Dynamical mass inference of galaxy clusters with machine learning

Doogesh Kodi Ramanah

DARK Research Fellow, Niels Bohr Institute

arXiv: 2003.05951 (Neural flow mass estimator)
arXiv: 2009.03340 (Simulation-based inference)

In collaboration with **Radek Wojtak**, **Nikki Arendse**, Zoe Ansari, Christa Gall, Jens Hjorth

Galaxy clusters

1. Introduction

Galaxy clusters are the most massive gravitationally bound structures in the universe. Clusters are complex, dark-matter-dominated systems of mass $\gtrsim 10^{14} h^{-1} M_{\odot}$. Galaxy clusters

Galaxy clusters

1. Introduction

Galaxy clusters are the most massive gravitationally bound structures in the universe. Clusters are complex, dark-matter dominated systems of mass $\gtrsim 10^{14} h^{-1}$

1. Introduction

Galaxy clusters are the most massive bound systems in the universe and are uniquely powerful cosmological probes. Cluster dynamical parameters, such as line-of-sight velocity

1. INTRODUCTION

Galaxy clusters are the most massive gravitationally bound systems in the universe. They are dark matter dominated, and have halos of mass $\gtrsim 10^{14}\, M_{\odot}\, h^{-1}$. The majority of multiple-

1. INTRODUCTION

Galaxy clusters are the most massive gravitationally bound systems in the universe, consisting of hundreds of luminous galaxies and hot gas embedded in dense

1 INTRODUCTION

Galaxy clusters are the most massive bound structures in the Universe and are often complex. This is in part due to their tendencies to be triaxial, to lie in an anisotropic environmen

1. Introduction

Clusters and rich groups of galaxies are the most massive gravitationally bound objects in the Universe. Many global properties such as luminosities, X-ray temperatures or velocity disper-

1 INTRODUCTION

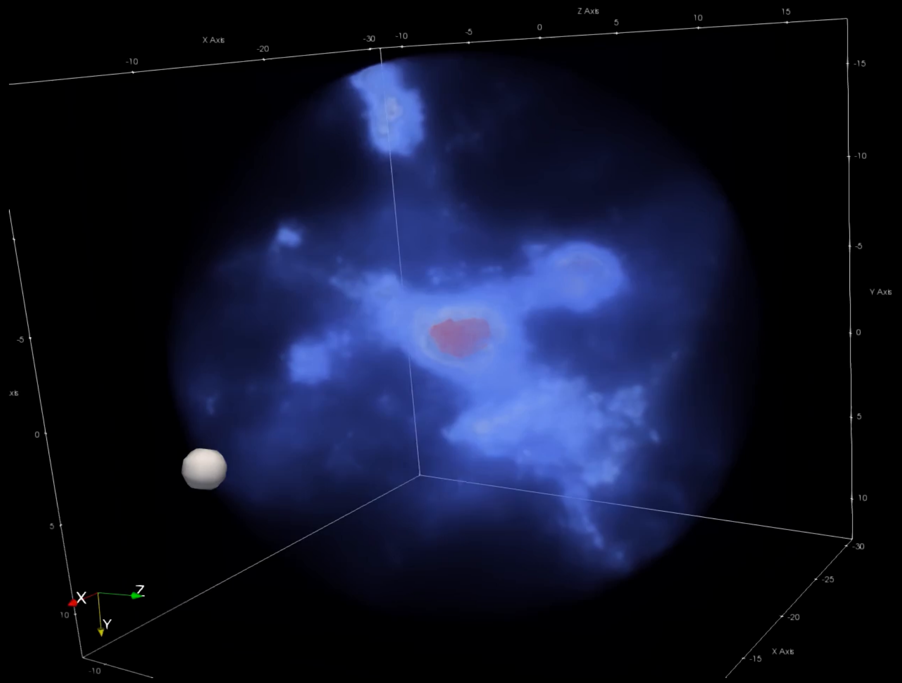
Galaxy clusters are massive, rare objects which form from high peaks in the underlying density field and whose population characteristics are sensitive to the expansion history of the Universe and

Galaxy clusters

1. Introduction

Galaxy clusters are the most massive gravitationally bound structures in the universe. Clusters are complex, dark-matter-dominated systems of mass $\gtrsim 10^{14} h^{-1} M_{\odot}$. Galaxy clusters

