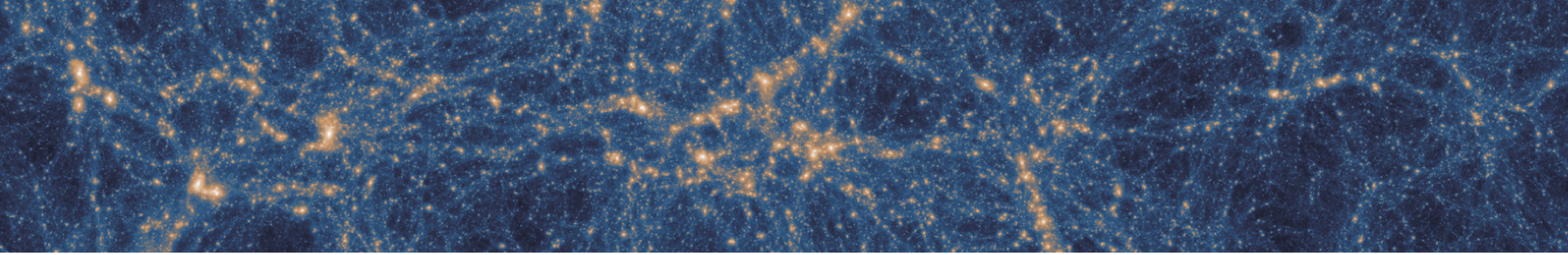


Astro/Cosmo Highlights

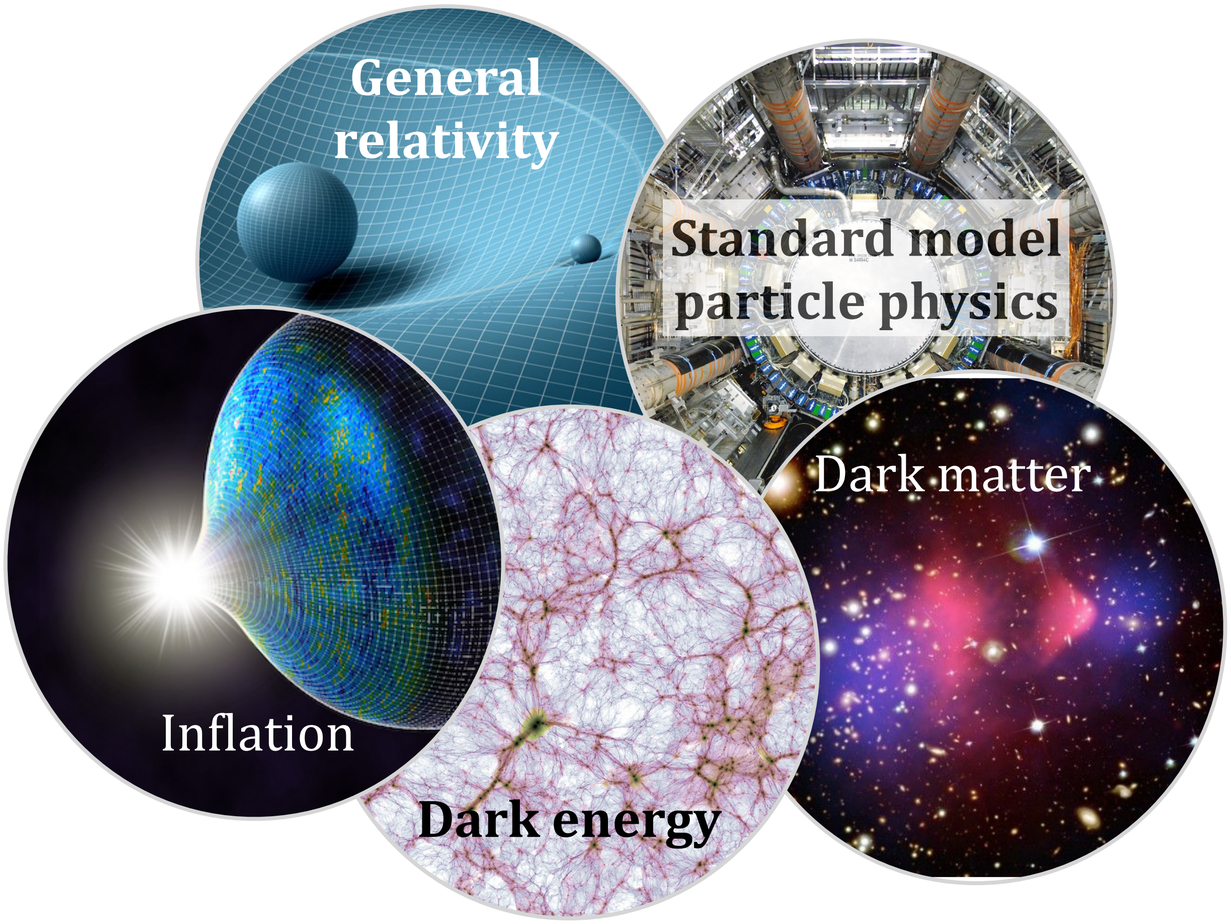
Luisa Lucie-Smith
Postdoctoral Research Fellow @ Max-Planck-Institute for Astrophysics, Garching

PHYSTAT Workshop on “Statistics meets Machine Learning”
Imperial College London, 12th September 2024

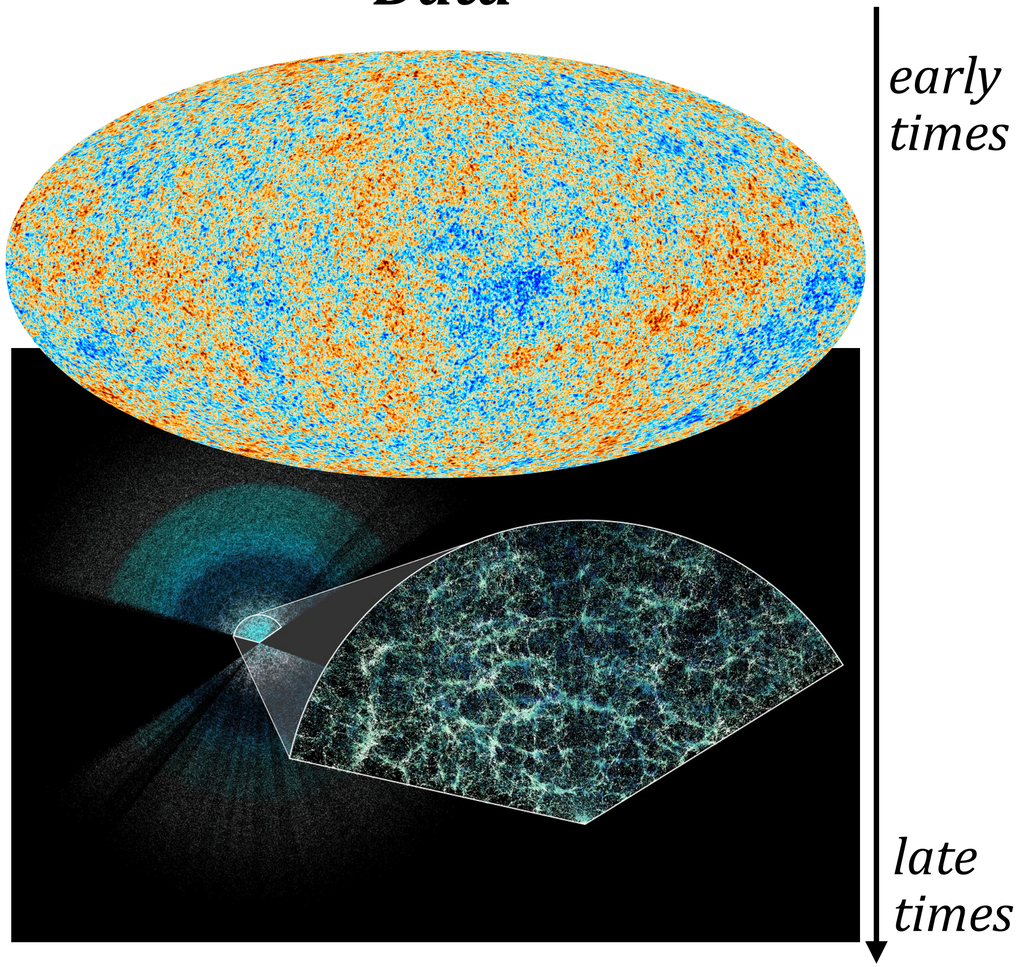
MAX PLANCK INSTITUTE
FOR ASTROPHYSICS

Theory vs Data

Theory



Data



Theoretical challenges in connecting theories to data

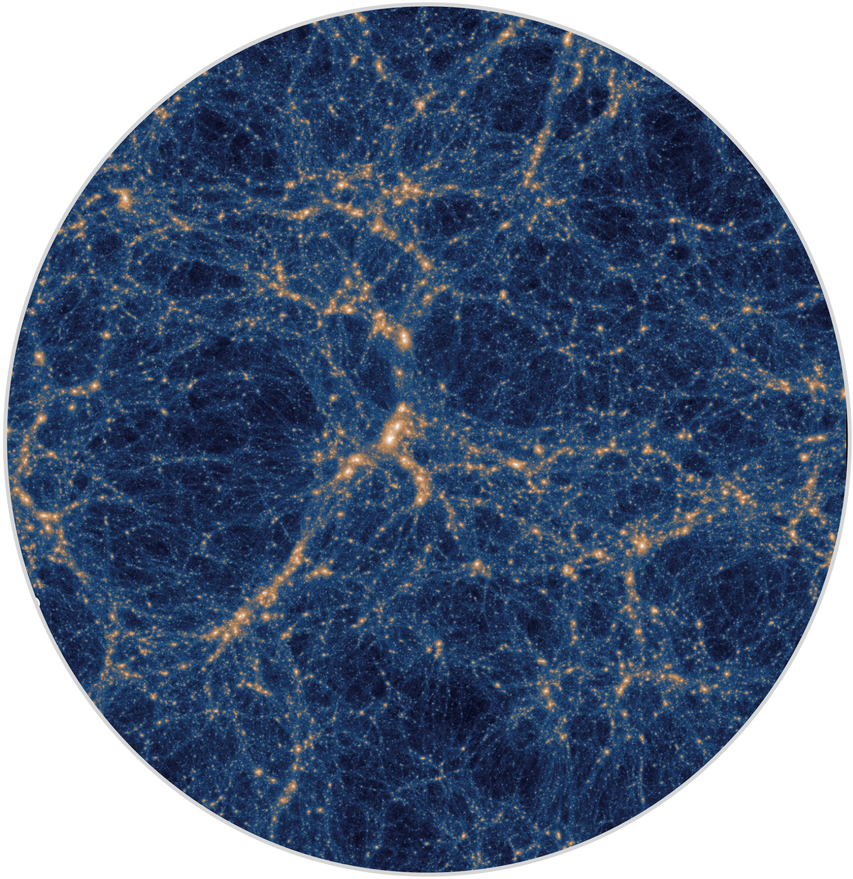
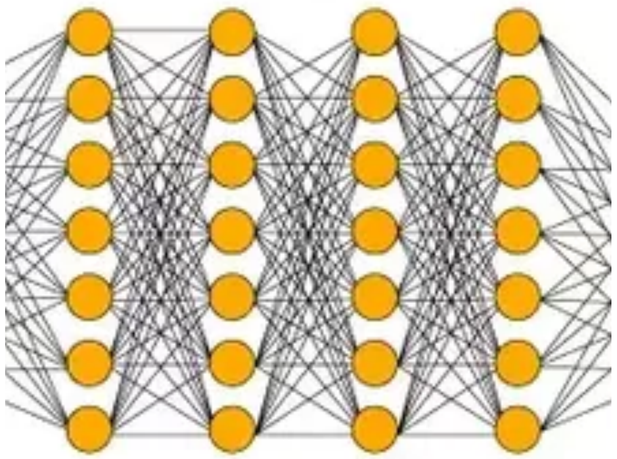


Image credit: IllustrisTNG team

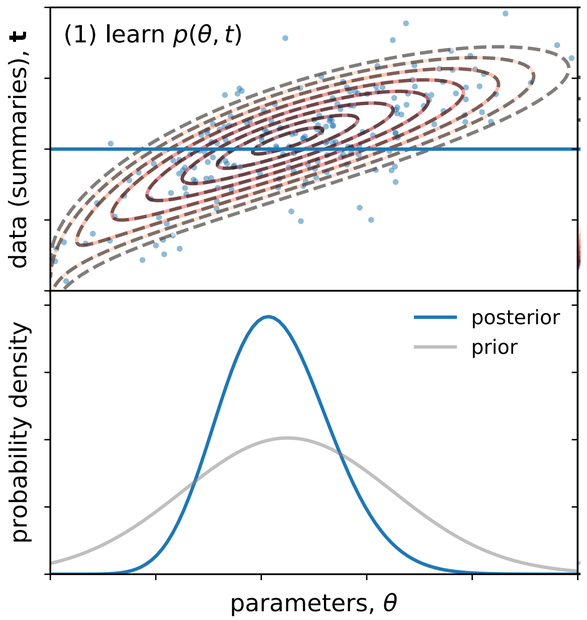
- ***Computationally expensive** cosmological simulations with large volumes & high resolution*
- *Highly **non-linear** gravitational evolution*
- *Uncertainties in **galaxy formation** process*
- *High-dimensional, **intractable likelihoods***
- *Connection to **fundamental physics***
  Fitting cosmological parameters does not mean understanding…

PHYSTAT "ML meets Statistics": many successful applications of AI aimed at solving these challenges

