Title: Explainable AI in (Astro)physics

Speakers: Luisa Lucie-Smith

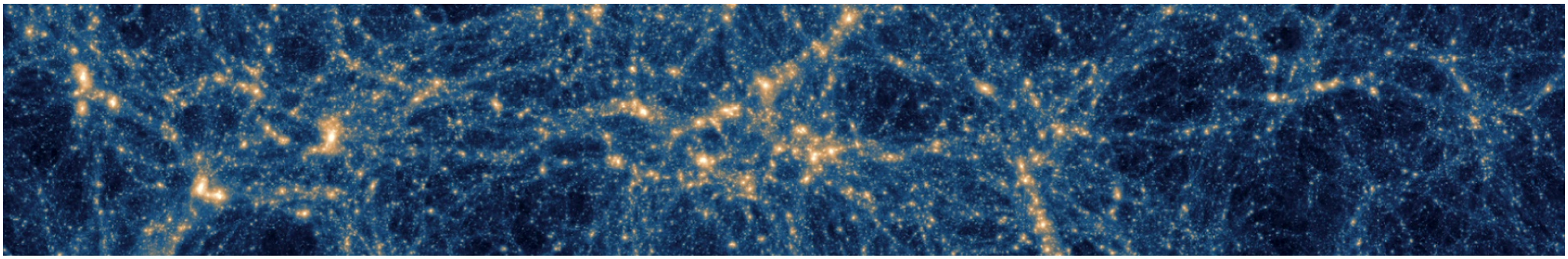
Collection/Series: Theory + AI Workshop: Theoretical Physics for AI

Date: April 11, 2025 - 9:45 AM

URL: https://pirsa.org/25040098

Abstract:

Machine learning has significantly improved the way scientists model and interpret large datasets across a broad range of the physical sciences; yet, its "black box" nature often limits our ability to trust and understand its results. Interpretable and explainable AI is ultimately required to realize the potential of machine-assisted scientific discovery. I will review efforts toward explainable AI focusing in particular in applications within the field of Astrophysics. I will present an explainable deep learning framework which combines model compression and information theory to achieve explainability. I will demonstrate its relevance to cosmological large-scale structures, such as dark matter halos and galaxies, as well as the cosmic microwave background, revealing new physical insights derived from these explainable AI models.



Explainable AI in (Astro)physics

Luisa Lucie-Smith
University of Hamburg (UHH)

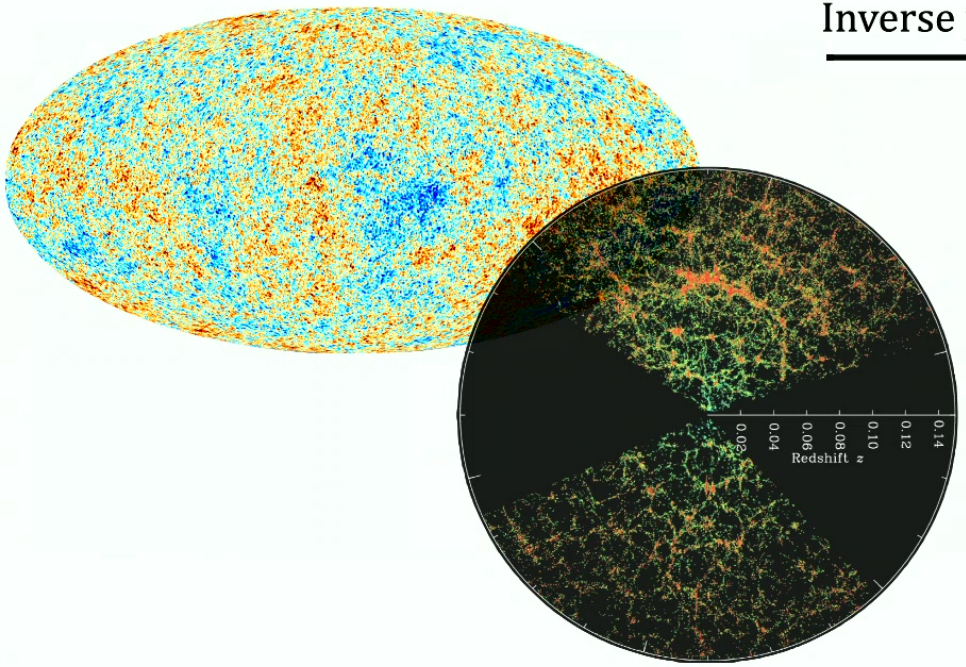
AI+Theory workshop @ Perimeter Institute, 11th April 2025

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

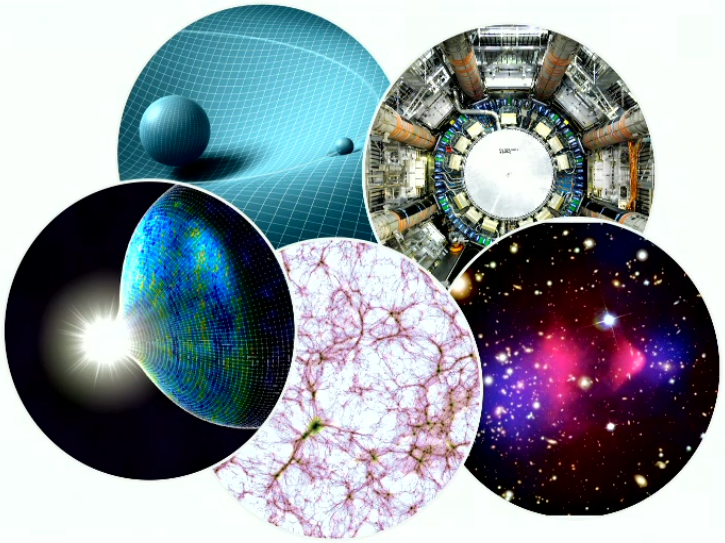
Theory vs data

Data



Inverse problem

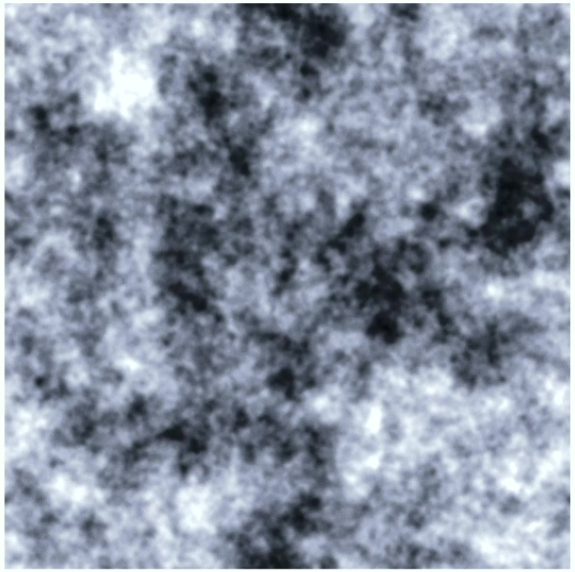
Theory

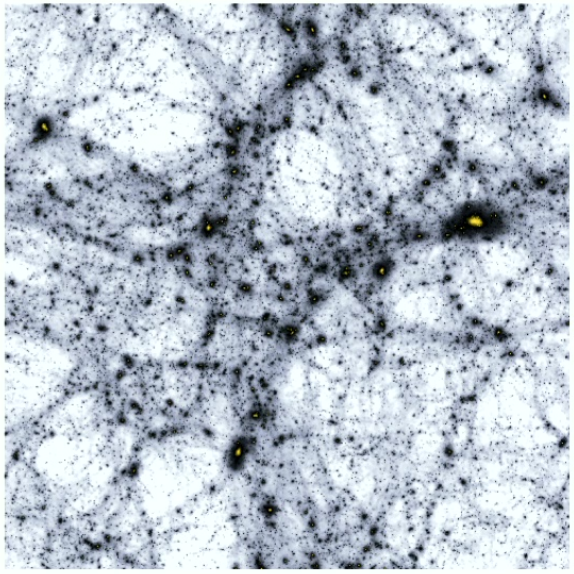


6 cosmological parameters

Theoretical challenge in cosmology

Cosmological interpretation increasingly reliant on evaluating
computationally-costly, non-linear models with many parameters.





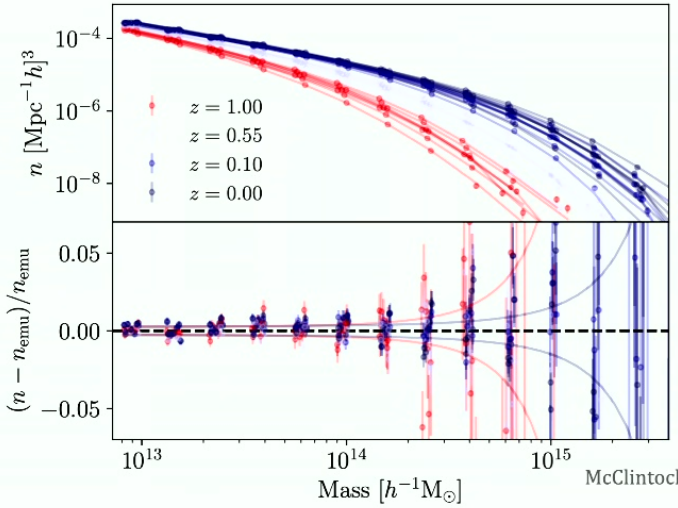
Simple, linear initial conditions

Complex, non-linear large-scale structure

AI for Physics: emulation/acceleration

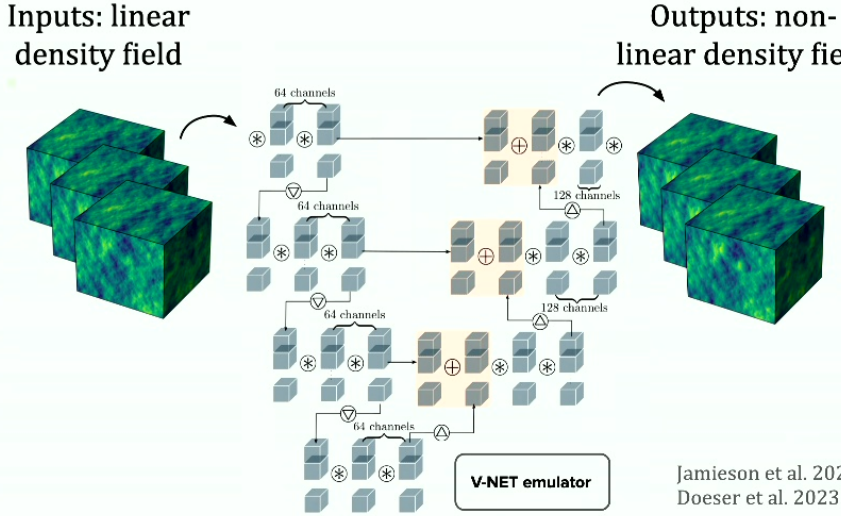
- *accelerate computationally expensive models of structure formation*

Theory emulation of summary statistics



McClintock et al. (2019)

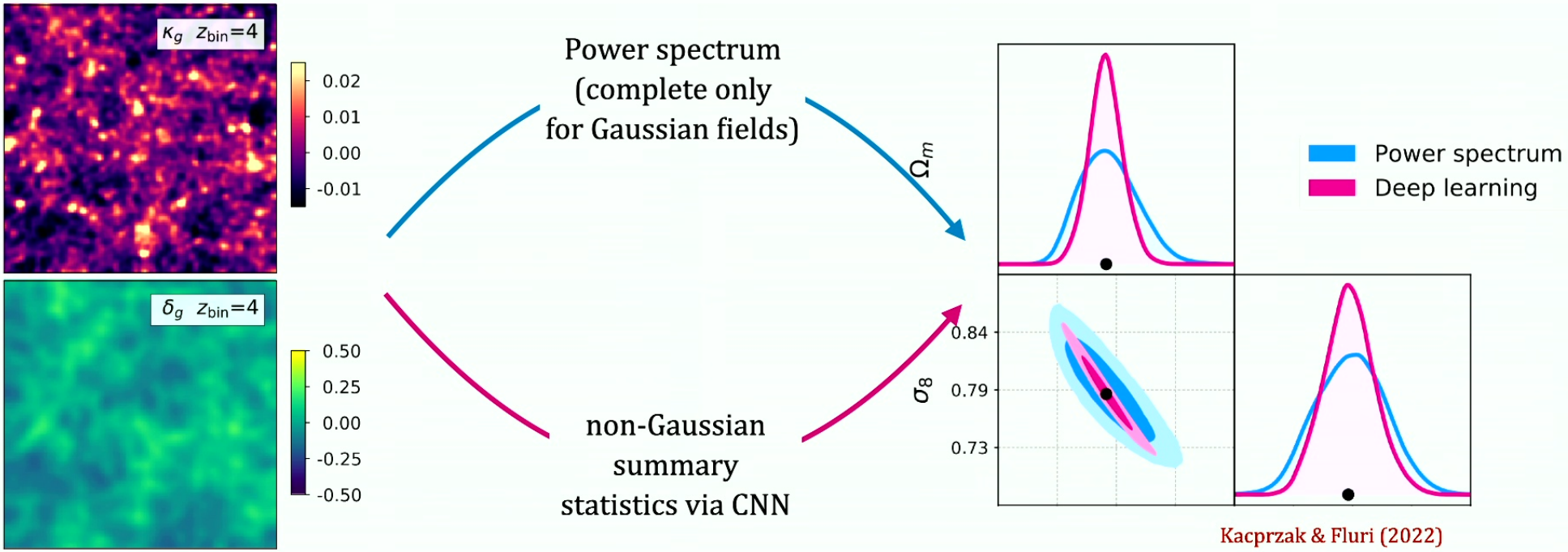
From linear inputs to non-linear outputs



