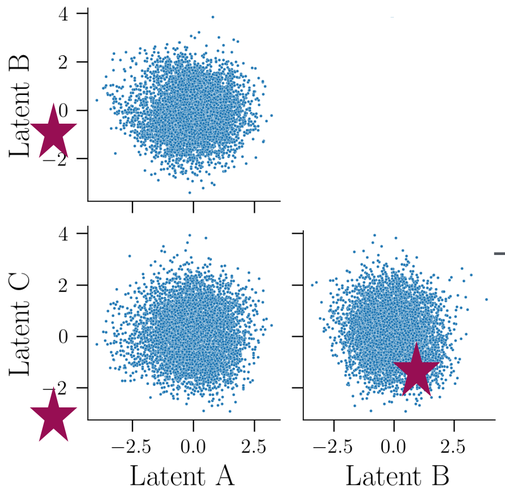
'Inner shape' *latent sensitive to half-mass* ***formation time***

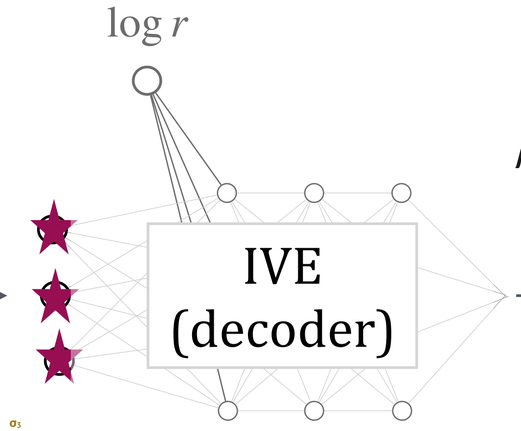
Lucie-Smith, Despali, Springel (MNRAS, 2024)

A physically interpretable subhalo density profile for strong gravitational lensing

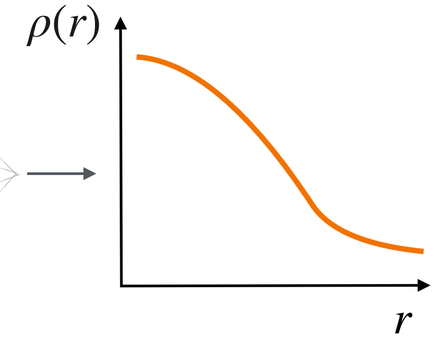


Sample from the latent space





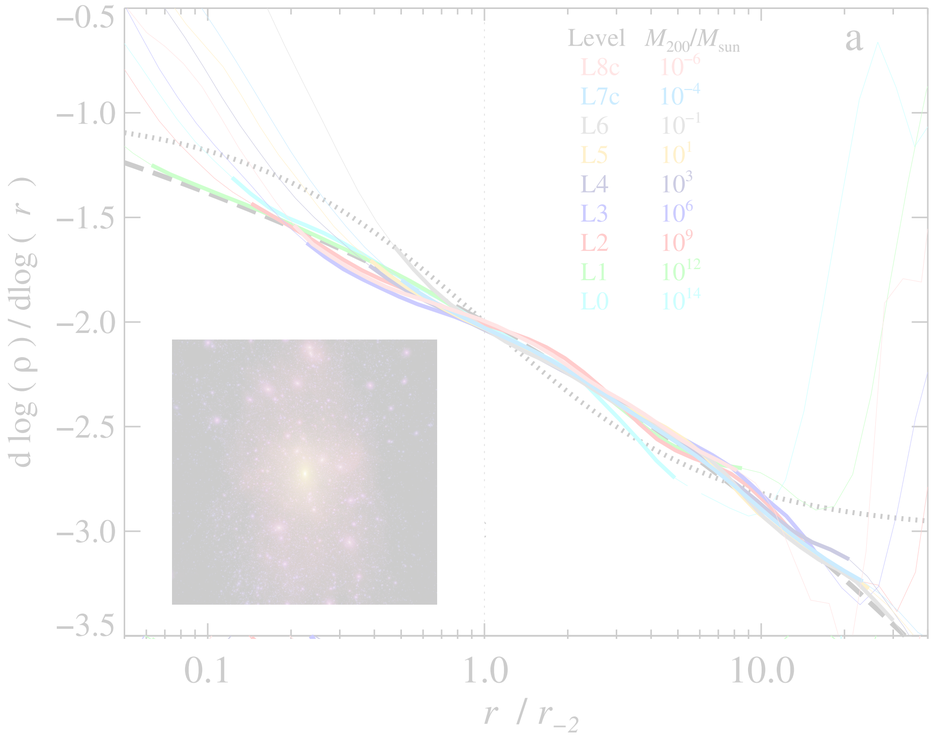
Infer subhalo density profile



Next step: Integrate IVE model within strong gravitational lensing pipeline

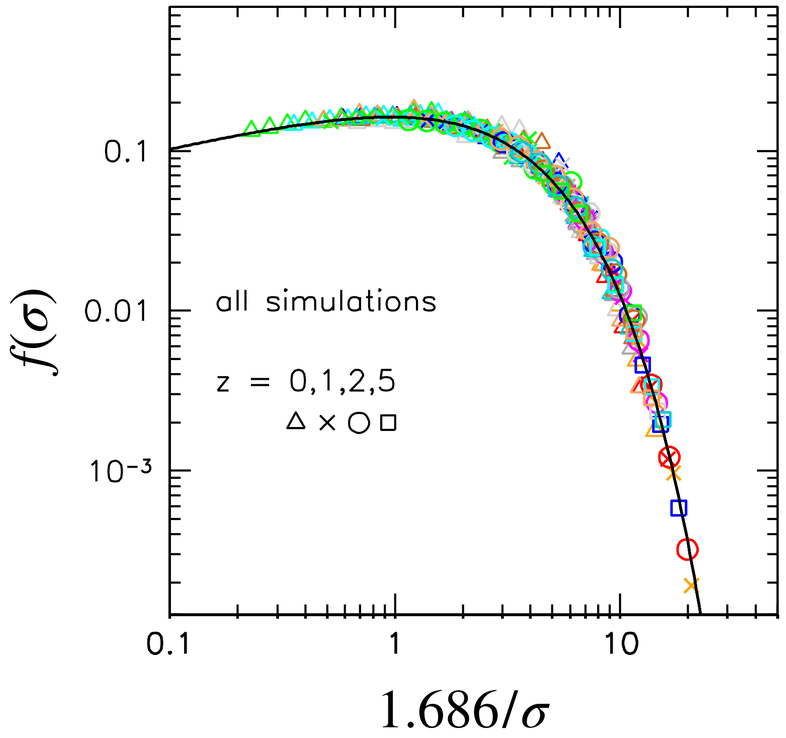
'Emergent' universal properties of the large-scale structure

Halo density profiles



Wang et al. (2020)