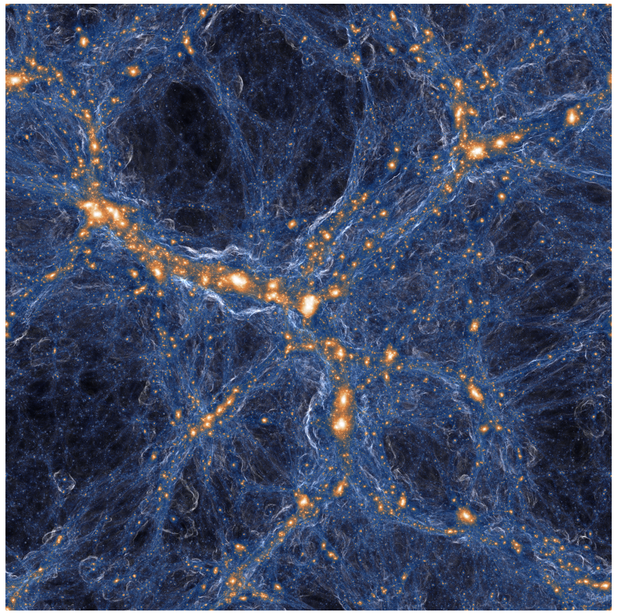


Generative models:
emulating forward models

Alsing+2018

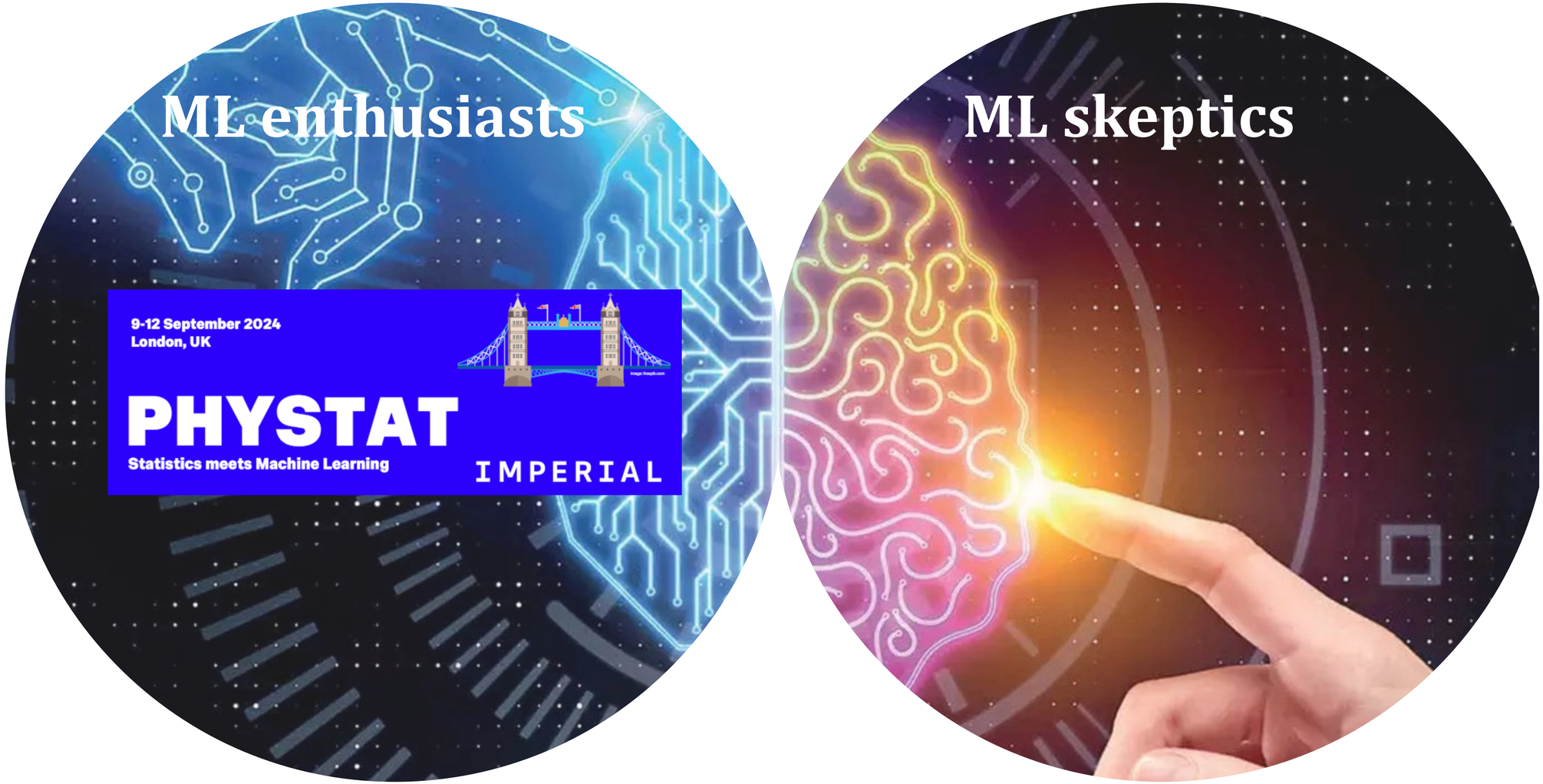


Implicit likelihood inference:
new parameter inference paradigm

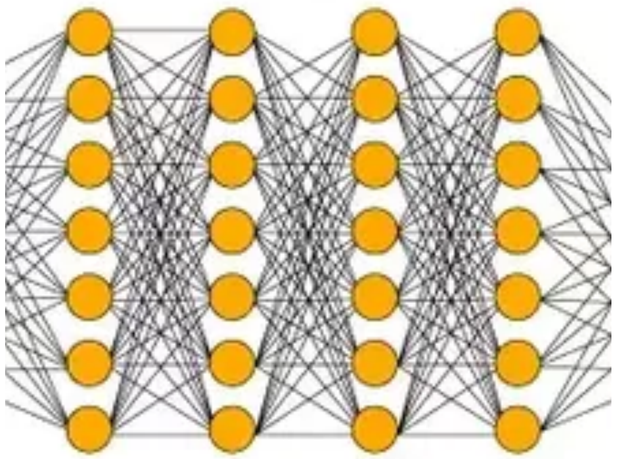


Explainable AI:
ML-enabled scientific discoveries

PHYSTAT meeting predominantly made of ML enthusiasts…

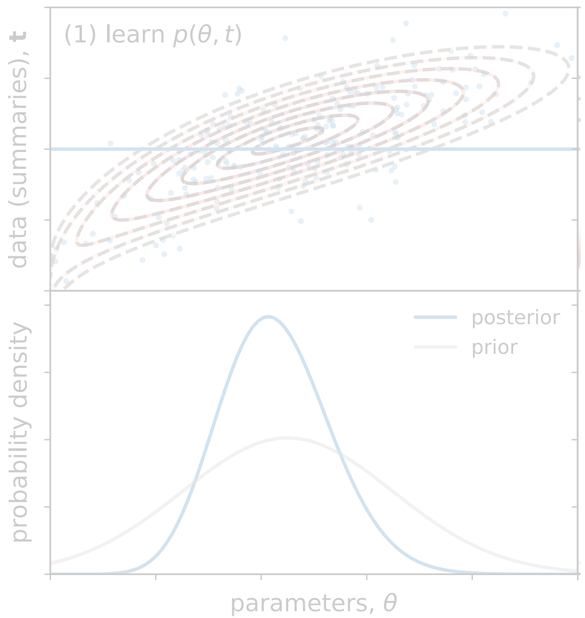


PHYSTAT "ML meets Statistics": many successful applications of AI aimed at solving these challenges



Generative models:
emulating forward models

Alsing+2018



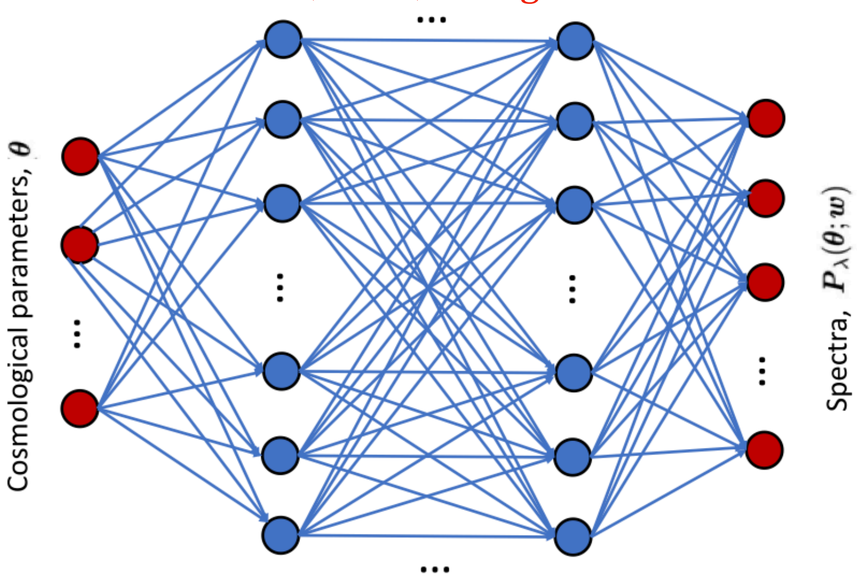
Implicit likelihood inference:
new parameter inference paradigm

Explainable AI:
ML-enabled scientific discoveries

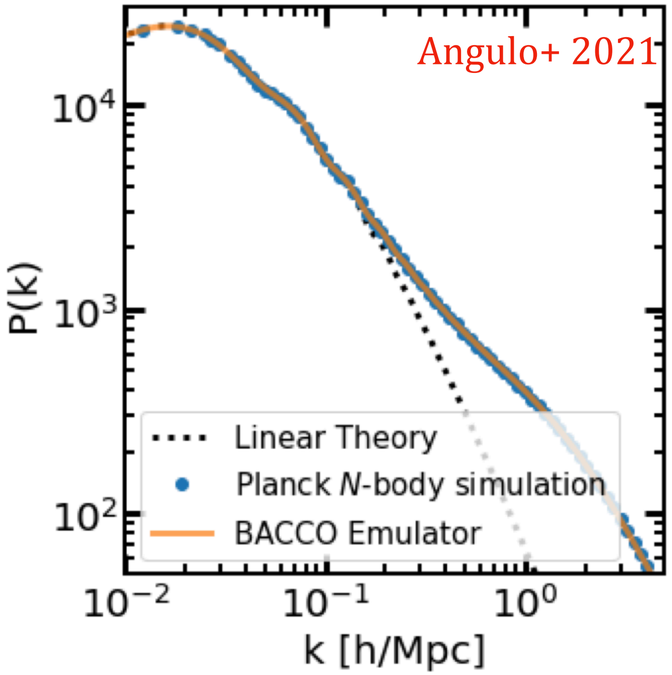
Emulators for summary statistics

***COSMOPOWER** suite of LSS/CMB spectra*

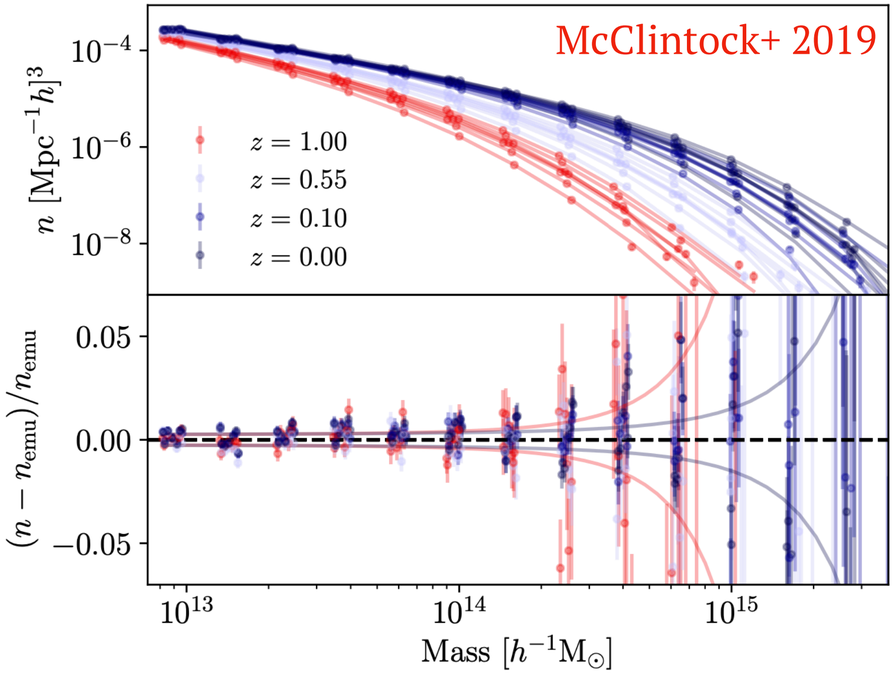
Mancini, Piras, Alsing et al. 2022



***BACCO** emulator for matter P(k)*



*Emulator **Aemulus** for halo mass function*



Work by Alessio Spurio Mancini

Accelerating High-Dimensional Cosmological Inference with COSMOPOWER *Alessio Spurio Mancini*

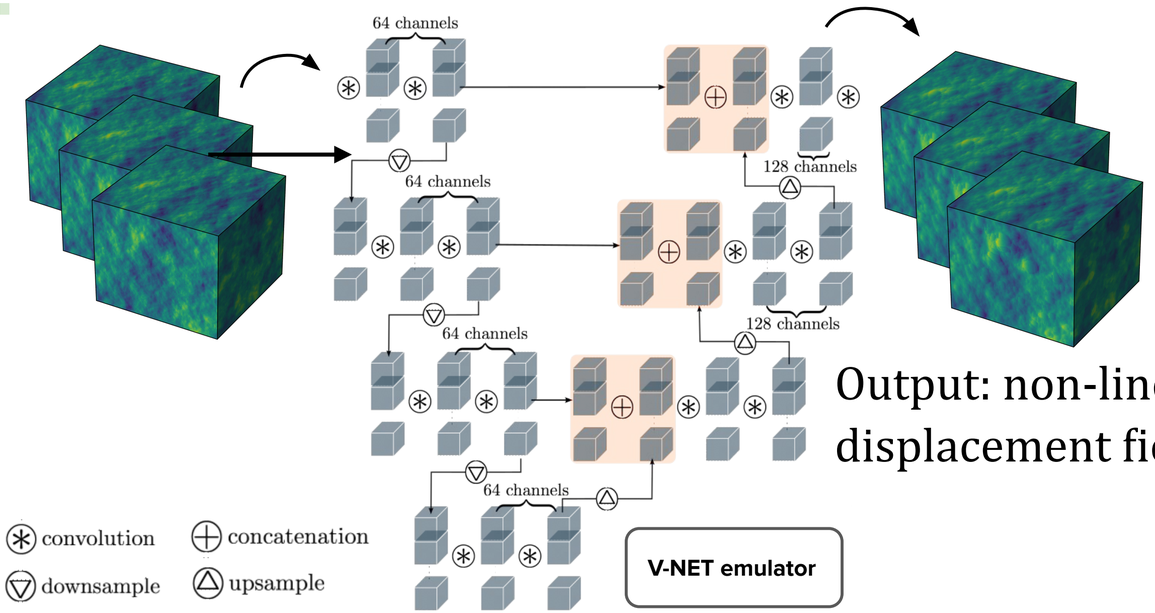
Lecture Theatre 2, Blackett Laboratory, Imperial College London 18:16 - 18:17

+ many more..