Jamieson et al. 2023; Doeser et al. 2023

AI for Physics: extracting non-Gaussian information from data

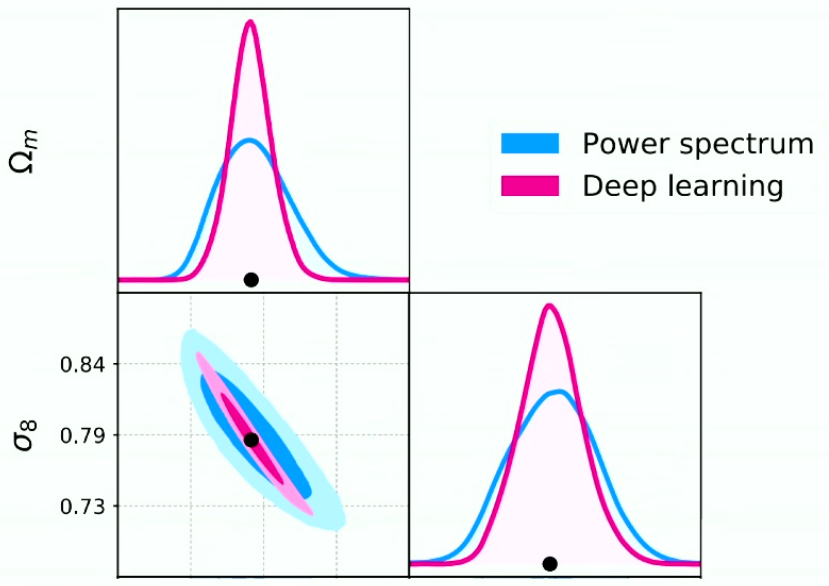
- *Non-Gaussian summary statistics extracted with deep learning*



Kacprzak & Fluri (2022)

ML's potential and applicability limited by "black box" nature

Case study: cosmological parameter inference



Kacprzak & Fluri (2022)

- **Potential**: Deep learning yields tighter cosmological parameter constraints than traditional techniques
- **Limitation**: Where is the additional information coming from?
- Can we claim a new discovery (e.g. nature dark energy/neutrino masses) if we cannot explain ***how*** that discovery came about?

The need for interpretability

Interpretability for building ***understanding*** and **trust**, with potential of new machine-learning assisted scientific discoveries

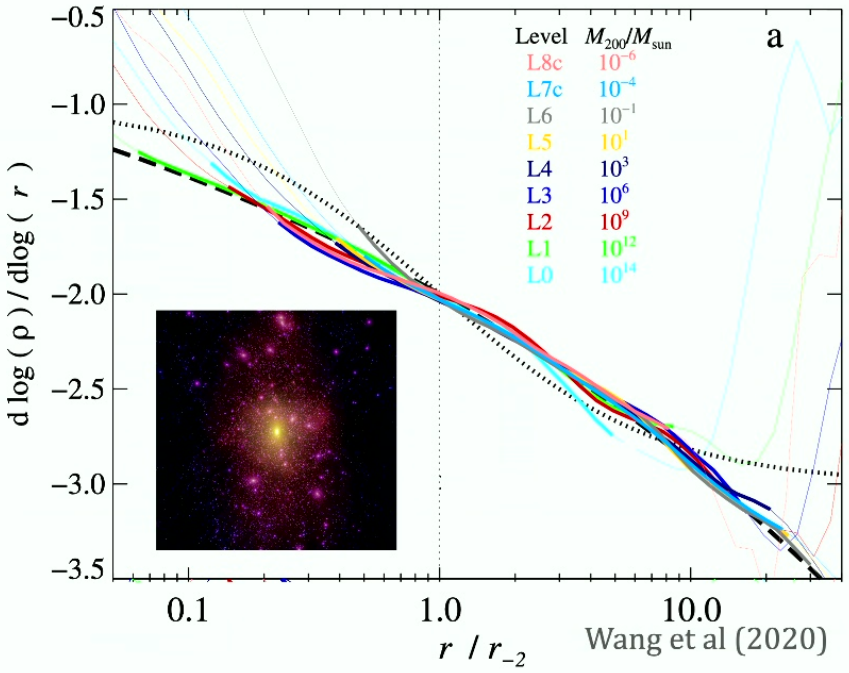
1. ***Interpretability:*** account for why the ML model reaches its predictions

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

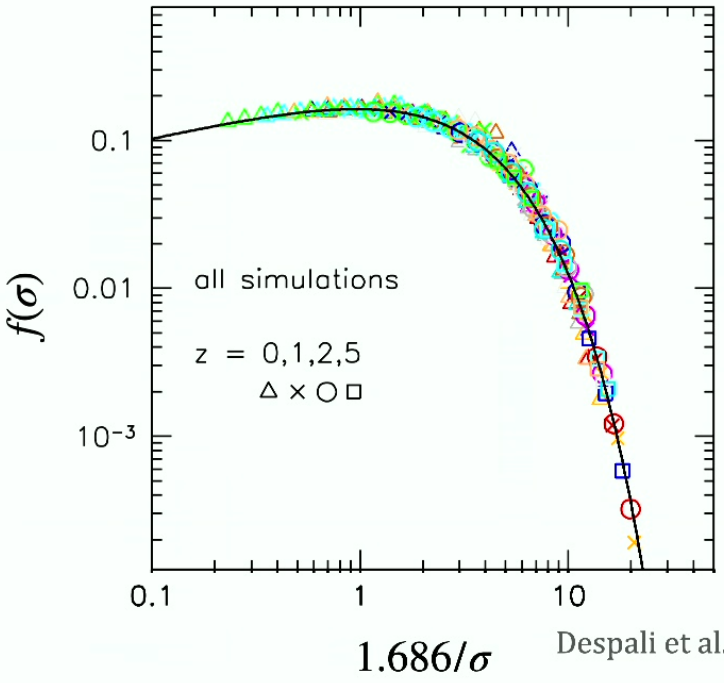
Recent review on Interpretable ML in Physics (Wetzel+ 2025, arXiv:2503.23616)

Explainability serious concern for many 'physical' models in Astro

The density structure of dark matter halos

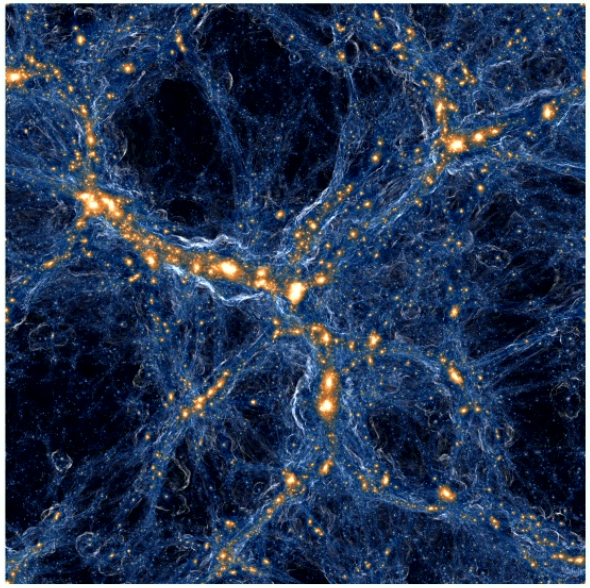


The abundance of dark matter halos



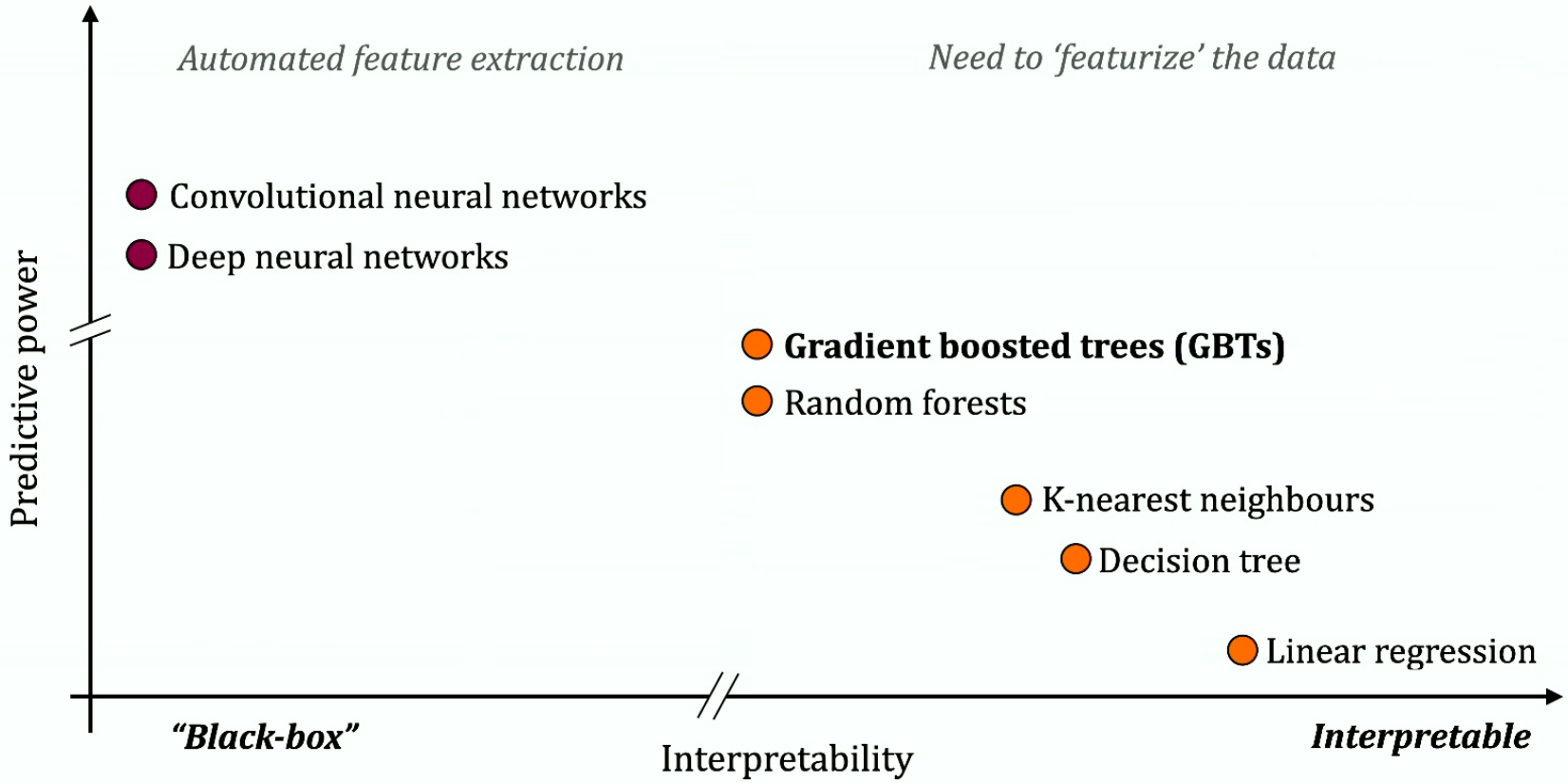
Our models are **empirical fitting functions** calibrated to simulations which **lack explainability**