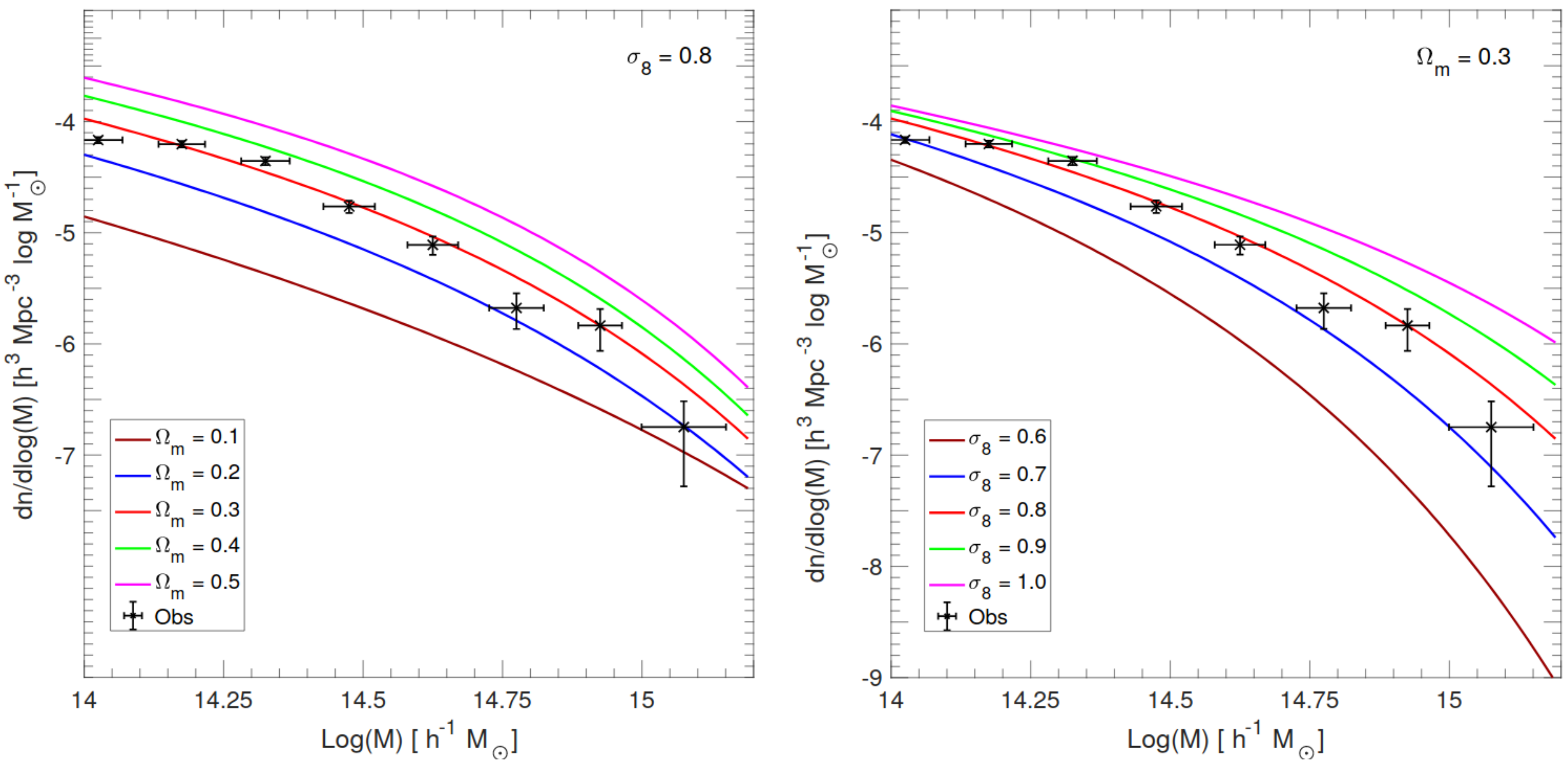
Virgo Cluster
(Image credit: NASA/ESA)



BORG 2M++ reconstruction
(Movie by Guilhem Lavaux)