Field-level emulators for structure formation

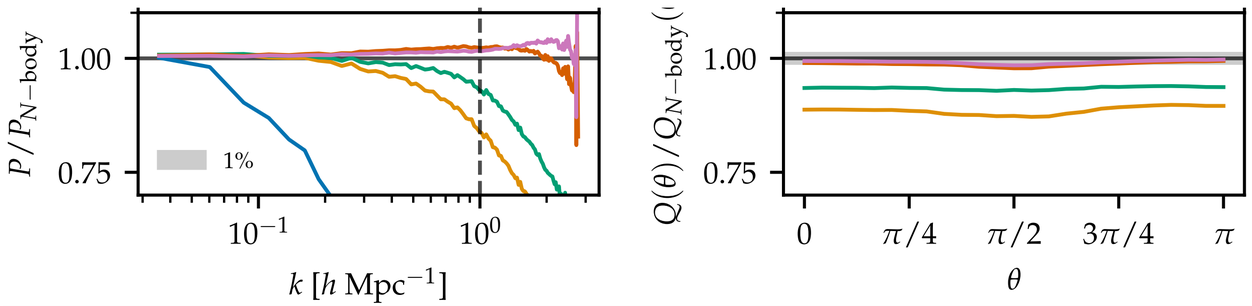
Input: linear displacement field + Ω_m

Jamieson et al. 2023; Doeser et al. 2023



Output: non-linear displacement field z=0

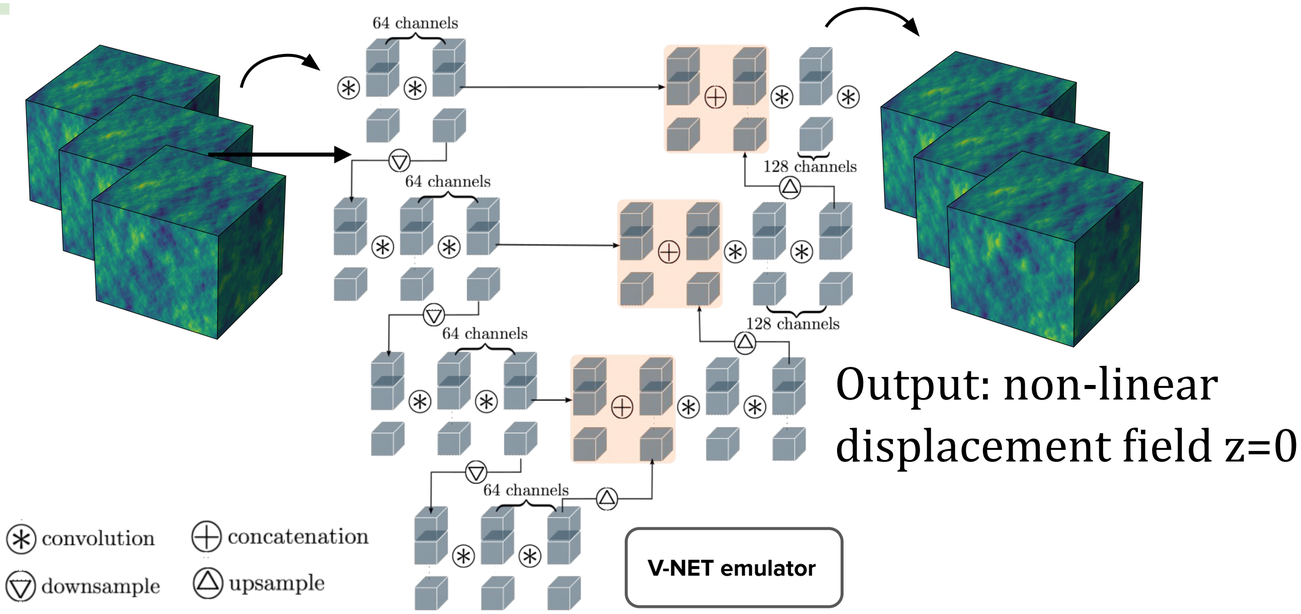
Emulator



Field-level emulators for structure formation

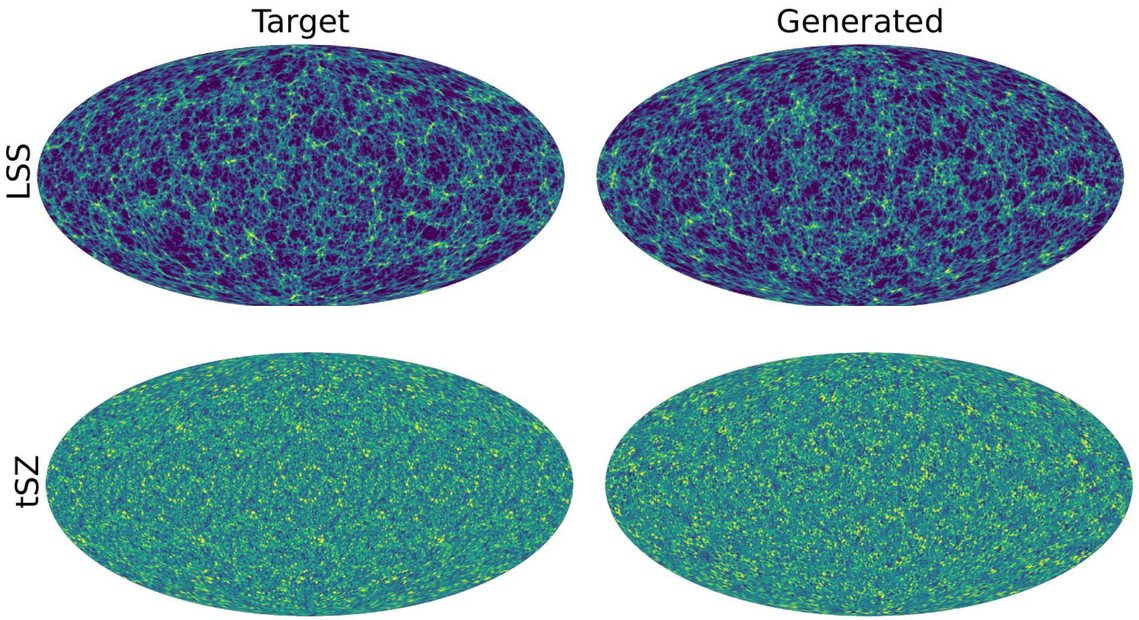
Input: linear displacement field + Ω_m

Jamieson et al. 2023; Doeser et al. 2023

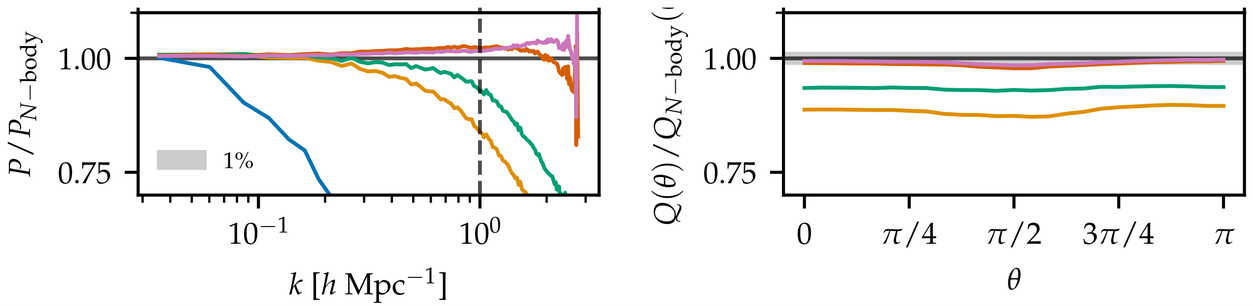


Output: non-linear displacement field z=0

Also on the sphere using scattering transforms!



Emulator



Work by Matt Price

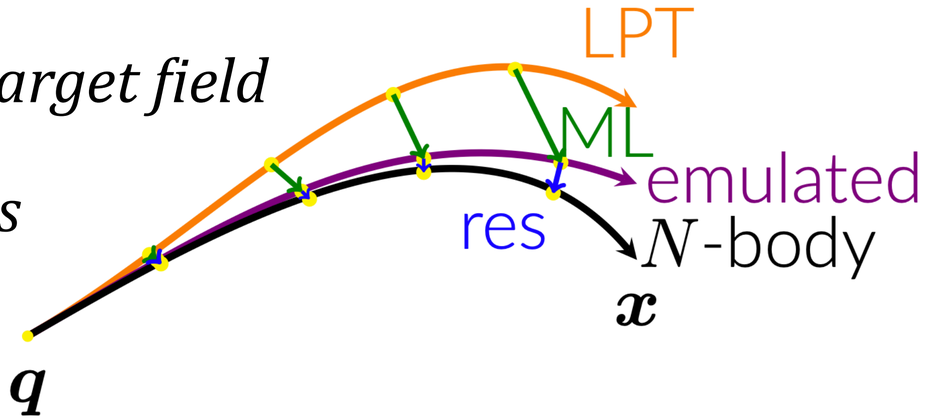
Generative models of astrophysical fields with scattering transforms on the sphere *Matt Price*

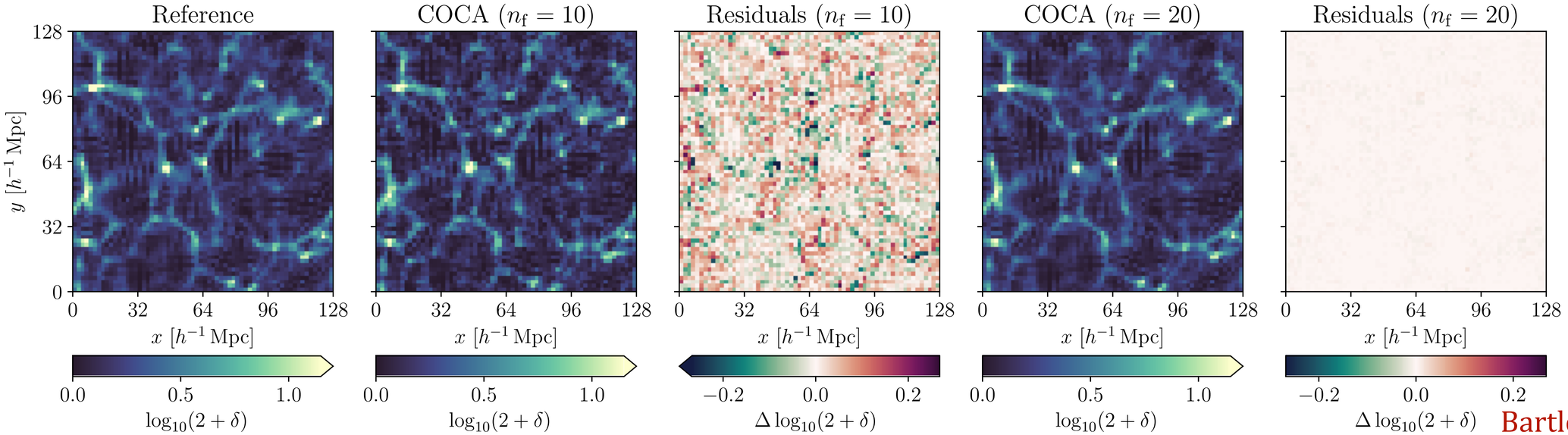
Lecture Theatre 2, Blackett Laboratory, Imperial College London 18:08 - 18:09

COCA: Correct for emulator errors to improve accuracy

Step I: *Use ML surrogate to get as close as possible to target field*

Step II: *Force calculation to correct for emulator errors*



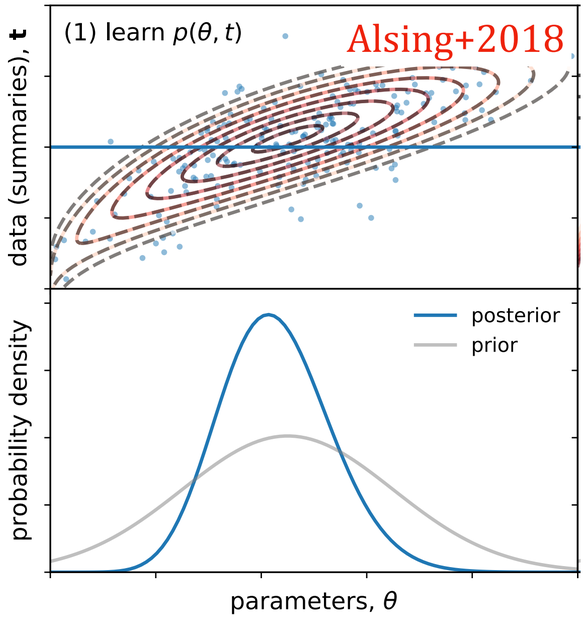


Bartlett et al. 2024

Work by Deaglan J. Bartlett

COmoving Computer Acceleration (COCA): Correcting Emulation Errors for Trustworthy N-Body Simulations
Deaglan Bartlett

PHYSTAT "ML meets Statistics": many successful applications of AI aimed at solving these challenges



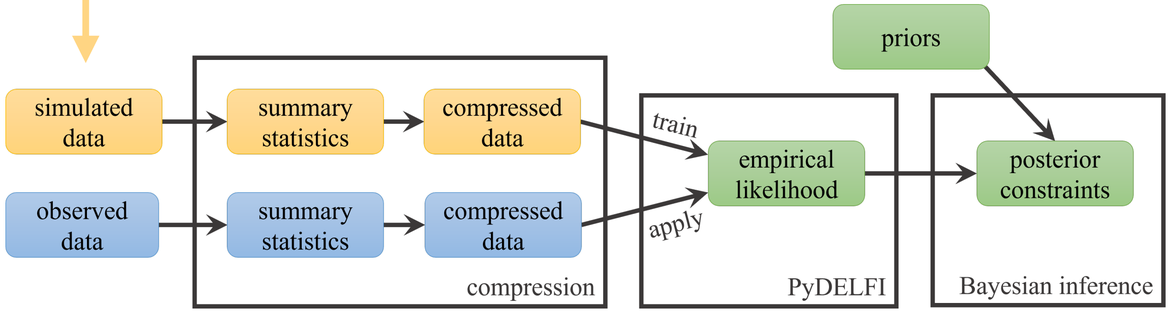
Generative models:
emulating forward models

Implicit likelihood inference:
new parameter inference paradigm

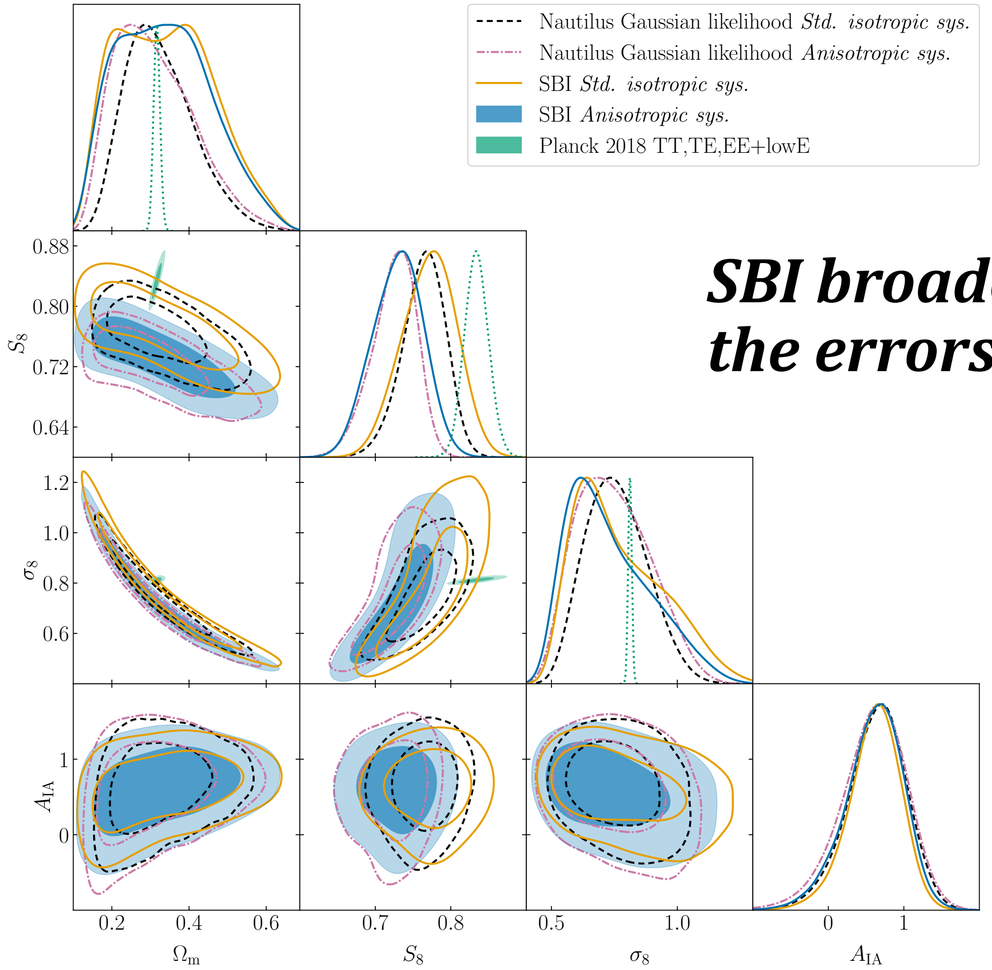
Explainable AI:
ML-enabled scientific discoveries

Simulation-based inference using weak lensing summary statistics

Realistic forward simulations to produce 2-point summary statistic



SBI pipeline



SBI broadens the errors

Work by Kiyam Lin

SBI for wide field weak lensing *Kiyam Lin*

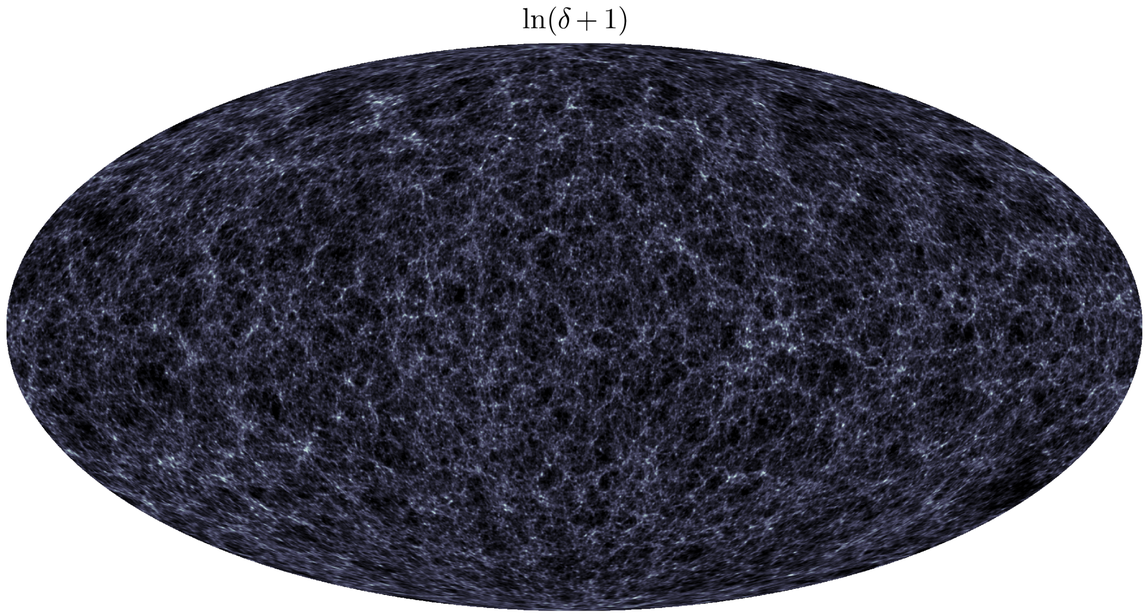
Lecture Theatre 2, Blackett Laboratory, Imperial College London 18:10 - 18:11