New frontier of explainable AI for knowledge extraction

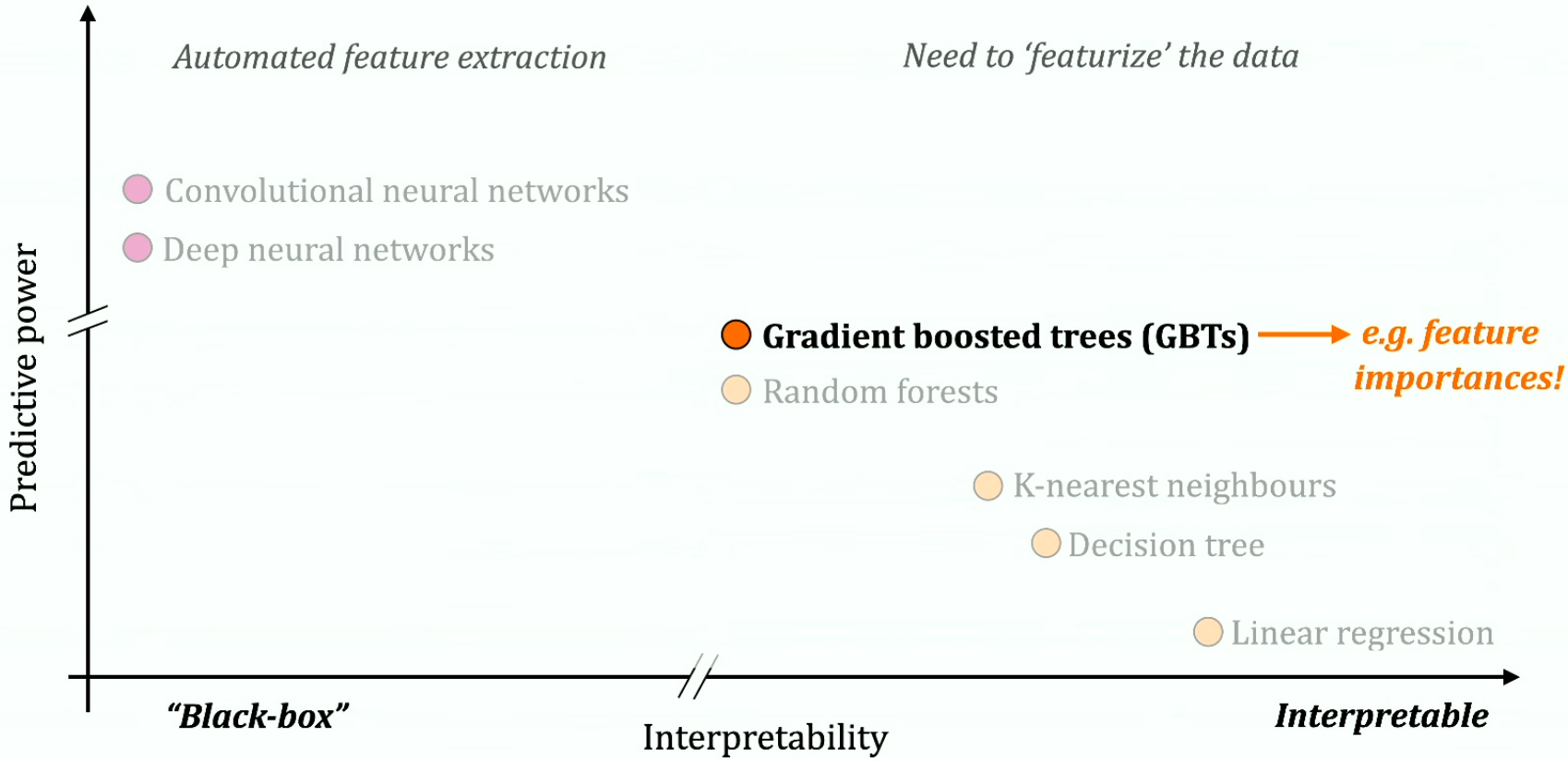


*Can we **extract new knowledge** about the underlying physics from deep learning models by interpreting their outputs?*

Is interpretability an unrealistic goal for current "black box" models?

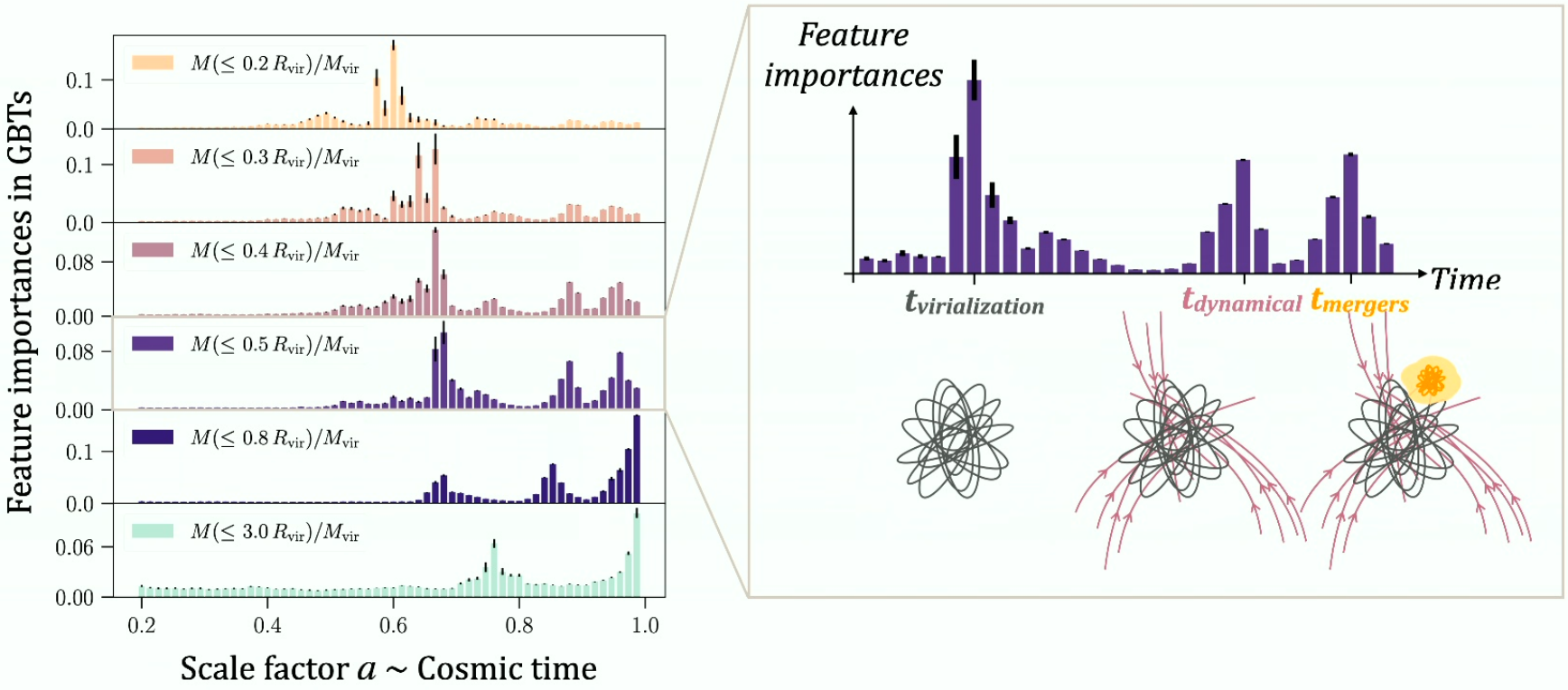


Is interpretability an unrealistic goal for current "black box" models?



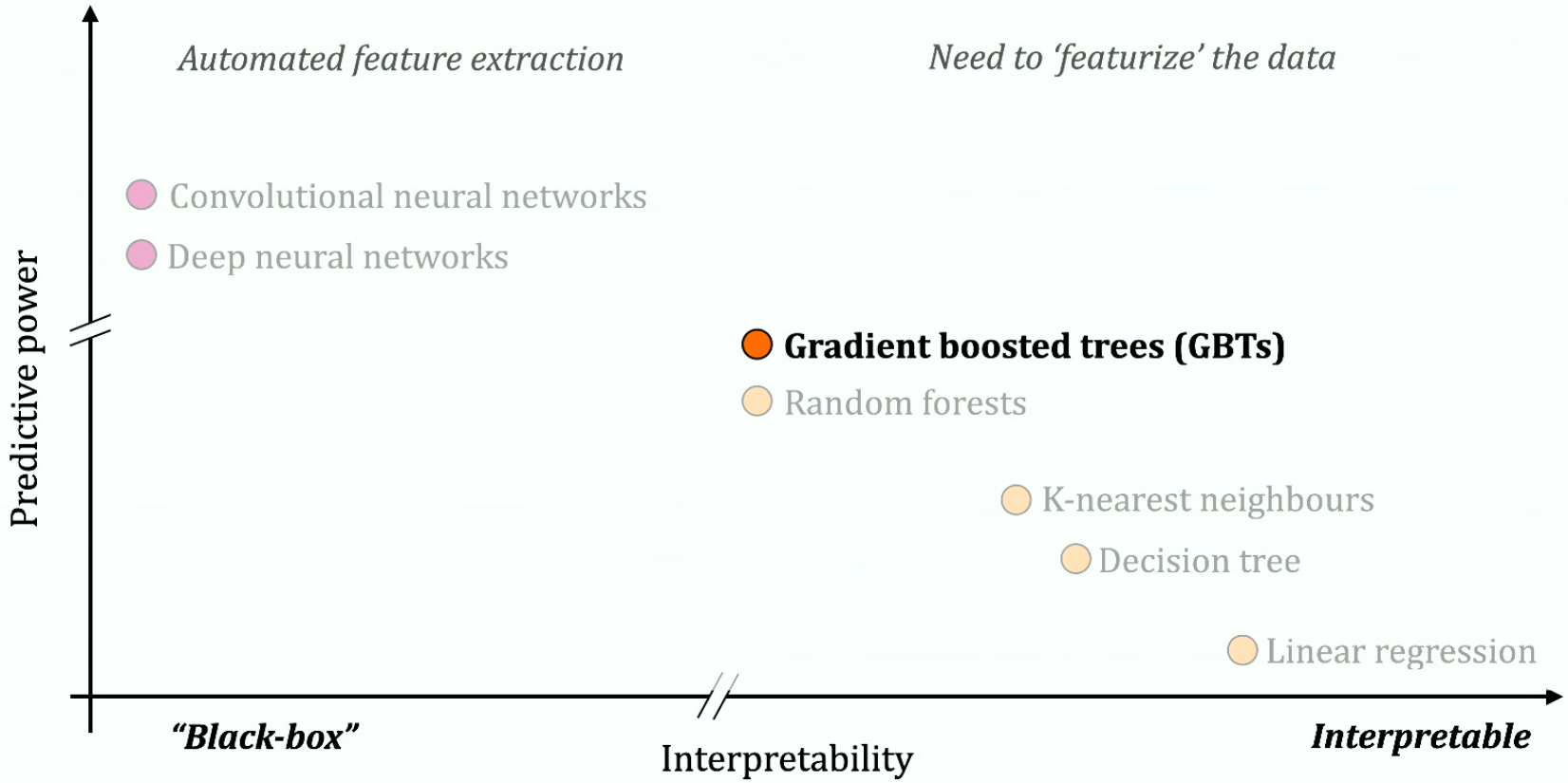
Case study: the origin of the final mass distribution of galaxies

What are the key epochs in the evolution history of a galaxy which determine its final mass distribution?

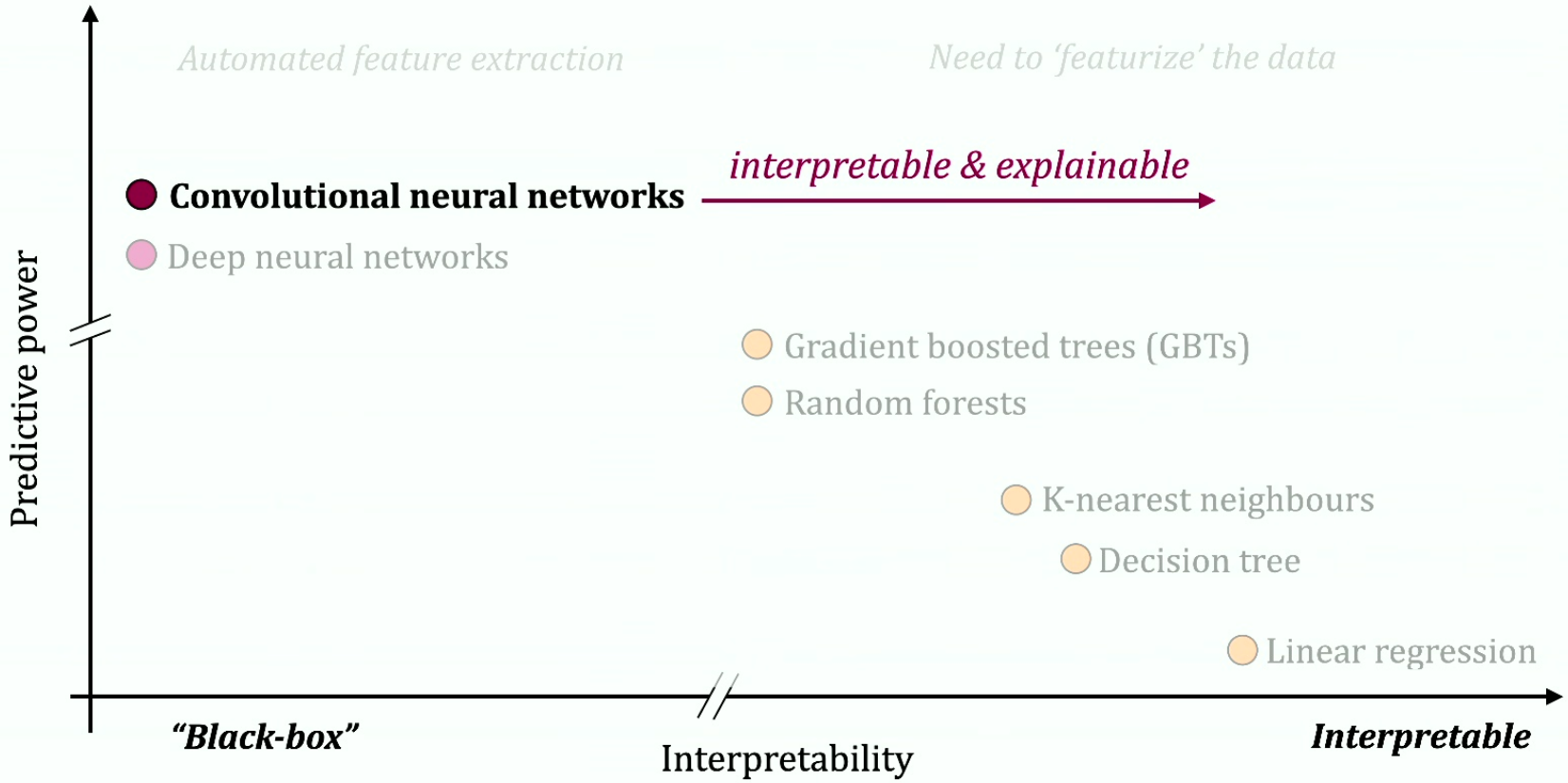


Lucie-Smith, Adhikari, Wechsler (MNRAS, 2022)

