Explainable deep learning models for cosmological structure formation

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

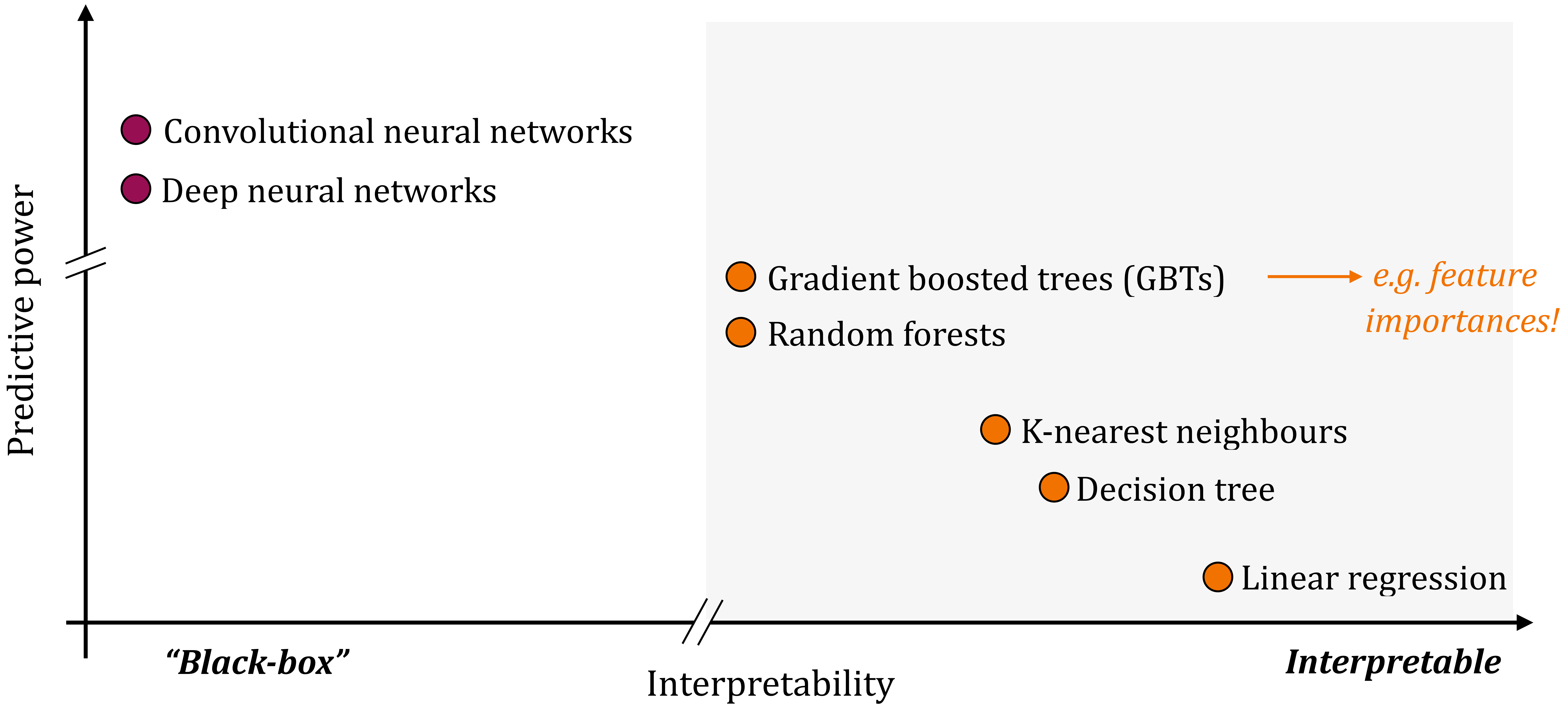
European AI for Fundamental Physics Conference
Amsterdam, 1st May 2024

MAX PLANCK INSTITUTE
FOR ASTROPHYSICS

Machine learning in (Astro)physics

- *ML successful at automating/accelerating known physical models (e.g. emulators)*
- *Can we* ***extract new knowledge*** *about the underlying physics from deep learning models by interpreting their outputs? Requires* ***explainable AI***

Current landscape for explainable AI



Requirements for explainable AI

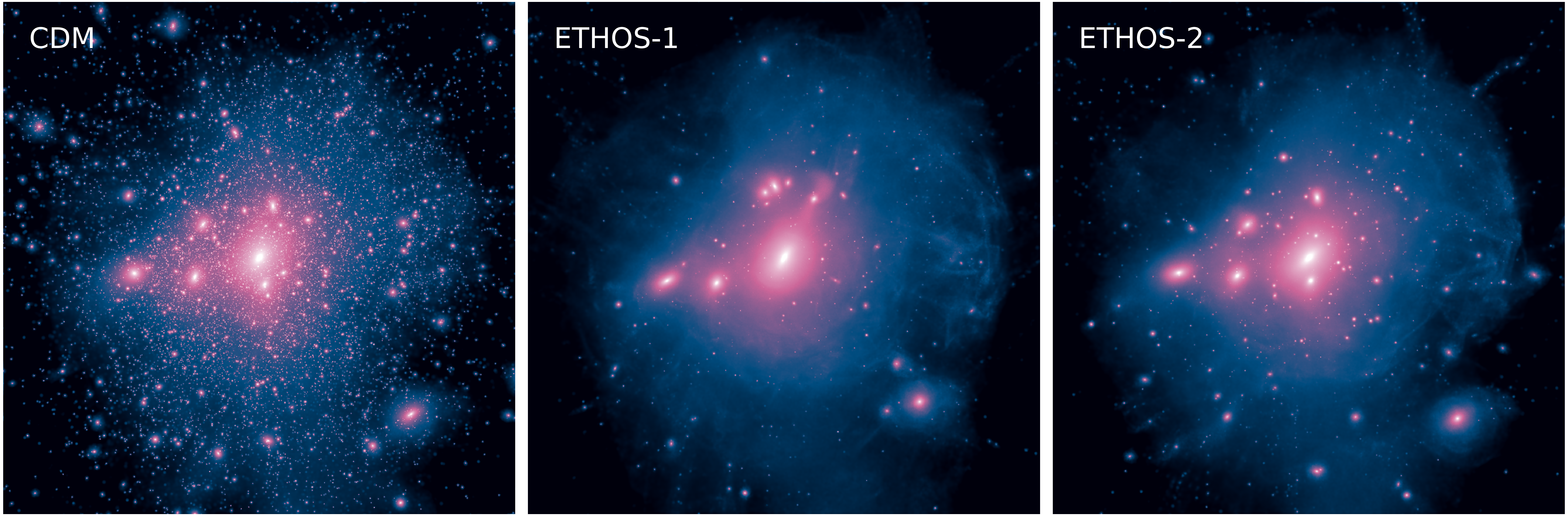
1. ***Interpretability:*** account for why the ML model reaches its predictions

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

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

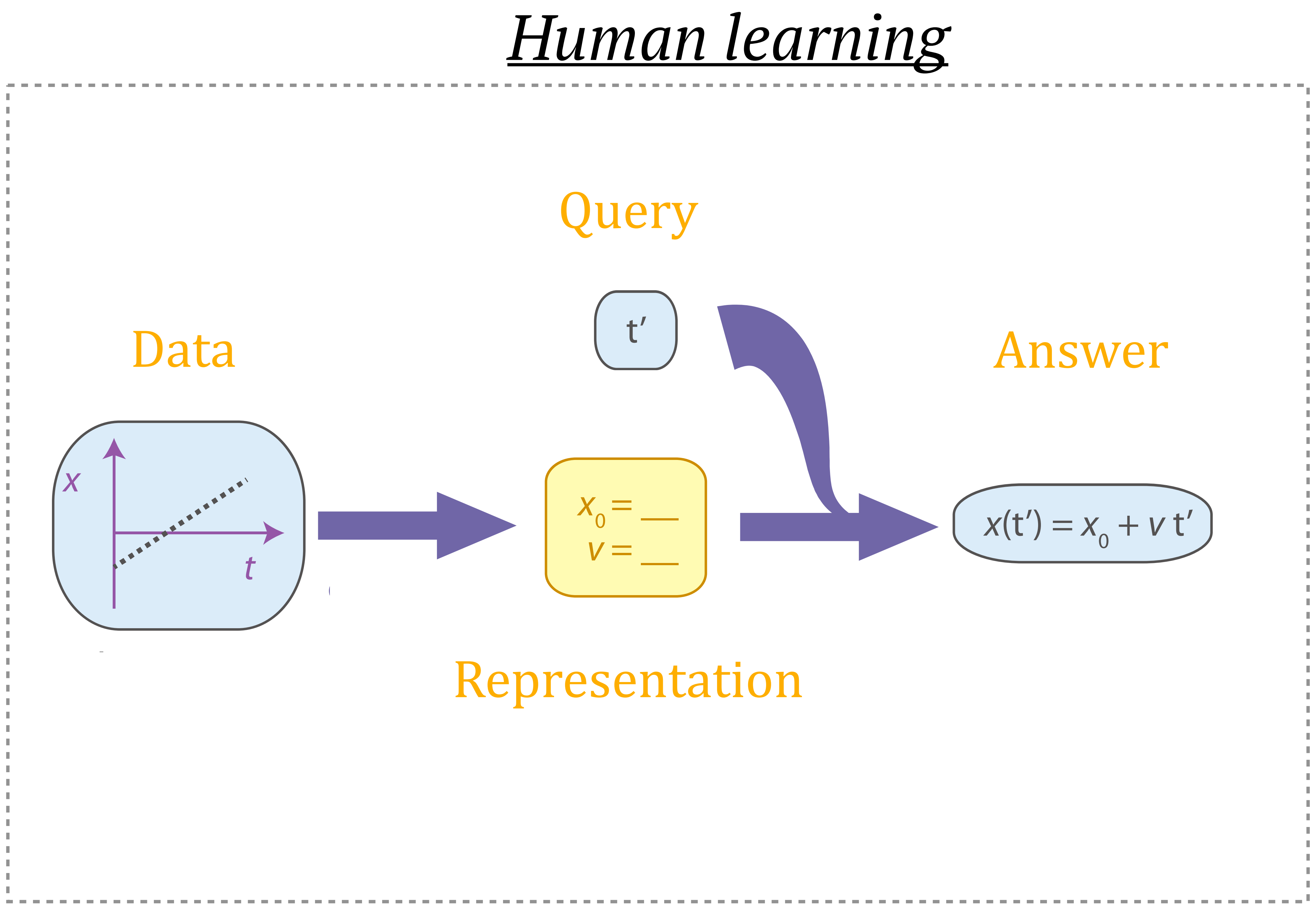
Dark matter halo structure contains signatures of nature of dark matter

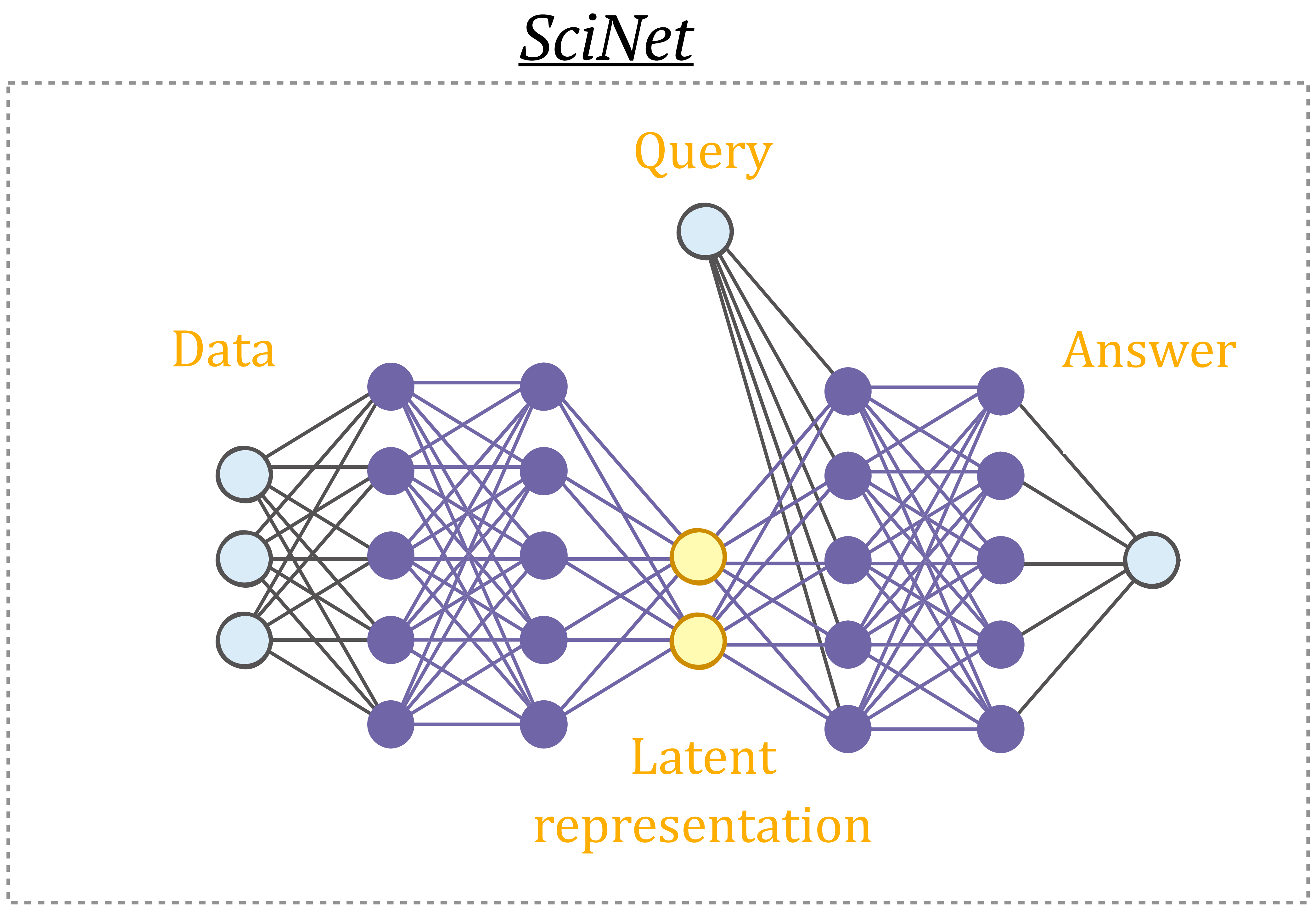
Vogelsberger et al. (2016)