Lin , Von Wietersheim-Kramsta et al. 2023

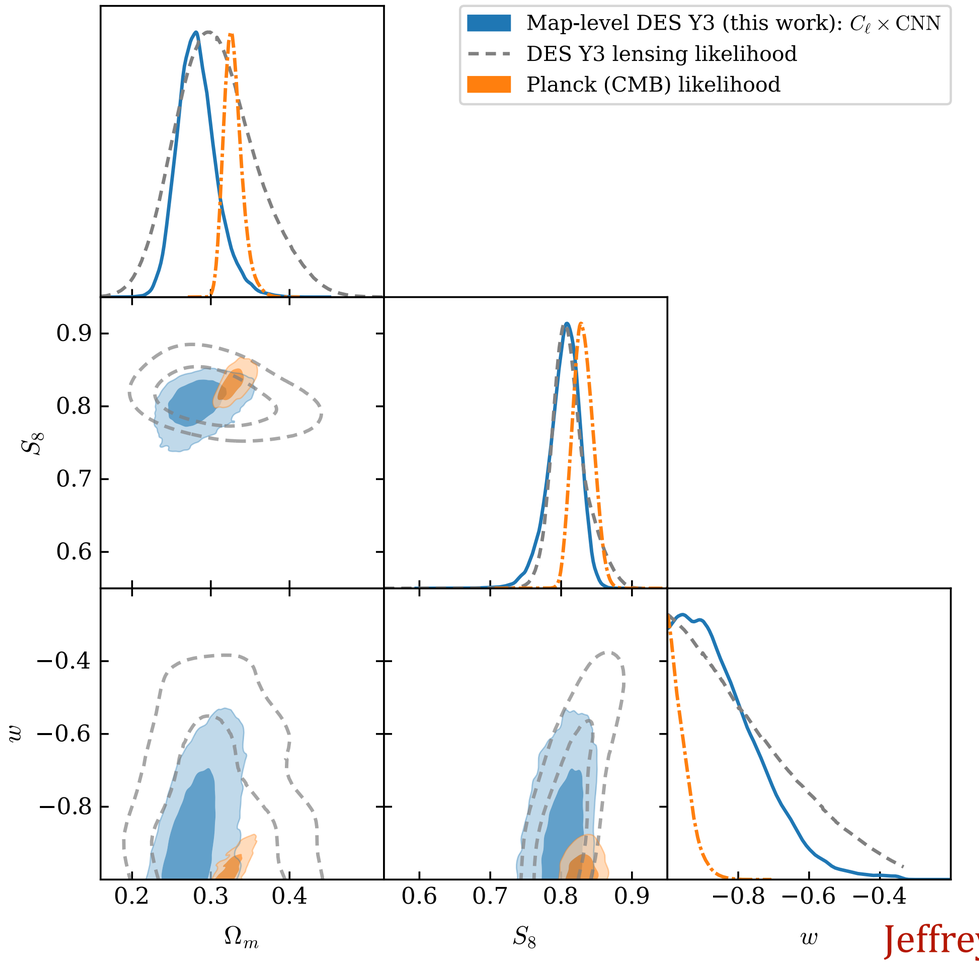
Von Wietersheim-Kramsta, Lin et al. 2024

Simulation-based inference using weak lensing maps

Realistic forward model weak lensing maps → *Neural compression (CNN)* → *SBI*



Dark matter overdensity map

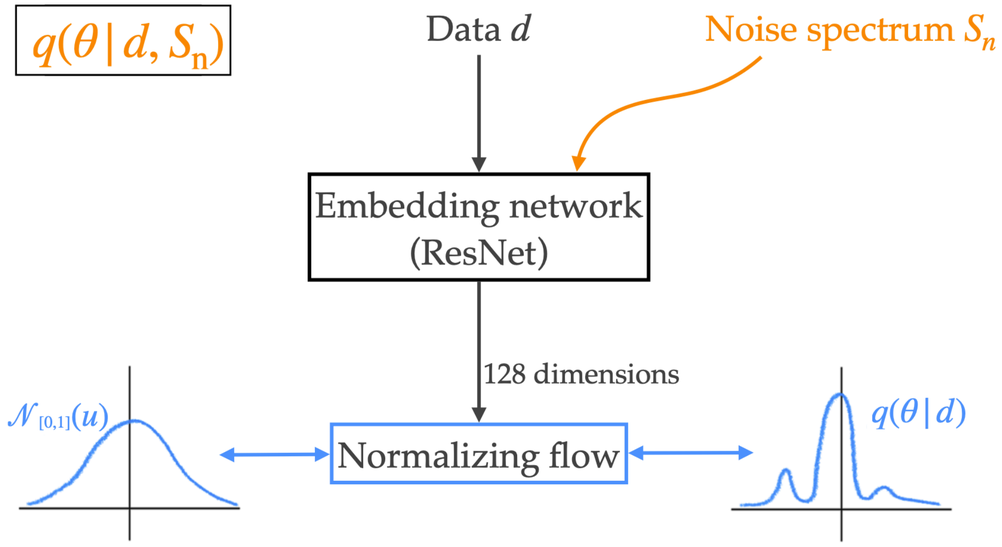


Jeffrey et al. 2024

Work by Niall Jeffrey

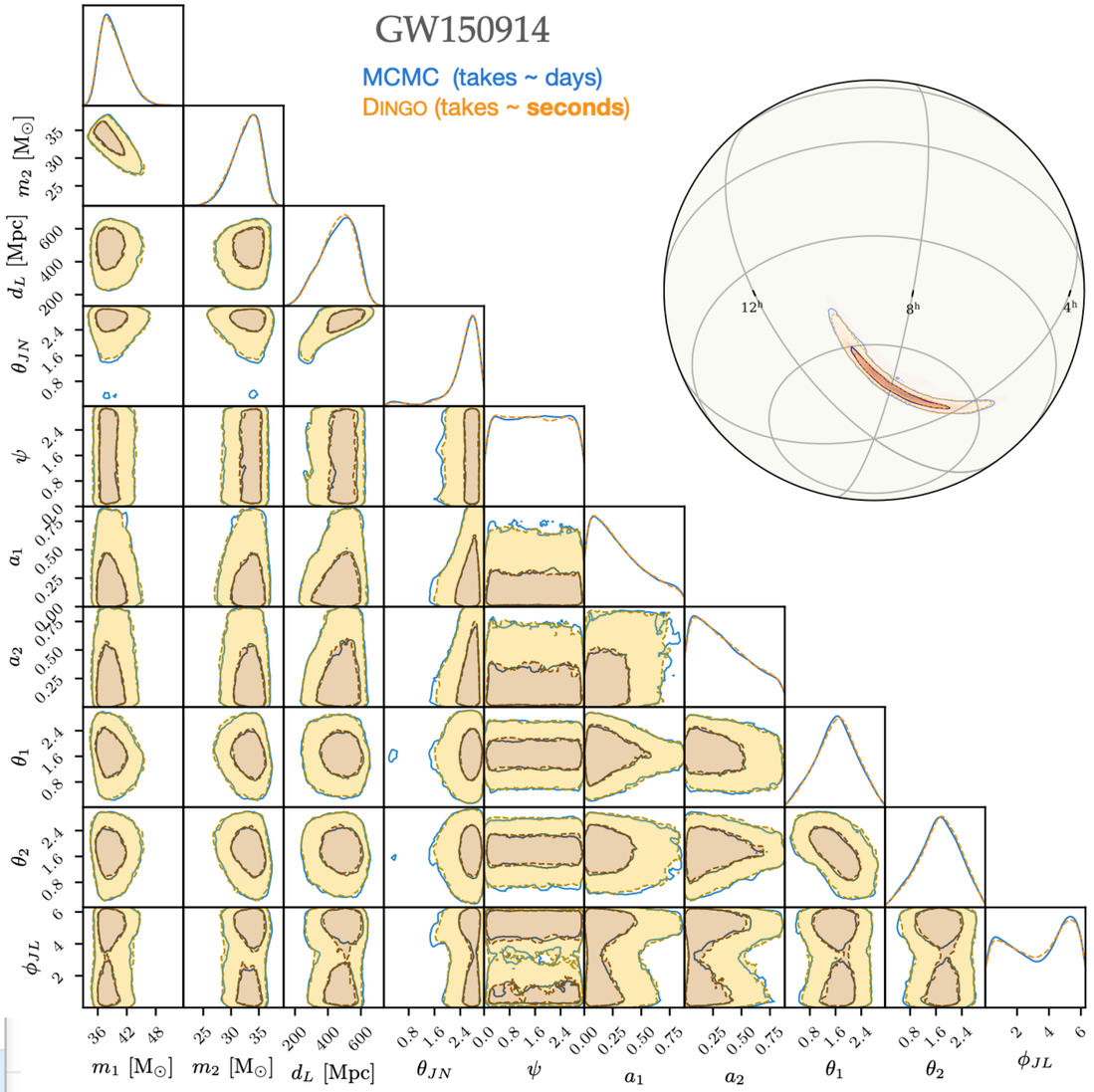
Types of ML in Astro/Cosmology — Dr Niall Jeffrey

Simulation-based inference for gravitational waves



- Group-equivariant NPE
  enforcing symmetries within NPE
- Importance sampling + NPE
  when likelihood is tractable

Dax+ (PRL 2021)



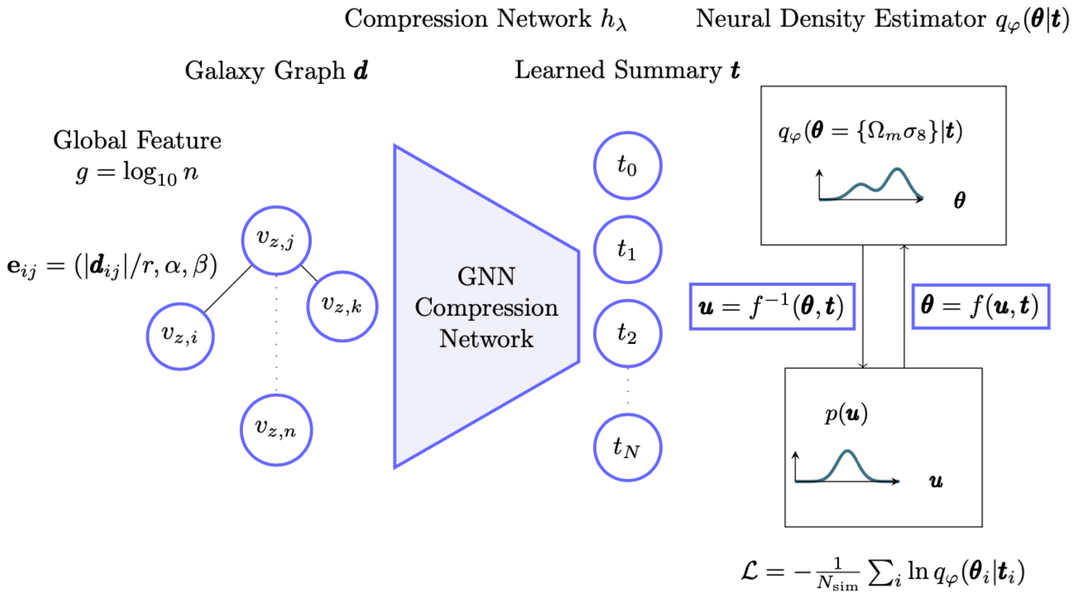
Work by Max Dax

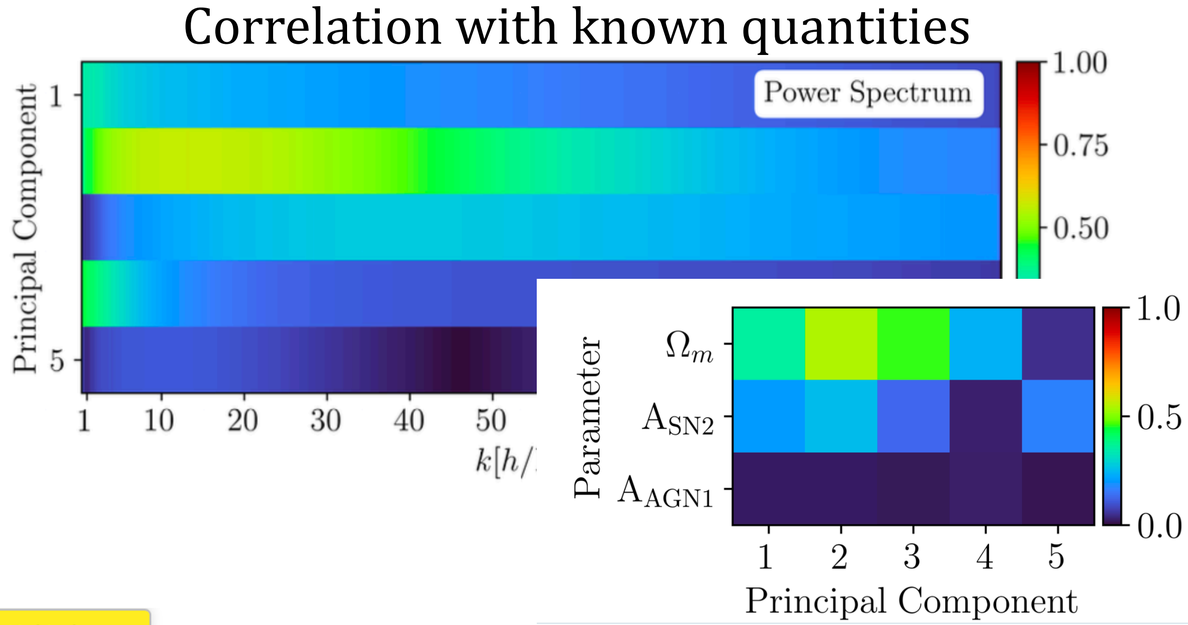
Simulation-based machine learning for gravitational-wave analysis *Maximilian Dax*

Lecture Theatre 2, Blackett Laboratory, Imperial College London 11:45 - 12:30

Simulation-based inference meets interpretability…

- *Can we interpret the summaries learnt in the neural compression step?*





Learning Optimal and Interpretable Summary Statistics of Galaxy Catalogs with SBI *Kai Lehman*

Lecture Theatre 2, Blackett Laboratory, Imperial College London 18:17 - 18:18

Work by Kai Lehman

- *Do we even need 'black-box' ML for SBI? Linear regression is good enough…*

lsbi: linear simulation based inference *Dr William Handley*

Lecture Theatre 2, Blackett Laboratory, Imperial College London 14:25 - 14:50

Work by Will Handley

If so successful, why still so much skepticism towards ML in Astro?

Debate #1: what would it take for the community to accept ML findings?

Debate #2: is there truth in latent space?

Are machine learning and physics fundamentally different?

arXiv > stat > arXiv:2405.18095

Sear

Help

Statistics > Machine Learning

[Submitted on 28 May 2024 (v1), last revised 31 May 2024 (this version, v2)]

Is machine learning good or bad for the natural sciences?

David W. Hogg (NYU, MPIA, Flatiron), Soledad Villar (JHU, Flatiron)

Machine learning (ML) methods are having a huge impact across all of the sciences. However, ML has a strong ontology – in which only the data exist – and a strong epistemology – in which a model is considered good if it performs well on held-out training data. These philosophies are in strong conflict with both standard practices and key philosophies in the natural sciences.