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

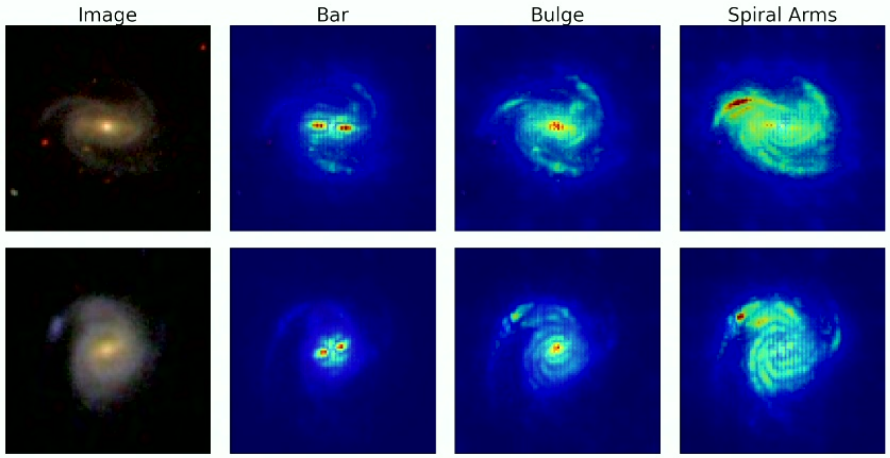


Goal for machine-assisted scientific discovery: knowledge extraction with deep learning



Visualization techniques: saliency maps

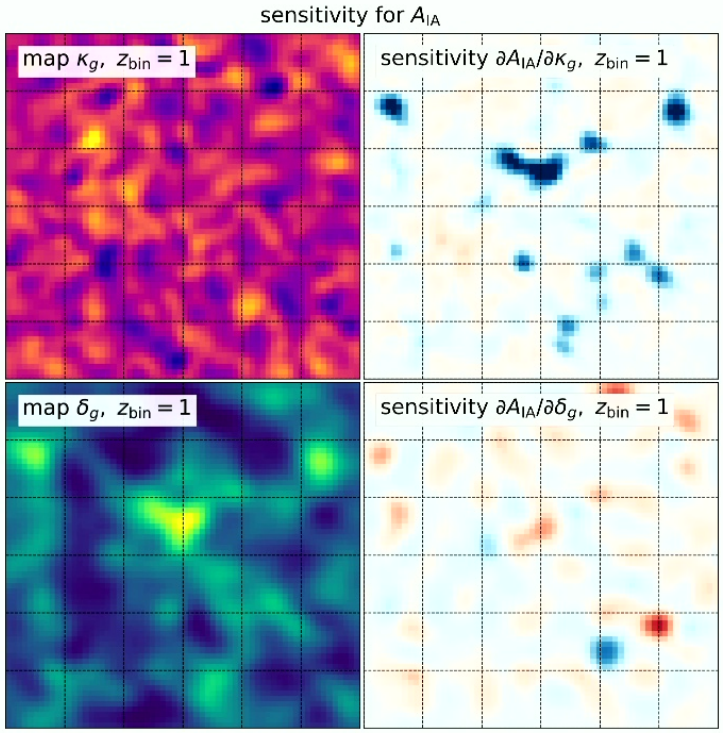
Predicting galaxy properties



Bhambra+ 2022

Qualitative robustness check but difficult to pull out quantitative new physical insights

Cosmological parameter inference

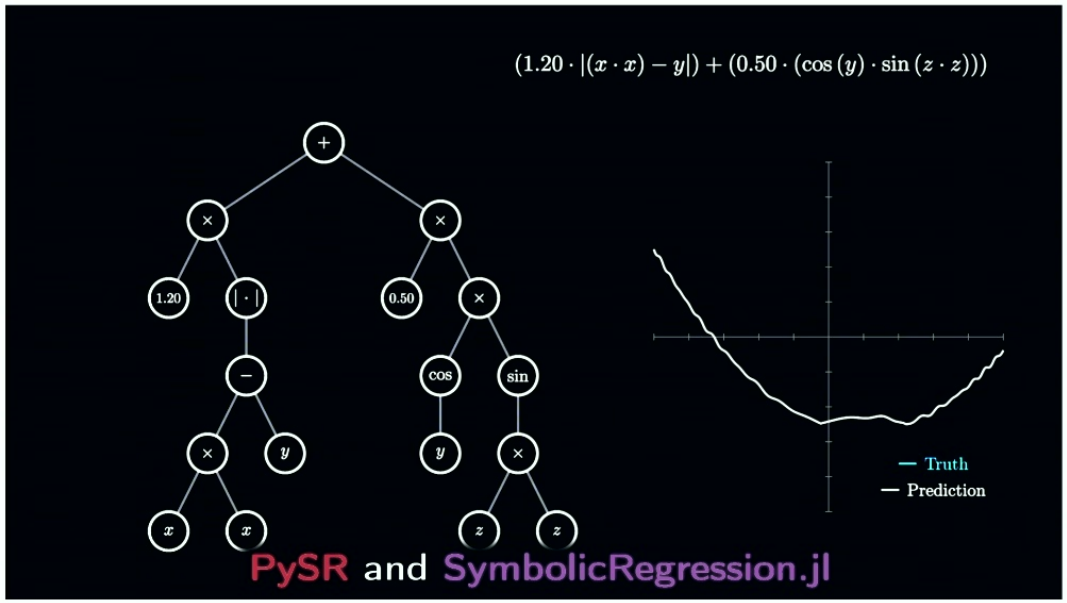


Kacprzak & Fluri 2022

Symbolic regression

Symbolic regression: goal of the machine learning task is to find ***analytic expressions*** that optimize some objective.

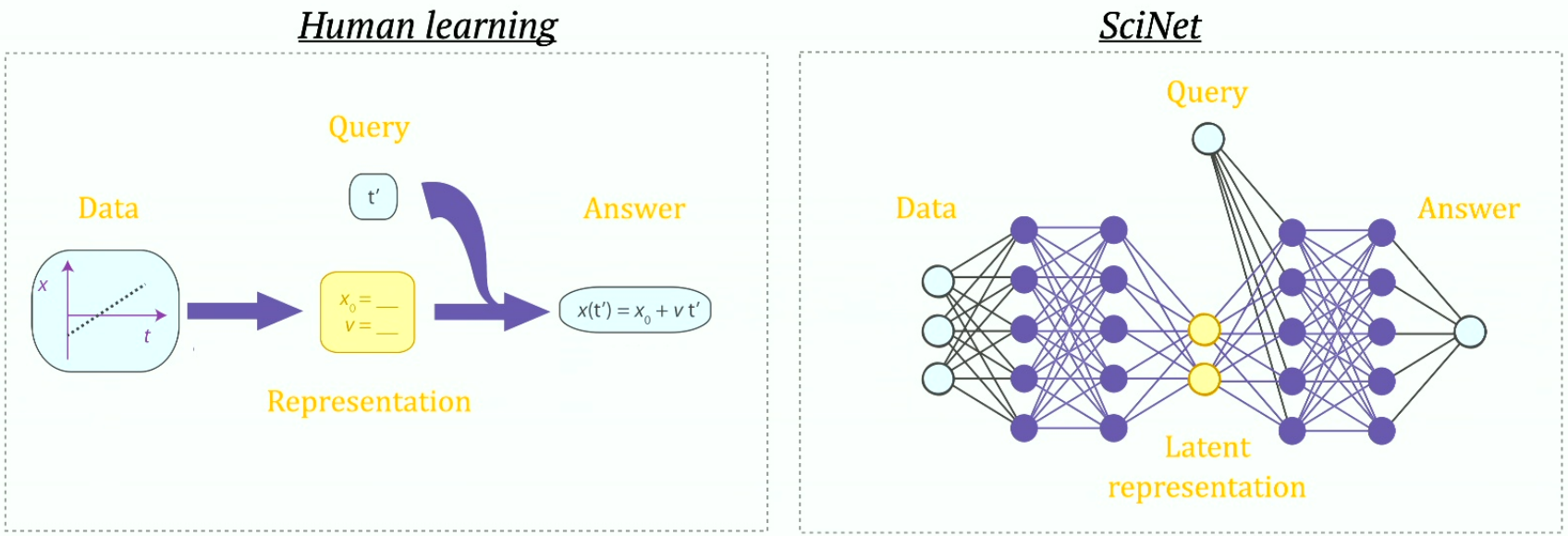
Credit: Miles Cranmer



Langley et al. 1980s; Cranmer et al. 2019, 2020; Lemos et al. 2022; Bartlett et al. 2023
github.com/MilesCranmer/PySR/

- Are analytic expressions necessarily more "interpretable" than neural networks?
- Complicated expression can be **difficult to relate to any physics** in the relevant science domain

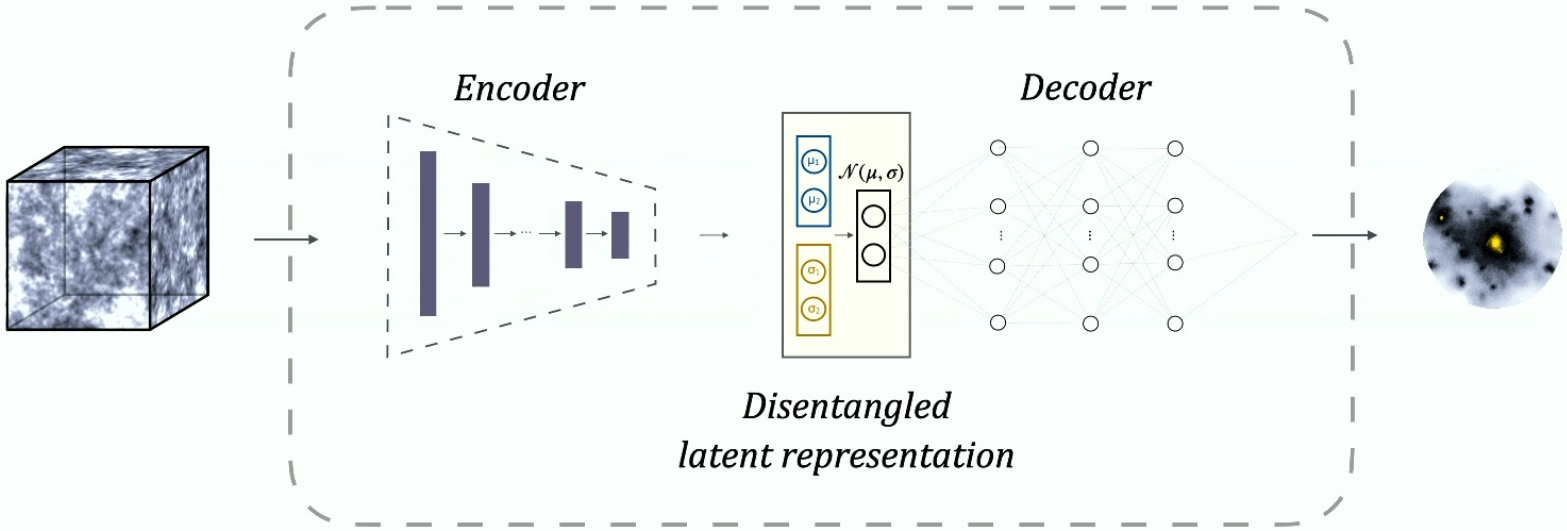
SciNet model



- 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

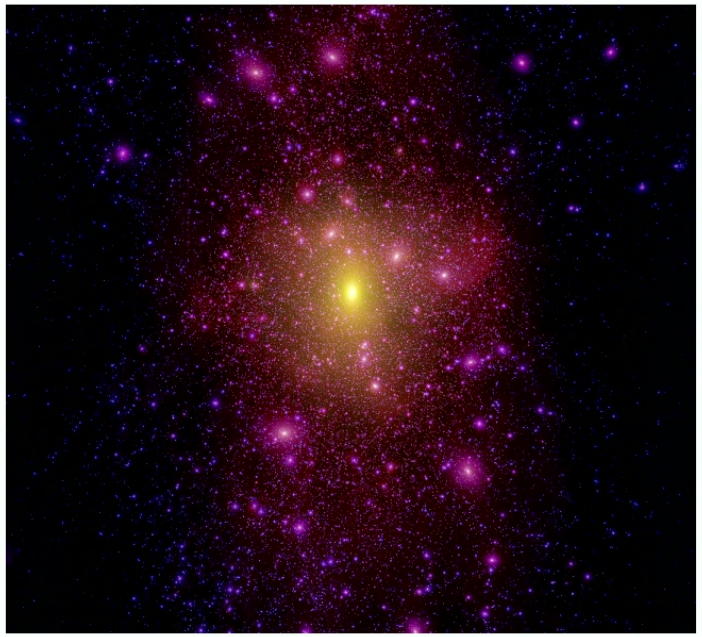


Model compression enables "explainability"