Importance for cosmology

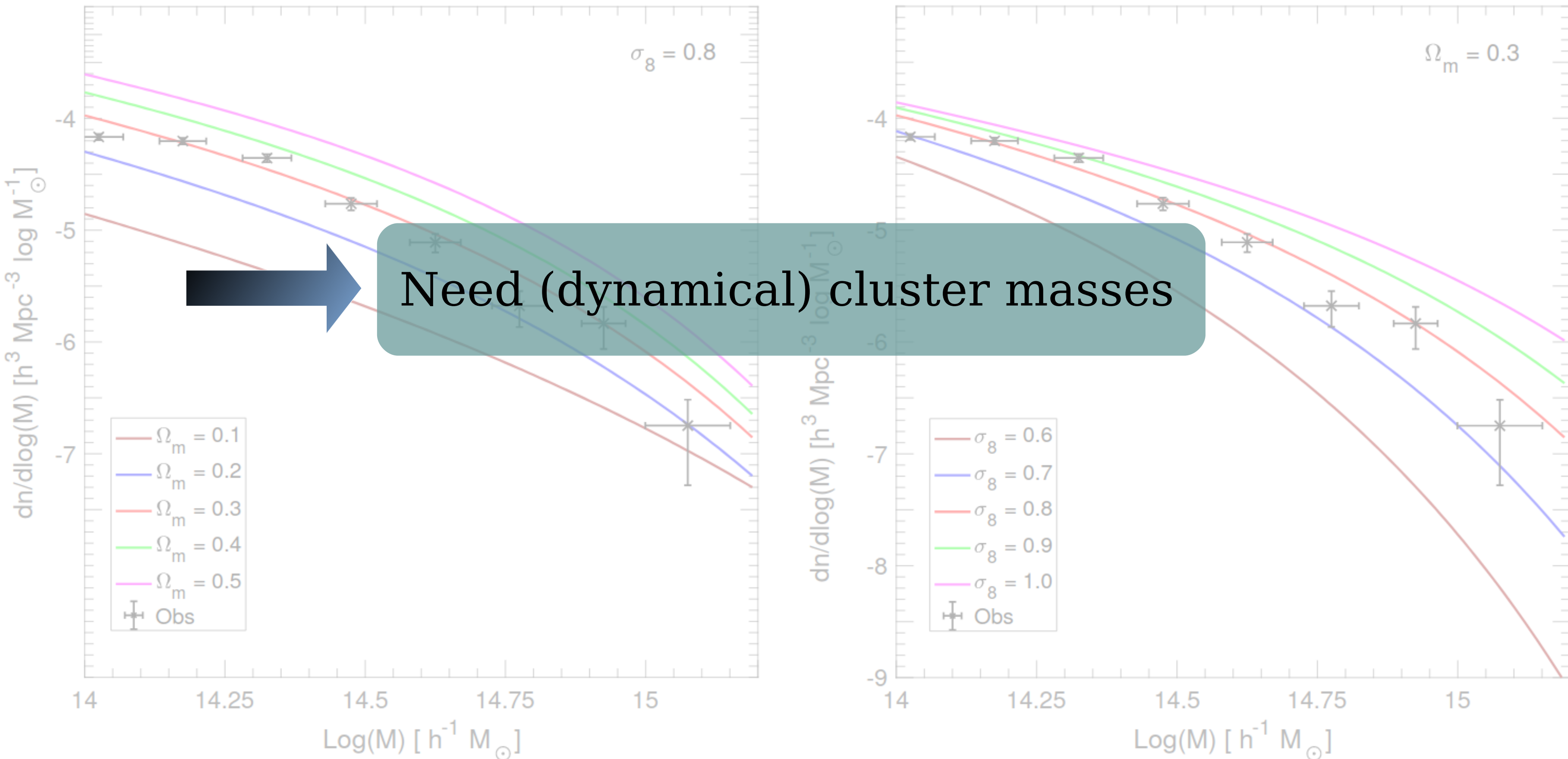
- Cosmological information encoded in abundance of galaxy clusters
- **Cluster mass function (CMF)** – variation of number density of clusters with mass
- CMF particularly sensitive to matter density and amplitude of fluctuations, $\{\Omega_{\rm m}, \sigma_8\}$



Abdullah+ 2020 (ApJ) - arXiv: 2002.11907

Importance for cosmology

- Cosmological info encoded in abundance of galaxy clusters
- **Cluster mass function (CMF)** – variation of number density of clusters with mass
- CMF particularly sensitive to matter density and amplitude of fluctuations, $\{\Omega_\mathrm{m}, \sigma_8\}$



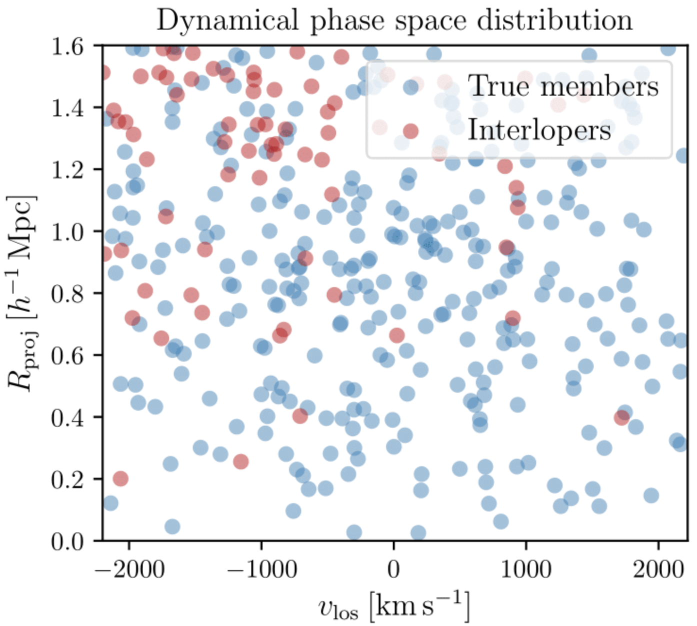
Abdullah+ 2020 (ApJ) - arXiv: 2002.11907

Neural flow mass estimator

Observables & challenges

- **What are our observables?**
  Projected radial distance from cluster centre → R_{proj}
  Galaxy line-of-sight velocity → v_{los}

- **Series of classical methods:**
  1) ***M*-*σ*$_v$** scaling relation
  2) Virial mass estimator
  3) Jeans analysis
  4) Distribution function



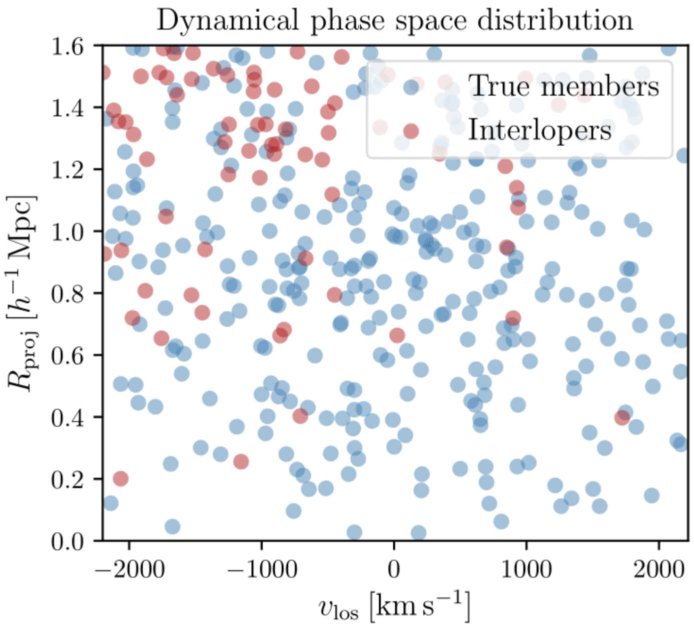
Observables & challenges

- **What are our observables?**
  Projected radial distance from cluster centre → R_{proj}
  Galaxy line-of-sight velocity → v_{los}