Current models based on 'universal' empirical fitting functions (e.g. NFW) lack ***explainability***

SciNet model

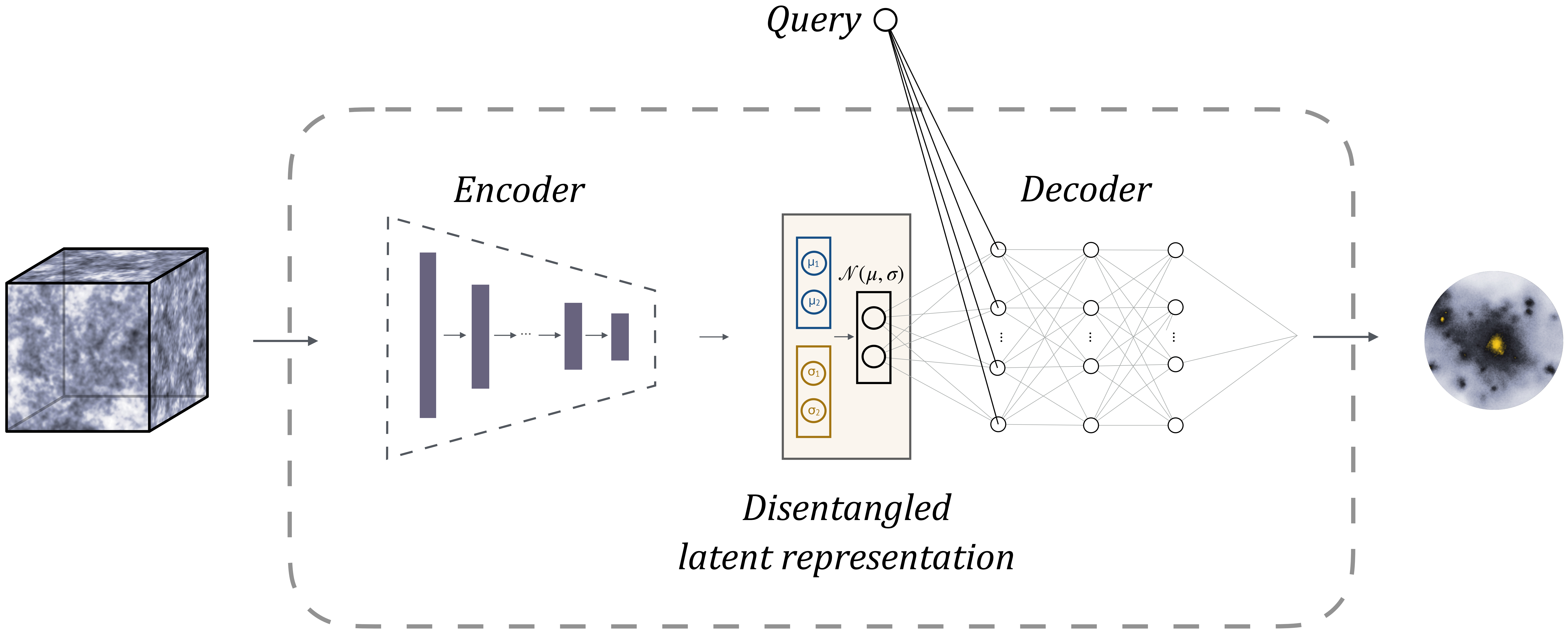




- SciNet learns relevant physical parameters in toy 1D problems
- Relies on comparing latents with already-known physical parameters

Iten et al. (PRL, 2020)

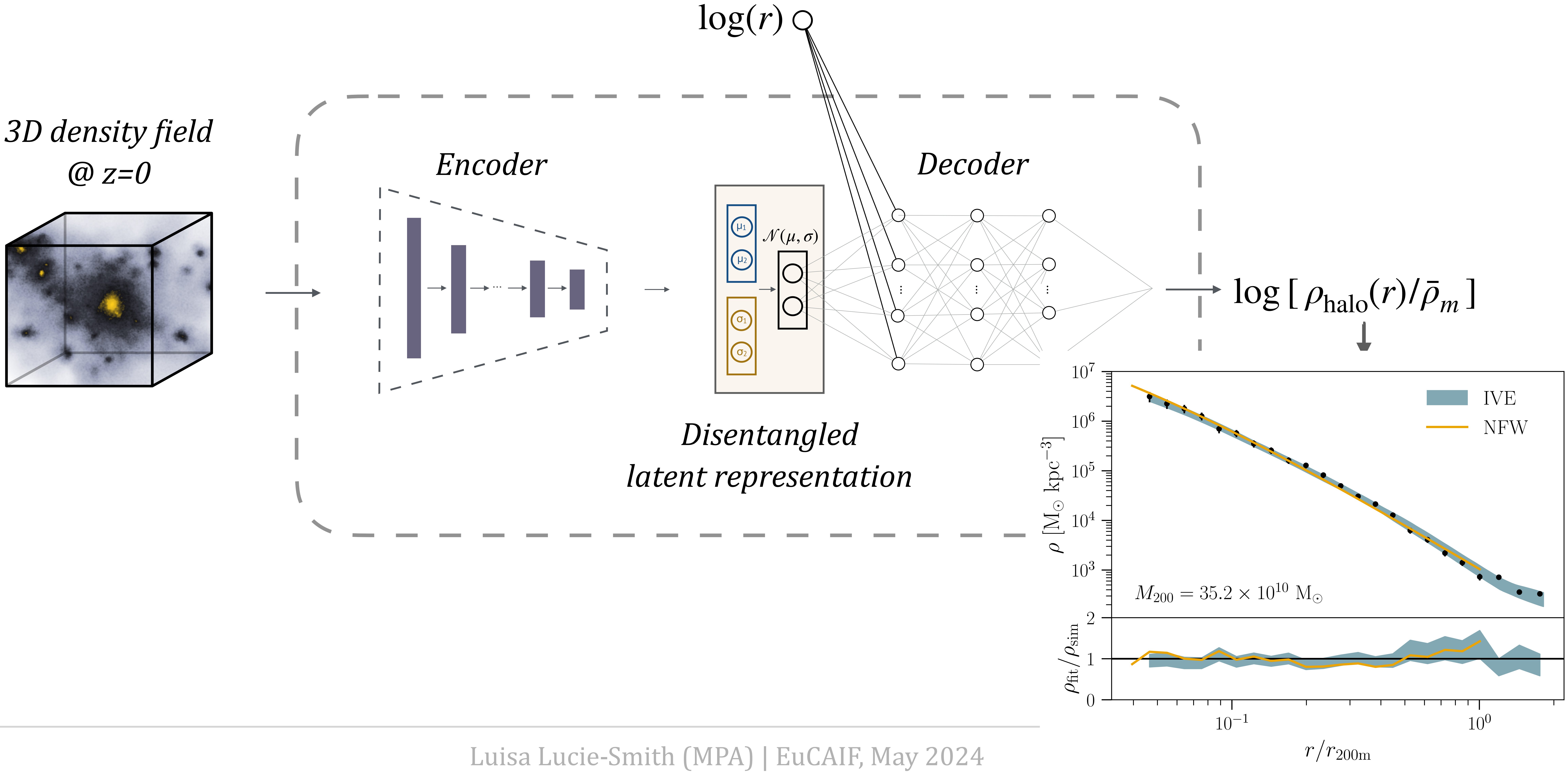
Interpretable Variational Encoder (IVE) for explainable AI



Model compression enables "explainability"

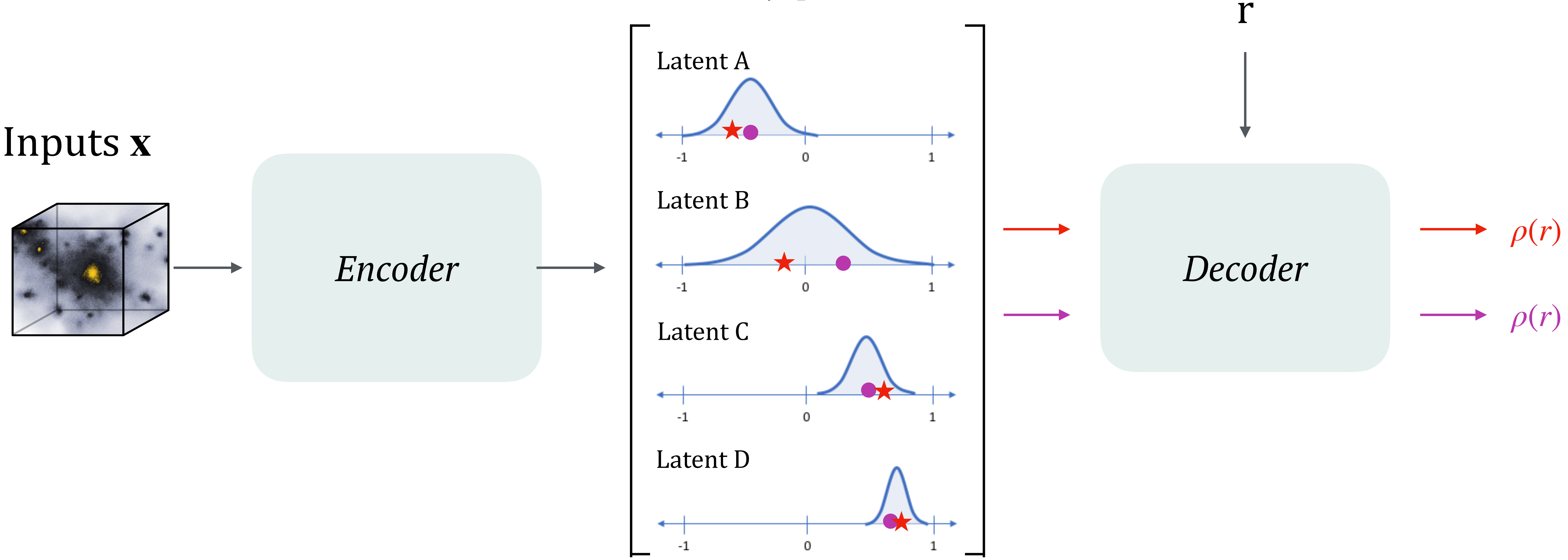
Iten et al. (PRL, 2020); Lucie-Smith et al. (PRD, 2022); Lucie-Smith et al. (PRL, 2024)

Discovering the building blocks of halo density profiles out to the halo outskirts



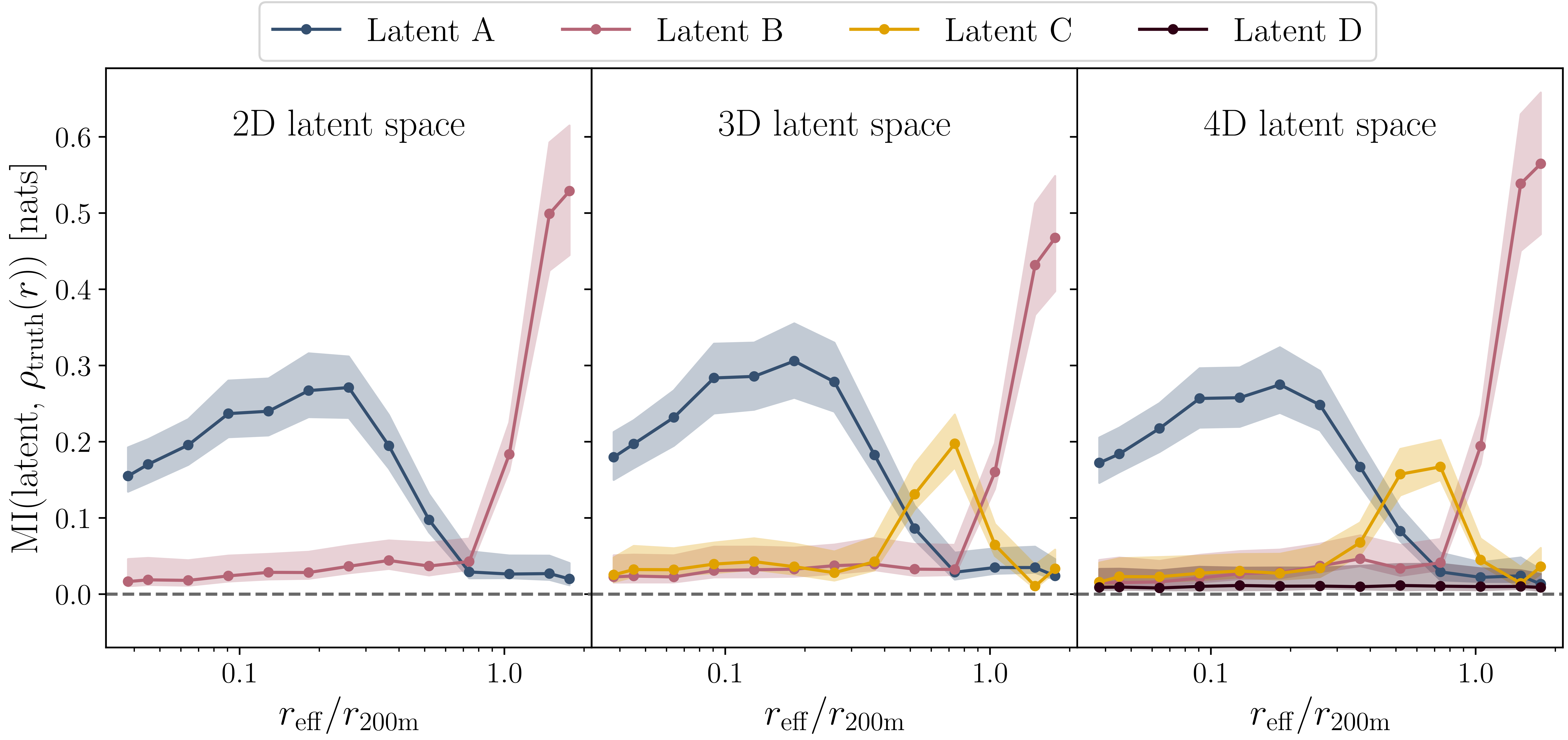
Desired latent representation properties for interpretability