Halo mass function



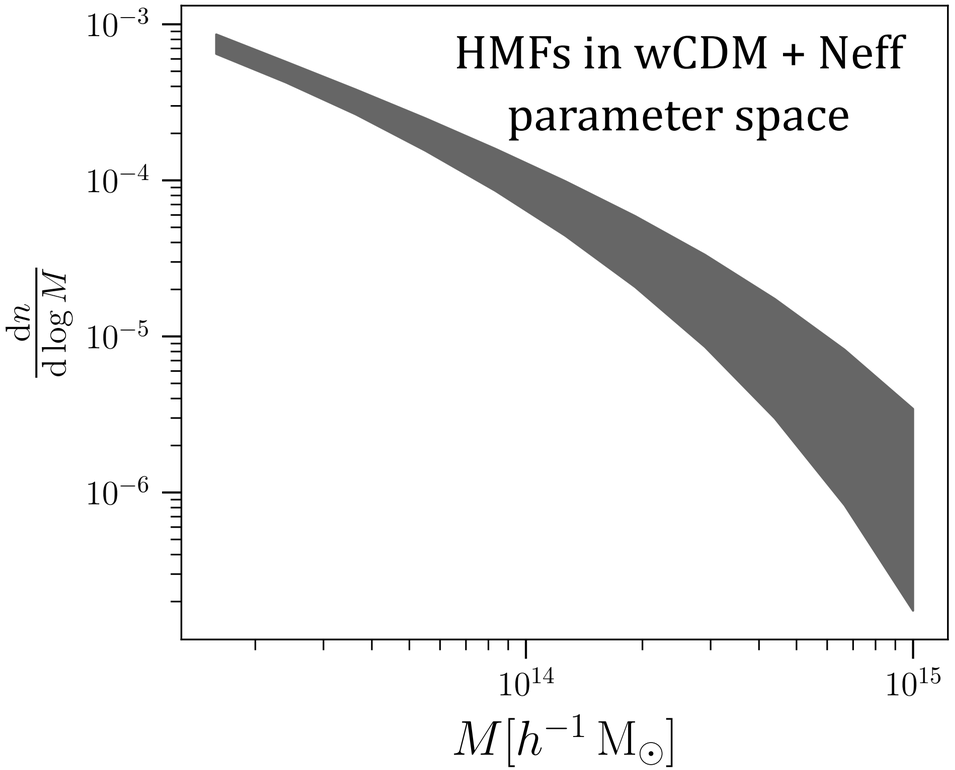
Despali et al. (2016)

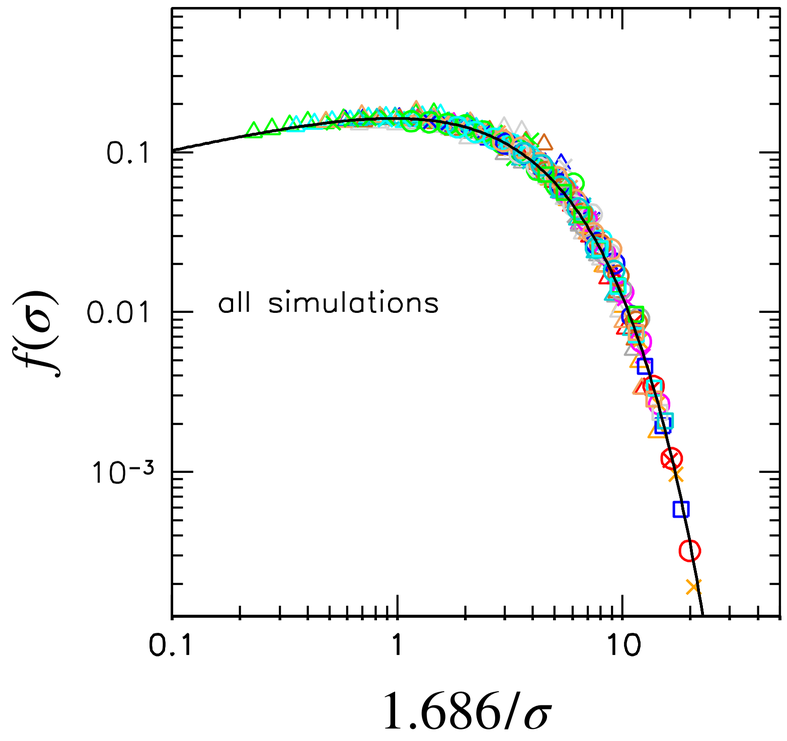
work led by PhD student Lillian Guo (UCL), with Hiranya Peiris and Andrew Pontzen

A 'universal' halo mass funcion

Theoretical framework where cosmological parameters + P(k) information encoded in $\sigma(M, z)$

$$\frac{\mathrm{d}n}{\mathrm{d}\log M} = \frac{\rho_b}{M}\frac{\mathrm{d}\ln\sigma^{-1}}{\mathrm{d}\log M} f(\sigma)$$

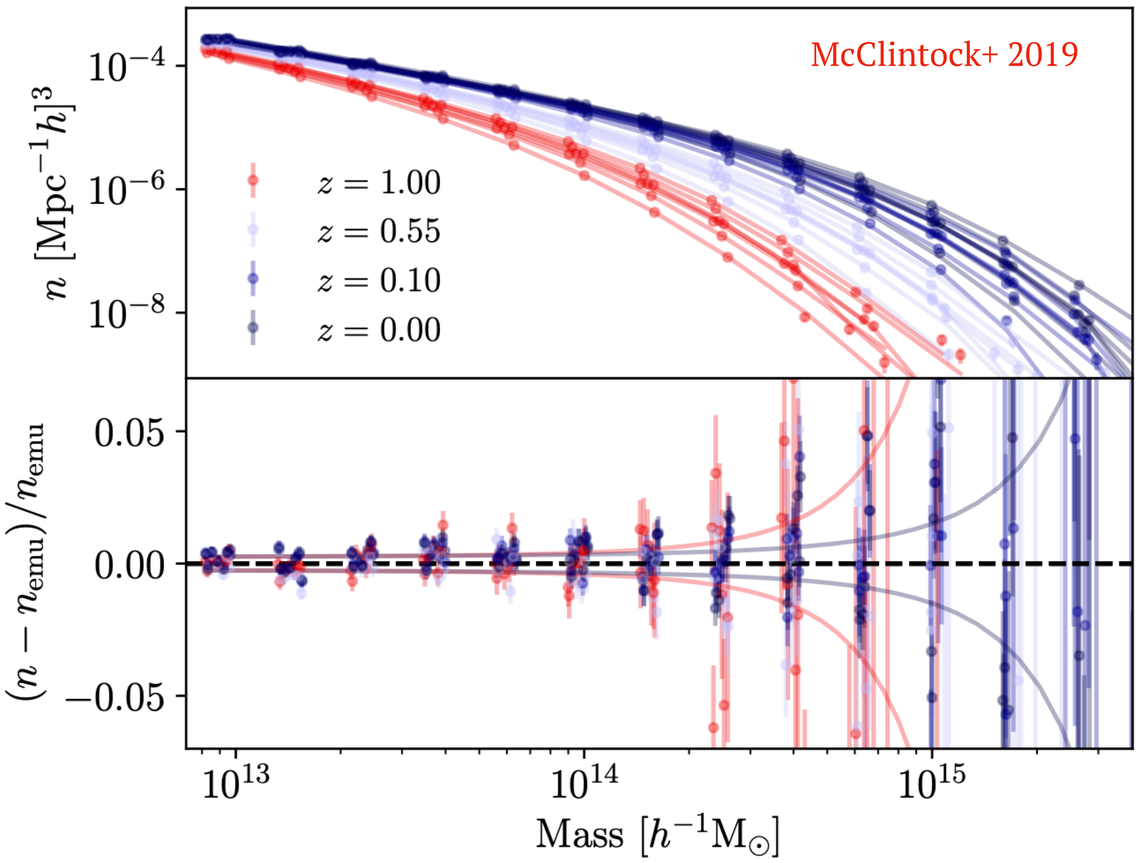




Universality holds up to 10% —> insufficient for cluster cosmology with upcoming galaxy surveys