- **Series of classical methods:**
  1) ***M*-*σ*$_v$** scaling relation
  2) Virial mass estimator
  3) Jeans analysis
  4) Distribution function

- **Physical effects breaking idealized assumptions:**
  1) Dynamical substructure
  2) Cluster triaxiality
  3) Halo environment
  4) Cluster mergers

- **Selection effects:**
  1) Incomplete cluster observations
  2) Interlopers (non-members)



Observables & challenges

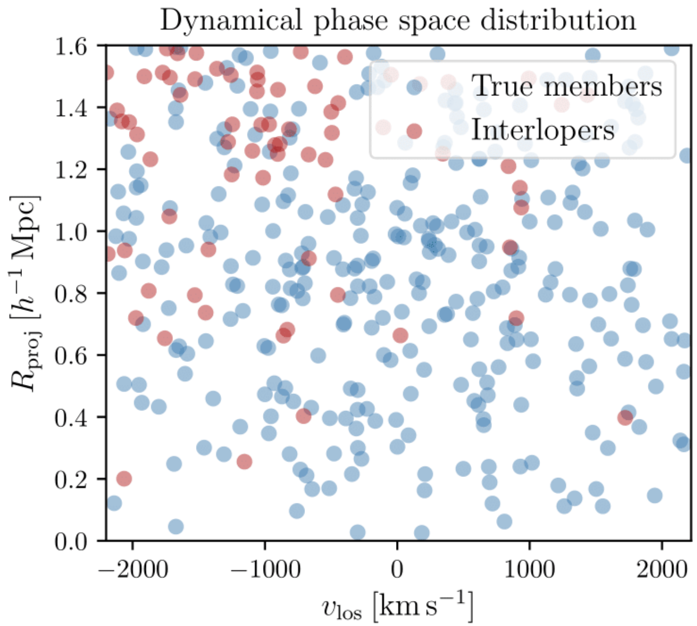
- **What are our observables?**
  Projected radial distance from cluster centre → R_{proj}
  Galaxy line-of-sight velocity → v_{los}

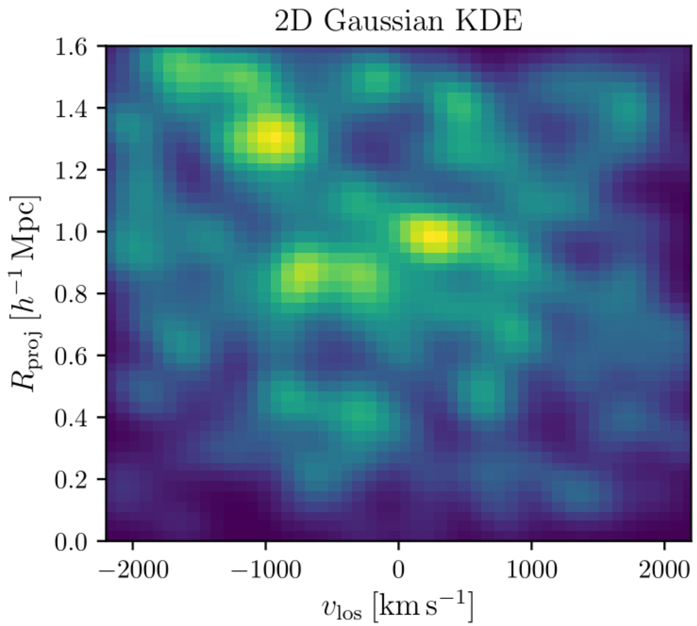
- **Series of classical methods:**
  1) ***M*-*σ*$_v$** scaling relation
  2) Virial mass estimator
  3) Jeans analysis
  4) Distribution function

- **Physical effects breaking idealized assumptions:**
  1) Dynamical substructure
  2) Cluster triaxiality
  3) Halo environment
  4) Cluster mergers

- **Selection effects:**
  1) Incomplete cluster observations
  2) Interlopers (non-members)

- **Inputs for neural network:**
  (Normalized) Gaussian KDE → smooth phase-space mapping





Mock cluster catalogue

Ho+ 2019 (ApJ) - arXiv: 1902.05950

- **MDPL2** - MultiDark (*N*-body) simulation (**`GADGET2`**)
- Halos → clusters, subhalos → galaxies (**`ROCKSTAR + UNIVERSEMACHINE`**)

Simulation box of 1 h^{-1} Gpc
Mass resolution of $1.51 \times 10^{9}\ h^{-1}\ M_{\odot}$