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



- **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

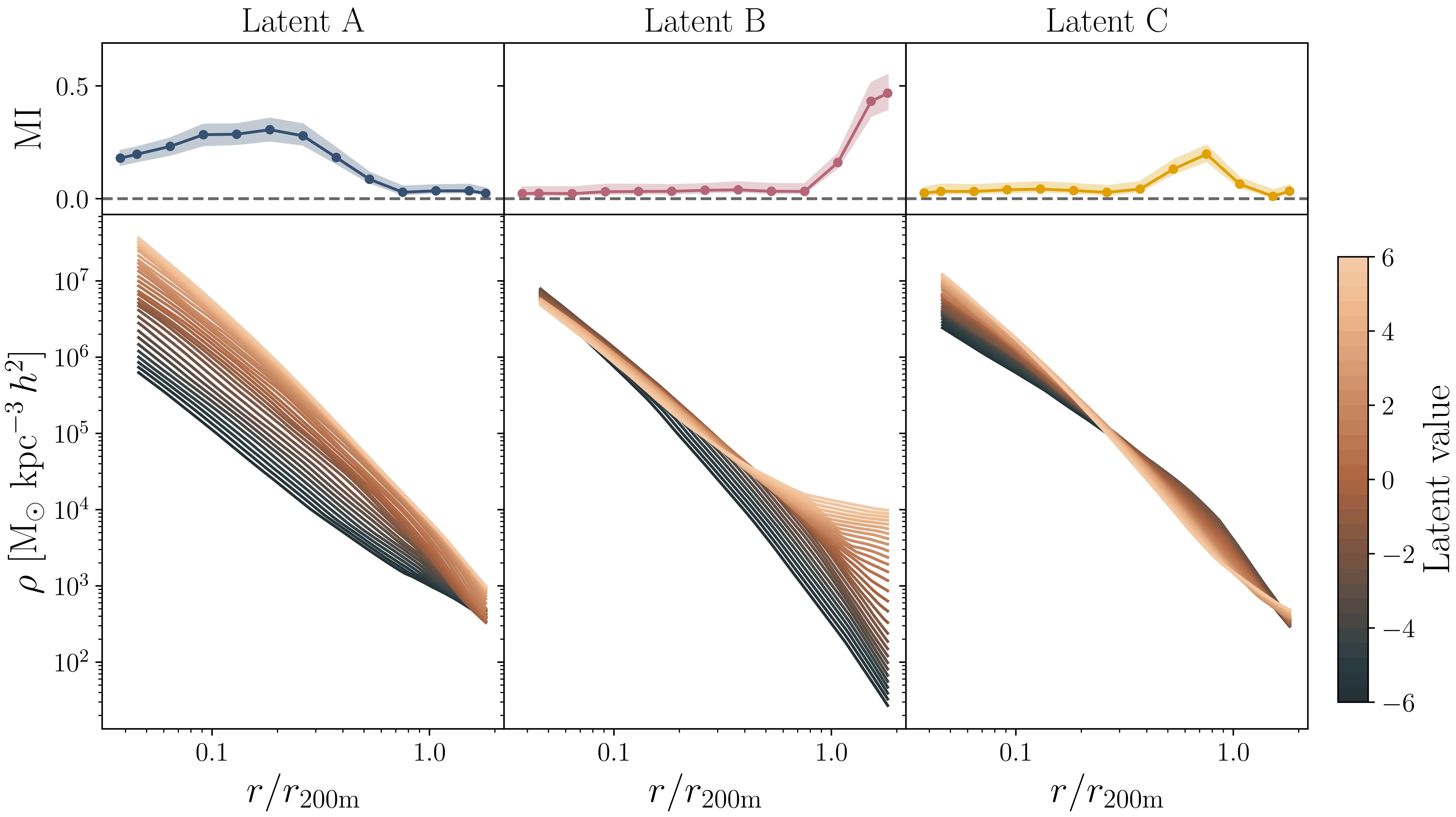
Interpreting the latent representation using mutual information



***Explainability** achieved by evaluating MI between latents and density profile*

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

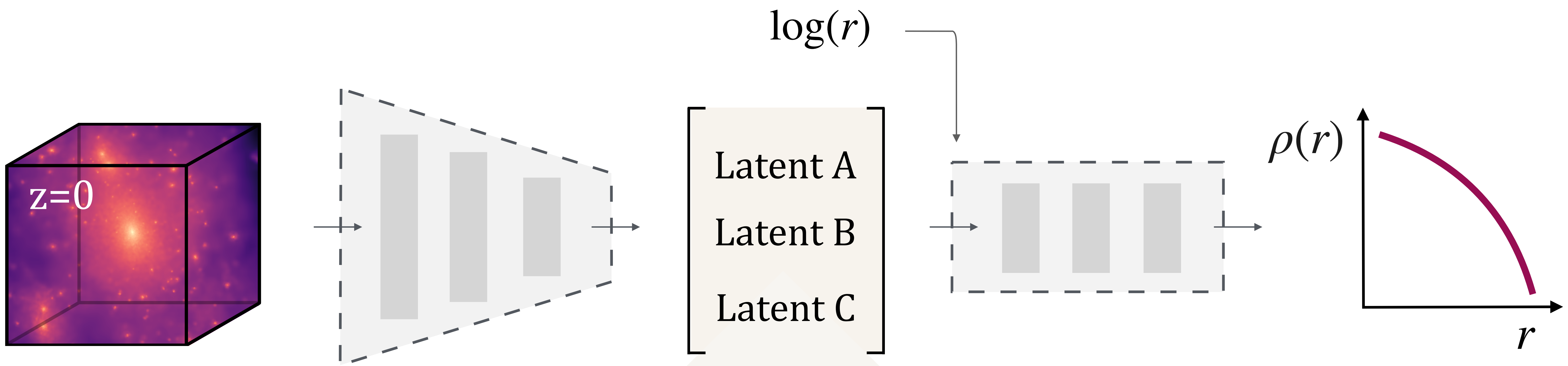
Systematically varying one latent at a time



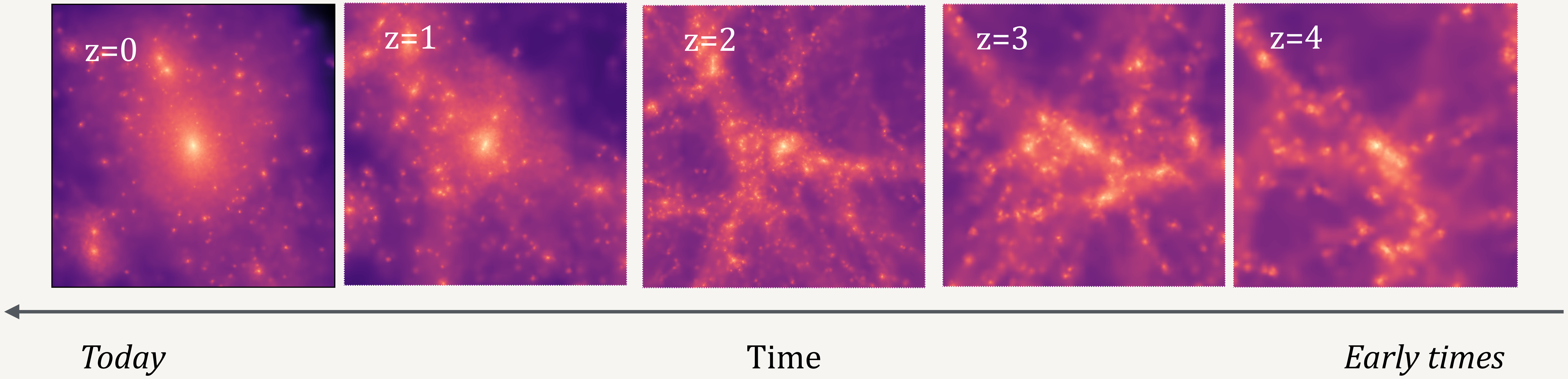
Latent A = ***normalisation***; Latent B = ***outer slope;*** Latent C = ***inner slope***