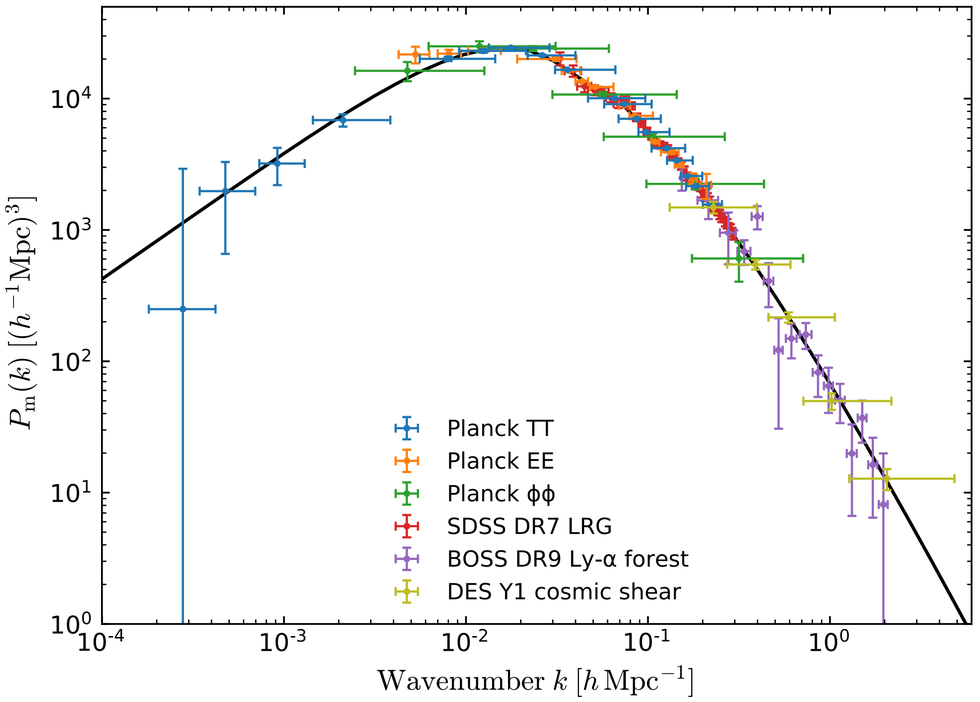
- Definitions used to define ML also applicable to many physical models in cosmology
  e.g. NFW model (i) fit to **data** and (ii) validated on **held-out halos** in simulations

- Definitions reflect basic usage of ML, not ML itself
  e.g. can validate beyond held-out training data, can include physics in ML model
  (see Peiris' talk)

Key issue: lack of explainability

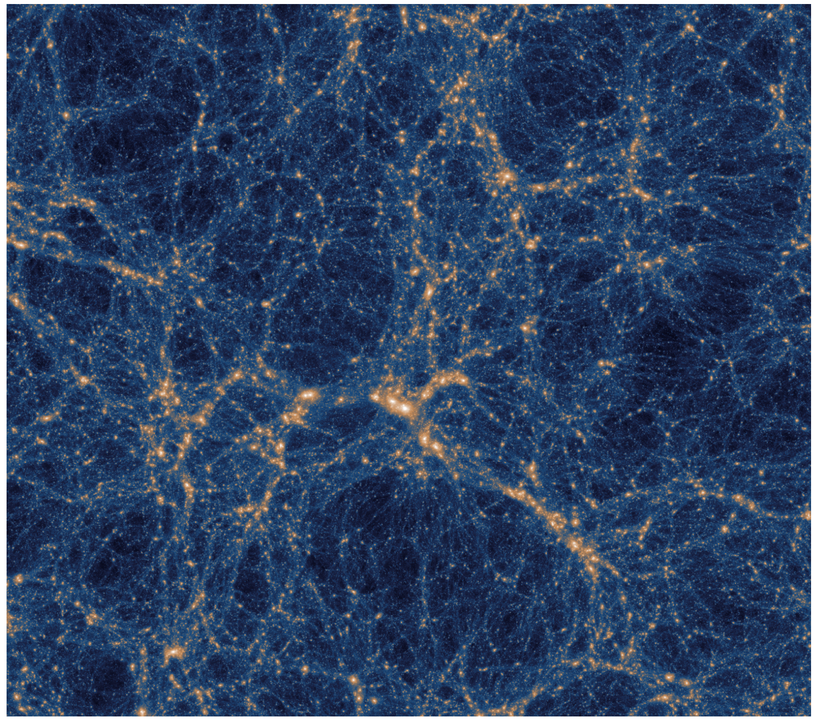
Don't need it → *Need it*

Physically understand the input and output



...but can explainability still help?

ML-enabled scientific discoveries

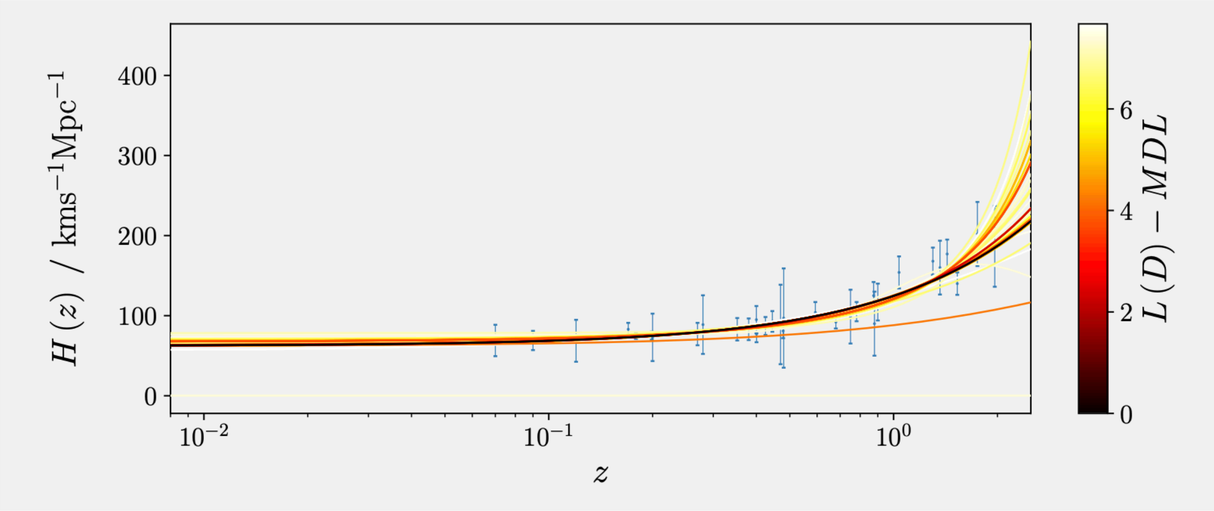


*No **explanation** for emergent properties of LSS from traditional methods*

Symbolic regression

find accurate & simple analytic expressions to fit data

Expansion rate of the Universe



- *Equations have higher portability than a NN*
- *Are equations always 'interpretable'?*
  Can be as complex as a neural network..
- *Are they 'explainable'? i.e. do they make sense within science domain?*
  Which theory predicts H(z) functional form?

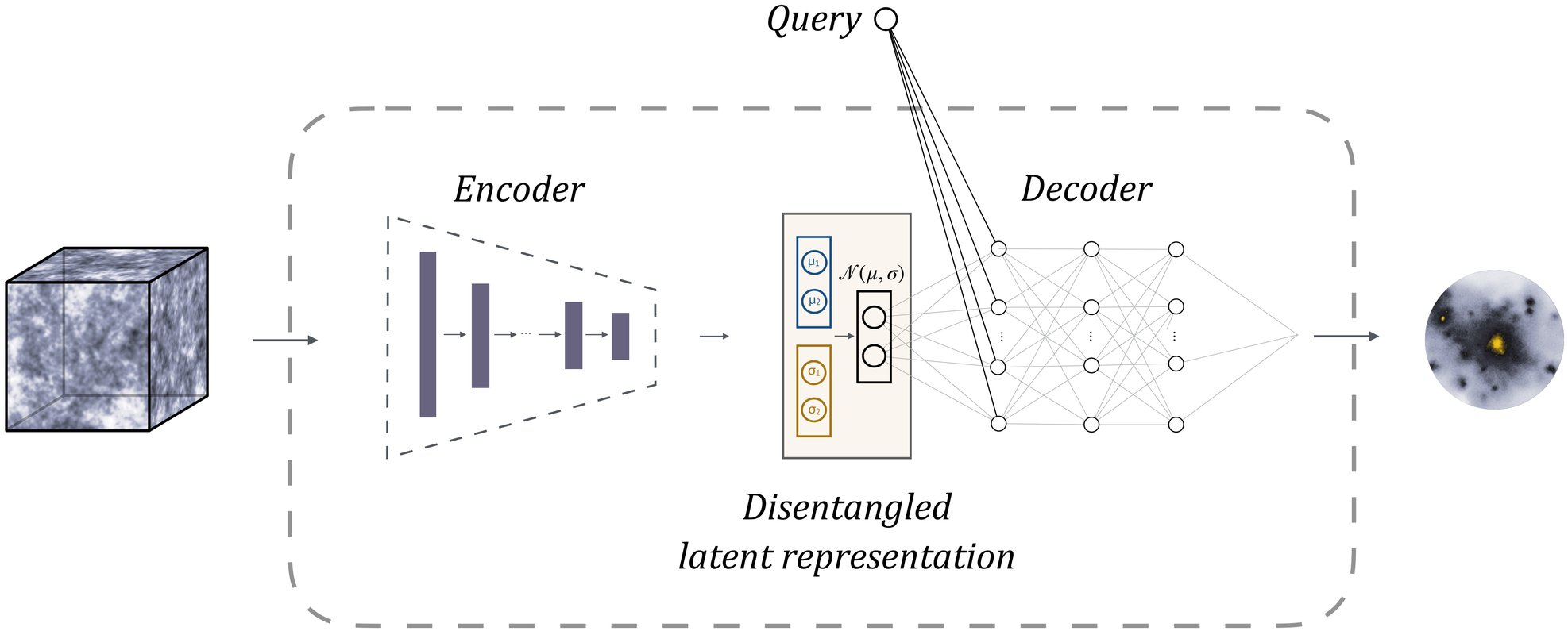
Work by Harry Desmond

Exhaustive Symbolic Regression: Learning Astrophysics directly from Data *Harry Desmond*

Lecture Theatre 2, Blackett Laboratory, Imperial College London 18:11 - 18:12

See also PySR from Cranmer et al. 2019, 2020 (github.com/MilesCranmer/PySR/)

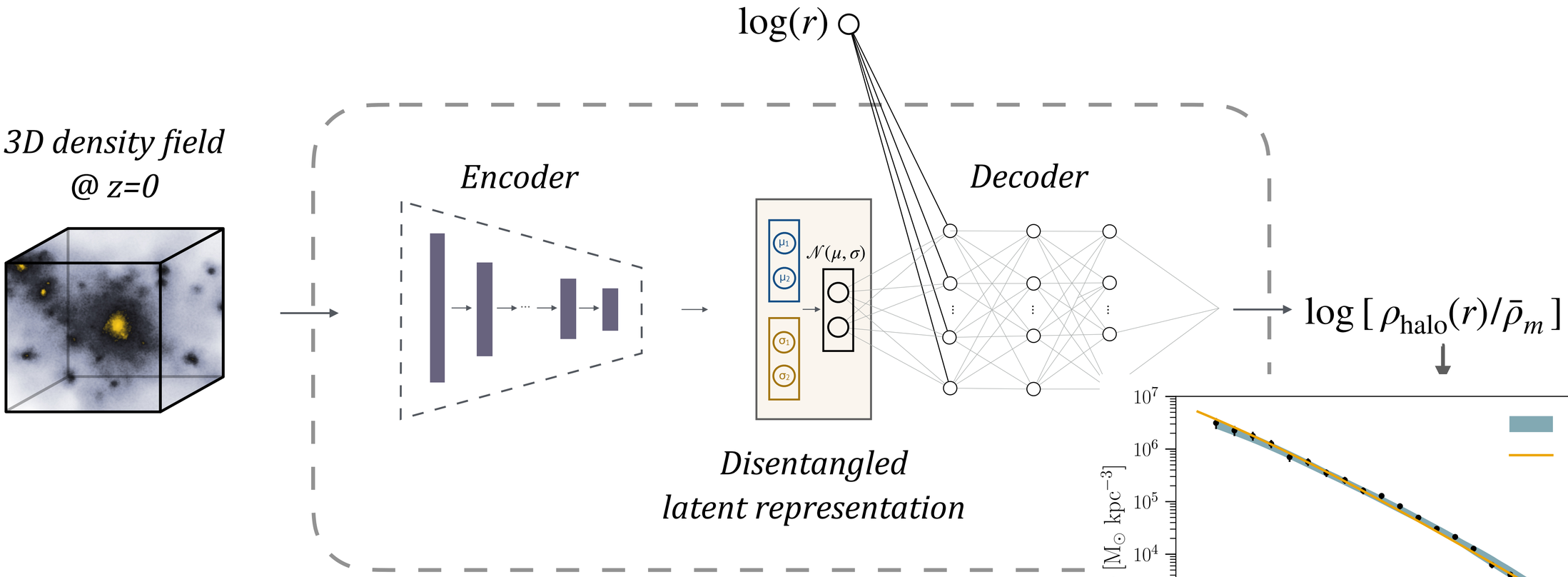
Interpretable Variational Encoder (IVE) for explainable AI



An information theoretic perspective:
Model compression enables "explainability"

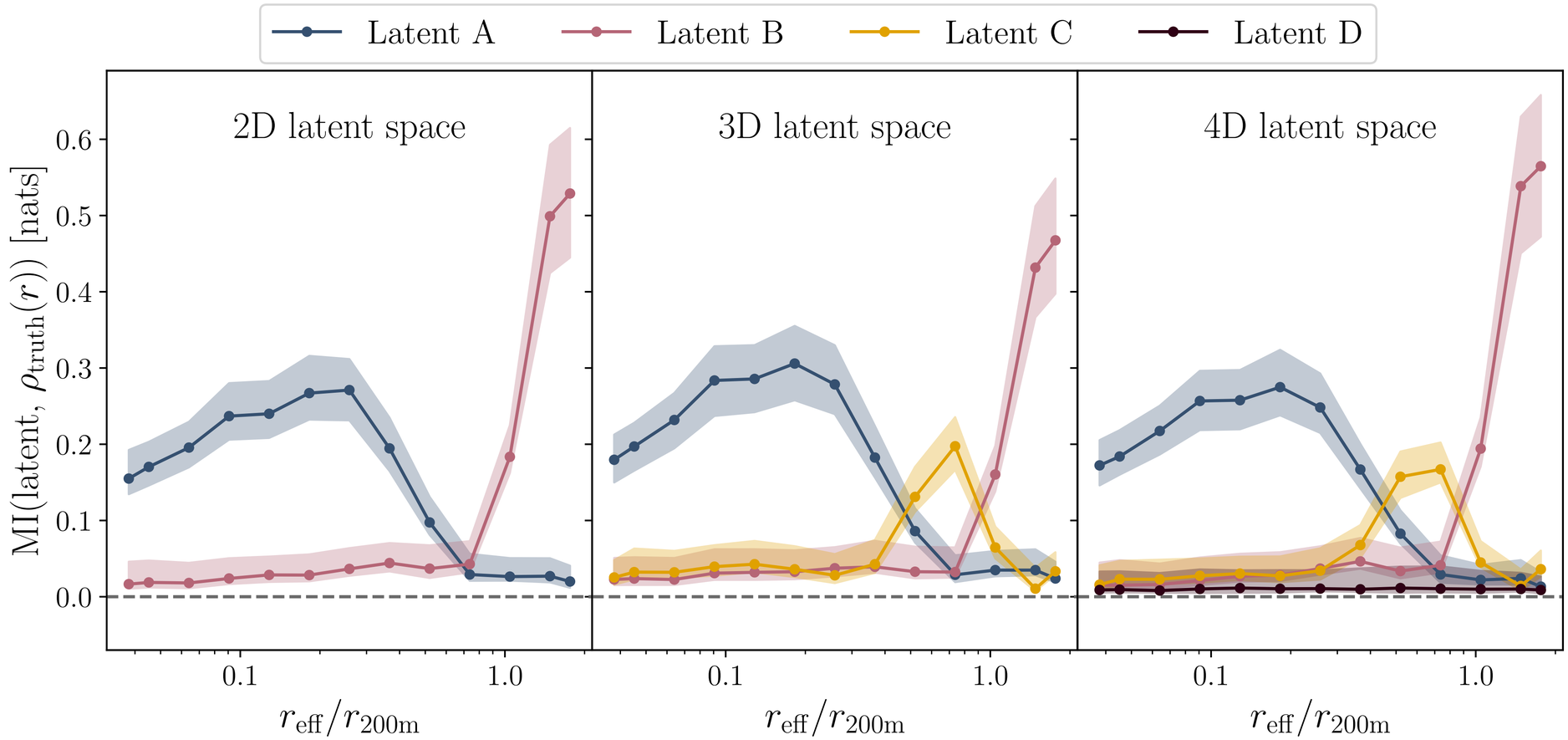
work with Hiranya Peiris (Cambridge), Andrew Pontzen (UCL)

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



Lucie-Smith, Peiris, Pontzen, Nord+ (PRD, 2022)
Lucie-Smith, Peiris, Pontzen (PRL, 2024)