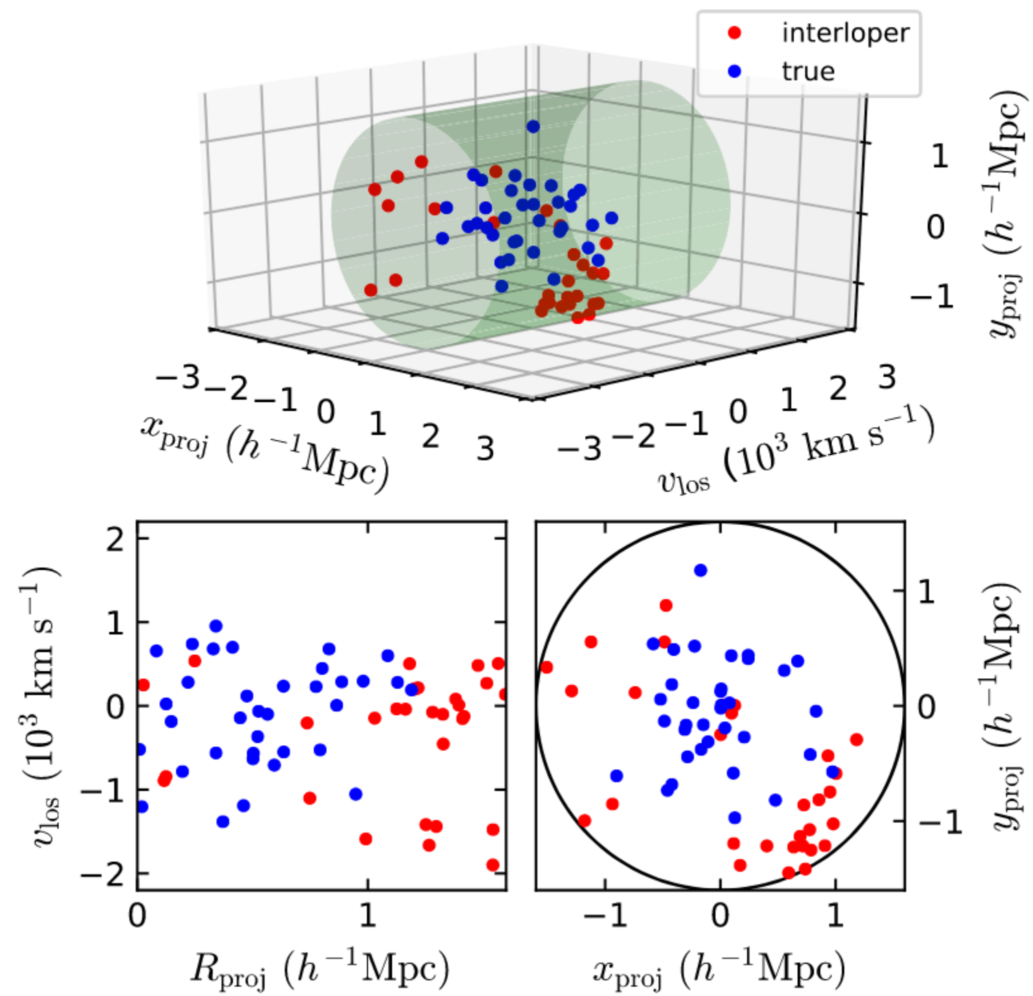
Matthew Ho

CNN → cluster mass (point estimates)

Mock cluster catalogue

Ho+ 2019 (ApJ) - arXiv: 1902.05950

- **MDPL2** - MultiDark (*N*-body) simulation (**`GADGET2`**)
- Halos → clusters, subhalos → galaxies (**`ROCKSTAR + UNIVERSEMACHINE`**)
- **Pure** v/s **contaminated** catalogue (interlopers)