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

Exploiting the latent representation beyond its original training task

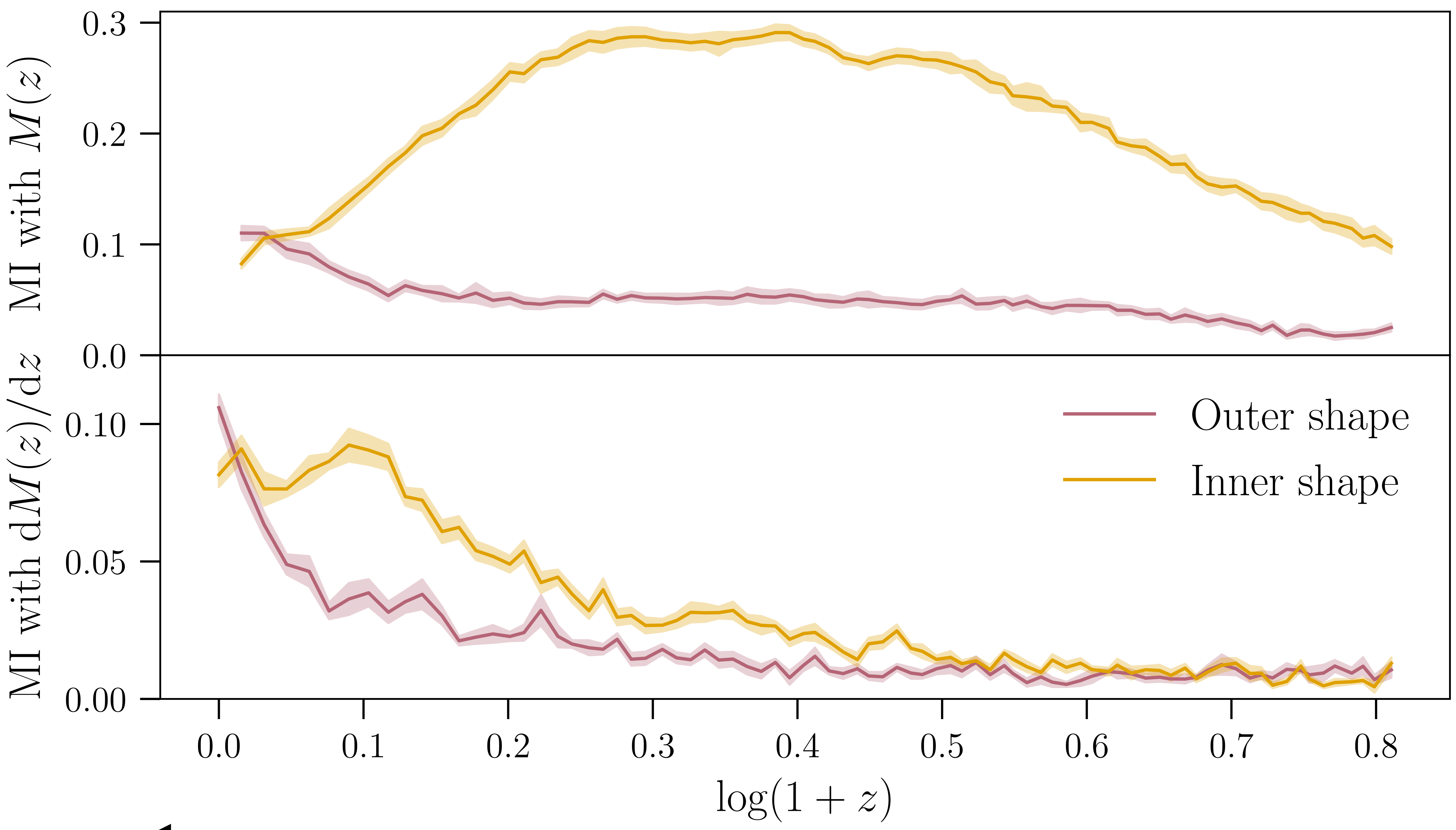


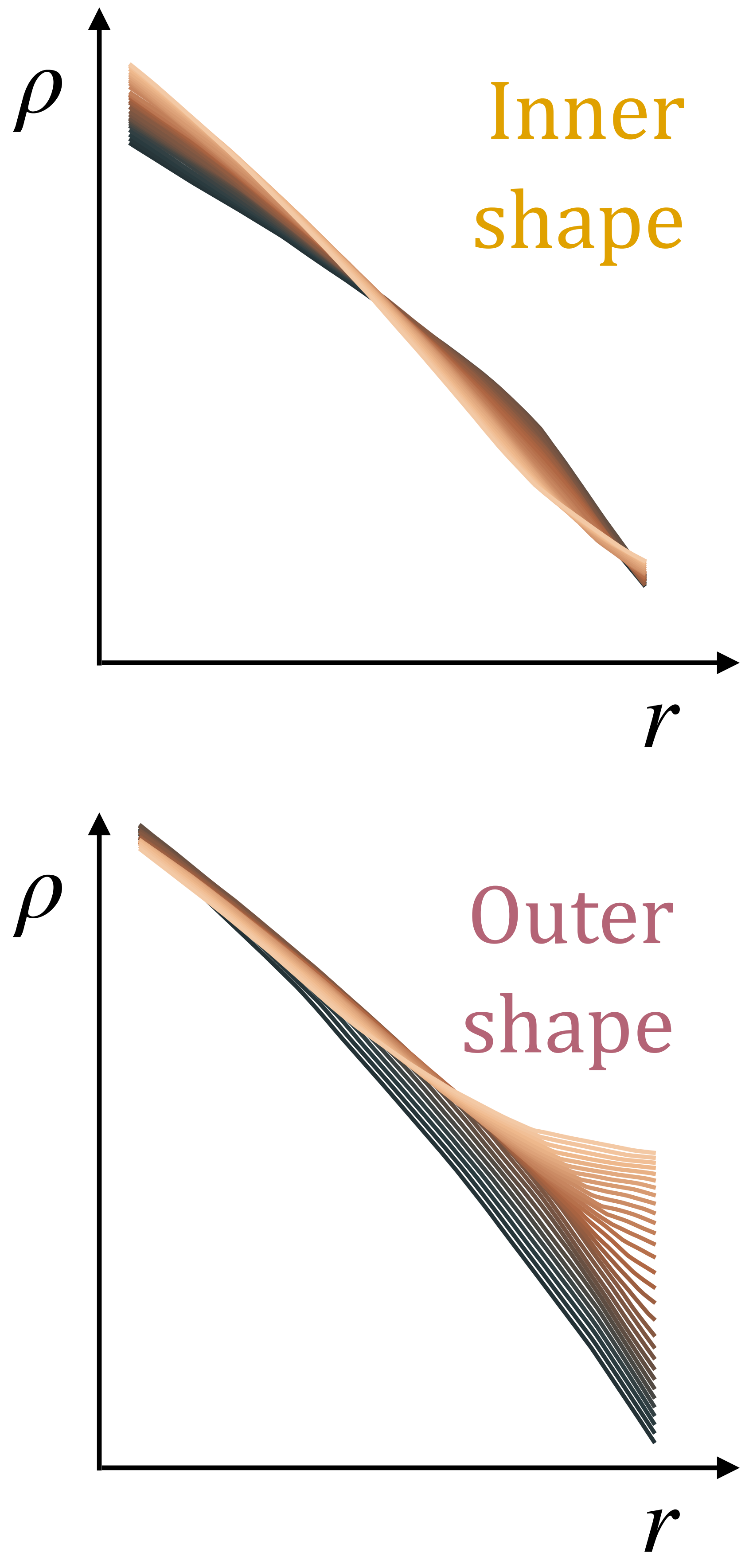
Does the latent space contain information about the origin of the halo density structure?



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

Connection between the latents and the halo evolution history



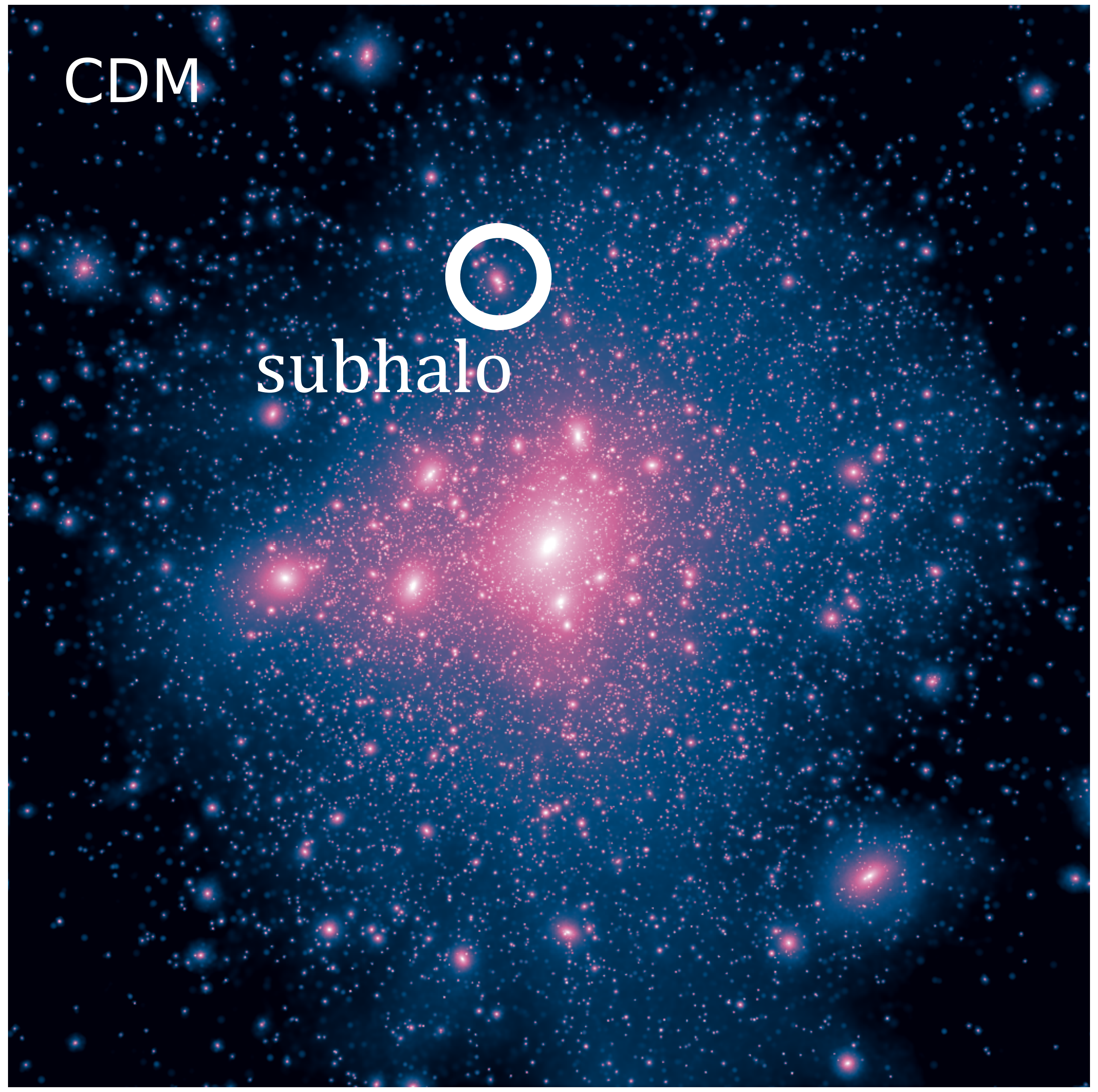


today Time *early times*

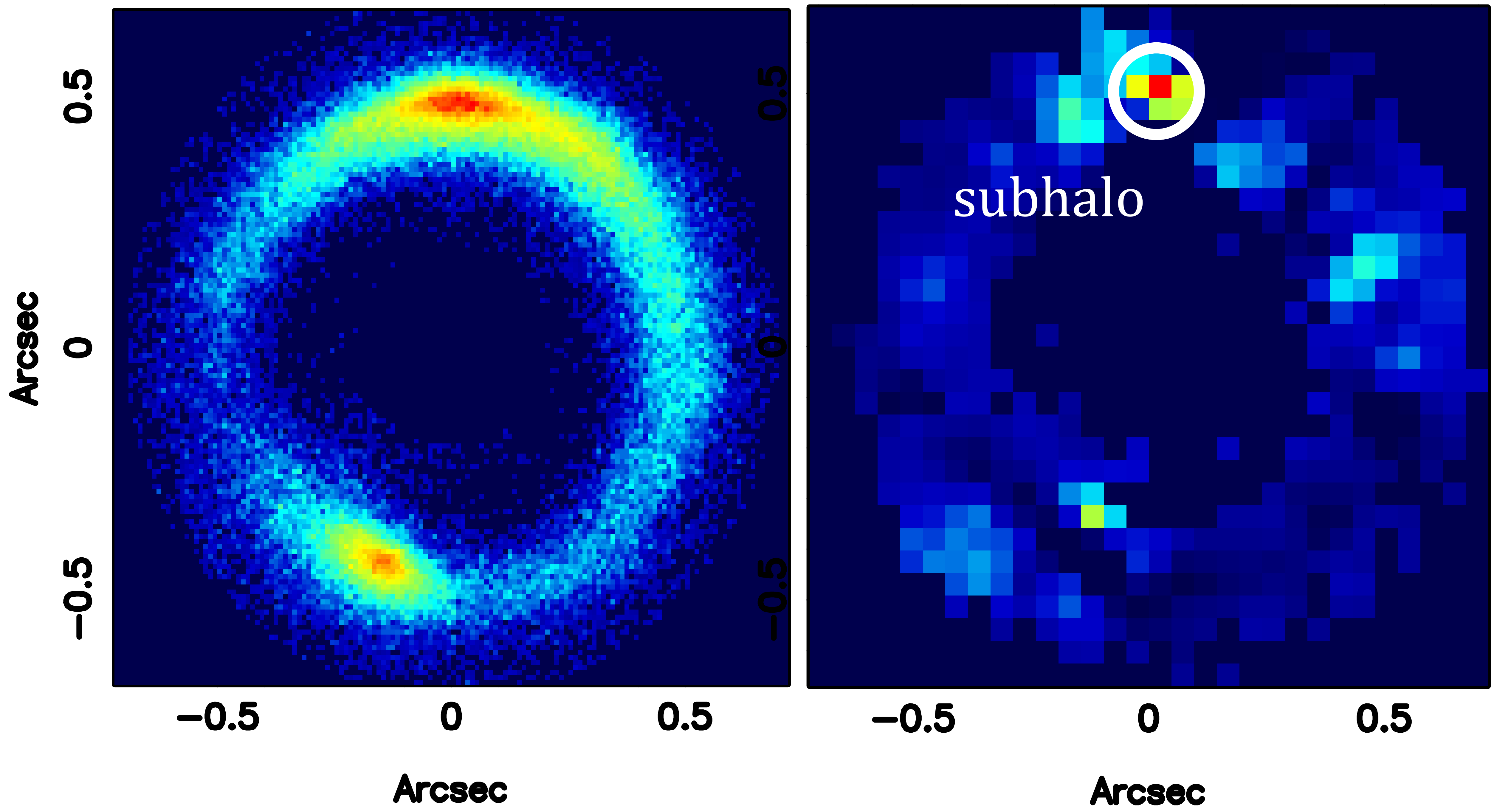
Lucie-Smith, Peiris, Pontzen (Phys. Rev. Lett., 2024)

What about dark matter substructures?