Interpreting the latent representation using mutual information



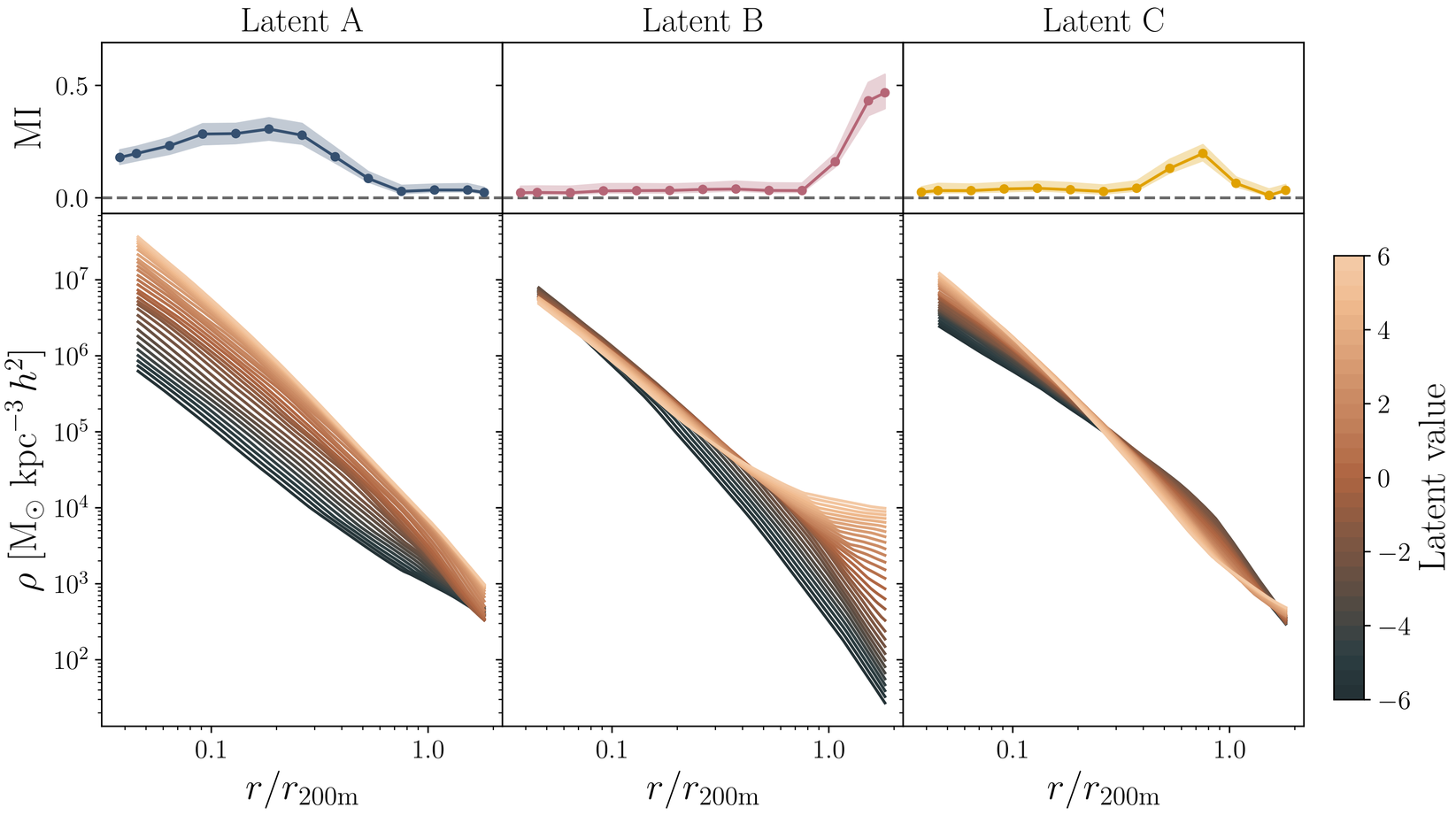
MI estimator:

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

Piras, Peiris, Pontzen, Lucie-Smith, Guo, Nord (MLST, 2023)

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

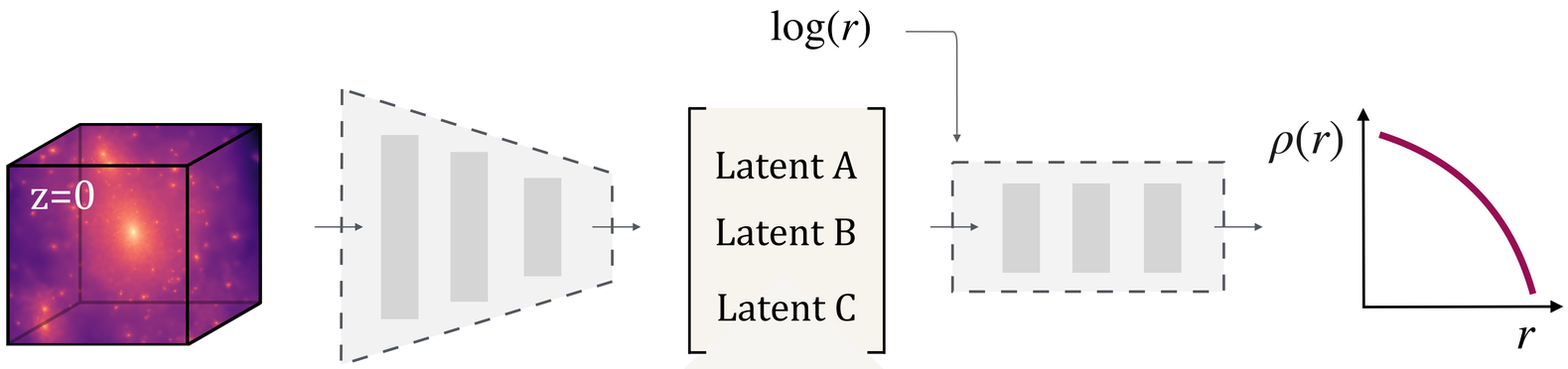
Systematically varying one latent at a time



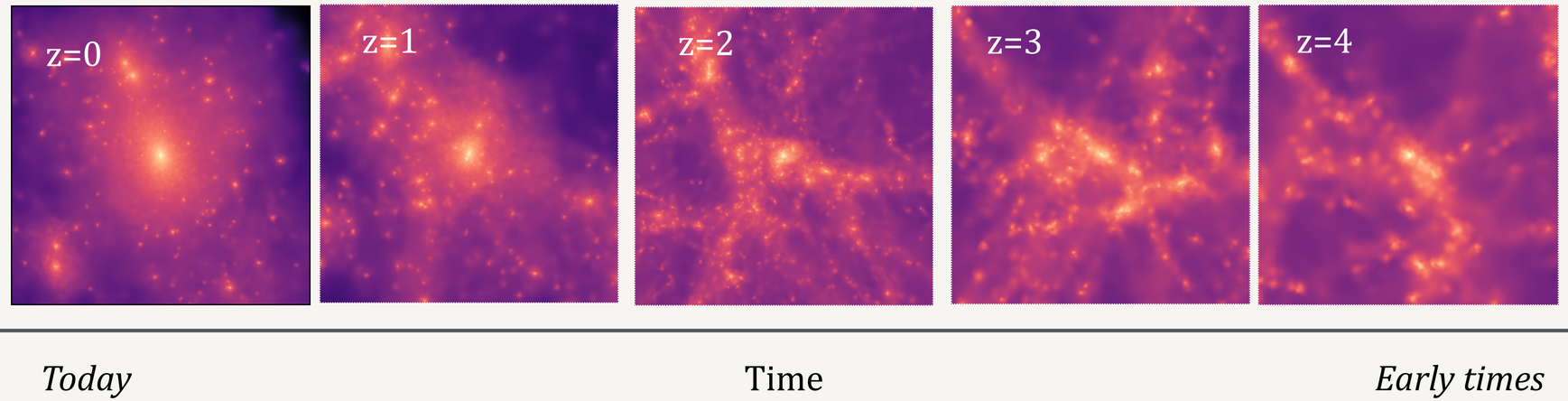
Latent A = ***normalisation***; Latent B = ***outer slope;*** Latent C = ***inner slope***

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

Exploiting the latent representation beyond its original training task

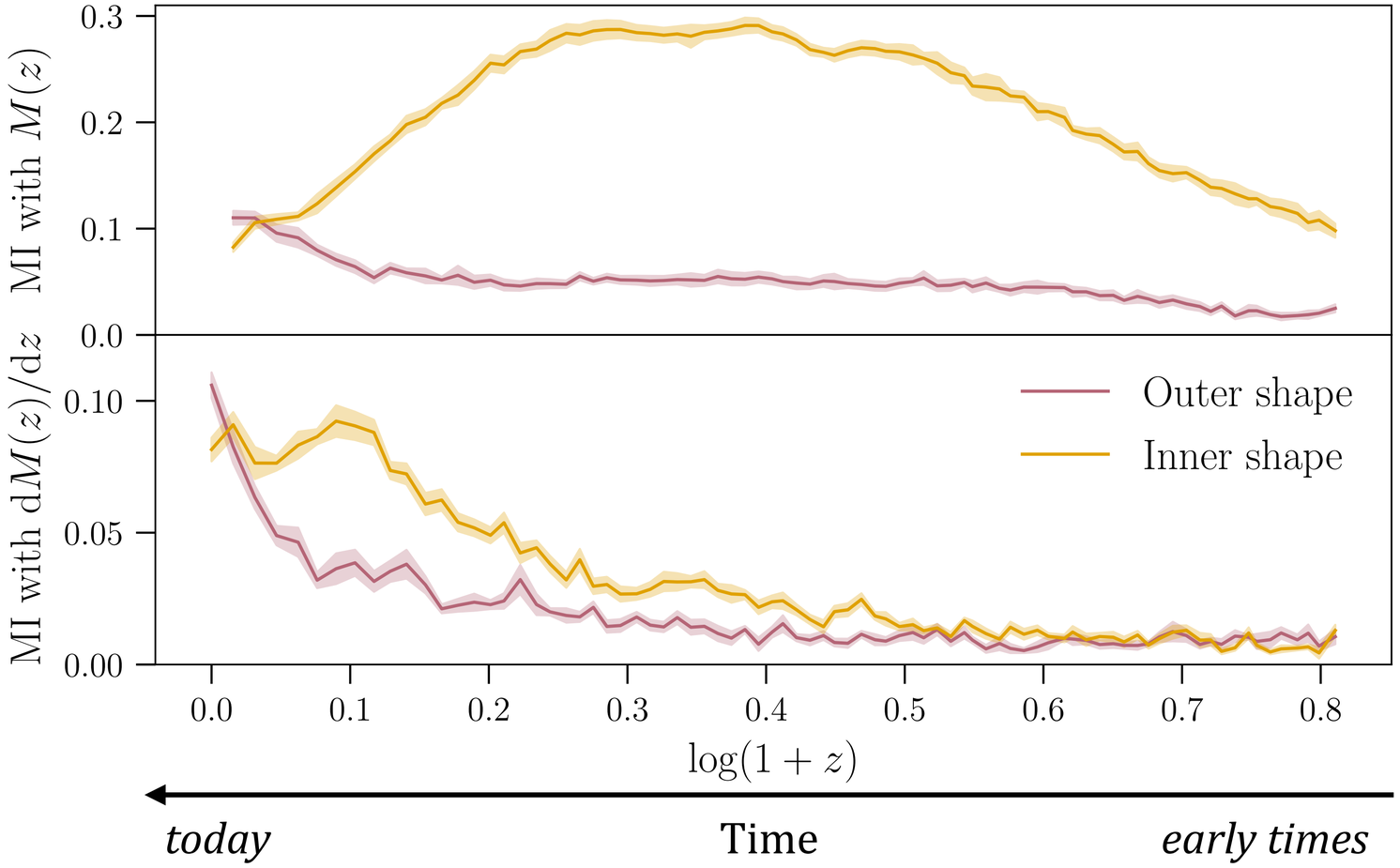


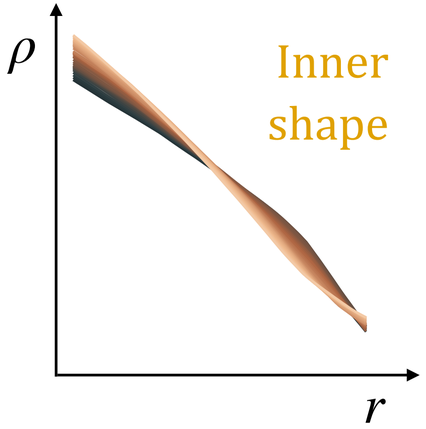
Does the latent space contain information about the origin of the halo density structure?

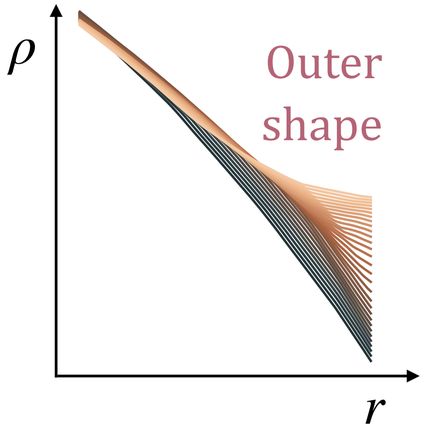


Lucie-Smith, Peiris, Pontzen (PRL, 2024)

Connection between the latents and the halo evolution history



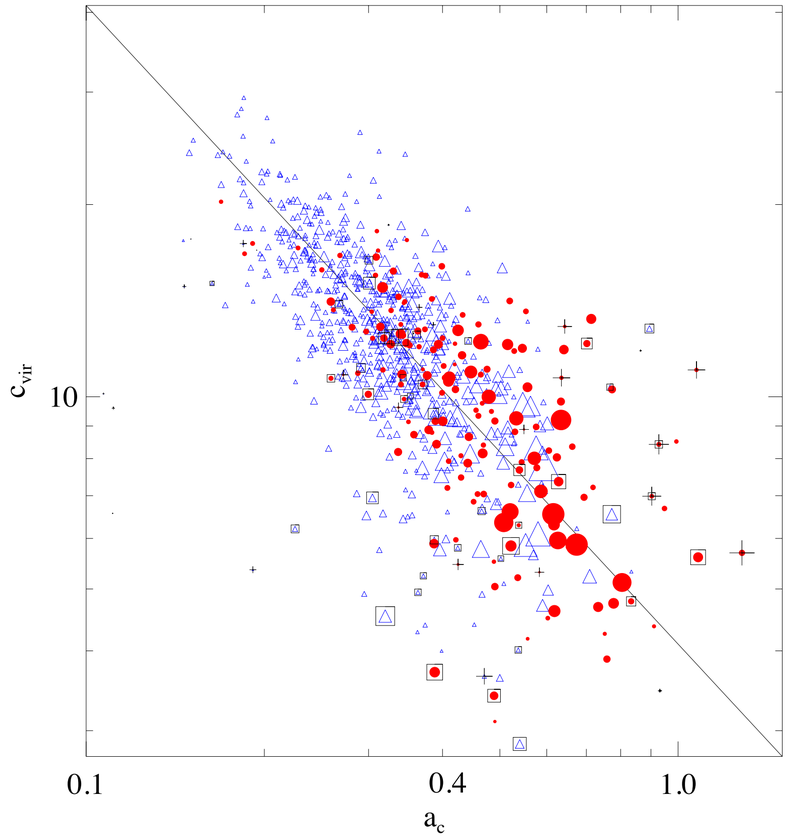




Lucie-Smith, Peiris, Pontzen (PRL, 2024)

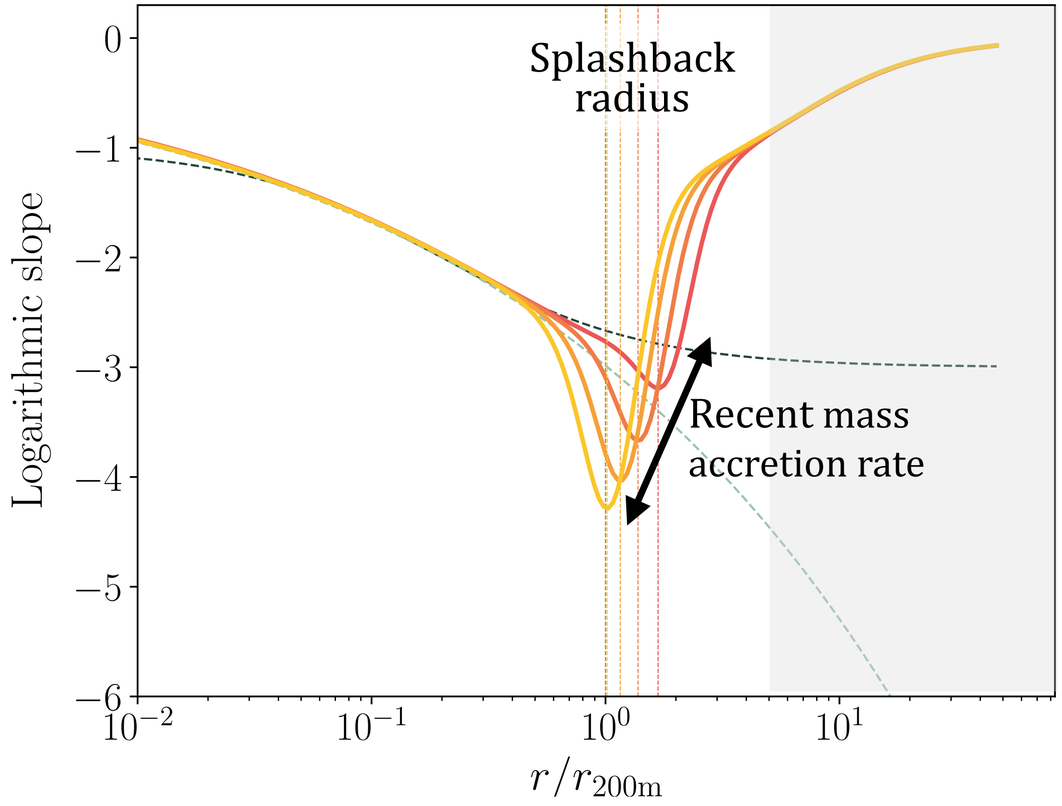
Have we learnt something new?