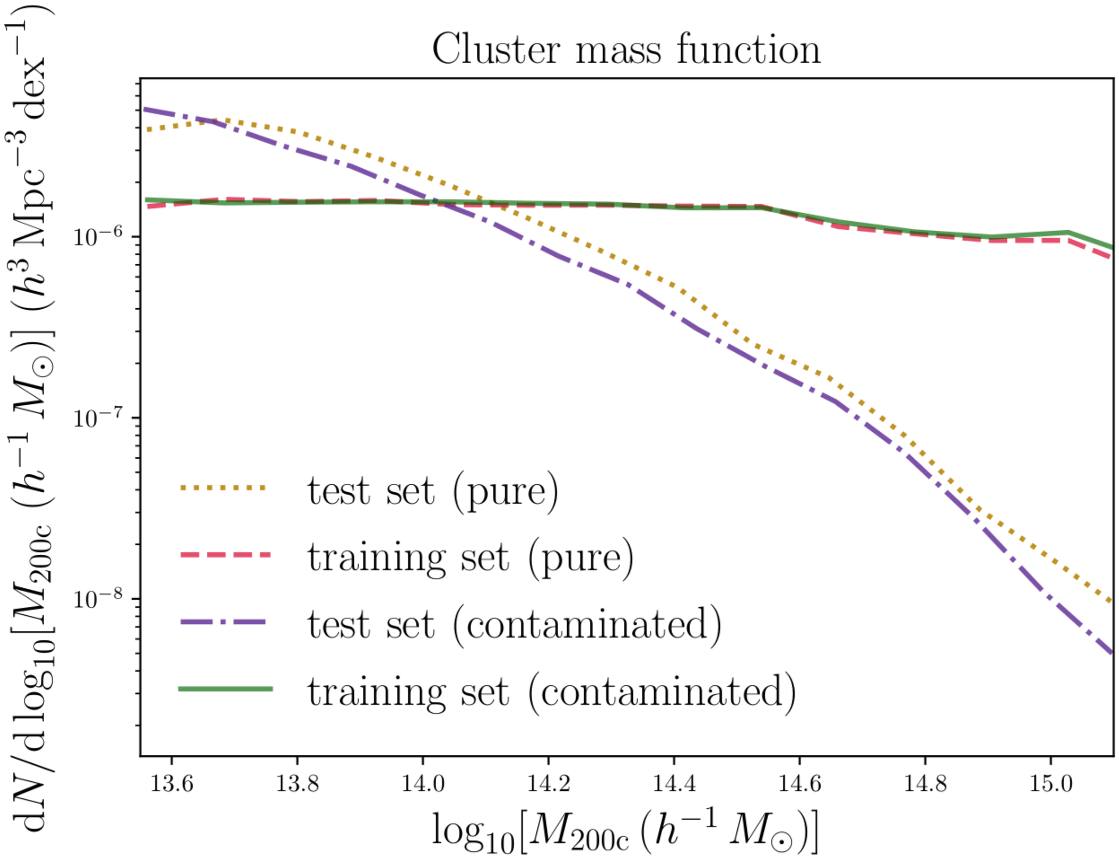
Matthew Ho

CNN → cluster mass
(point estimates)

Mock cluster catalogue

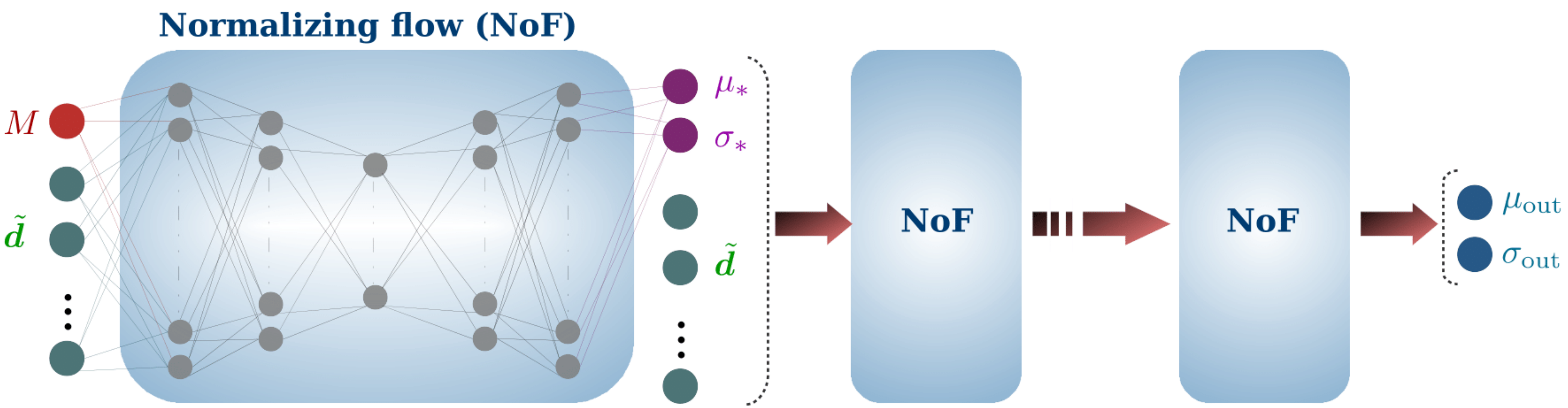
Ho+ 2019 (ApJ) - arXiv: 1902.05950

- **MDPL2** - MultiDark (N-body) simulation (`GADGET2`)
- Halos → clusters, subhalos → galaxies (`ROCKSTAR + UNIVERSEMACHINE`)
- **Pure** v/s **contaminated** catalogue (interlopers)
- **Flat mass function** for training so as not to encode cosmological info



Neural flow schematic

- **Normalizing flows** (neural density estimator)
- Model conditional density distribution → $\mathcal{P}(M|\tilde{\boldsymbol{d}})$, where $\tilde{\boldsymbol{d}} \equiv \{\boldsymbol{R}_{\rm proj}, \boldsymbol{v}_{\rm los}\}$
- In essence, neural network learns transformation from base (e.g. Gaussian) distribution
- Network trained using pairs of $\{M, \tilde{\boldsymbol{d}}\}$ - minimize negative log likelihood
- Train on pure & contaminated catalogues separately