Simulations



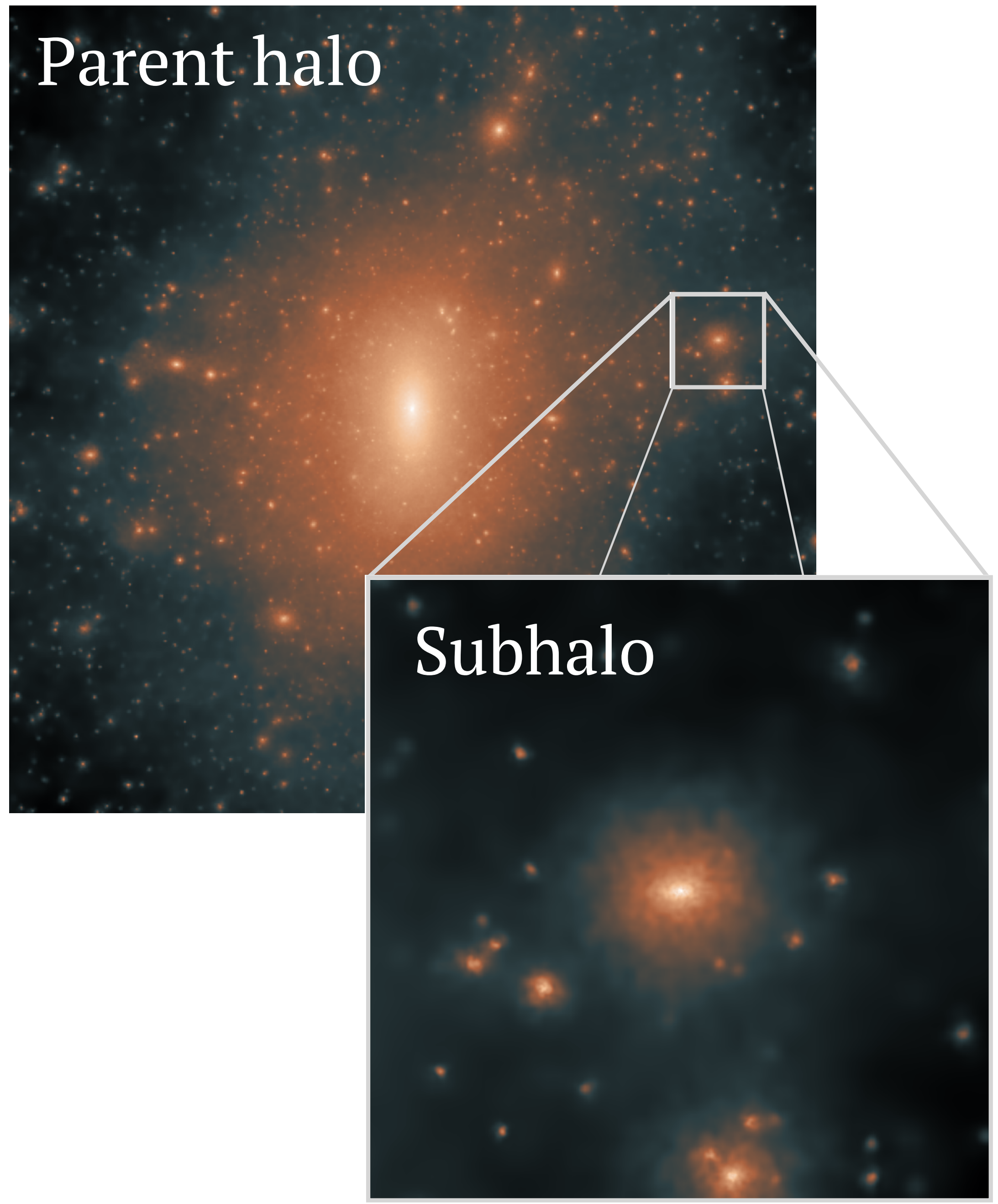
Strong lensing observations



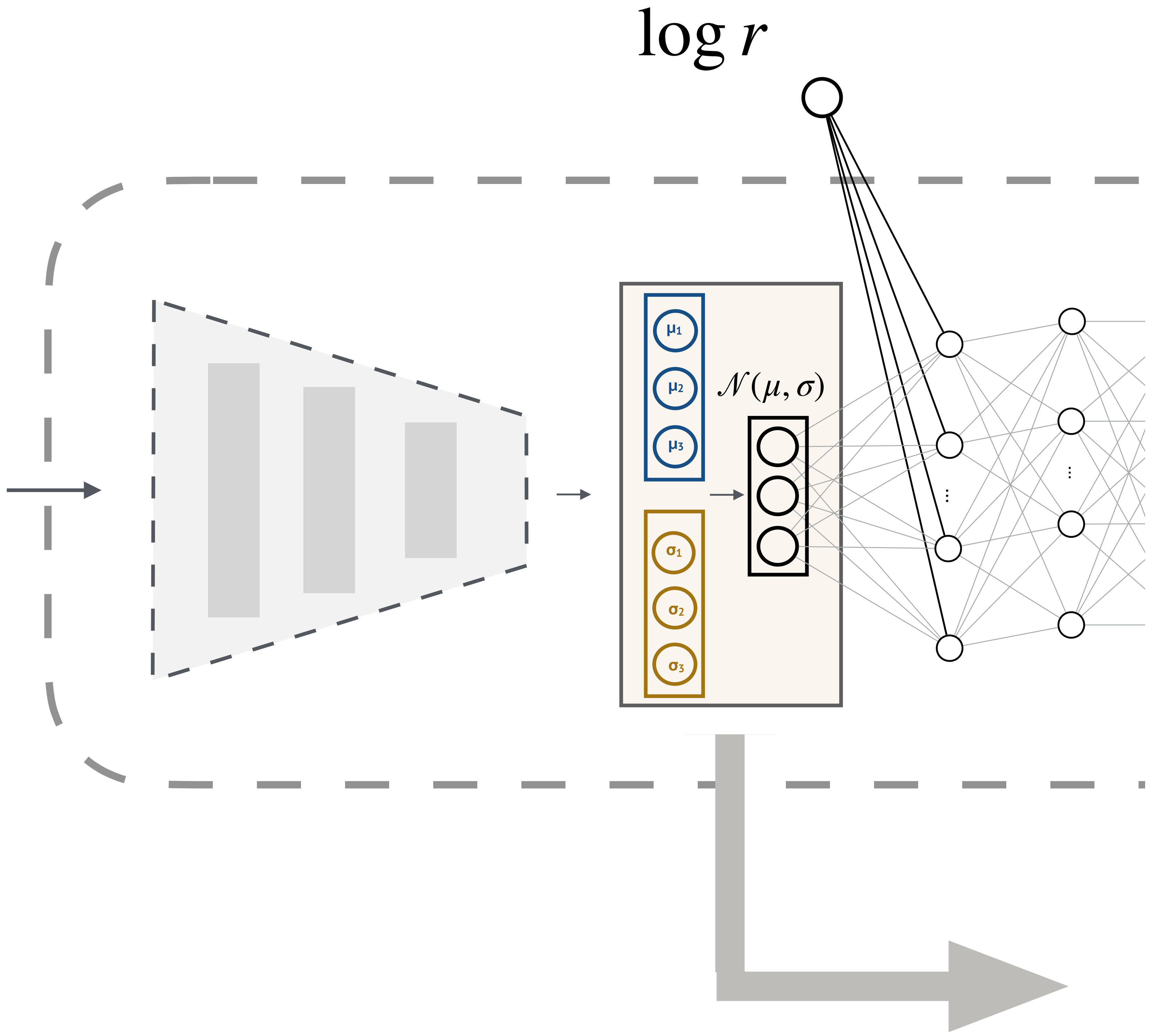
JVAS B1938+666; Vegetti et al. (2012)

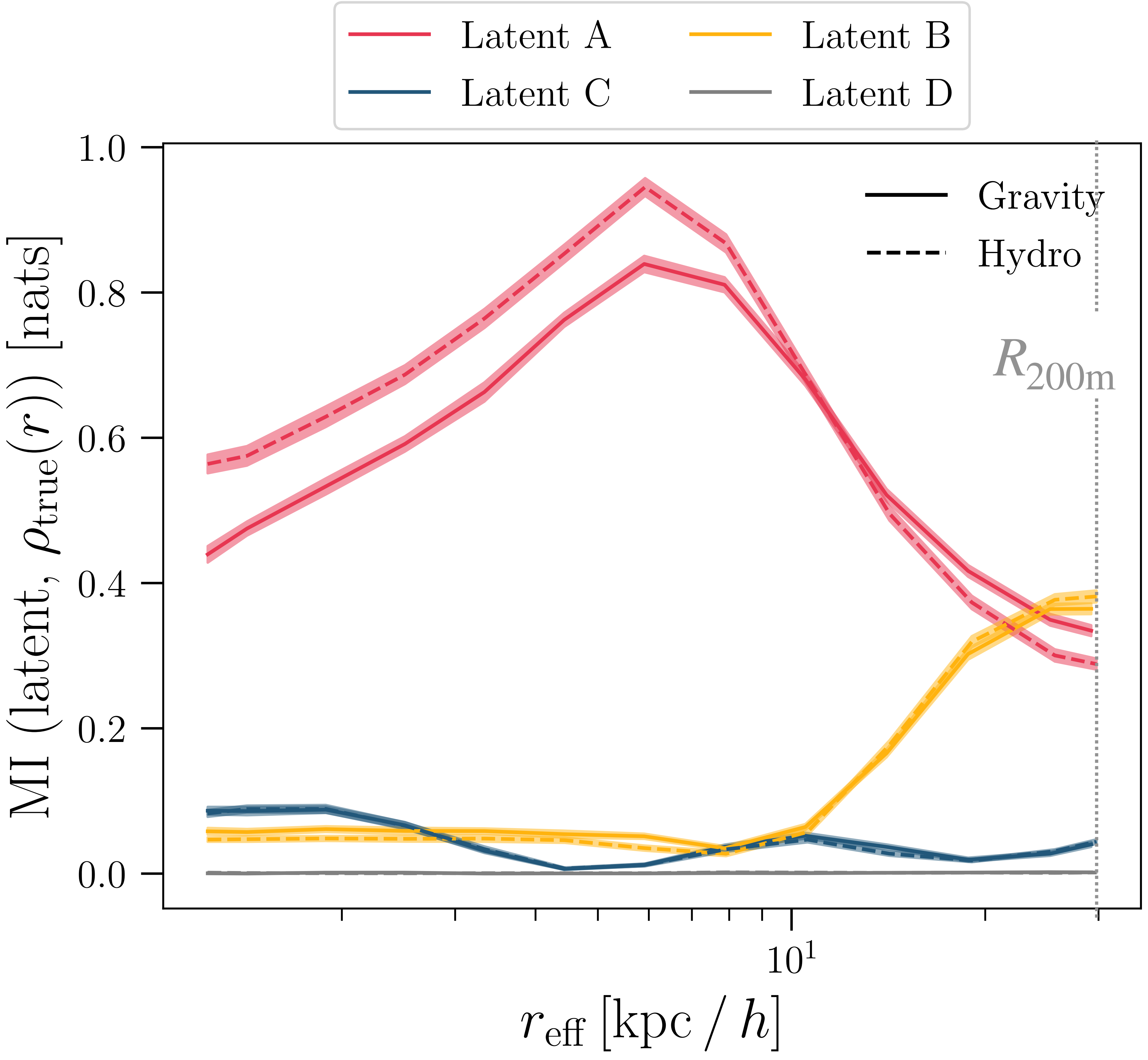
Inferred subhalo properties** depends strongly on **assumed density profiles

What about **subhalos** ($r < R_{200\mathrm{m}}$)?



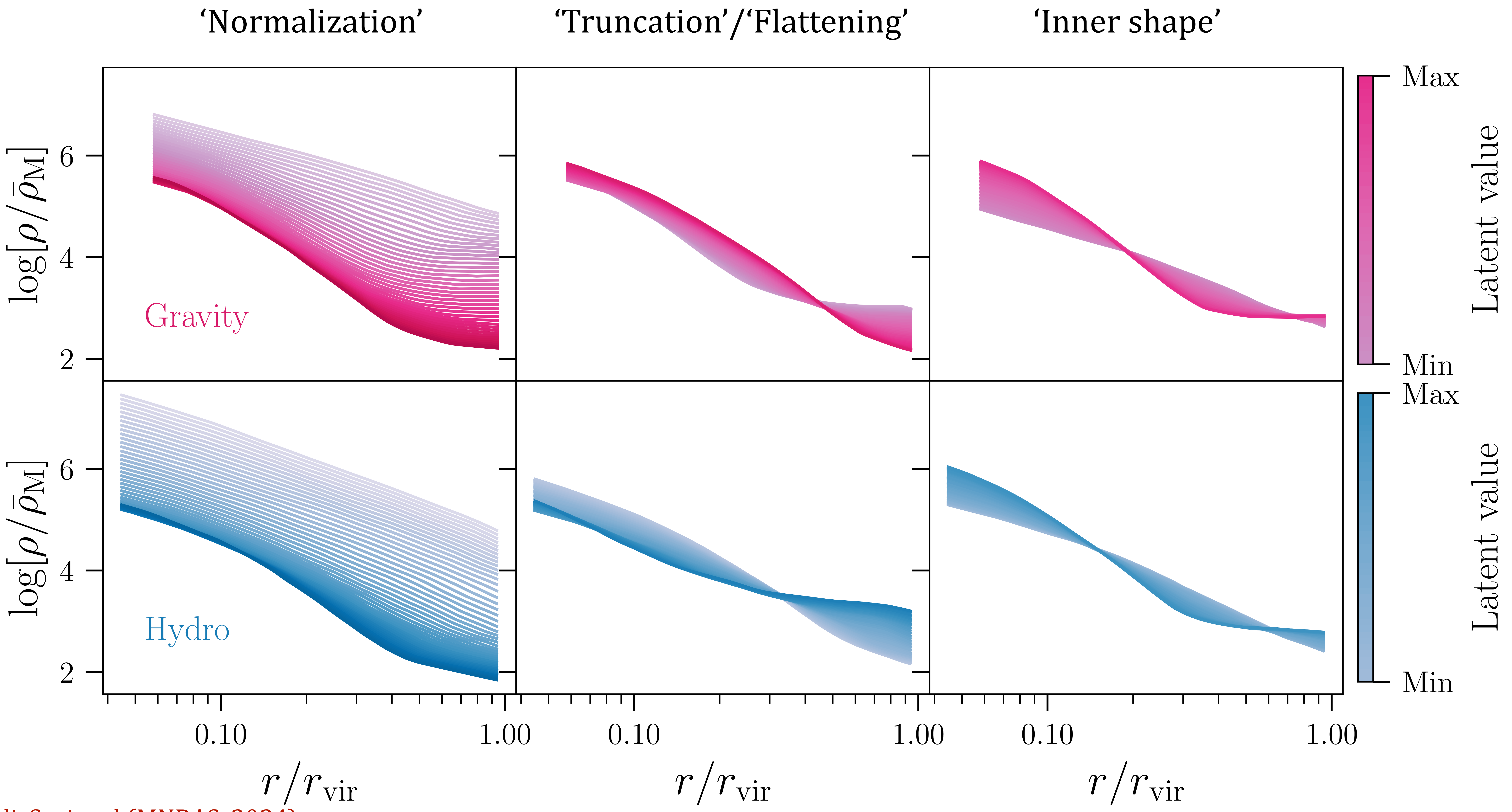
Illustris-TNG100
(Gravity & Hydro)