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



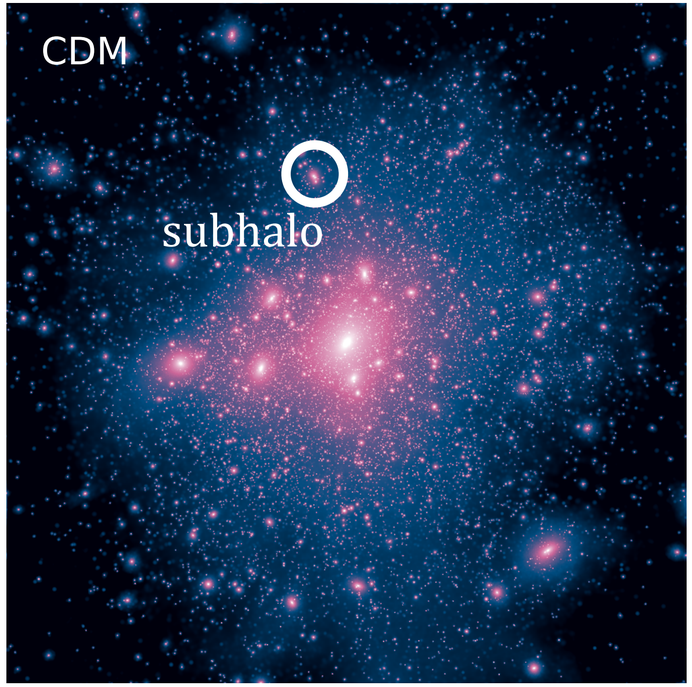
Wechsler et al. (2002)

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

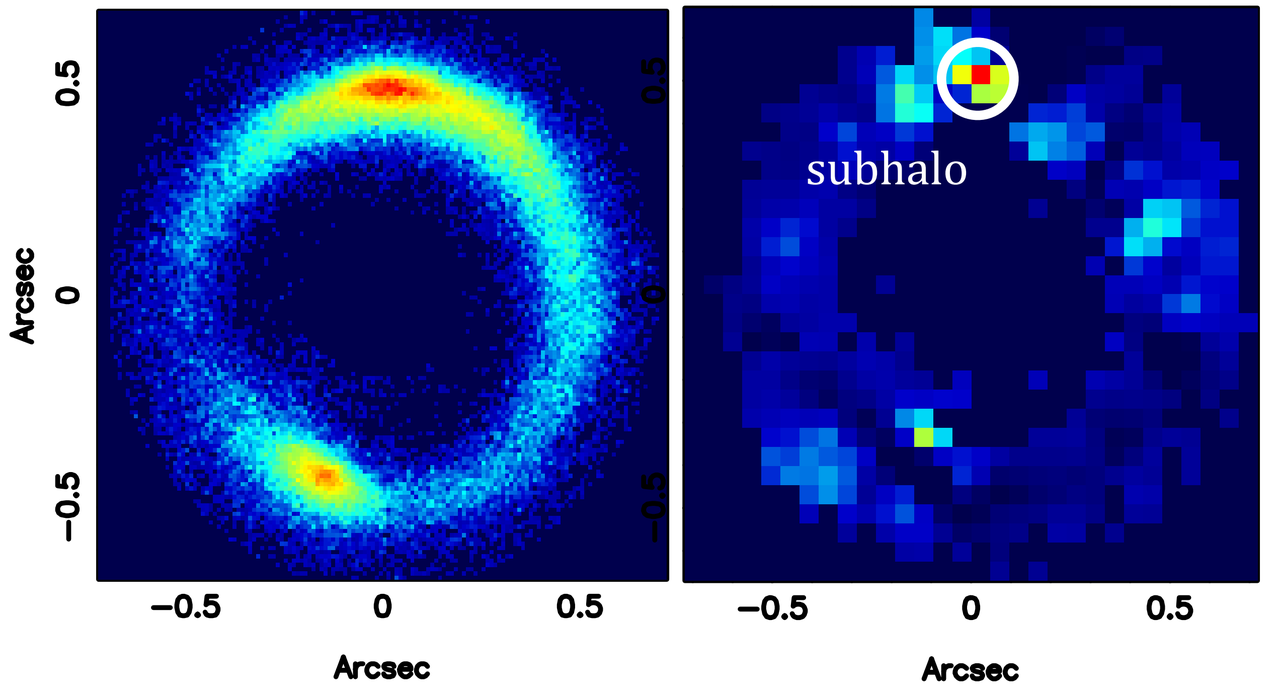


What about substructures?

Simulations



Strong lensing observations

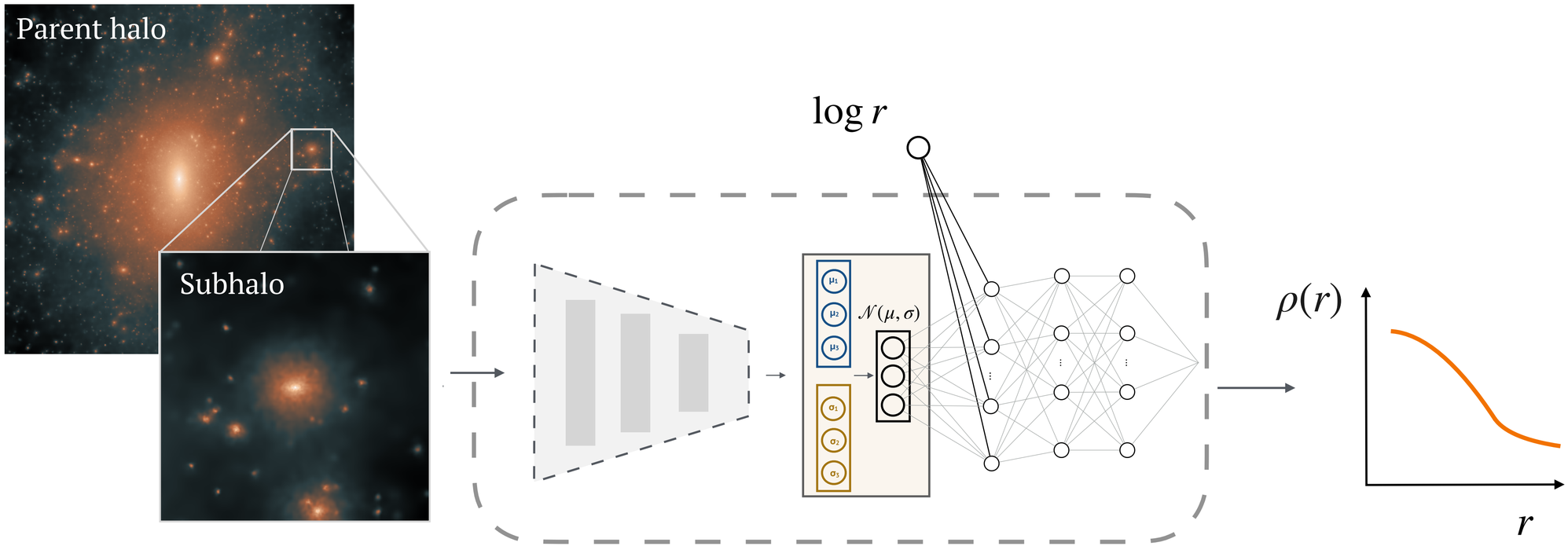


JVAS B1938+666; Vegetti et al. (2012)

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

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

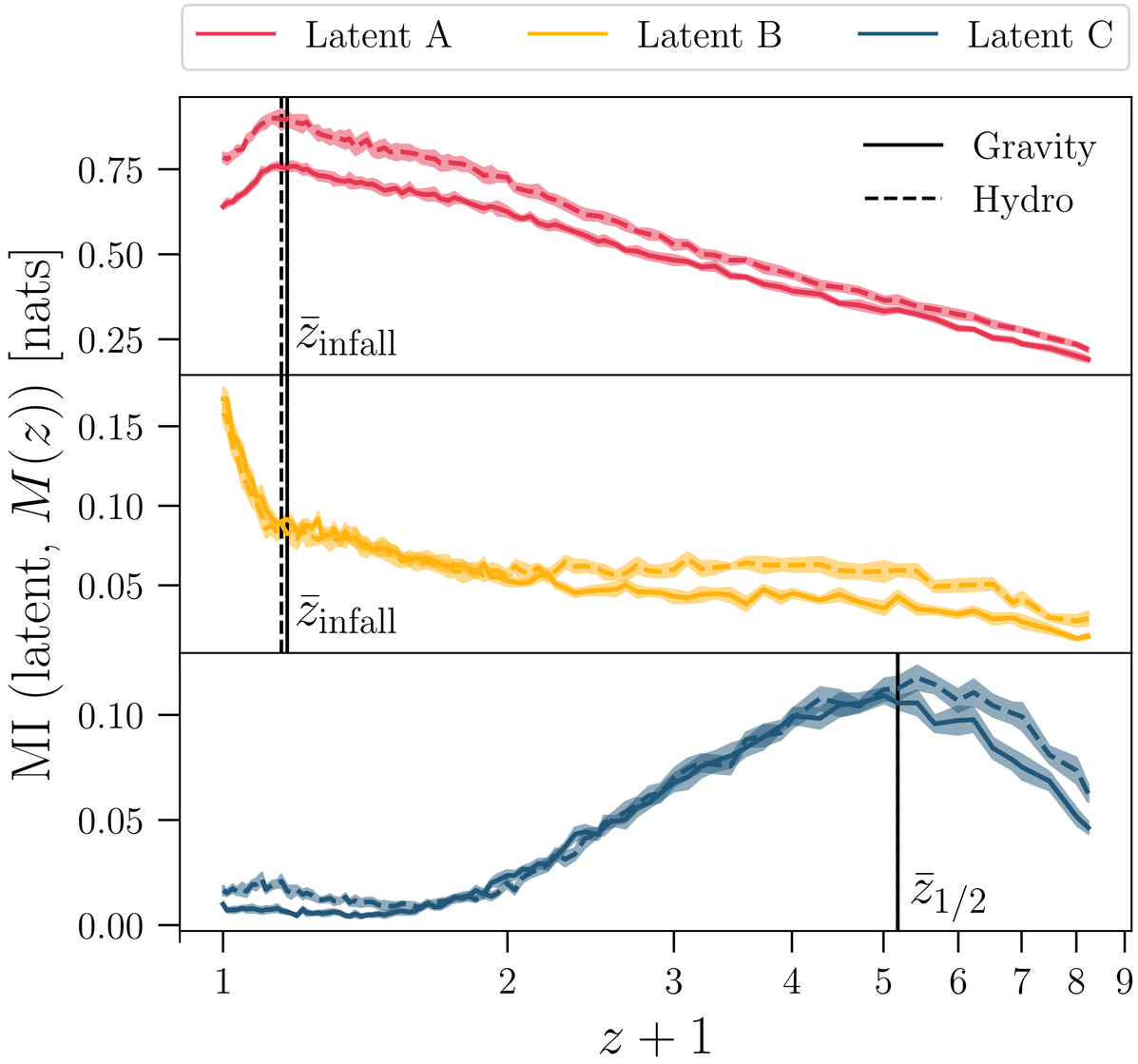
The subhalo density profile at $r < R_{200\mathrm{m}}$



Illustris-TNG100
(Gravity & Hydro)

Also accounting for the impact of galaxy formation on the profile

Three parameters with clear physical interpretation



'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*

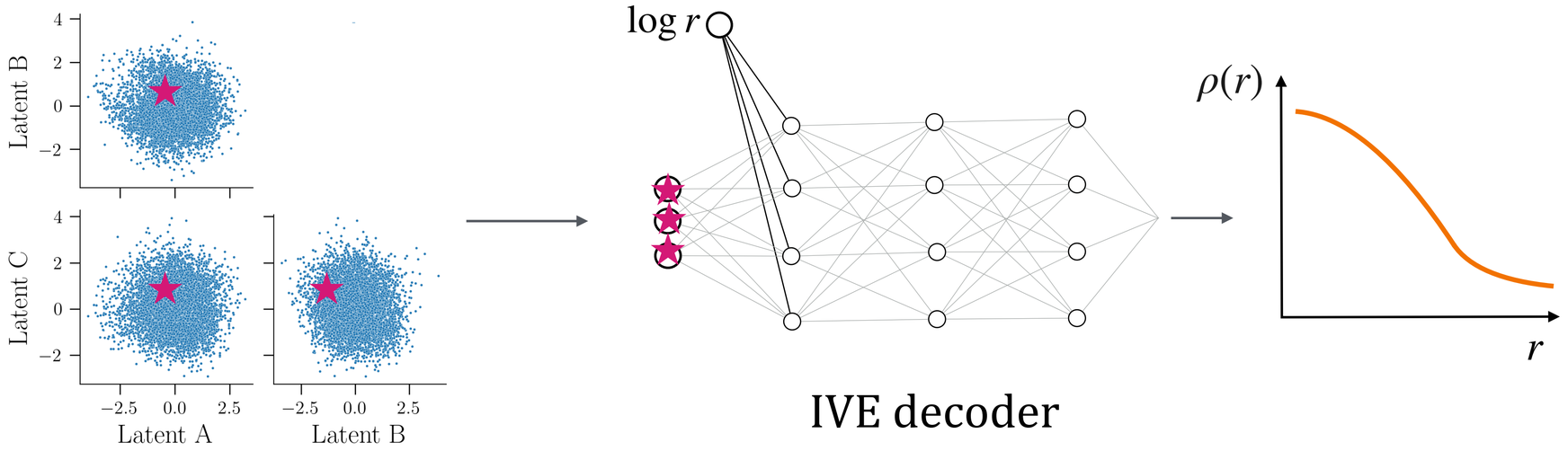
'Inner shape' *latent sensitive to half-mass* ***formation time***

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

New physically interpretable, data-driven subhalos density profile model for strong gravitational lensing