Beyond universal HMFs: emulators

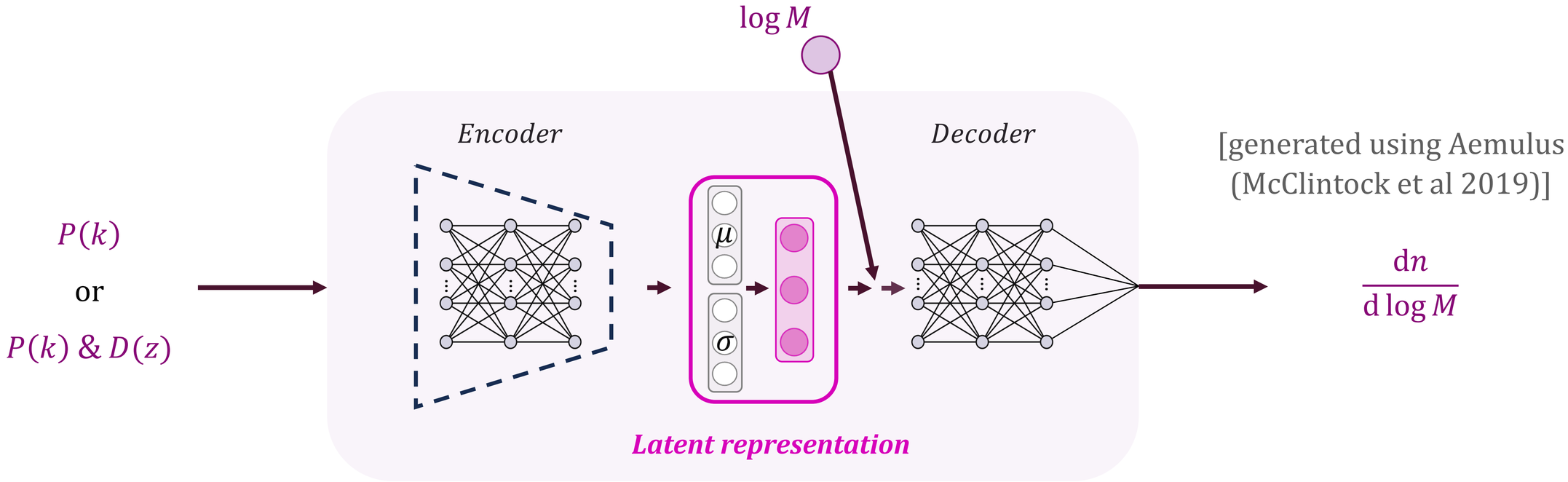
Aemulus HMF emulator



- *No theoretical understanding of the origin of non-universality*
- *Require sampling high dimensional cosmological parameter space*
- *Cannot generalize beyond its domain of validity*

What drives non-universality and can this inform emulators of more efficient training set designs?

IVE: new perspective on non-universality beyond $f(\sigma)$ functional form



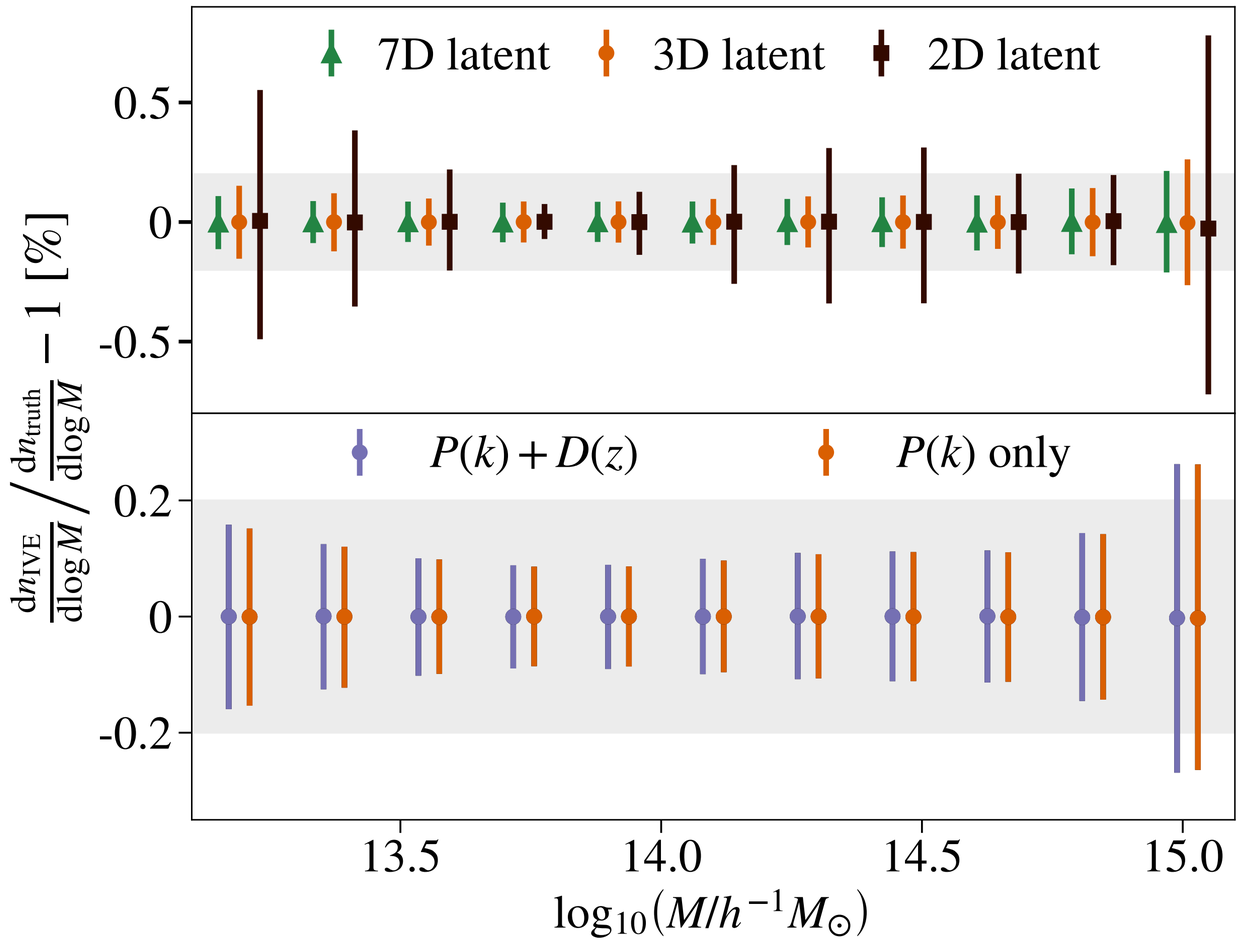
- *How many parameters are necessary and sufficient to describe the HMF?*

- *Can we isolate and quantify the universal and non-universal information?*

  Previous work suggests that linear growth is related to non-universality
  (Ondaro-Mallea et al 2021; Euclid collaboration et al. 2023)

Guo, Lucie-Smith, Peiris, Pontzen (in prep)