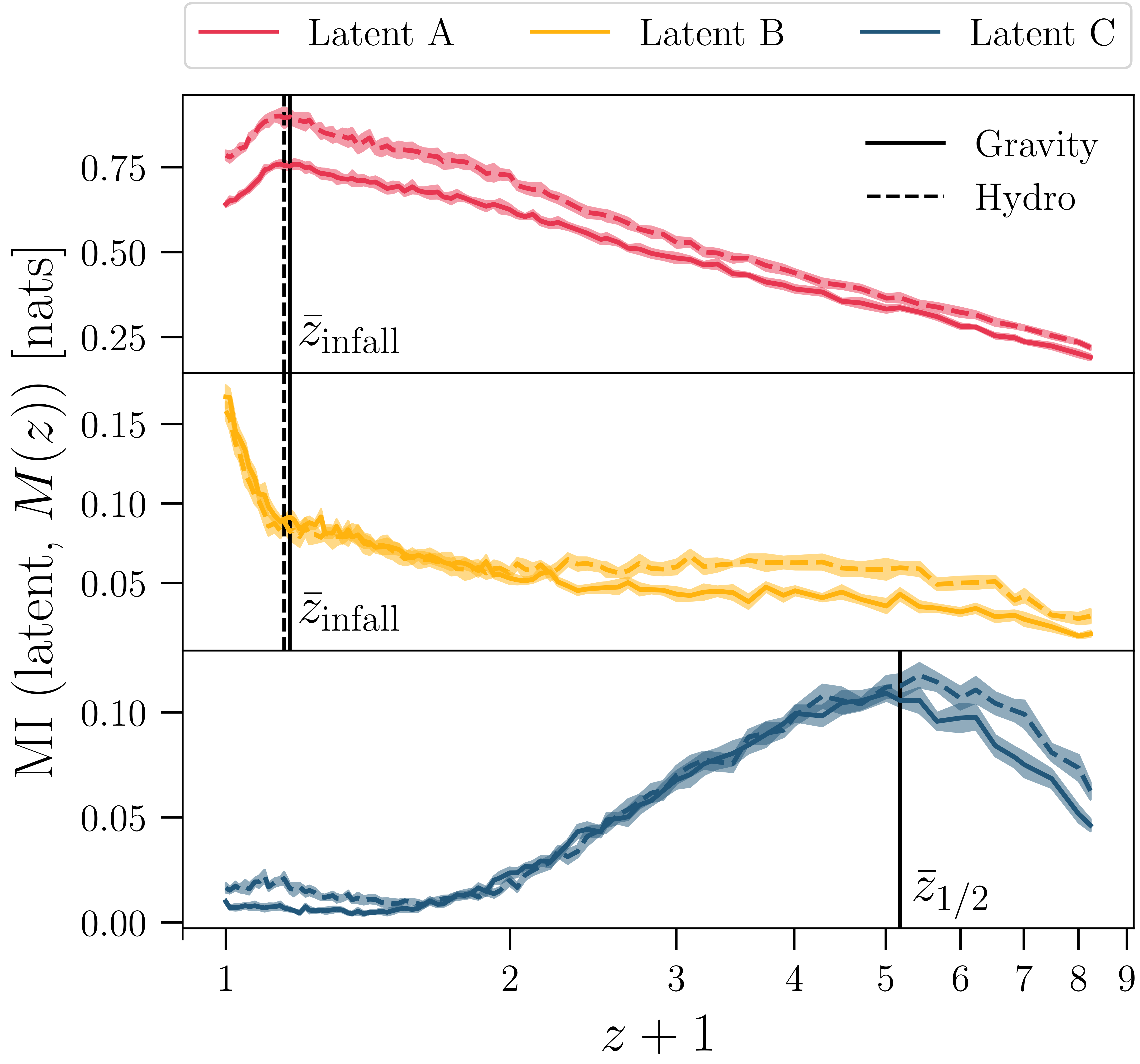
Lucie-Smith, Despali, Springel (MNRAS, 2024)

Subhalos require additional latent capturing tidal truncation



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

Mutual information between latents and M(z)

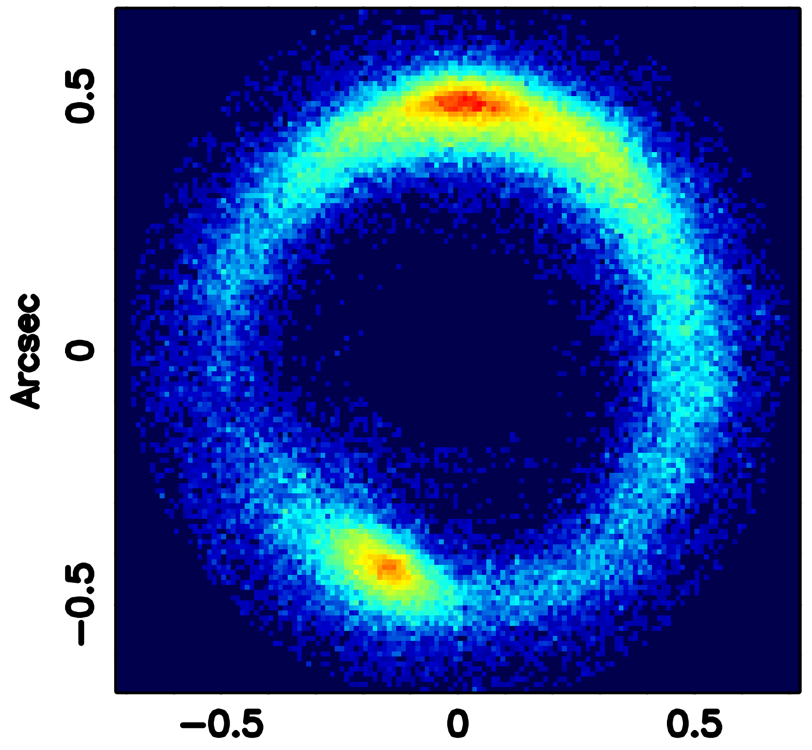


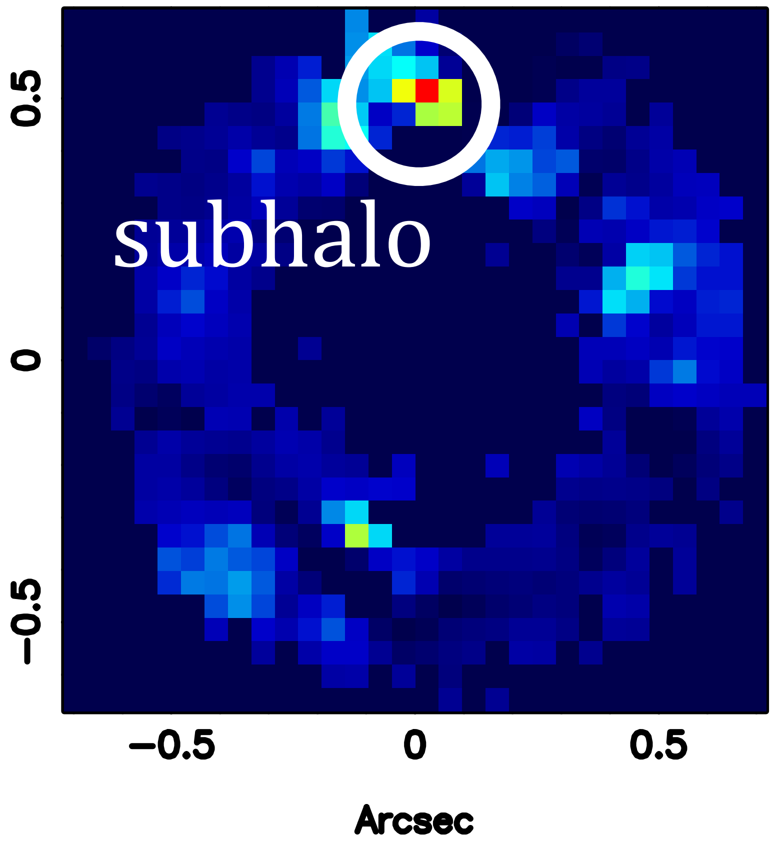
'Normalization' *latent sensitive to formation history* ***before infalling*** *into the main host halo*

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

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

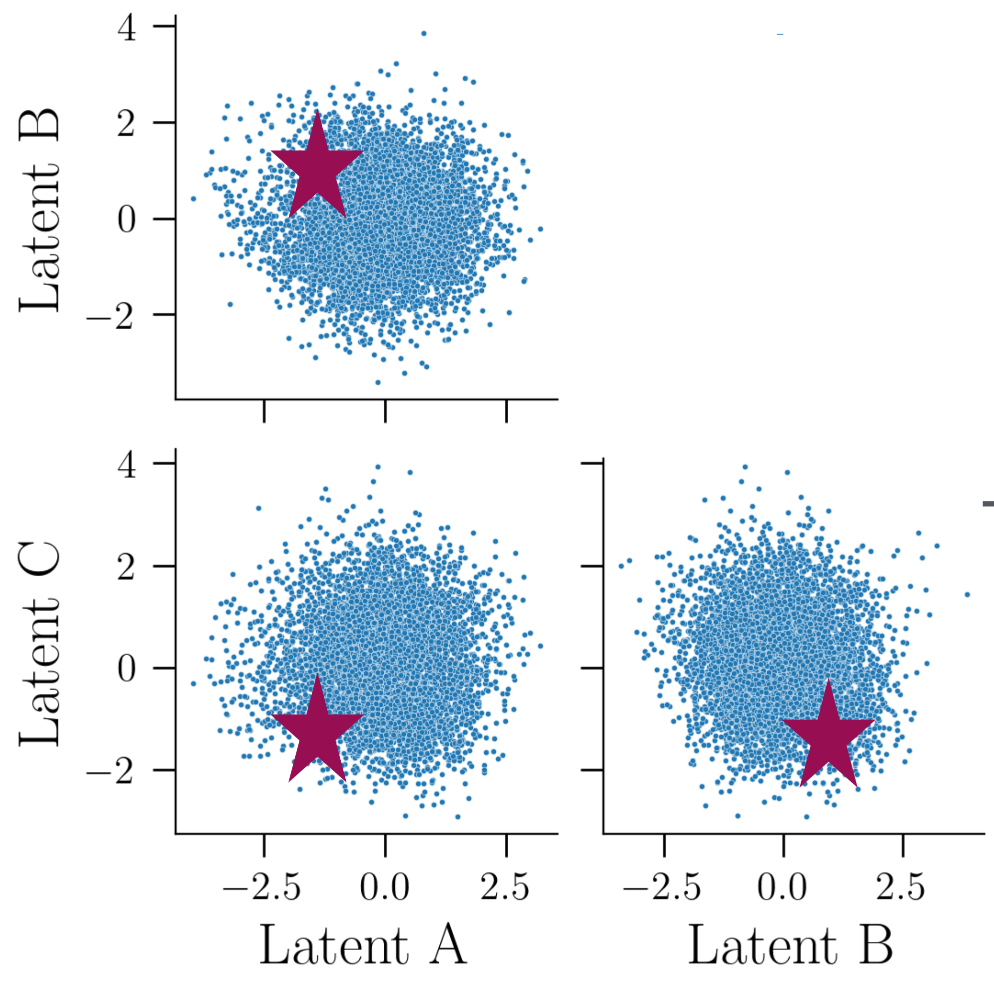
Lucie-Smith, Despali, Springel (MNRAS, 2024)