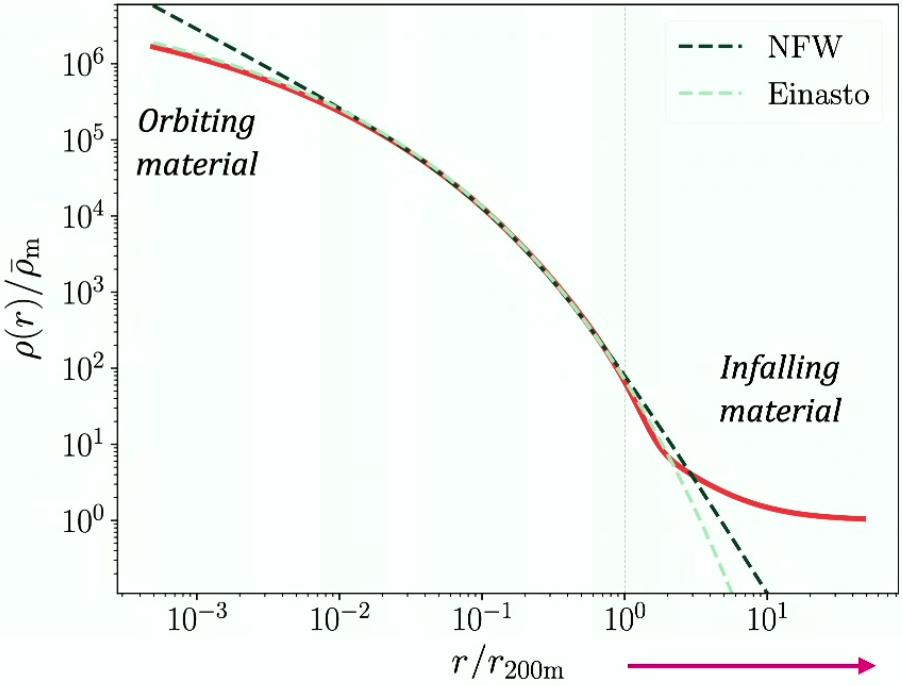
work with **H. Peiris (Cambridge), A. Pontzen (Durham),** D. Piras (Geneva), B. Nord (Fermilab) + follow-up papers with L. Guo (former student UCL), V. Springel (MPA), G. Despali (Bologna)

Case study: density profiles of dark matter halos

Full 3D phase-space distribution of a dark matter halo



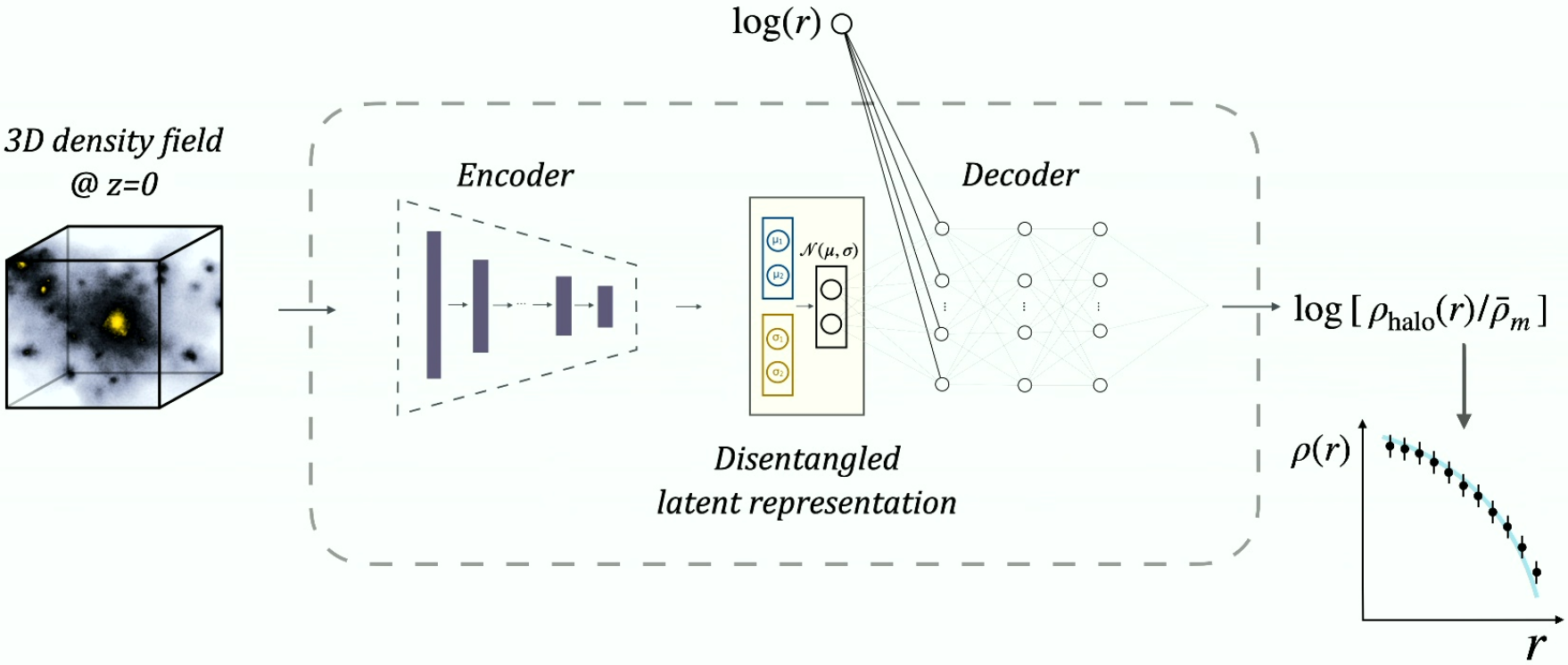
Spherically averaged density profile



Navarro, Frenk, White (1996, 1997); Einasto (1965); Diemer & Kravstov (2014); Adhikari et al. (2014); More et al. (2015) ; Diemer (2022, 2023)

Deviations in outskirts where infalling material dominates

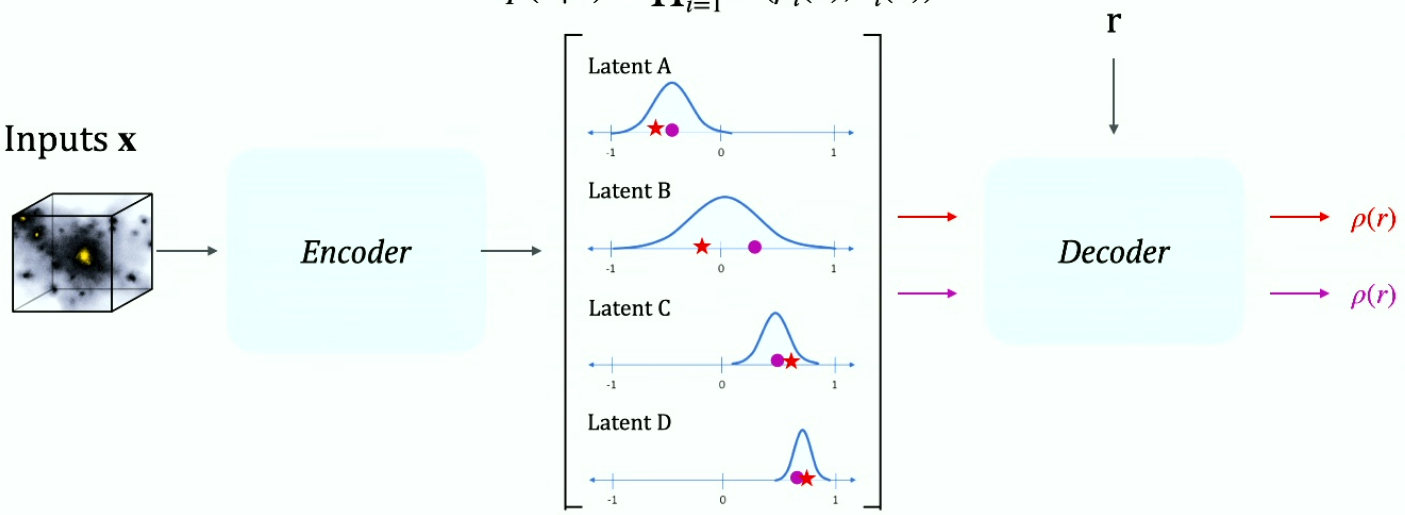
What are the building blocks of halo density profiles?



Lucie-Smith, Peiris, Pontzen, Nord+ (PRD, 2022);
Lucie-Smith, Peiris, Pontzen (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

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:

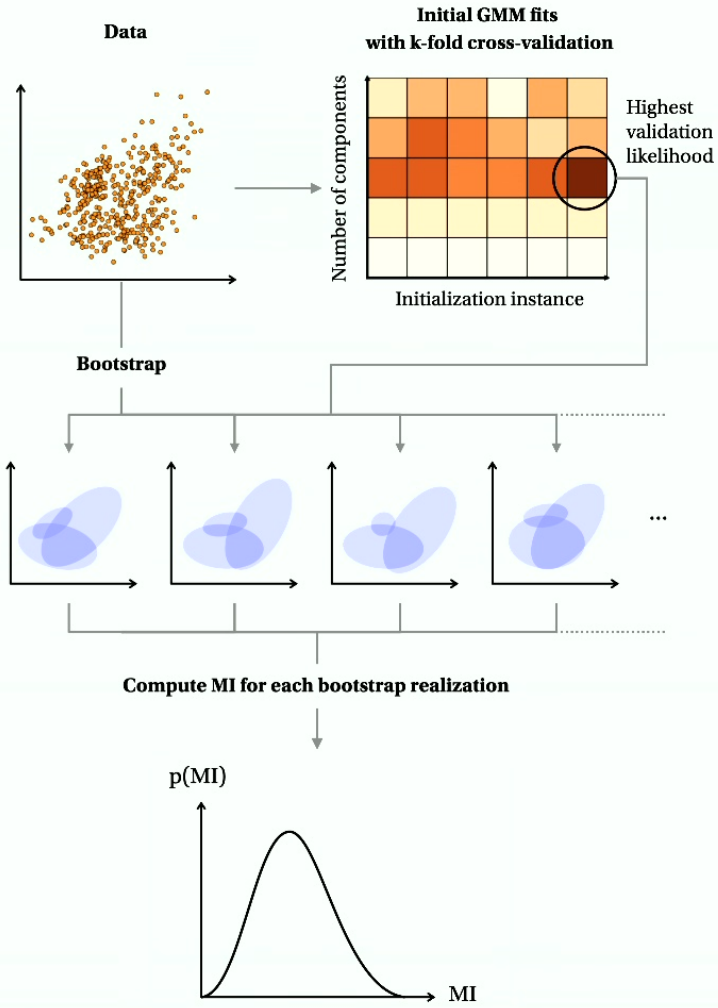
$$\text{MI}(z_i, z_j) = \int_{z_i}\int_{z_j} p(z_i, z_j)\log\left[\frac{p(z_i, z_j)}{p(z_i)p(z_j)}\right] dz_i dz_j$$