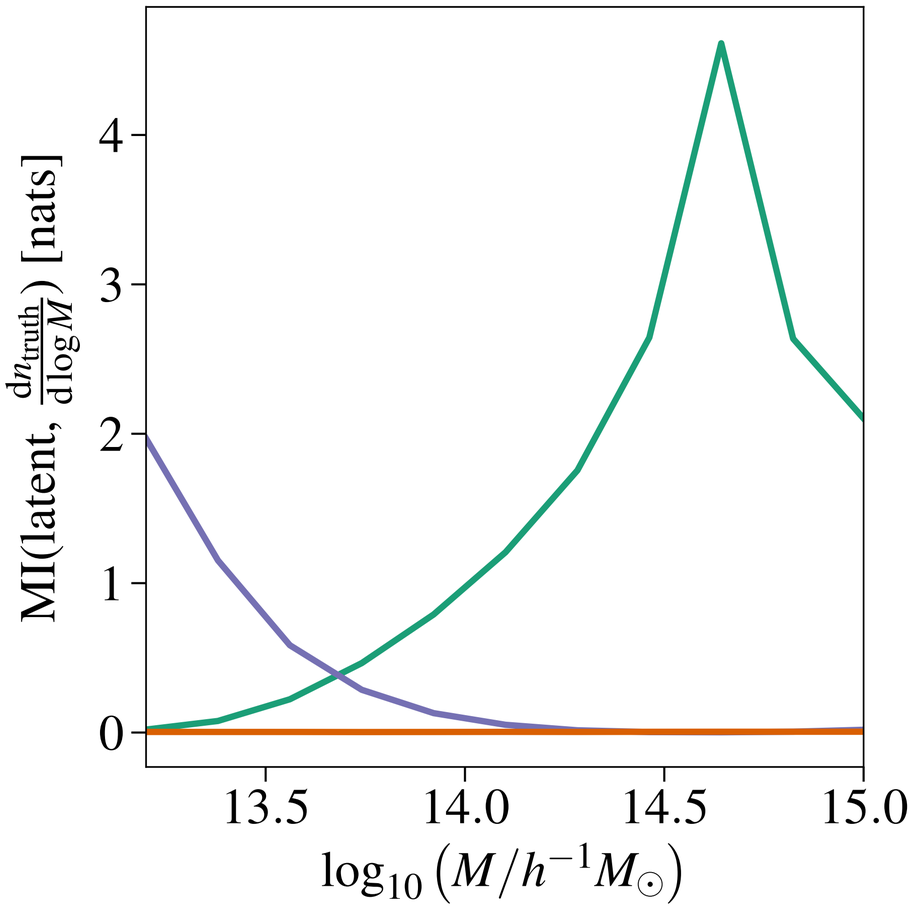
The dimensionality of the HMF and the role of growth



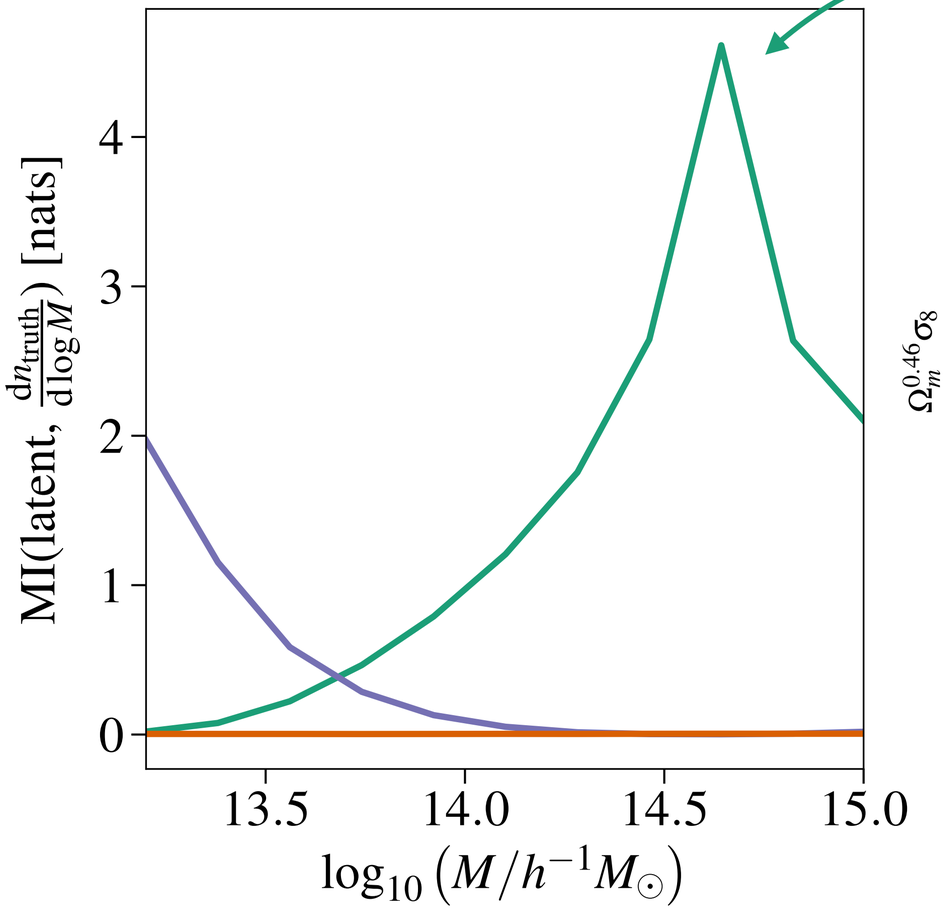
- ***Three latent parameters*** *required to describe the HMF*

- ***No improvement when adding D(z)*** *if P(k) is already an input*

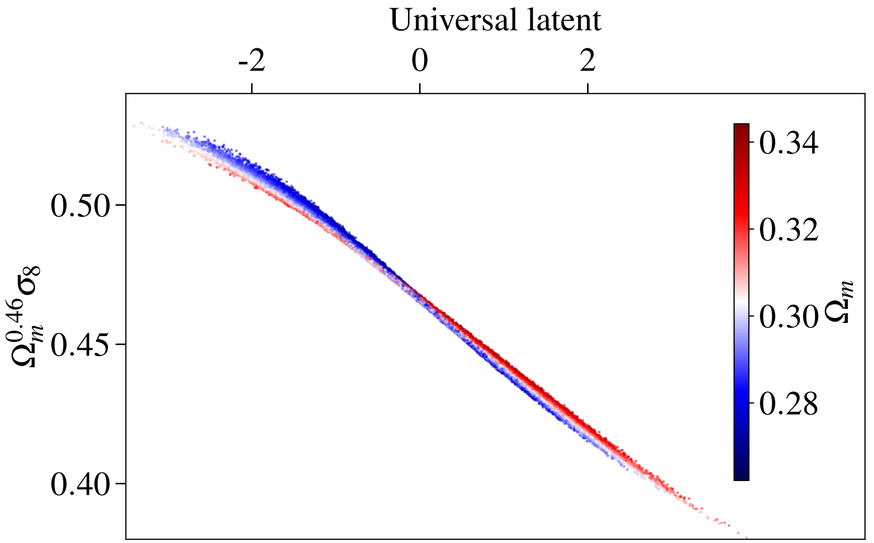
Three latents required to predict the (Aemulus) HMF