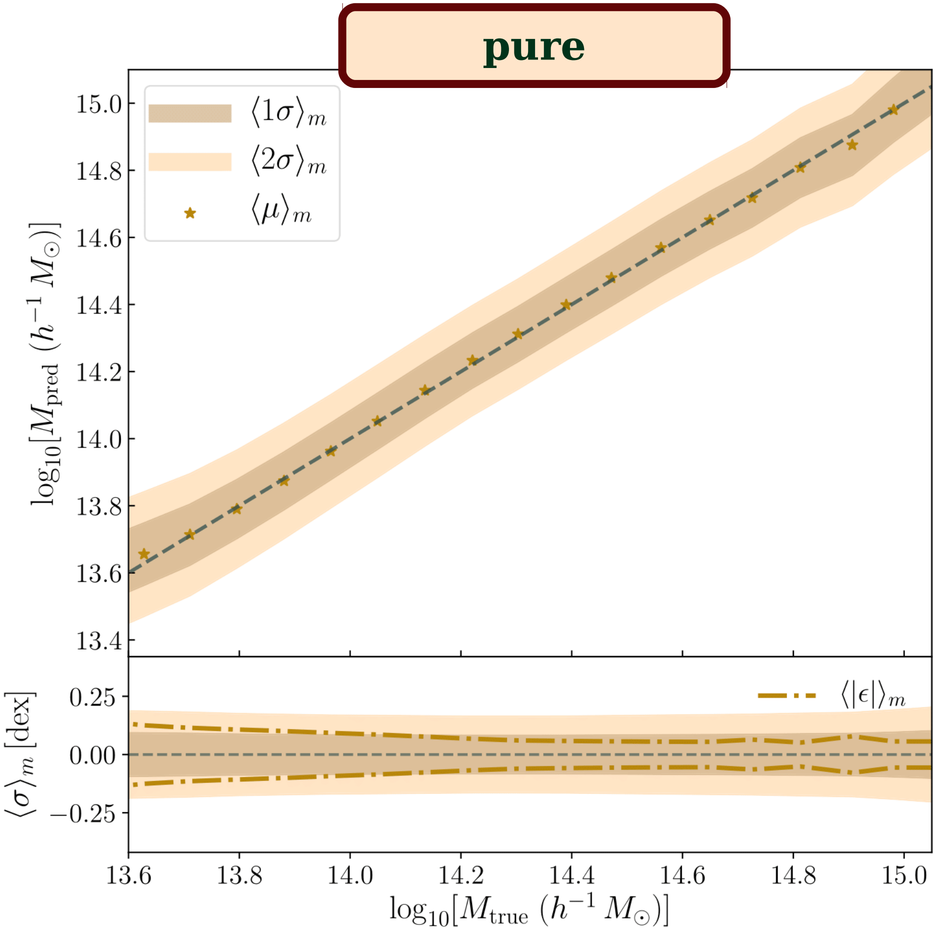


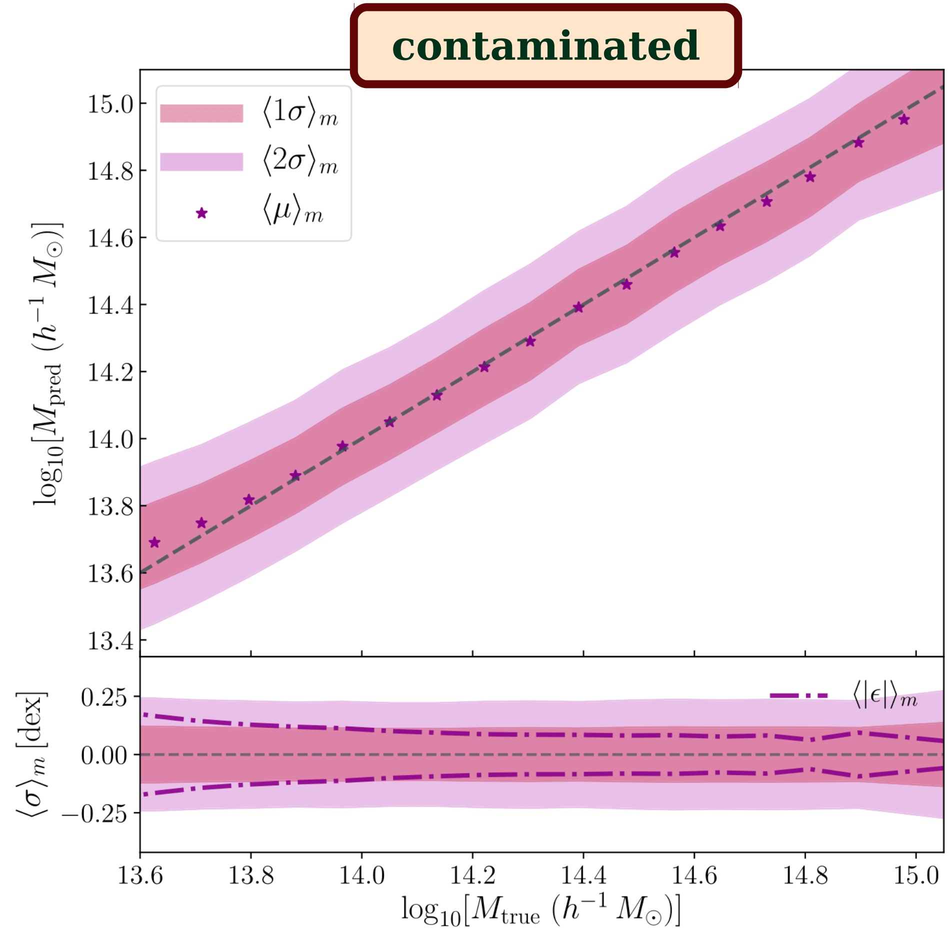
Performance validation

- Usual truth v/s predictions plot
- Larger uncertainties (& residual scatter ϵ) for contaminated set

$$\epsilon \equiv \log_{10}(M_{\rm true}/M_{\rm pred})$$

- Performance on contaminated set ~4 times improvement over classical (***M*-$\boldsymbol{\sigma_v}$**) relation