Gaussian mixture model $p(z_i, z_j)$

Mutual information

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

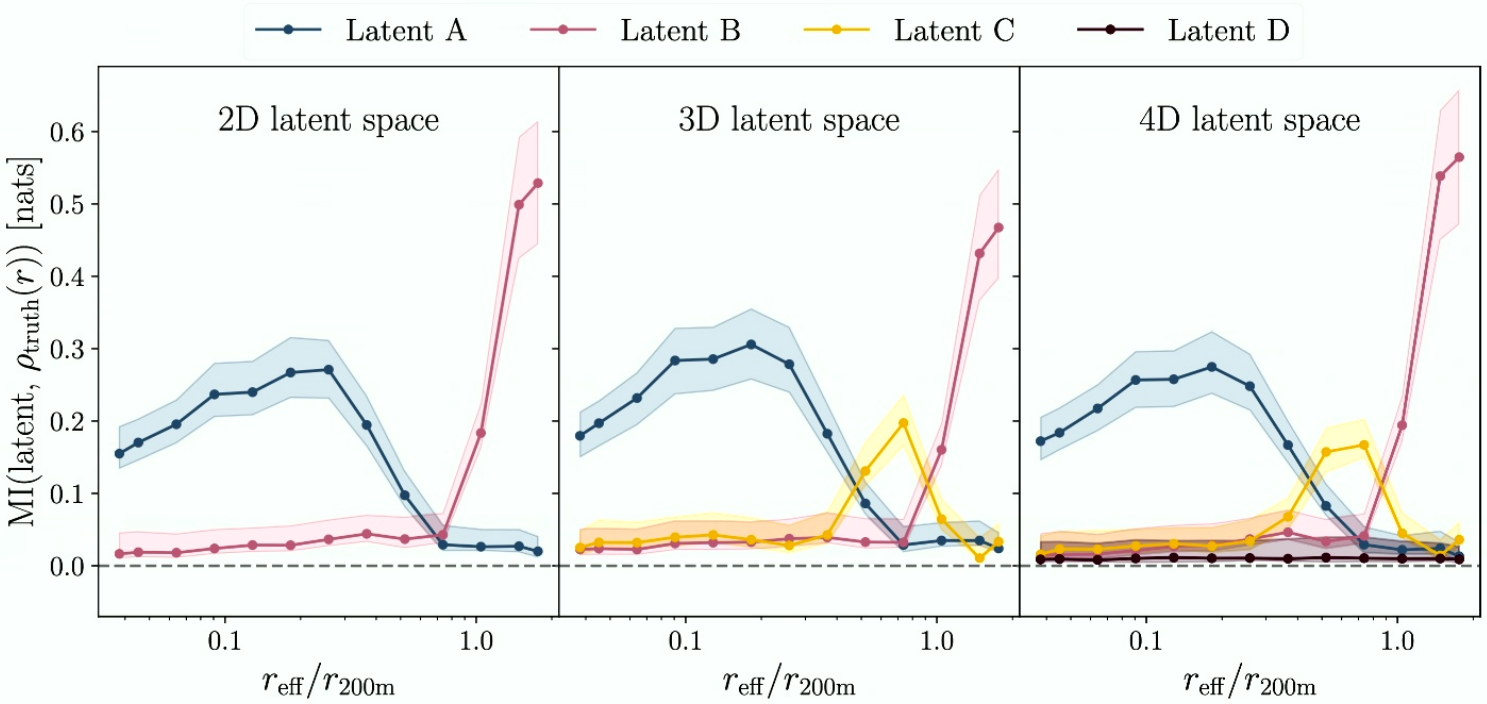
+ extension to **conditional** mutual information



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

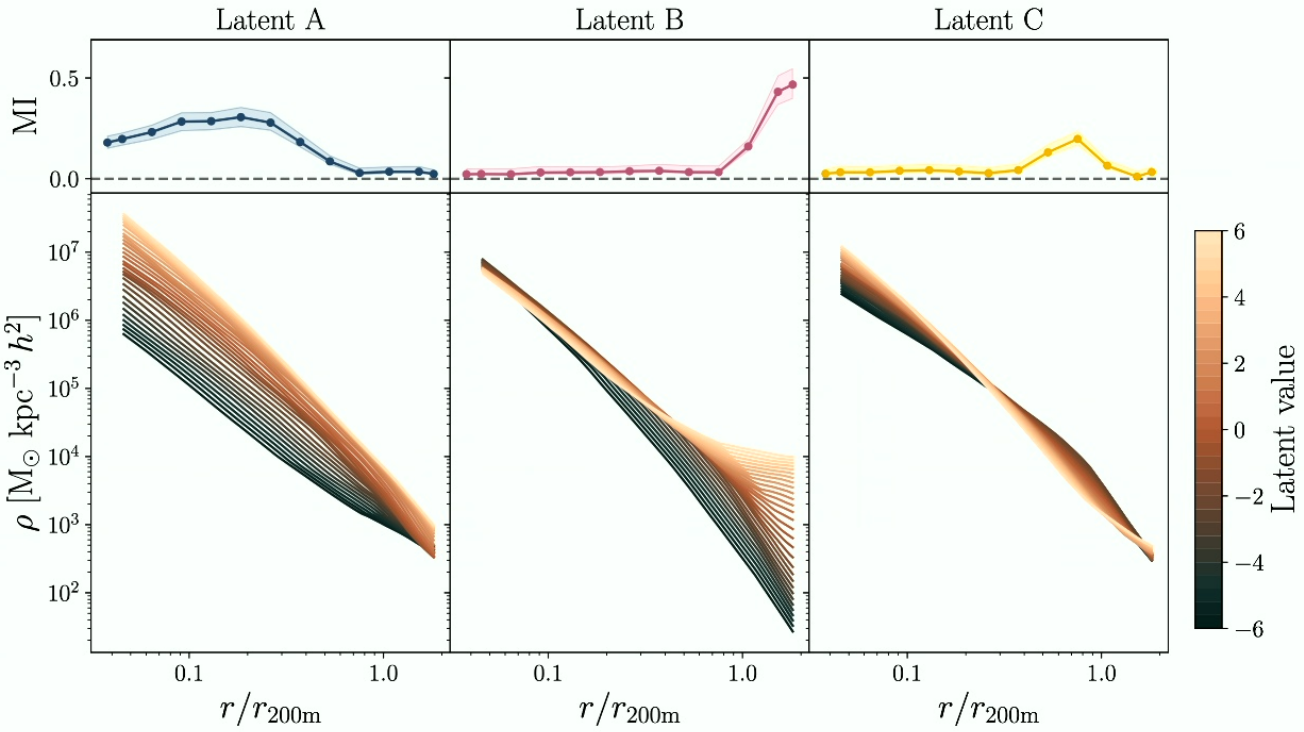
Piras, Peiris, Pontzen, Lucie-Smith et al. (2023, MLST)

Interpreting the latent representation using mutual information



Lucie-Smith, Peiris, Pontzen, Nord et al. (PRD, 2022)

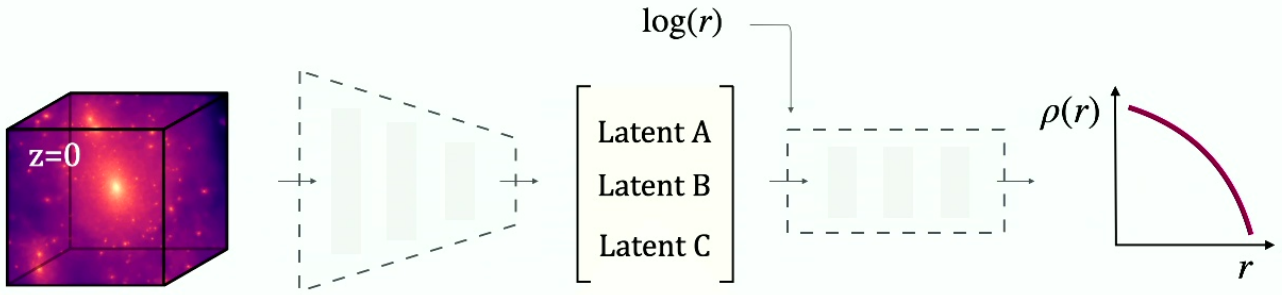
Systematically varying one latent at a time



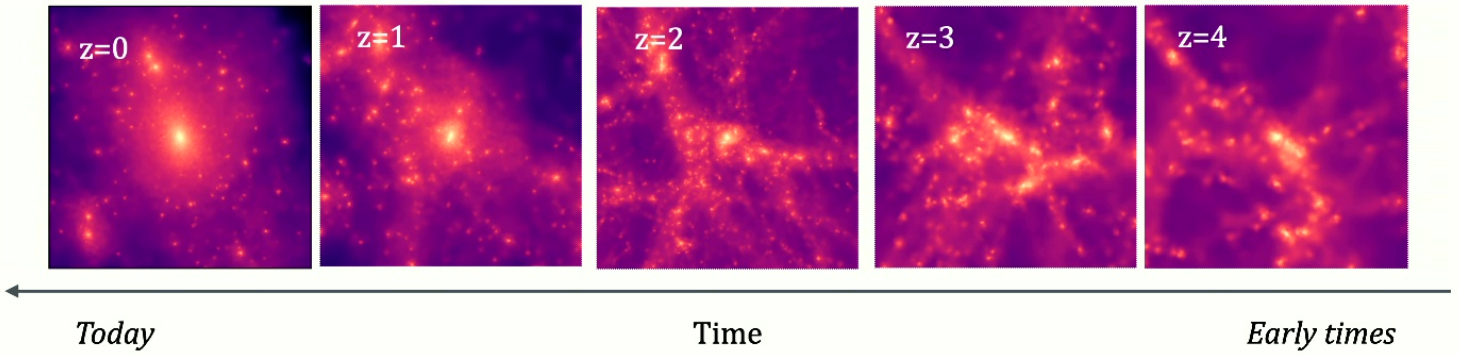
Latent A = ***normalisation***; Latent B = ***outer slope;*** Latent C = ***inner slope***

Lucie-Smith, Peiris, Pontzen, Nord et al. (PRD, 2022)

Are the discovered latents useful beyond original training task?

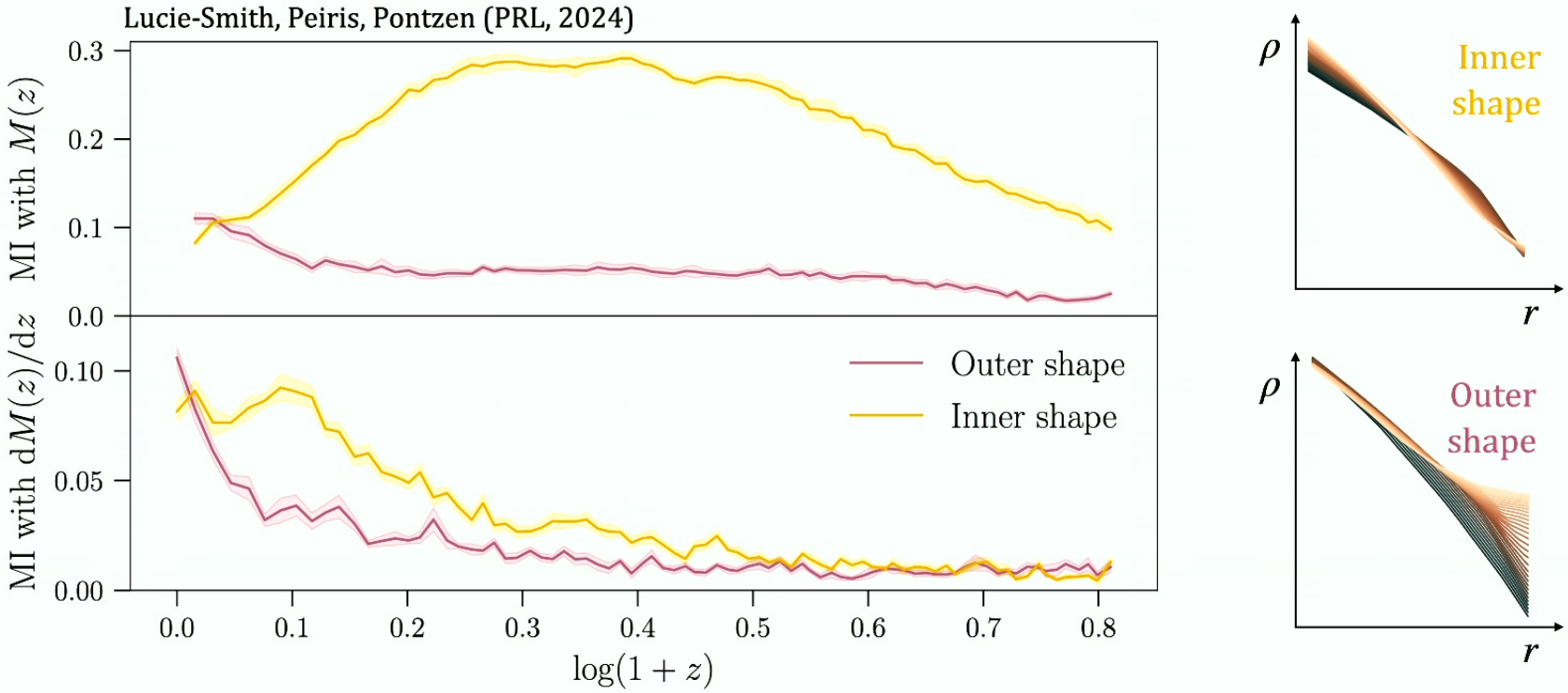


Does the latent space retain information about the evolution history of the halos?