Three latents required to predict the (Aemulus) HMF

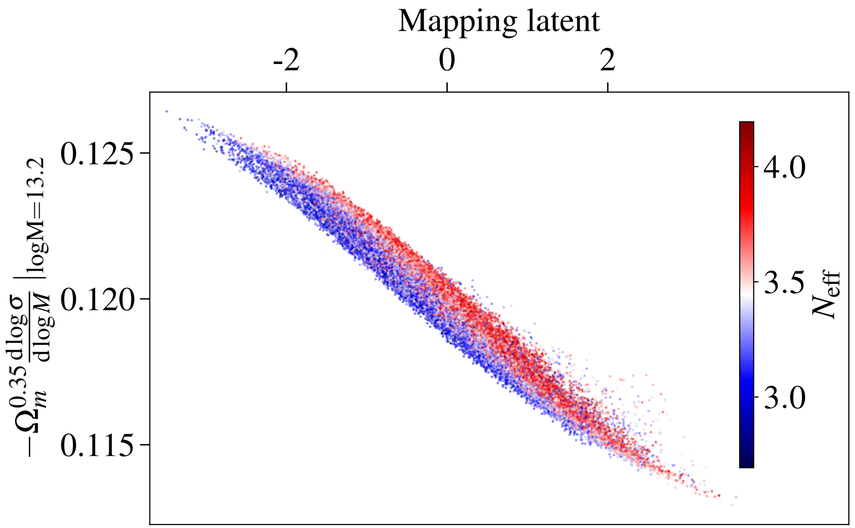


'Universal latent' ~ S_8



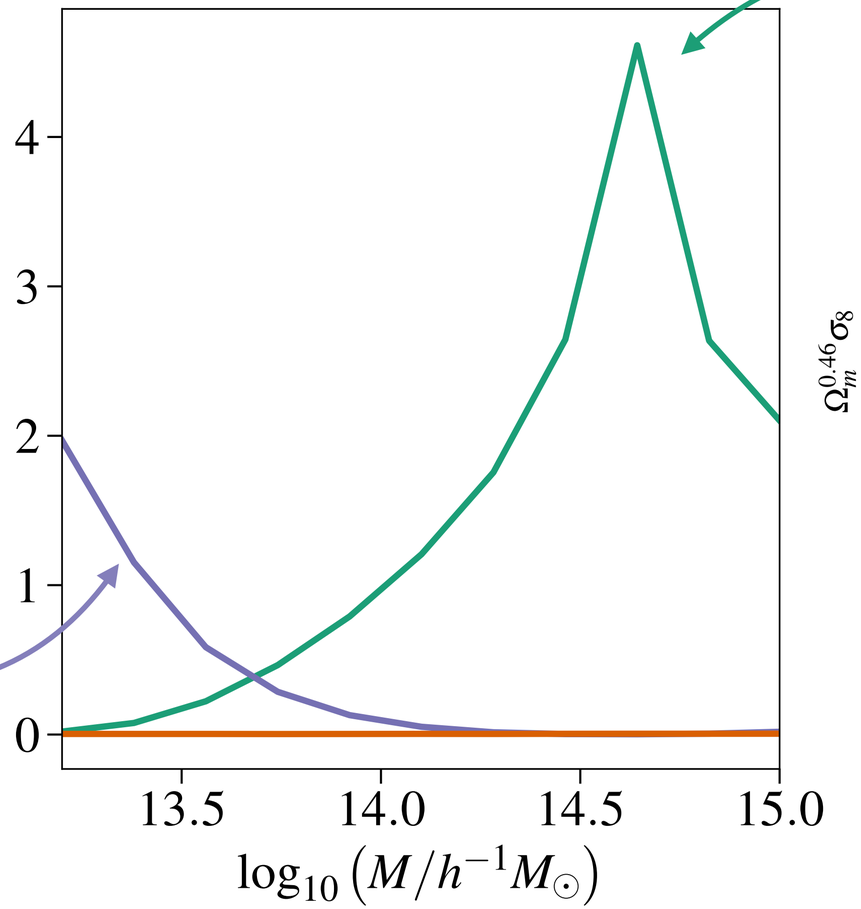
High-mass halos: $\Omega_m(z_{form}) \sim \Omega_m^{0.46}$

Three latents required to predict the (Aemulus) HMF

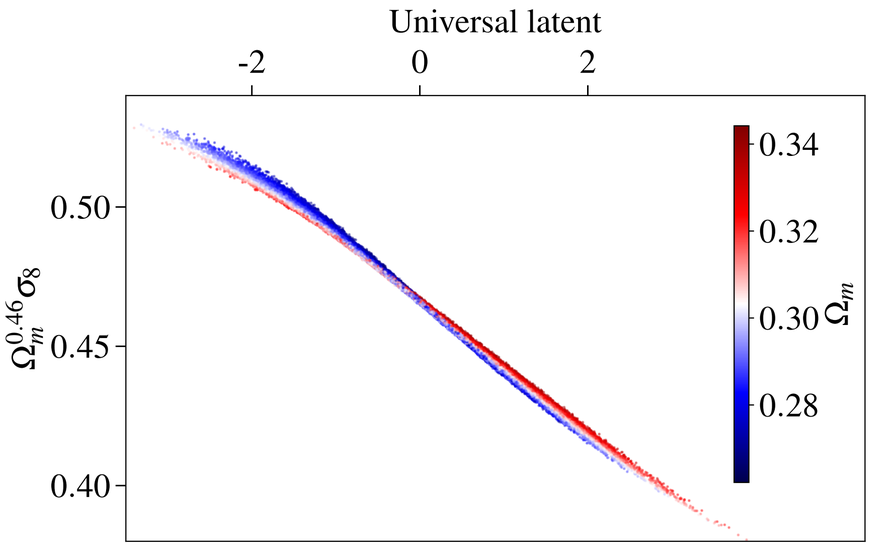


Low-mass halos: $\Omega_{\rm m}(z_{\rm form}) \sim \Omega_{\rm m}^{0.35}$

‘Mapping latent’