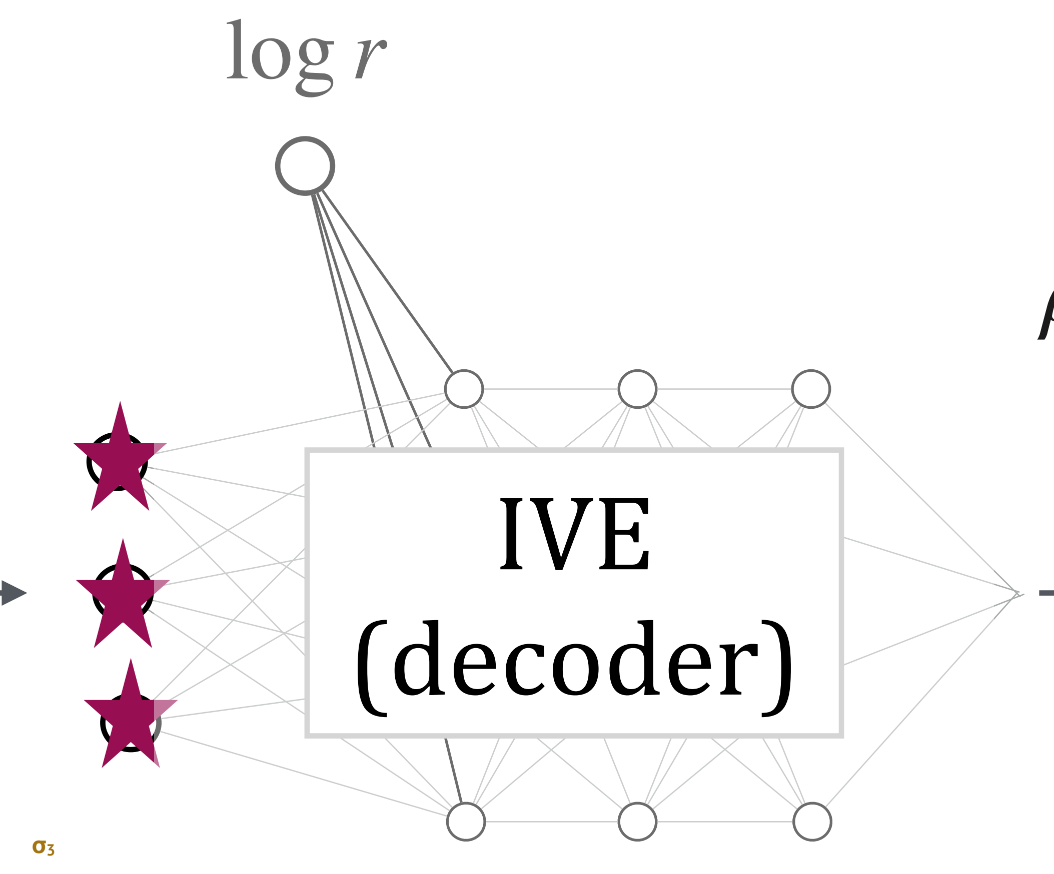




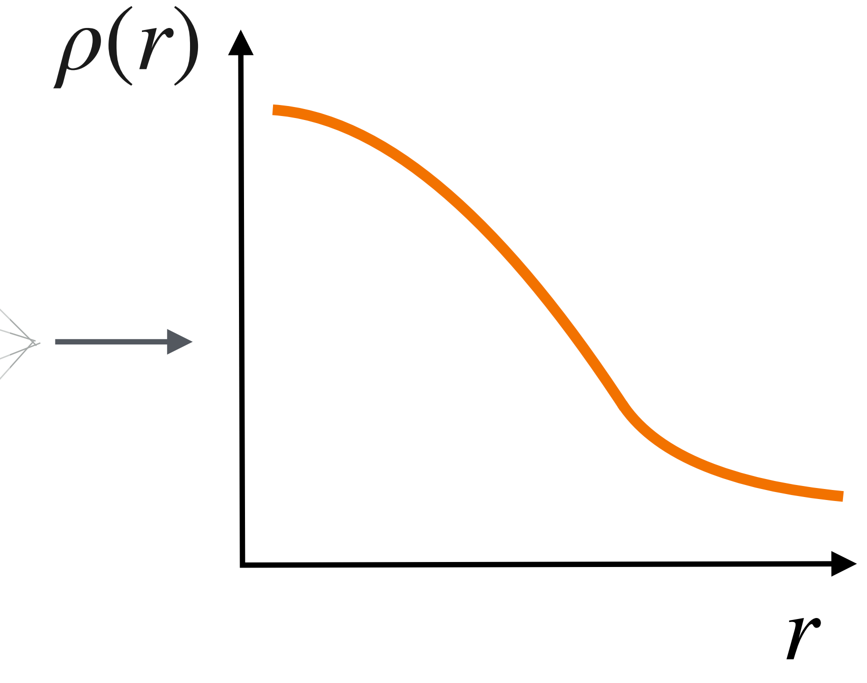
A physically interpretable subhalo density profile for strong gravitational lensing

Sample from the latent space





Infer subhalo density profile



Next step: Integrate IVE model within strong gravitational lensing pipeline

Conclusions

- Interpretable variational encoders (IVE) provide new avenue to provide robust, physically interpretable models
- IVE disentangles different physical effects in minimal set of ingredients
- Explainable AI shows promise in enabling new, data-driven scientific discoveries

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

MAX PLANCK INSTITUTE
FOR ASTROPHYSICS

IVE loss function

β must be carefully fine-tuned
to balance accuracy with disentanglement

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:

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

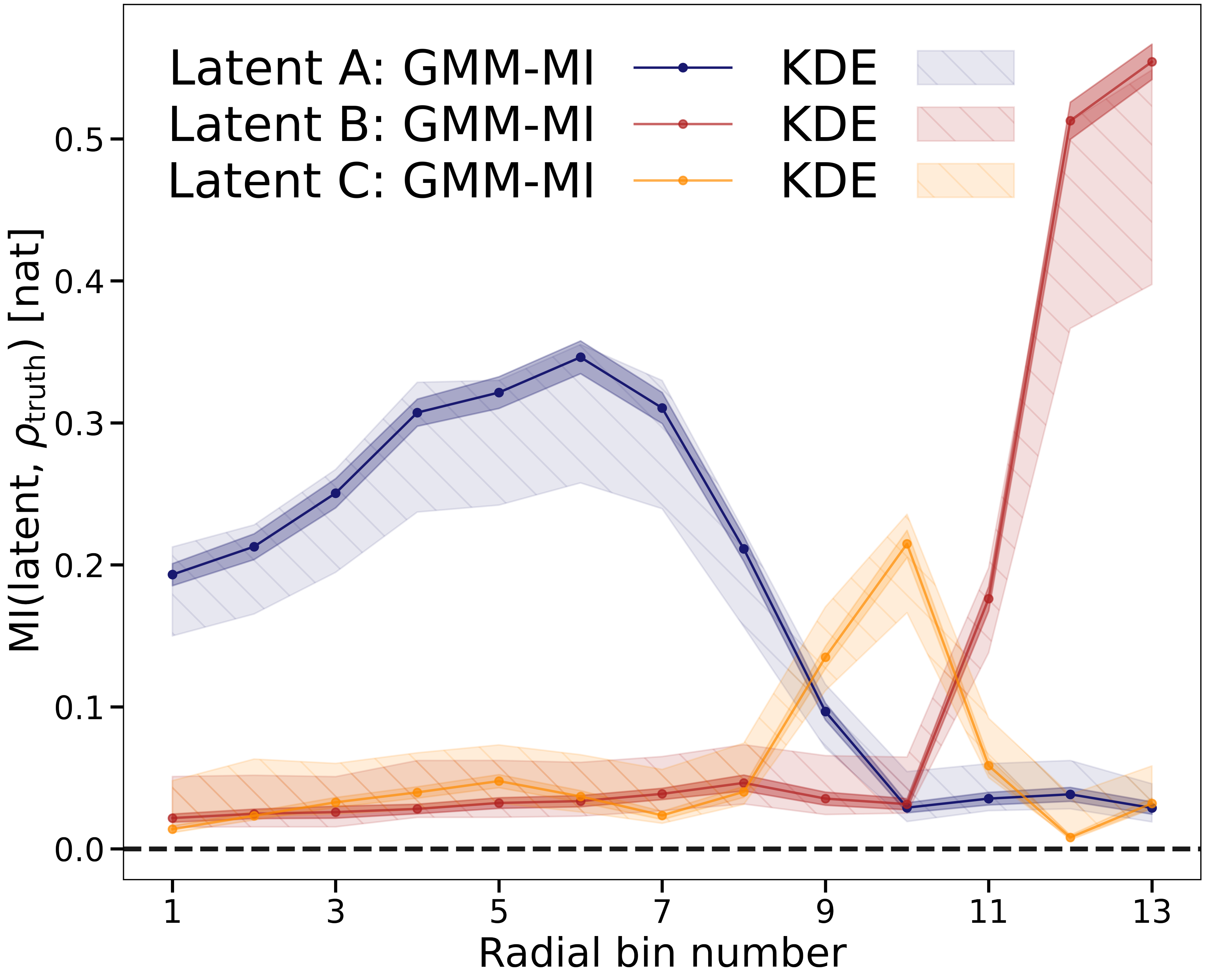
Prior:

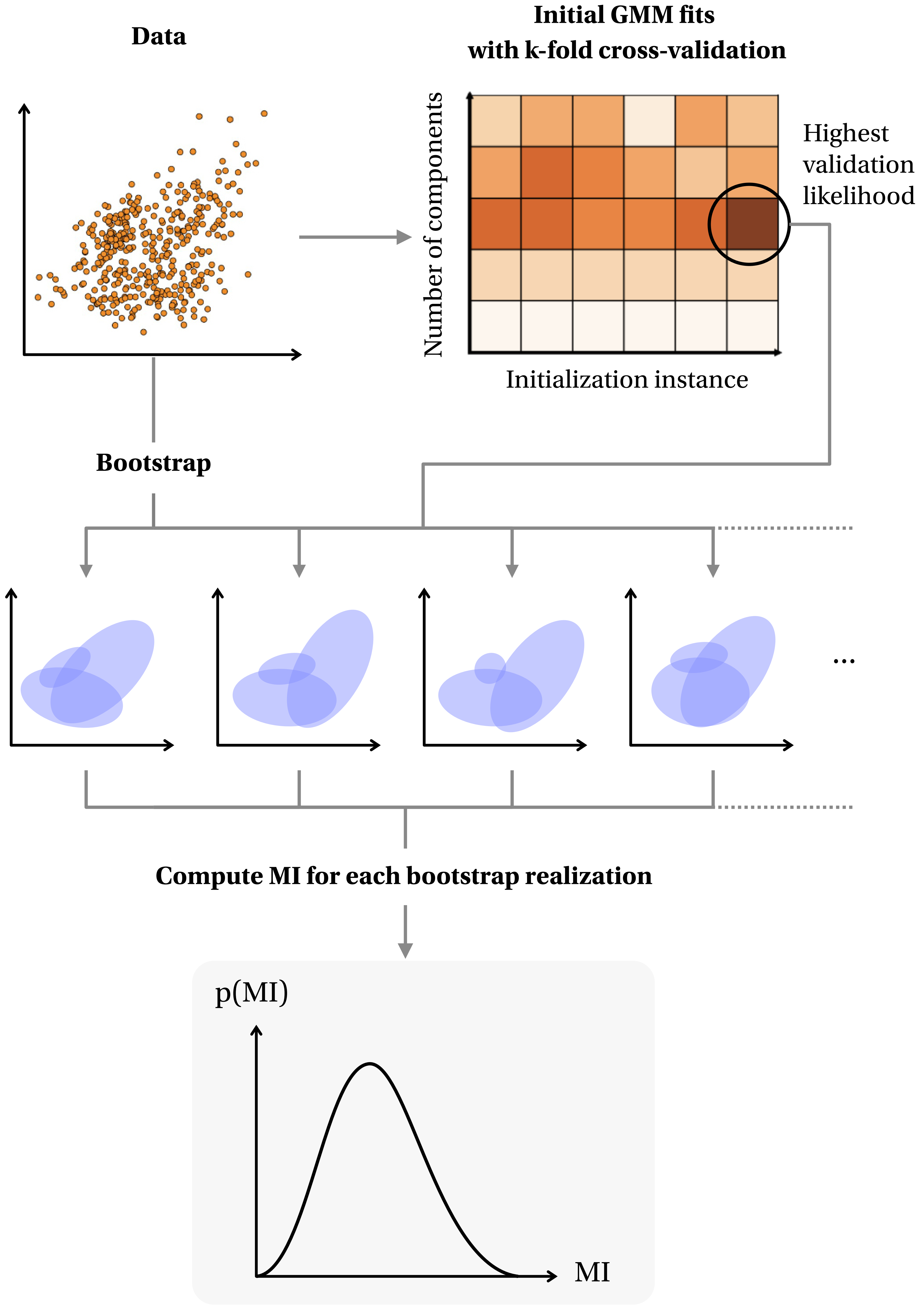
$$q(\mathbf{z}) = \prod_{i=1}^{L} \mathscr{N}(0,1)$$