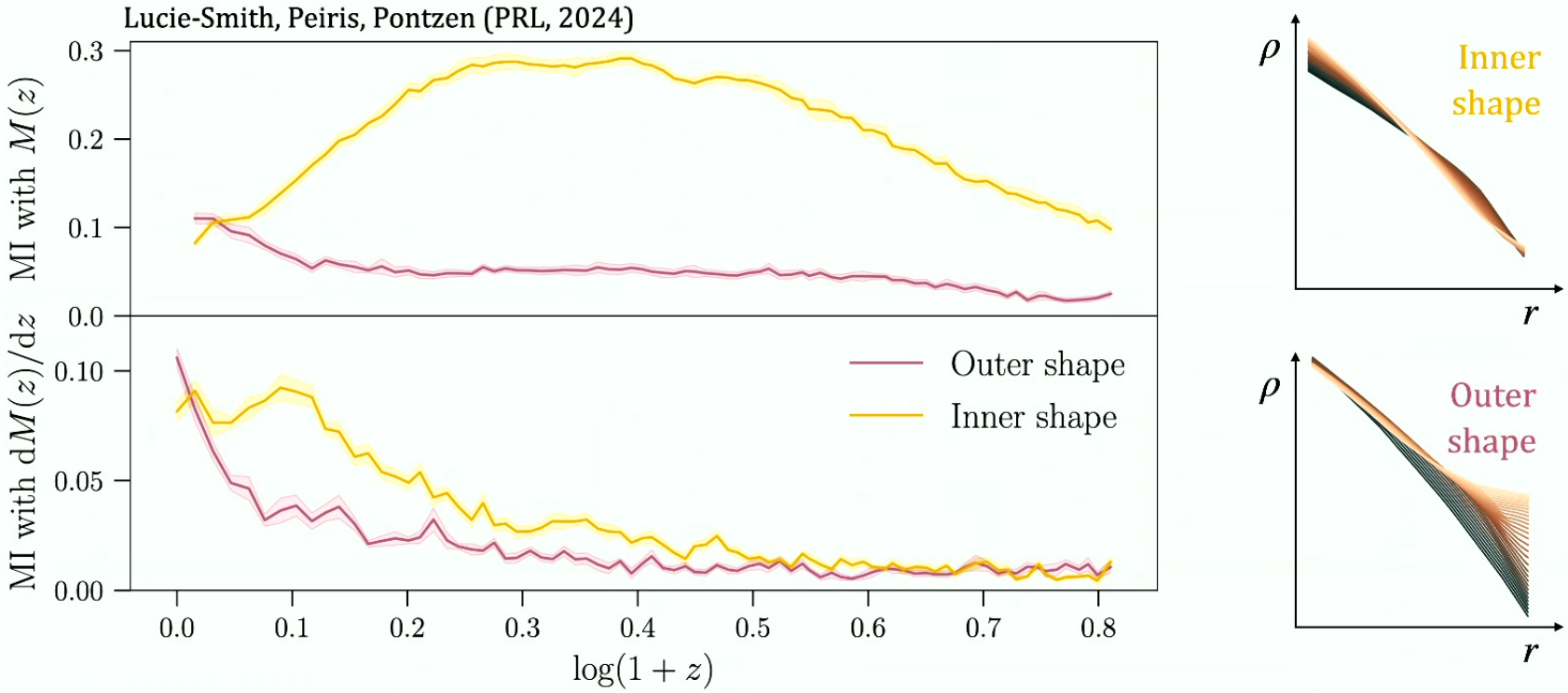
Lucie-Smith, Peiris, Pontzen (PRL, 2024)

Relation between the latents and the halo evolution history



New discovery: outer profile depends on ***only one component*** related to ***most recent accretion rate***

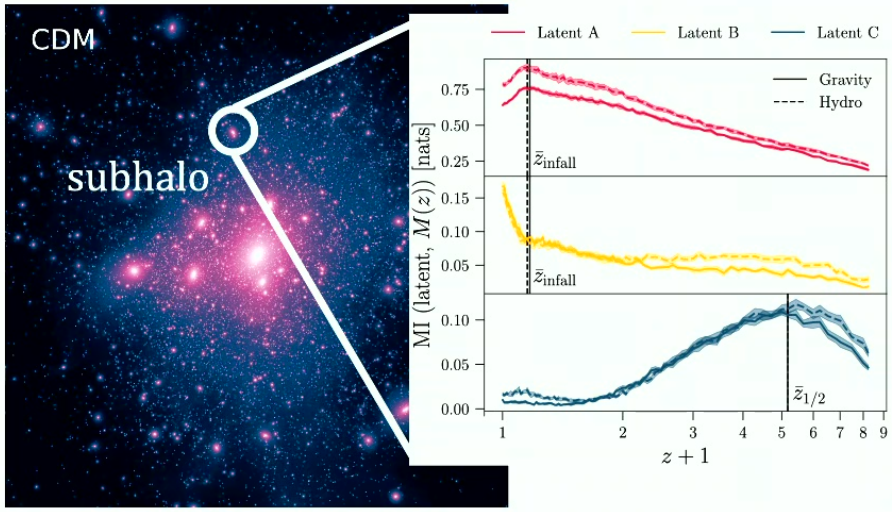
Relation between the latents and the halo evolution history



New discovery: outer profile depends on ***only one component*** related to ***most recent accretion rate***

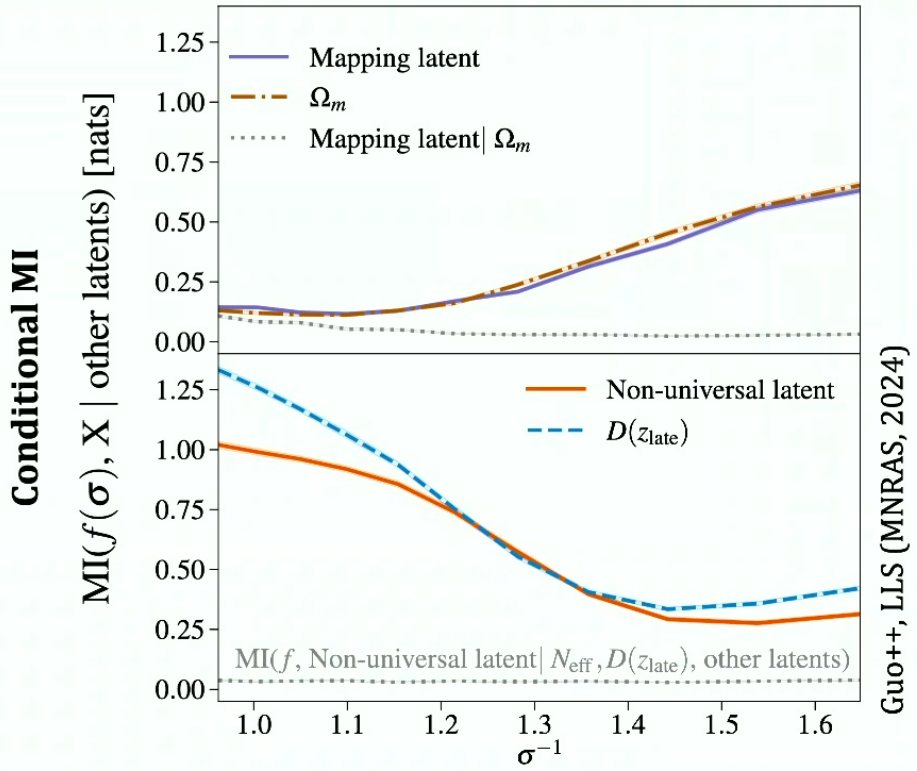
Applications to a variety of cosmological probes

Subhalo profiles for strong lensing



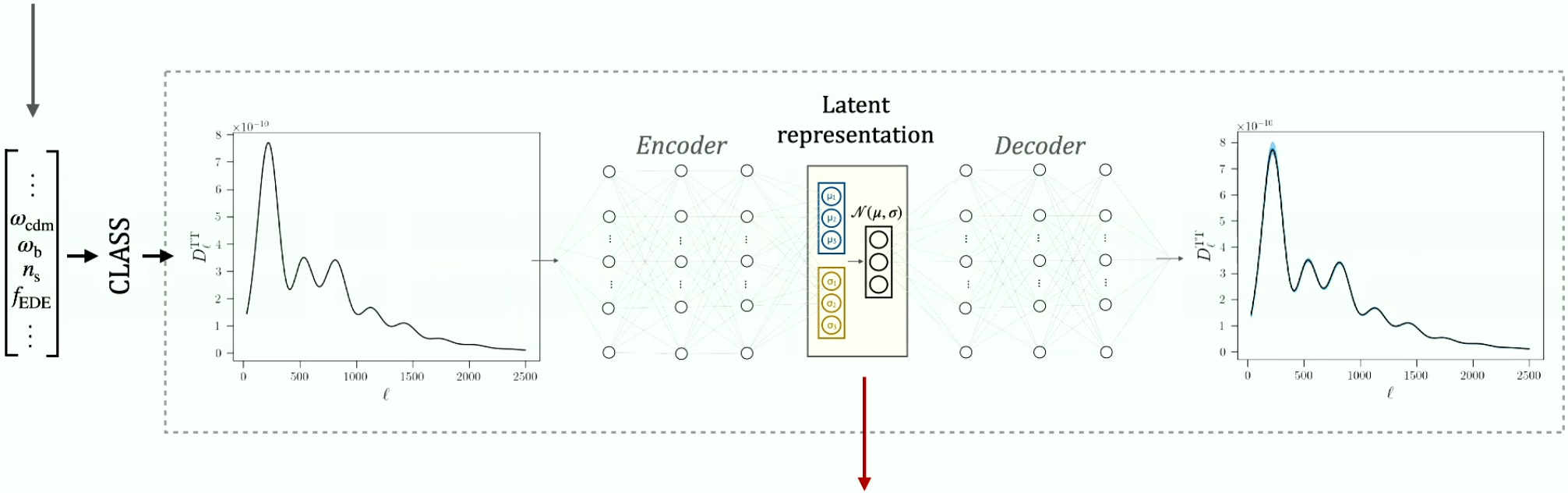
Lucie-Smith, Despali, Springel (MNRAS, 2024)

The abundance of dark matter halos: the origin of non-universality



I. Re-parametrize your data in terms of its degrees of freedom

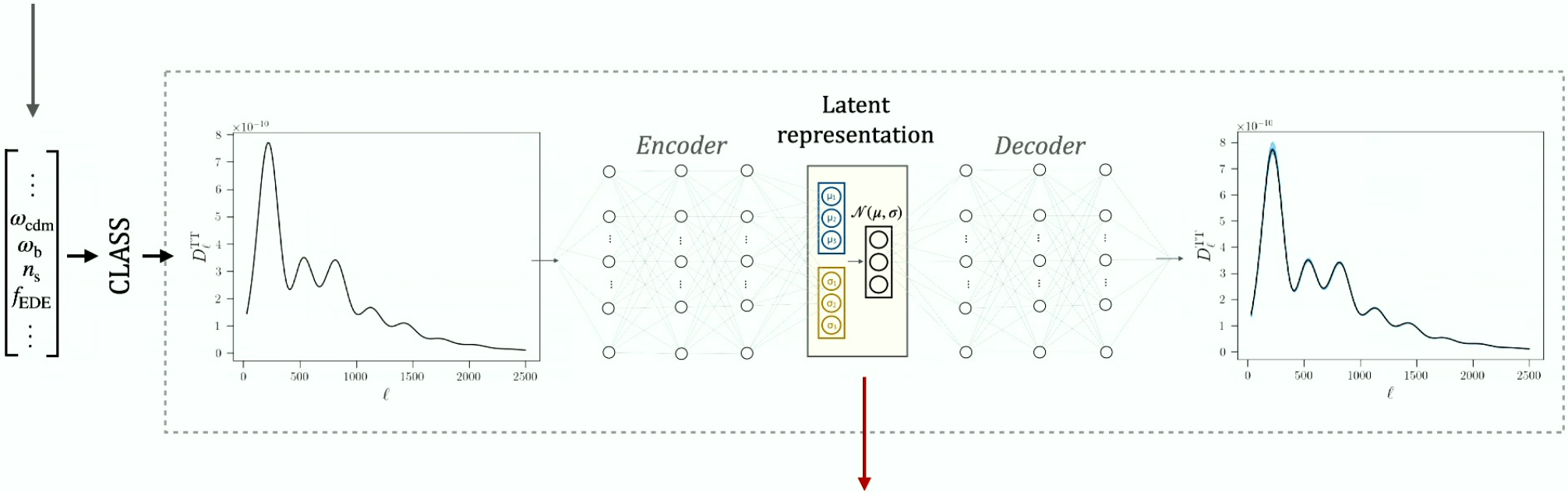
Consider cosmological model with an additional ‘early dark energy’