‘Universal latent’ ~ S_8



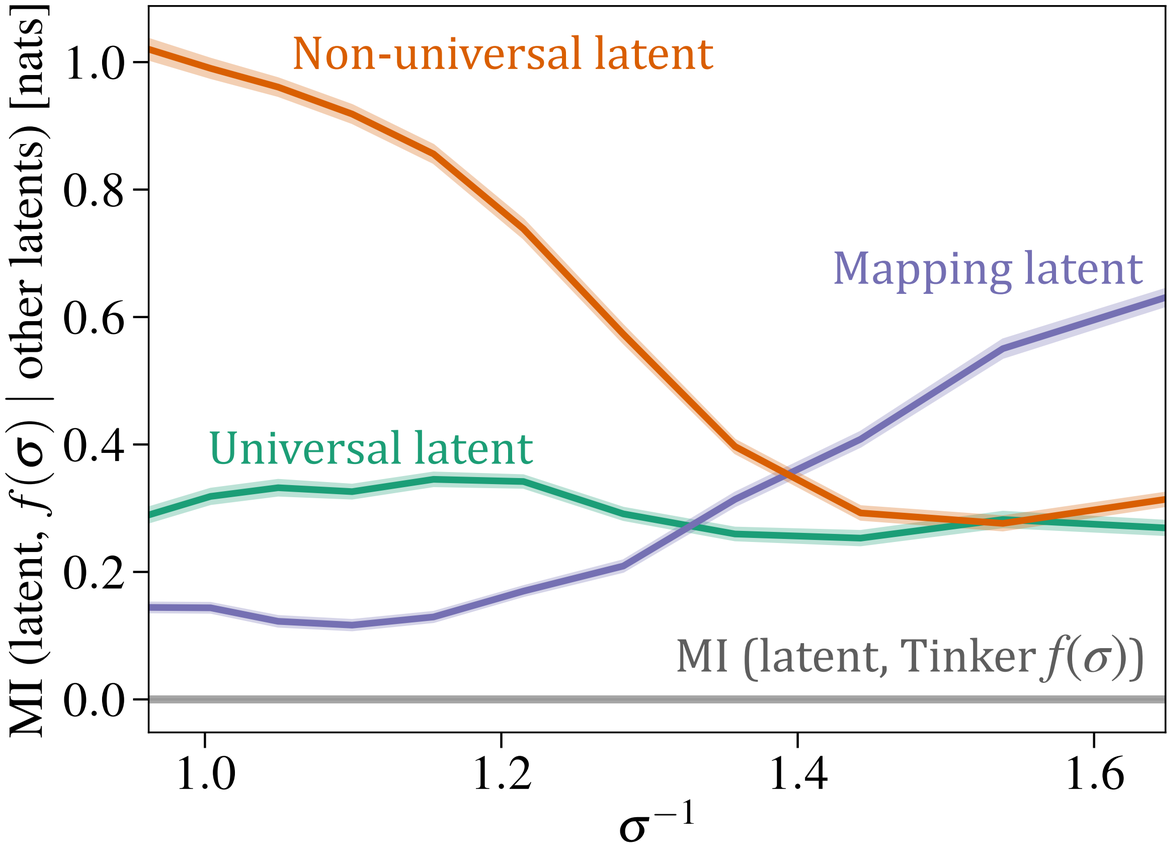
High-mass halos: $\Omega_{\rm m}(z_{\rm form}) \sim \Omega_{\rm m}^{0.46}$

The third, non-universal latent

Conditional MI:

$\mathrm{MI}(X, Y \mid Z)$ *is information shared between X, Y given information about Z is already known*

HTTPS://GITHUB.COM/DPIRAS/GMM-MI

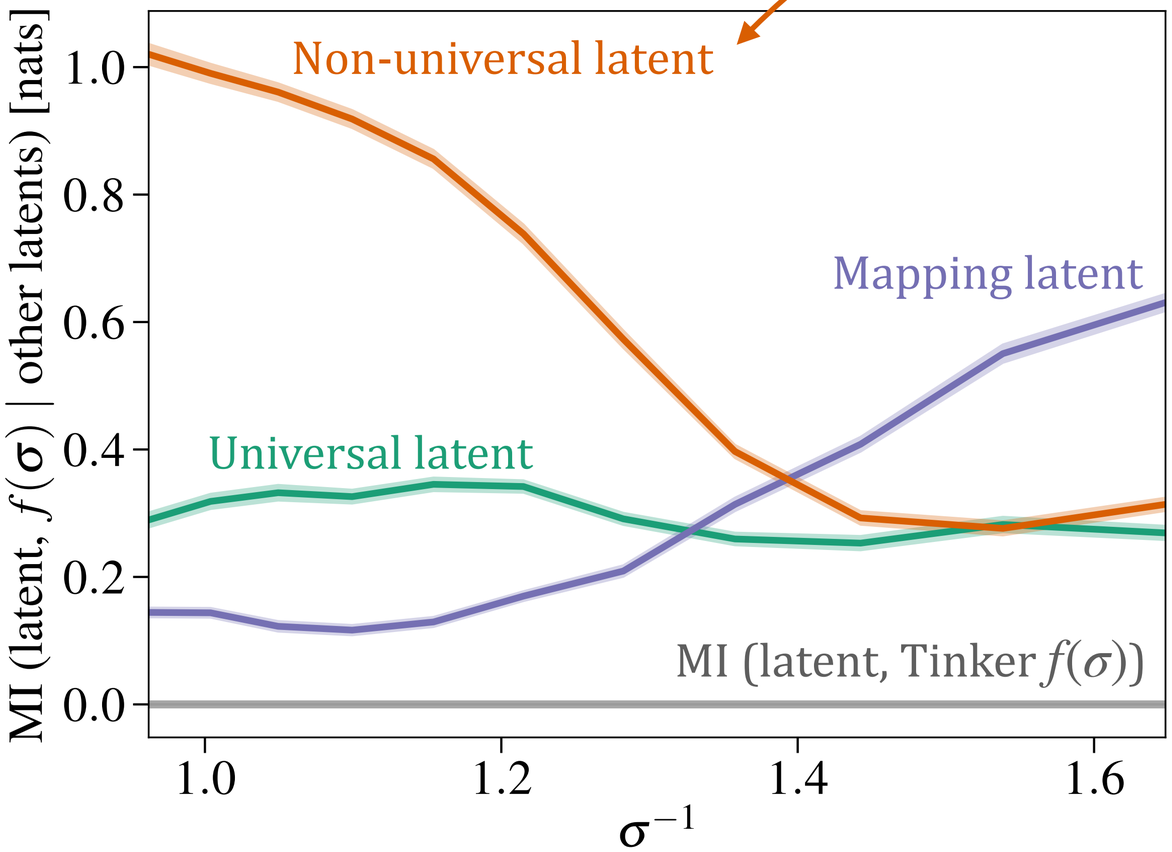


The third, non-universal latent

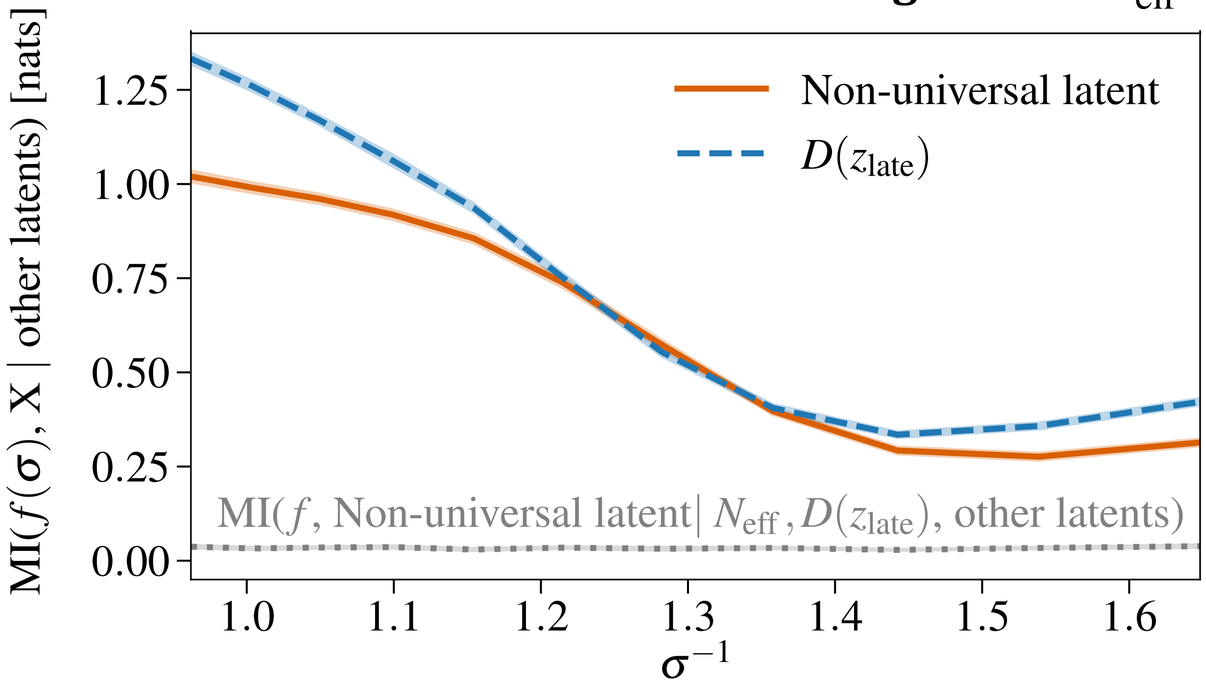
Conditional MI:

$\mathrm{MI}(X, Y \mid Z)$ *is information shared between X, Y given information about Z is already known*

HTTPS://GITHUB.COM/DPIRAS/GMM-MI



'Non-universal latent' ~ late-time growth + N_{eff}

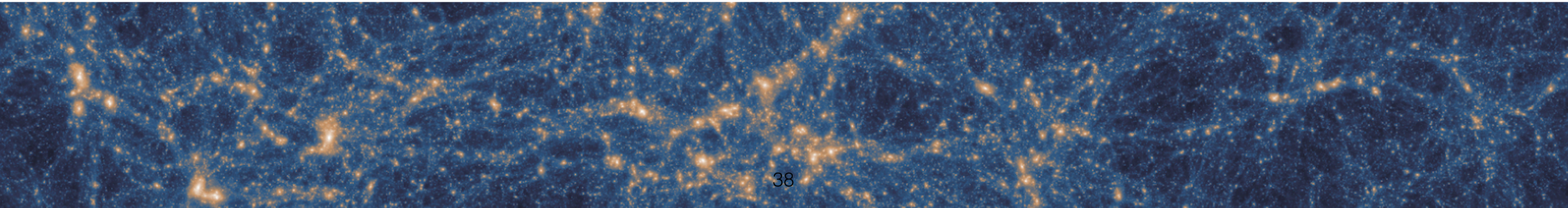


Conclusions

- Interpretability & explainability can be achieved via model compression + mutual information
- IVE disentangles different physical effects in minimal set of ingredients
- New insights into emergent large-scale structure properties such as density profiles and the halo mass function

Luisa Lucie-Smith, luisals@mpa-garching.mpg.de

MAX PLANCK INSTITUTE
FOR ASTROPHYSICS



IVE loss function

- **Loss function**

Predictive term *KL-divergence term*

$$\mathscr{L} = \mathscr{L}_{\text{pred}}(\rho_{\text{true}}, \rho_{\text{pred}}) + \beta\, \mathscr{D}_{\text{KL}}(p(\mathbf{z}\,|\,\mathbf{x}); q(\mathbf{z}))$$

(Higgins+, 2017)

MSE/Gaussian likelihood:

$$\mathscr{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N}\left[\log_{10}\rho_{i,\text{true}} - \log_{10}\rho_{i,\text{pred}}\right]^2$$

How close are the predictions to the ground truths

Learnt latent distribution:

$$p(\mathbf{z}\,|\,\mathbf{x}) = \prod_{i=1}^{L}\mathcal{N}(\mu_i(\mathbf{x}), \sigma_i(\mathbf{x}))$$

Prior:

$$q(\mathbf{z}) = \prod_{i=1}^{L}\mathcal{N}(0,1)$$

How close is the latent distribution to set of independent unit Gaussians

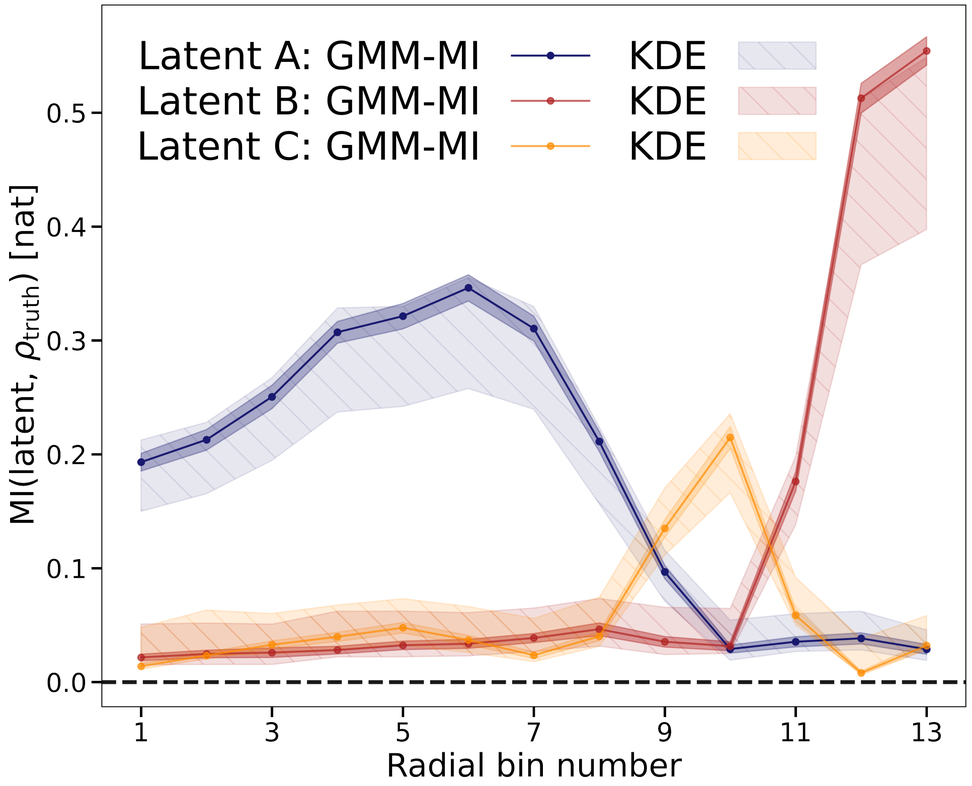
- **Mutual information** to measure the level of disentanglement:

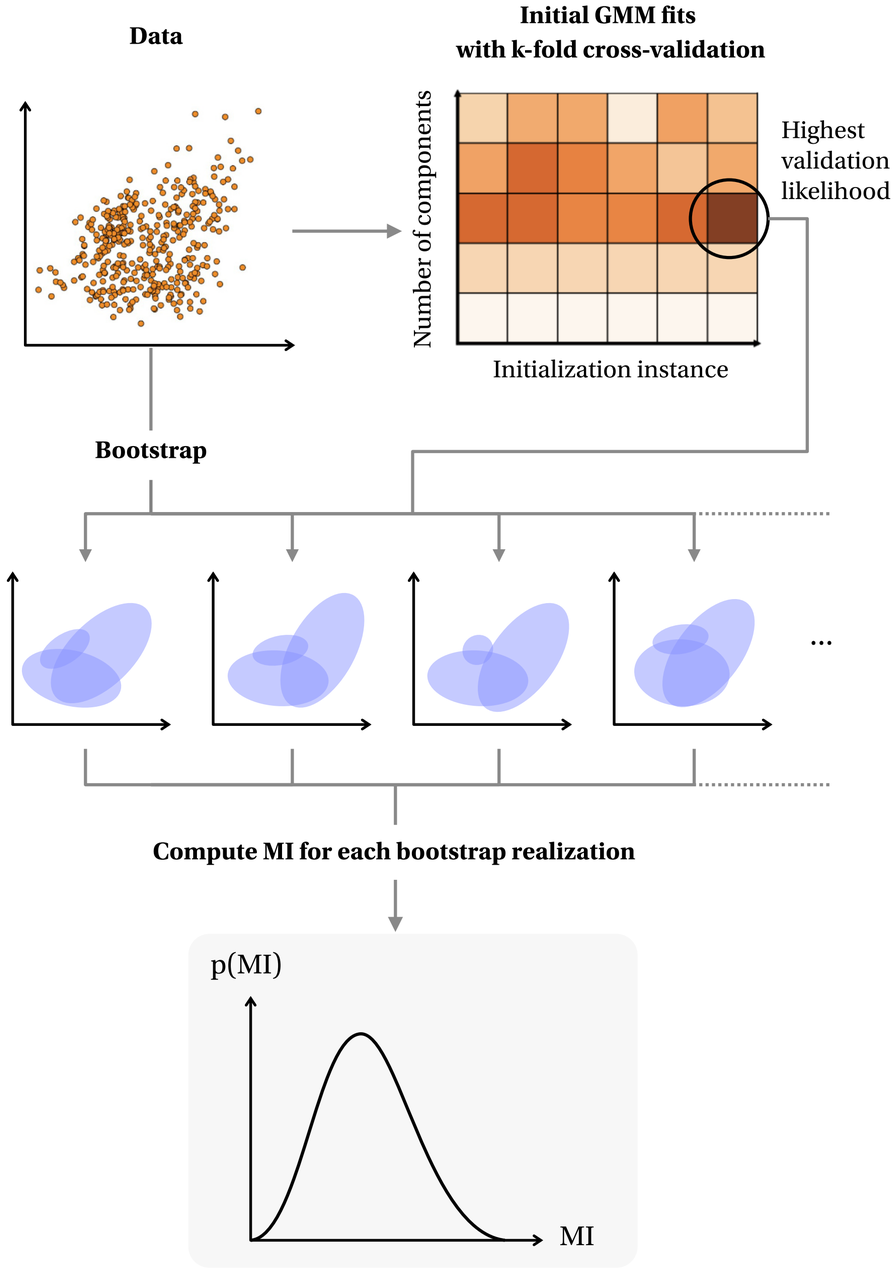
$$\text{MI}(z_i, z_j) = \int_{z_i}\int_{z_j} p(z_i, z_j)\log\left[\frac{p(z_i, z_j)}{p(z_i)p(z_j)}\right] dz_i dz_j$$

Gaussian mixture model $p(z_i, z_j)$

Mutual information

$$\mathrm{MI}(X, Y) = \int\int p(x, y) \log \left[\frac{p(x, y)}{p(x)p(y)} \right] \mathrm{d}x\mathrm{d}y$$



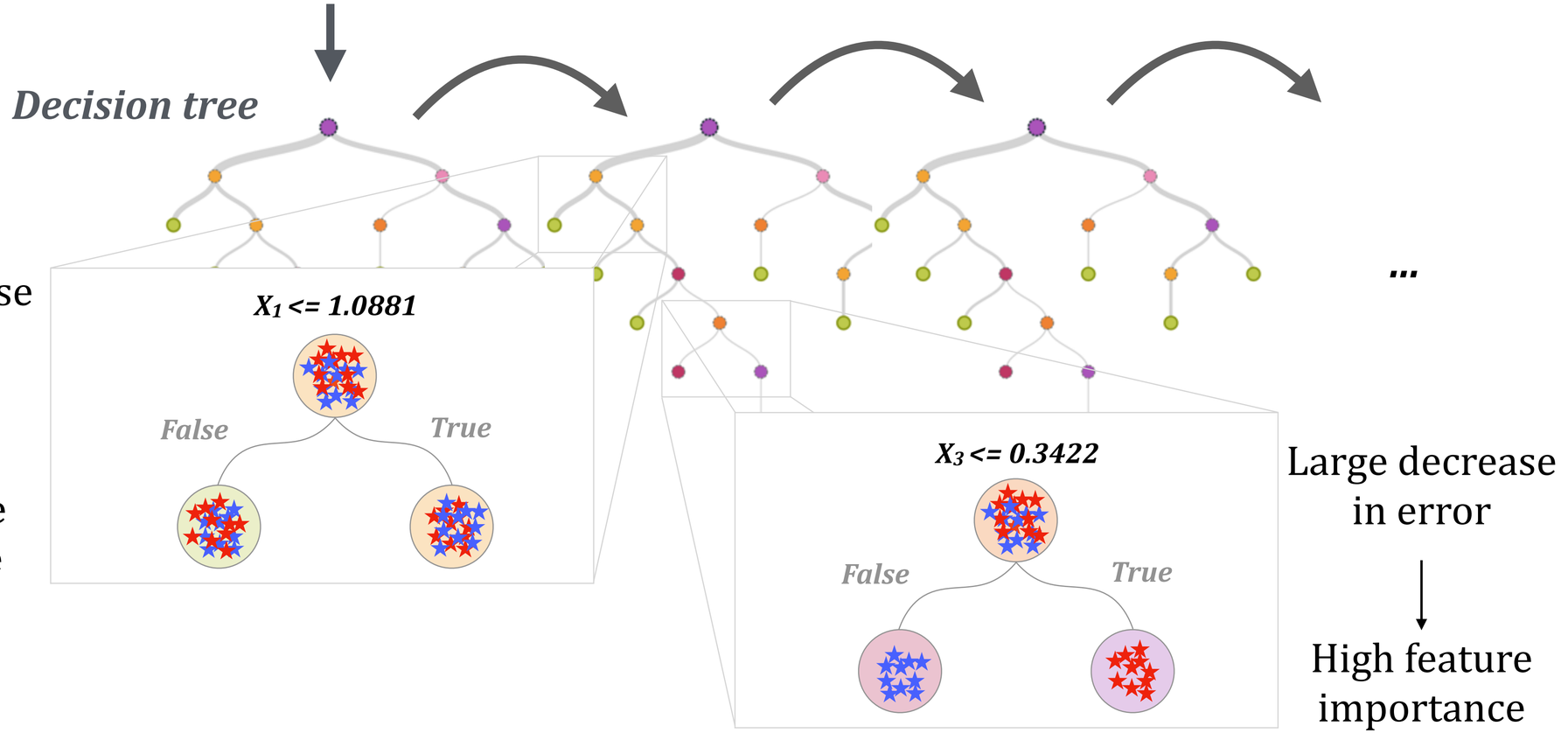


HTTPS://GITHUB.COM/DPIRAS/GMM-MI

Piras, Peiris, Pontzen, Lucie-Smith et al. (2023, MLST)

ML algorithm: gradient boosted trees (GBTs)

GBTs add new decisions trees to correct mistakes of previous trees



Feature importance ∝ *decrease in error due to splits made by feature*

Friedman, 2001; Friedman, 2002