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

Mutual information

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



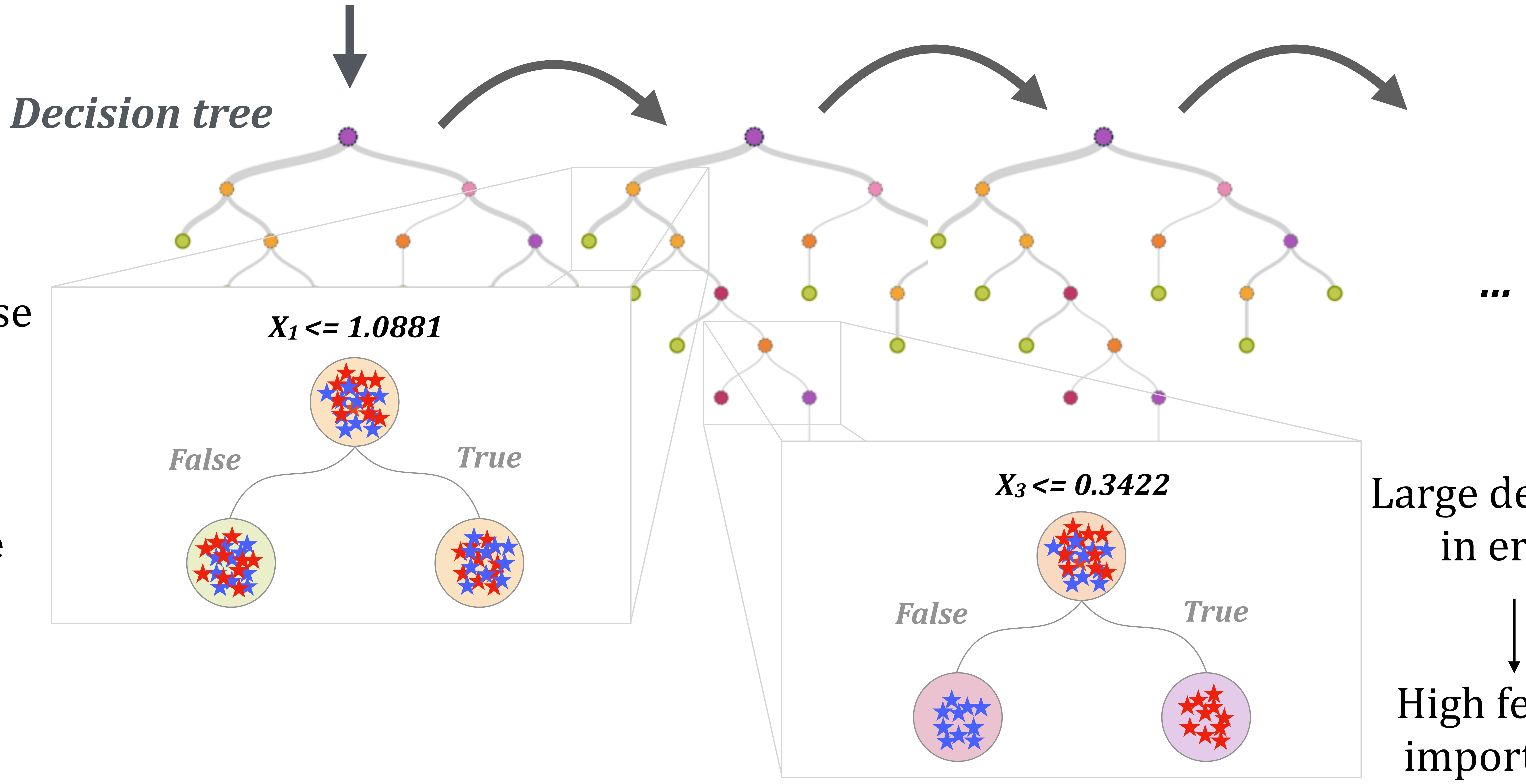


HTTPS://GITHUB.COM/DPIRAS/GMM-MI

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



Small decrease in error

↓

Low feature importance

Large decrease in error

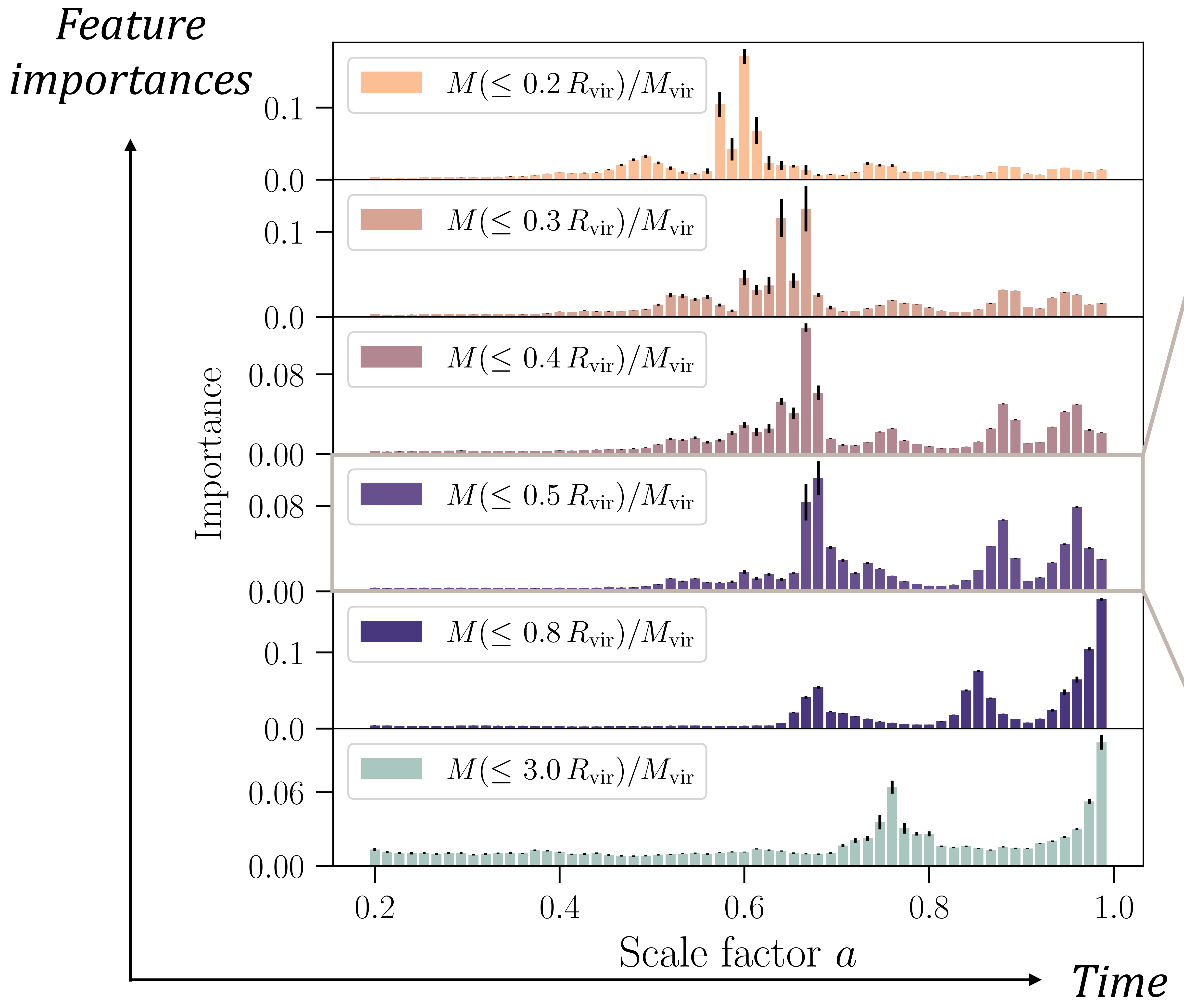
↓

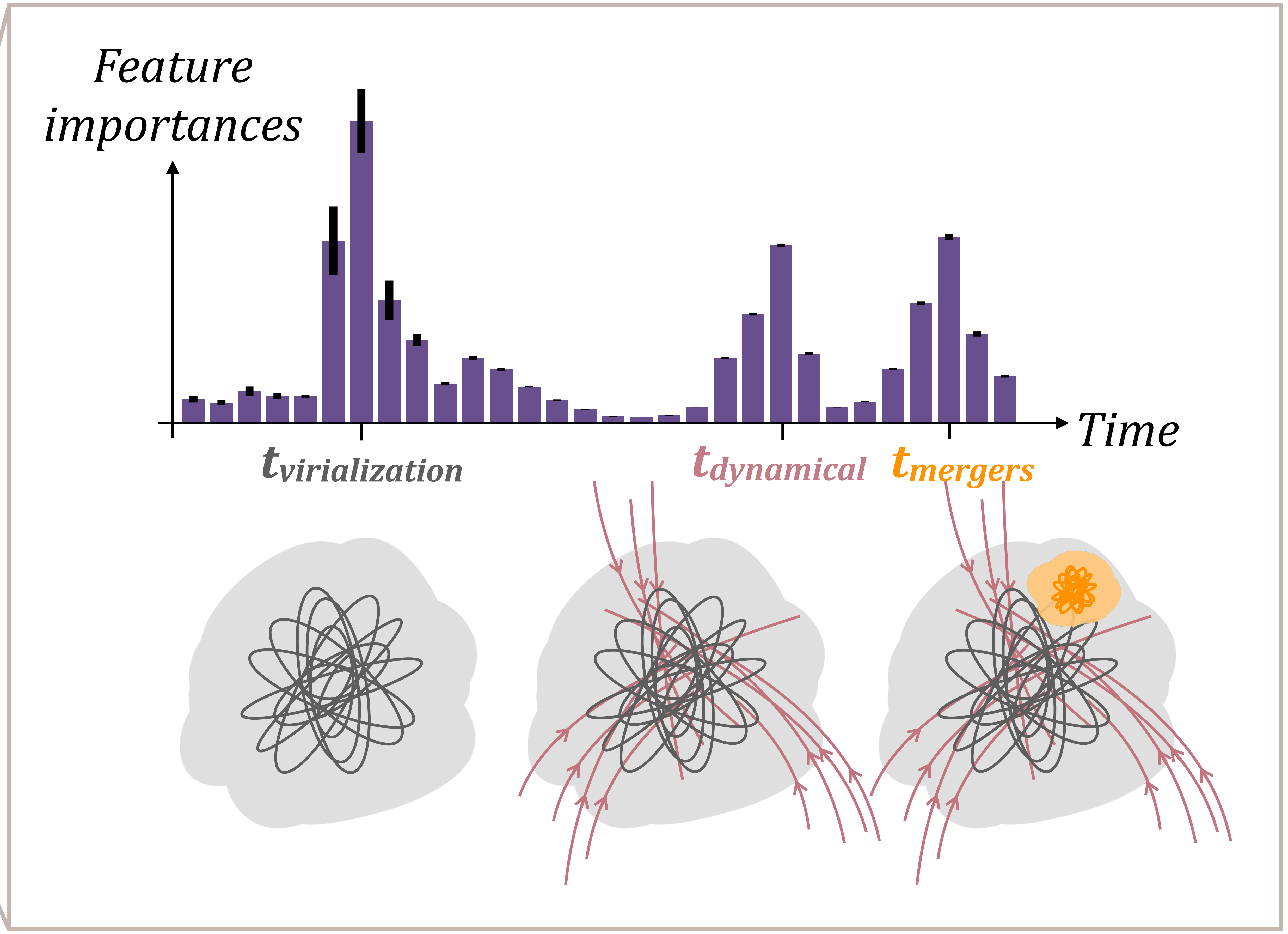
High feature importance

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

Friedman, 2001; Friedman, 2002

Importance of mass accretion history in predicting cluster mass profiles

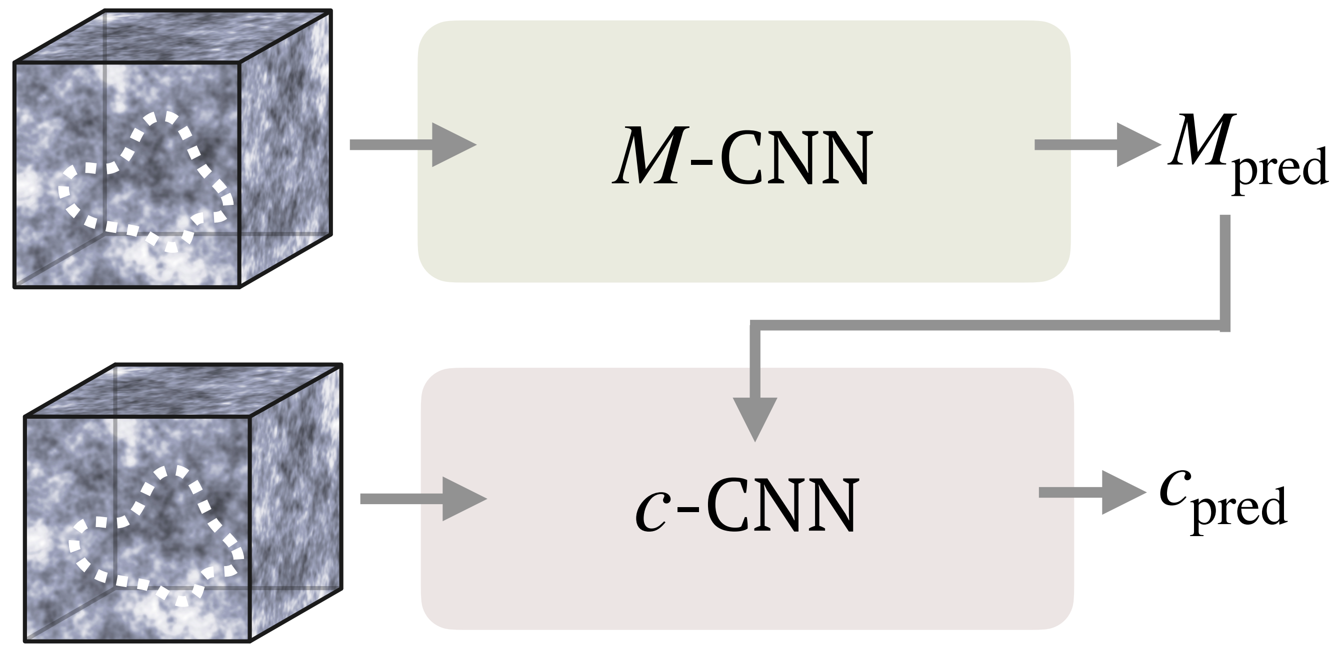




Lucie-Smith, Adhikari, Wechsler (MNRAS, 2022)

Interpretability methods for deep learning

- Saliency methods



My view: visualization methods give some insight but lack quantitative conclusions