*New parametrization = the **independent degrees of freedom in the data** vector itself*

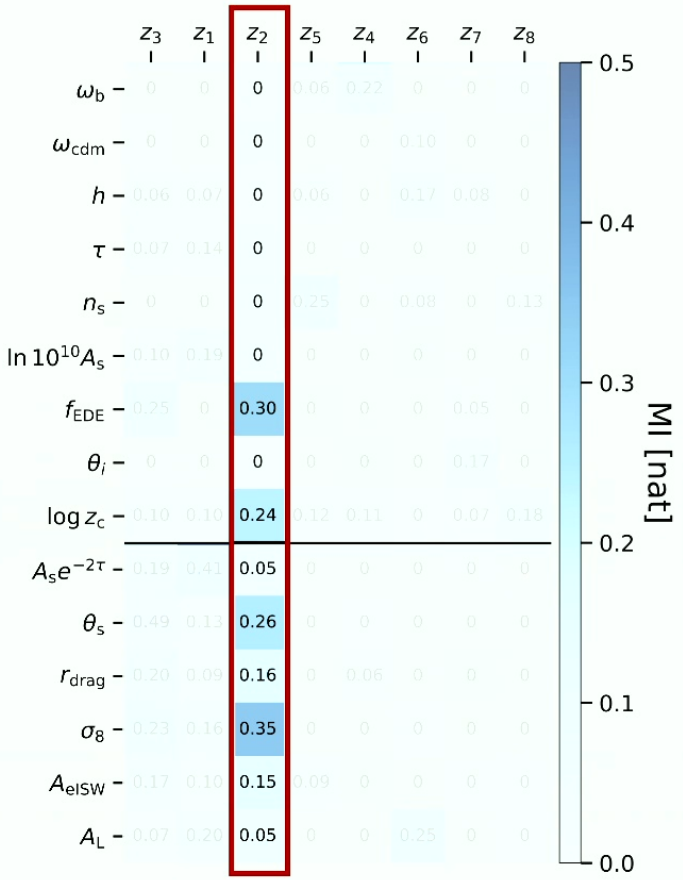
I. Re-parametrize your data in terms of its degrees of freedom

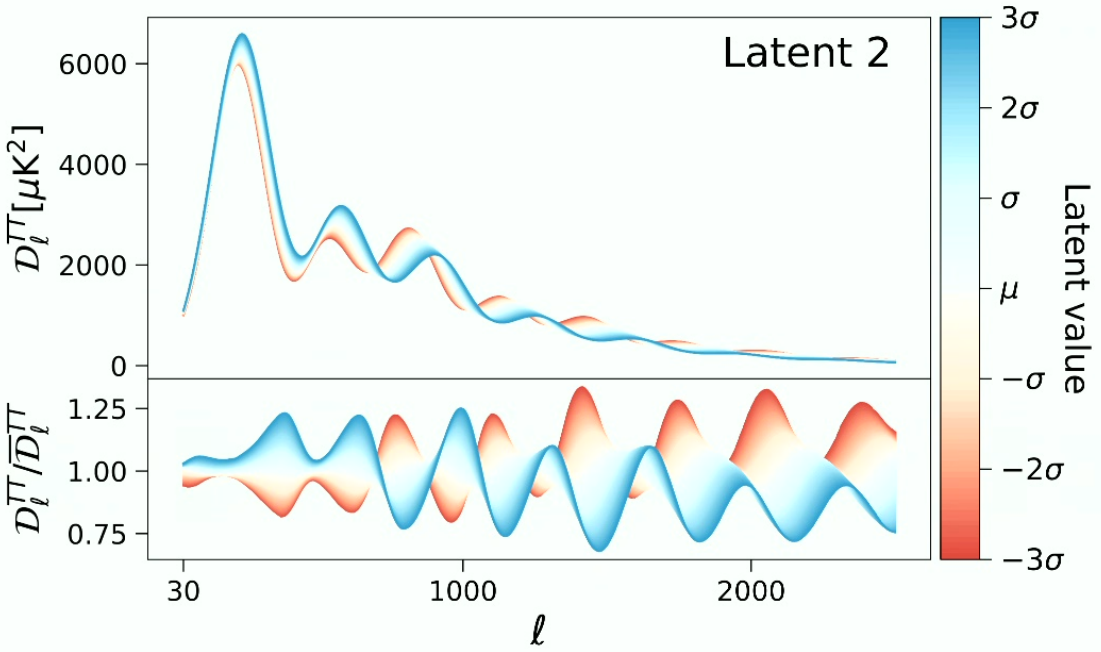
Consider cosmological model with an additional 'early dark energy'



*New parametrization = the **independent degrees of freedom in the data** vector itself*

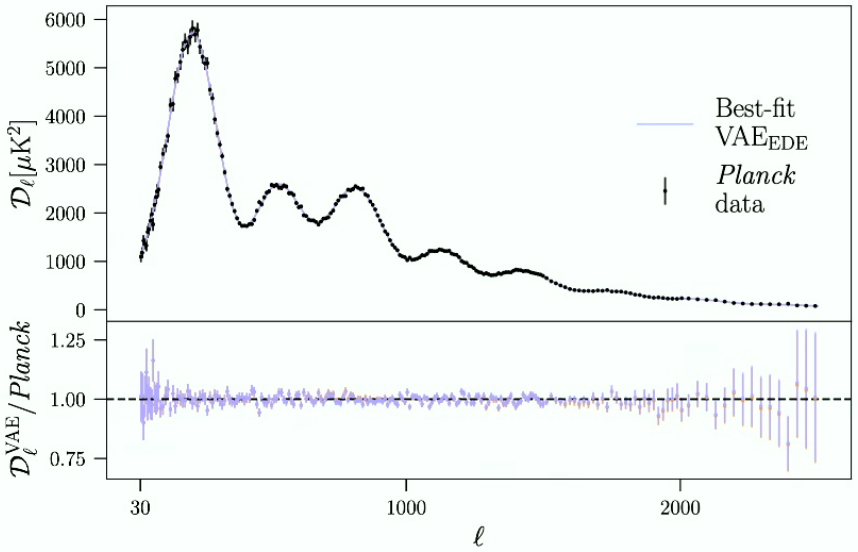
II. Quantify information content in the latents

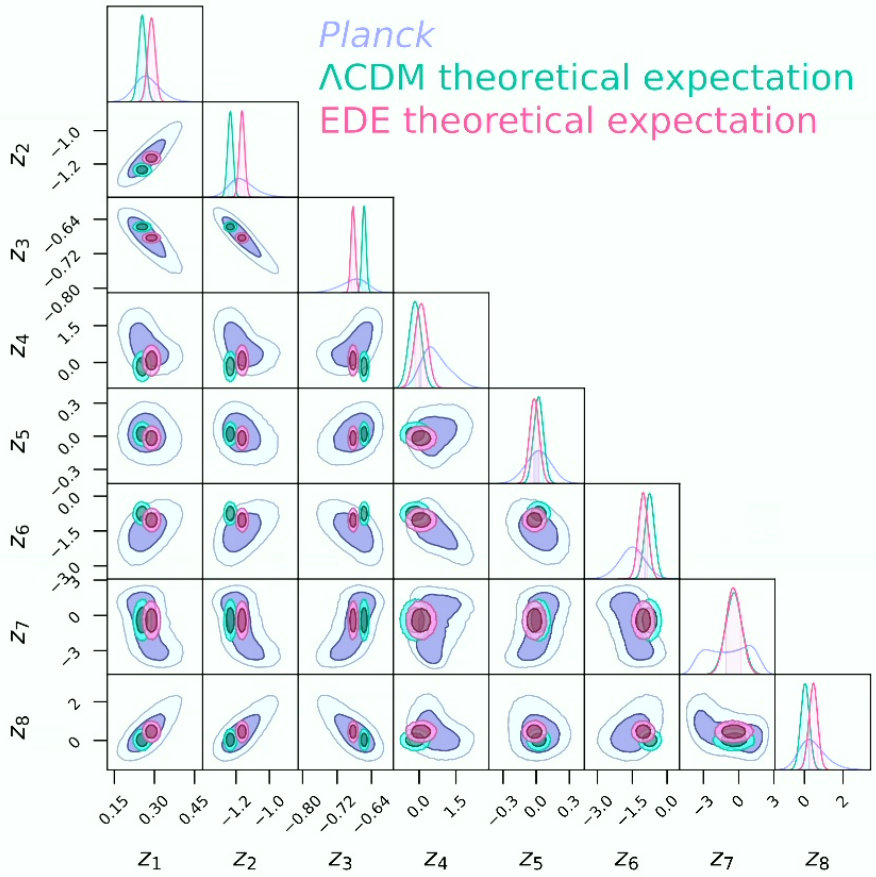




One latent parameter sensitive only to early dark energy effects

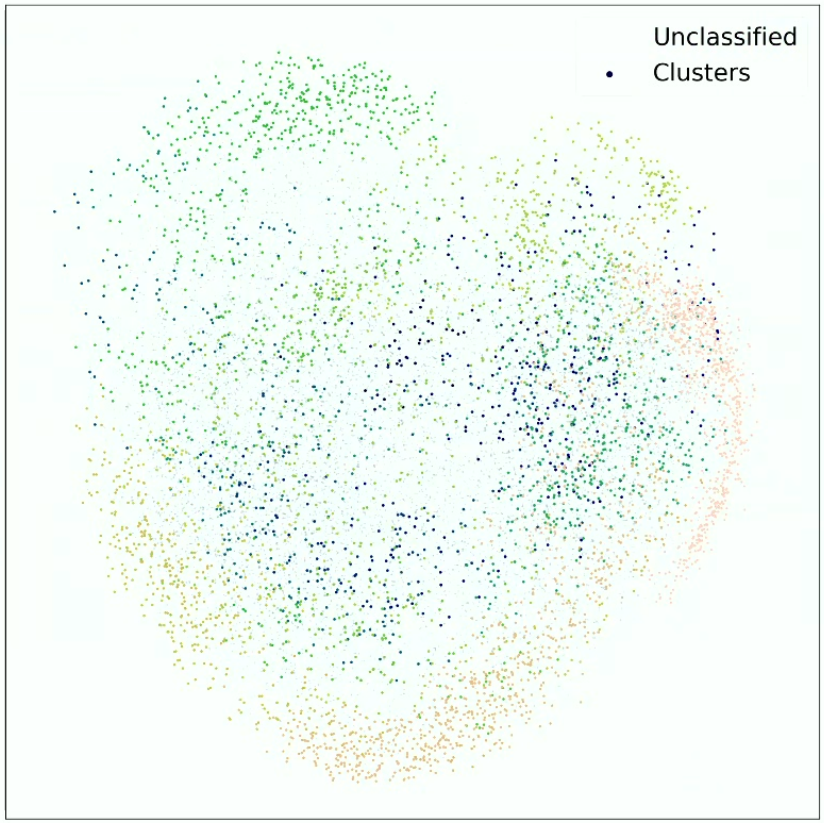
III. Bayesian inference on the latent parameters using Planck data





Piras, Herold, Lucie-Smith, Komatsu (arXiv:2502.09810)

Challenges ahead with high-dimensional latent spaces



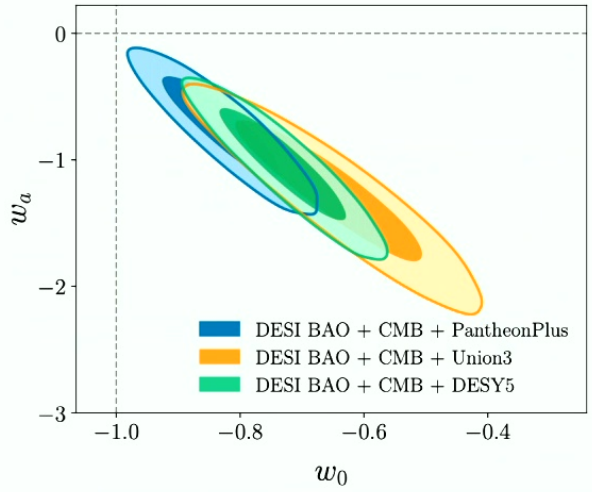
Baron Perez, Brüggen, Kasieczka, Lucie-Smith (2025)

Extensions to ***self-supervised models*** and ***foundation models*** with high-dim latent spaces

- How to interpret high-dimensional latent space? Can 'disentanglement' scale?
- Sparse priors? E.g. sparse autoencoders
- Hierarchical latent spaces?

What is 'cosmology in latent space' useful for?

- Identify **the independent degrees of freedom in the model data vector** (CMB and/or late-time probes such as BAO), especially for models which are poorly parametrized
- **Constrain** them using data from (one or many) experiments
- Potential: new way to reveal ***which features of the data drive tensions*** between two experiments



DESI collaboration

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 (sub)halo density profiles
- Data-driven parametrization of cosmological probes via neural-based model compression new way to reveal tensions between experiments

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Luisa Lucie-Smith, *luisa.lucie-smith@uni-hamburg.de*