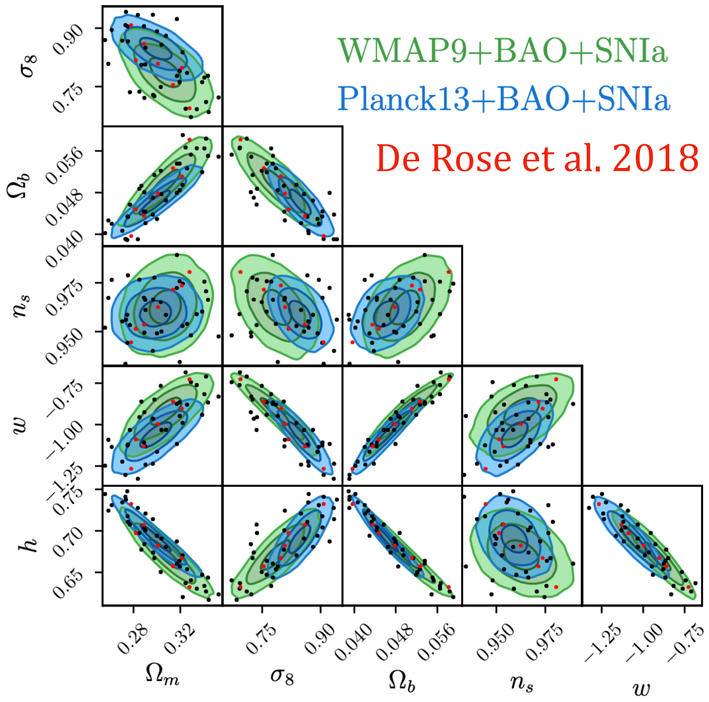
Sample from the latent space

Infer subhalo profile

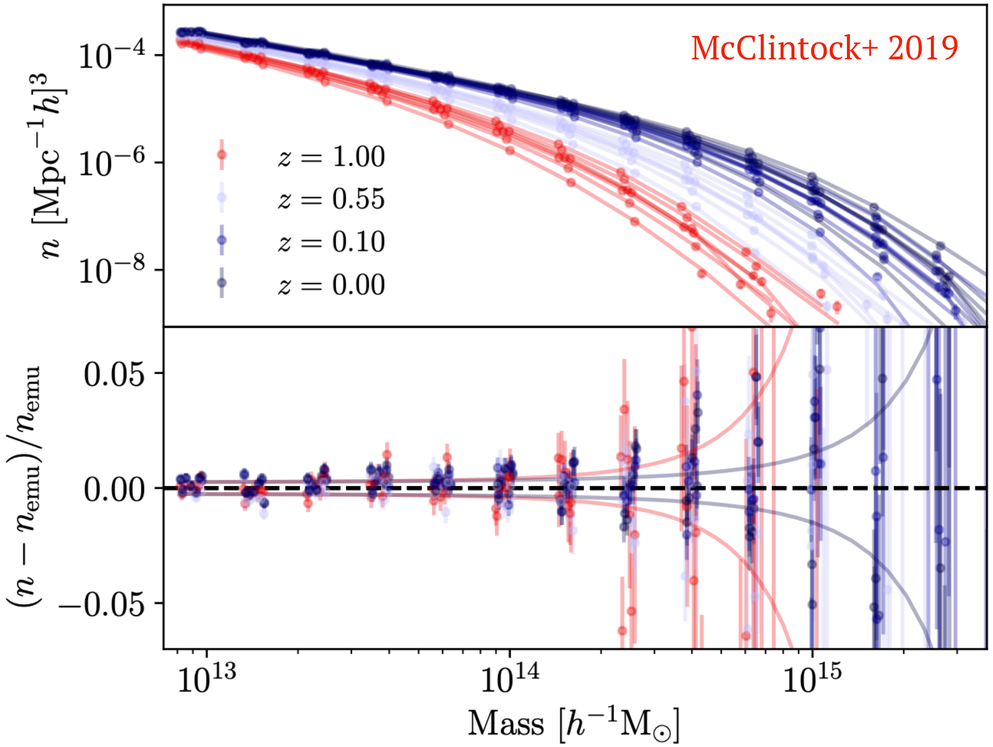


Next step: Integrate IVE model within strong gravitational lensing pipeline

Emulator goal is "accuracy", but can interpretability still be useful?



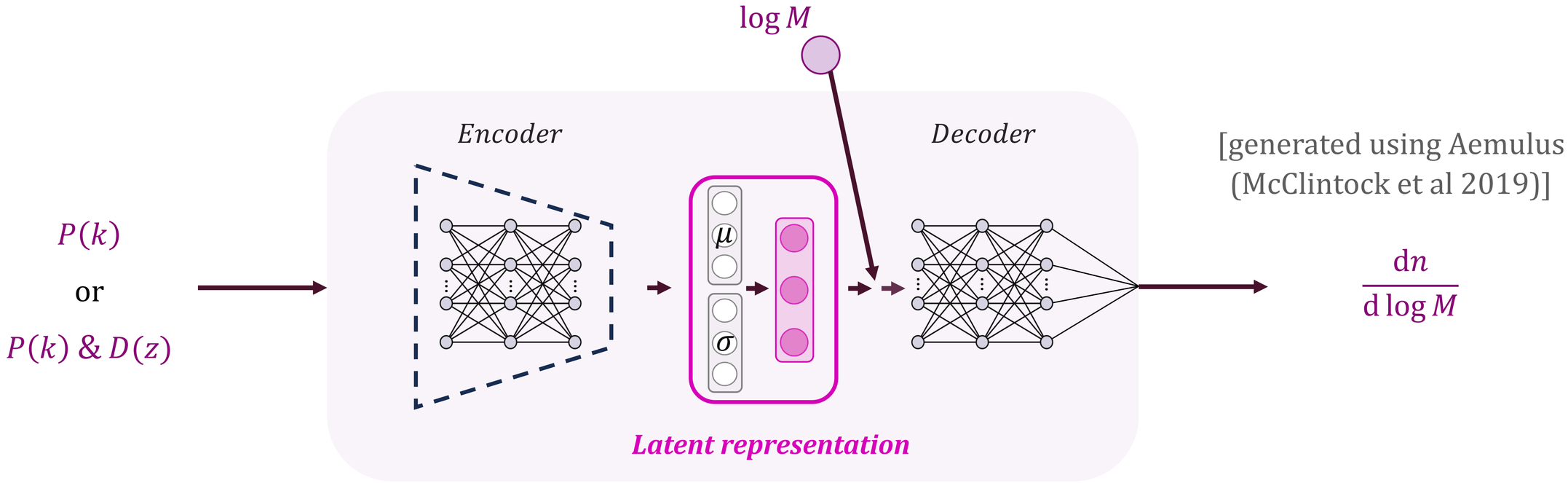
Aemulus emulator



- *Require sampling high dimensional cosmological parameter space*
- *Cannot generalize beyond its domain of validity*

Can ***interpreting this mapping*** inform emulators of more efficient training set designs?

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

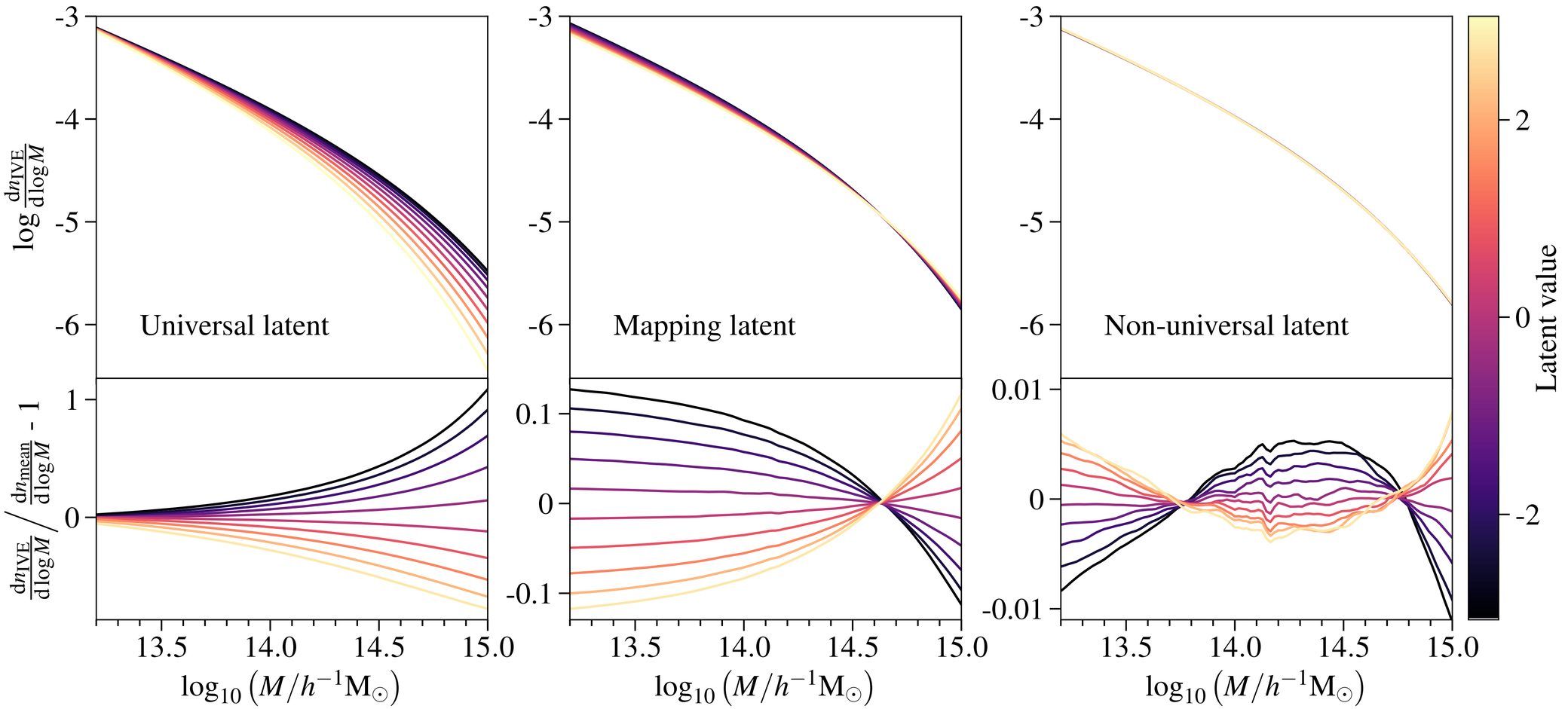


- *How many parameters are necessary and sufficient to describe the HMF?*
- *Latent space isolates universal and non-universal information in the HMF*

Work led by Dr. Lillian Guo (UCL)

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

From 7 cosmological parameters to 3 disentangled latents



Training emulators so that latent space is covered uniformly may result in more efficient and generalizable models from fewer training samples

Conclusions

- Machine learning in Astro has been successful at ***accelerating*** physical models & ***simulation-based inference***
- Interpretability key in bridging gap between ML enthusiasts and ML skeptics
- ***Explainable AI*** models promising for machine-assisted scientific discovery

Thank you to the organizers for a wonderful conference!

MAX PLANCK INSTITUTE
FOR ASTROPHYSICS

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

New group at the University of Hamburg starting Nov 2024

Hiring PhD students and postdocs in (explainable) AI x cosmology!

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

Extra slides

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:

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

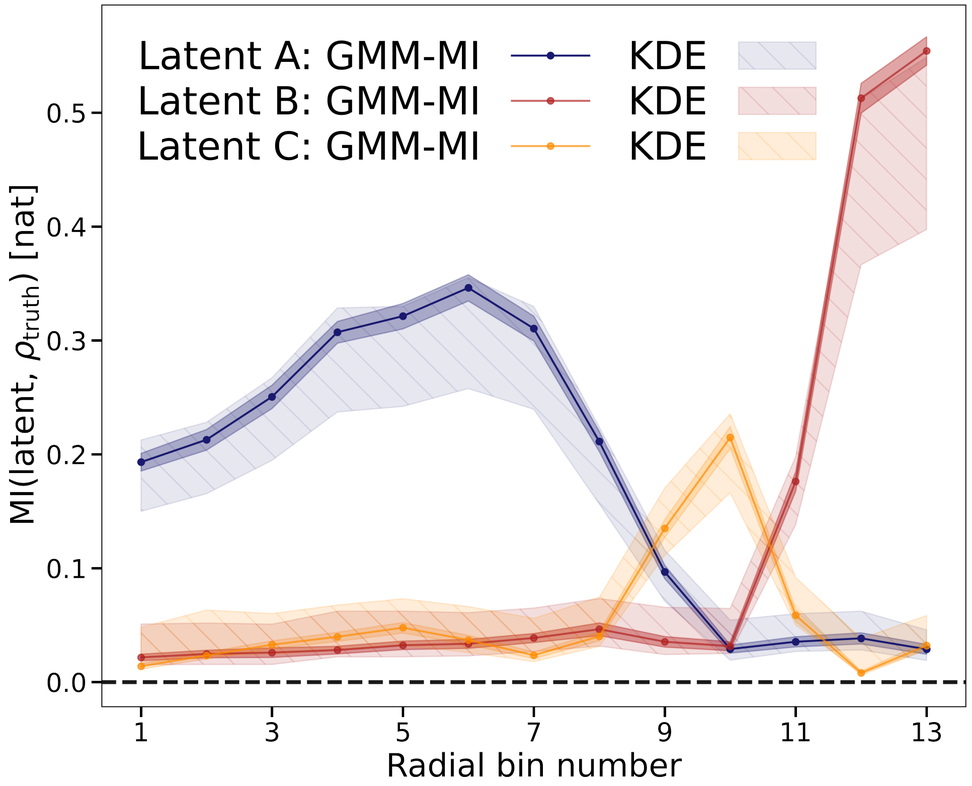
Prior:

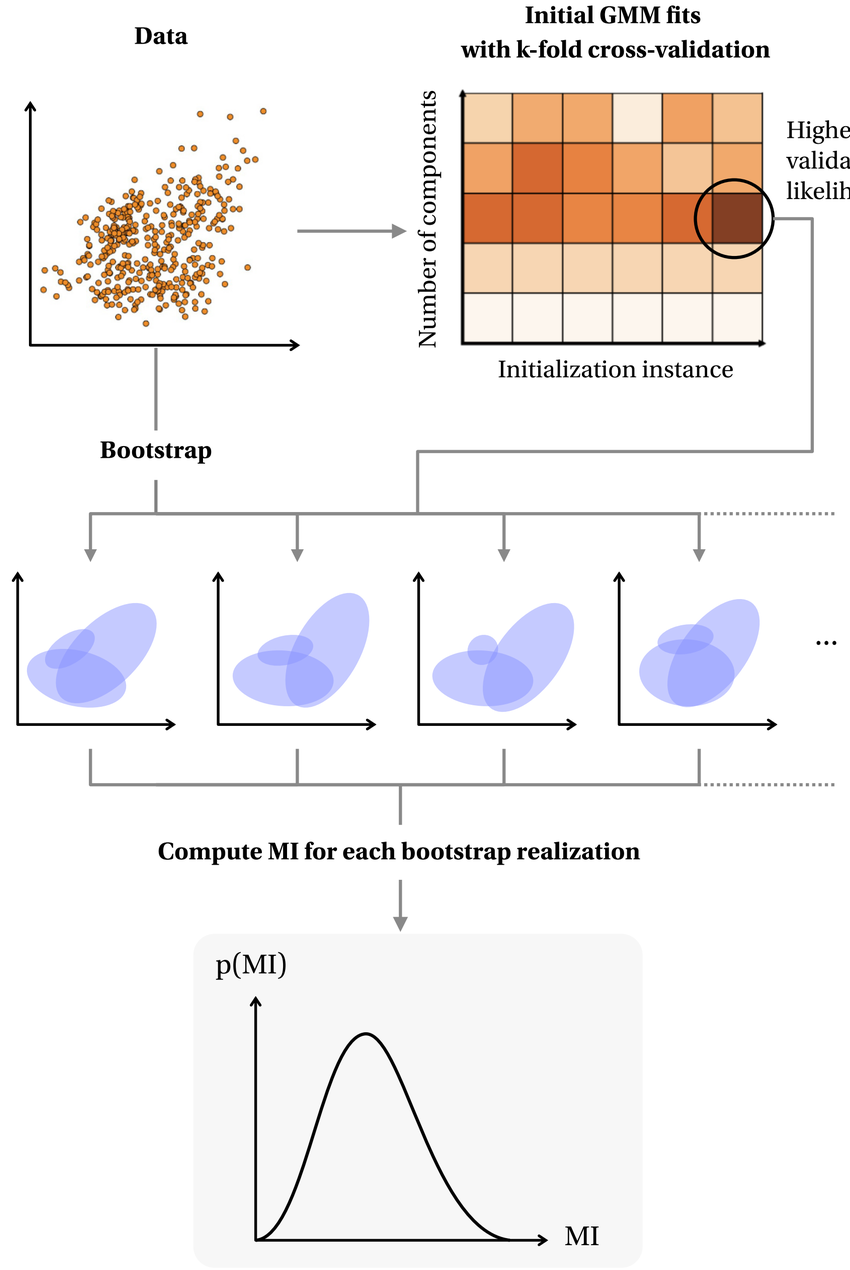
$$q(\mathbf{z}) = \prod_{i=1}^{L}\mathcal{N}(0,1)$$

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

Mutual information

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





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

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