Precision of cluster mass estimators

- Galaxy cluster people usually express total scatter as:

$$\boxed{\sigma^2 = \sigma_N^2 (N_{\rm members}/100)^{-1} + \sigma_0^2}$$

Richness-dependent component
("statistical error")

Richness-independent component
("systematic error")

- If negligible systematic errors, then:
  σ = statistical error given by Poisson noise with amplitude σ_N

Precision of cluster mass estimators

Galaxy Cluster Mass Comparison Project
Wojtak+ 2018 (MNRAS) - arXiv: 1806.03199

- Galaxy cluster people usually express total scatter as:

$$\boxed{\sigma^2 = \sigma_N^2 (N_{\mathrm{members}}/100)^{-1} + \sigma_0^2}$$

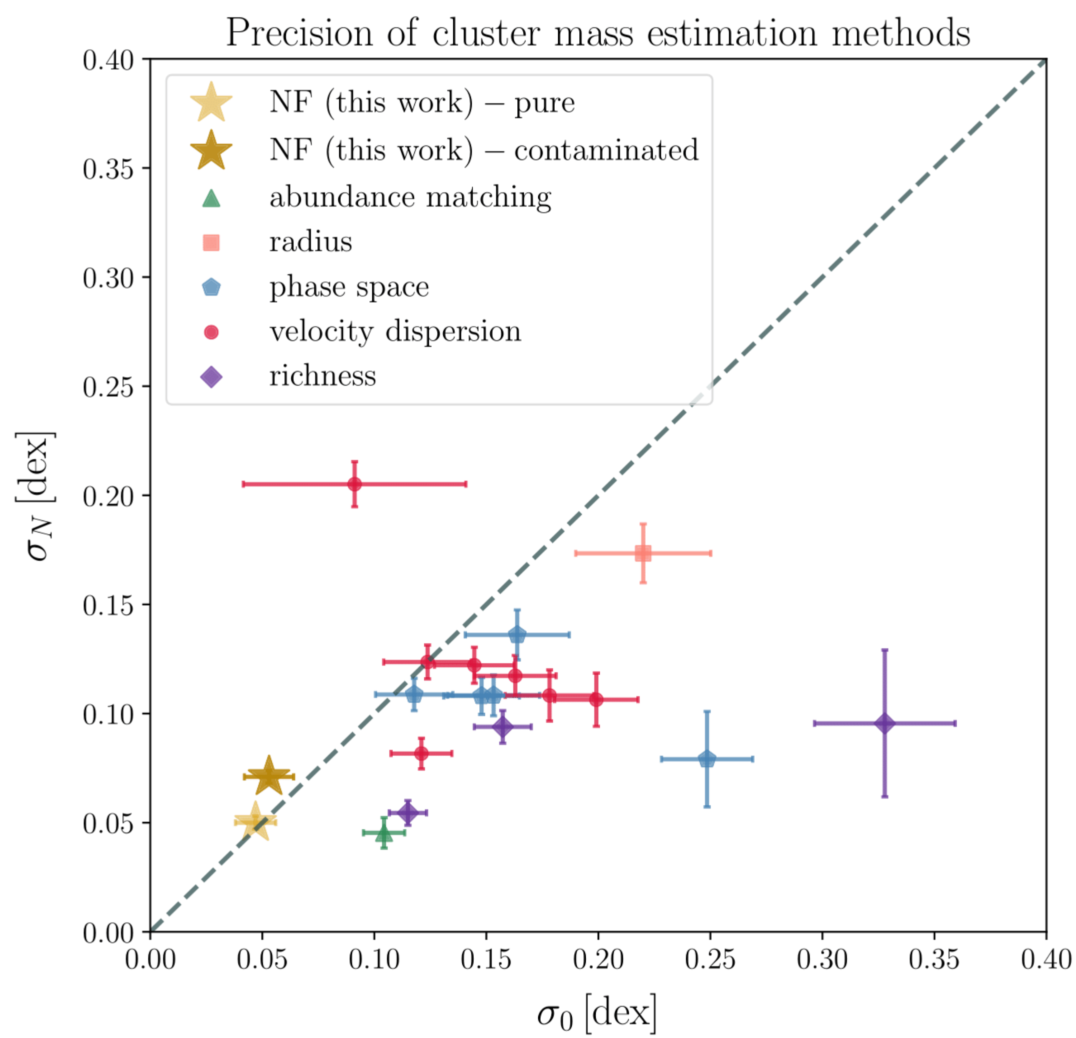
Richness-dependent component (“statistical error”)

Richness-independent component (“systematic error”)

- If negligible systematic errors, then:
  σ = statistical error given by Poisson noise with amplitude σ_N

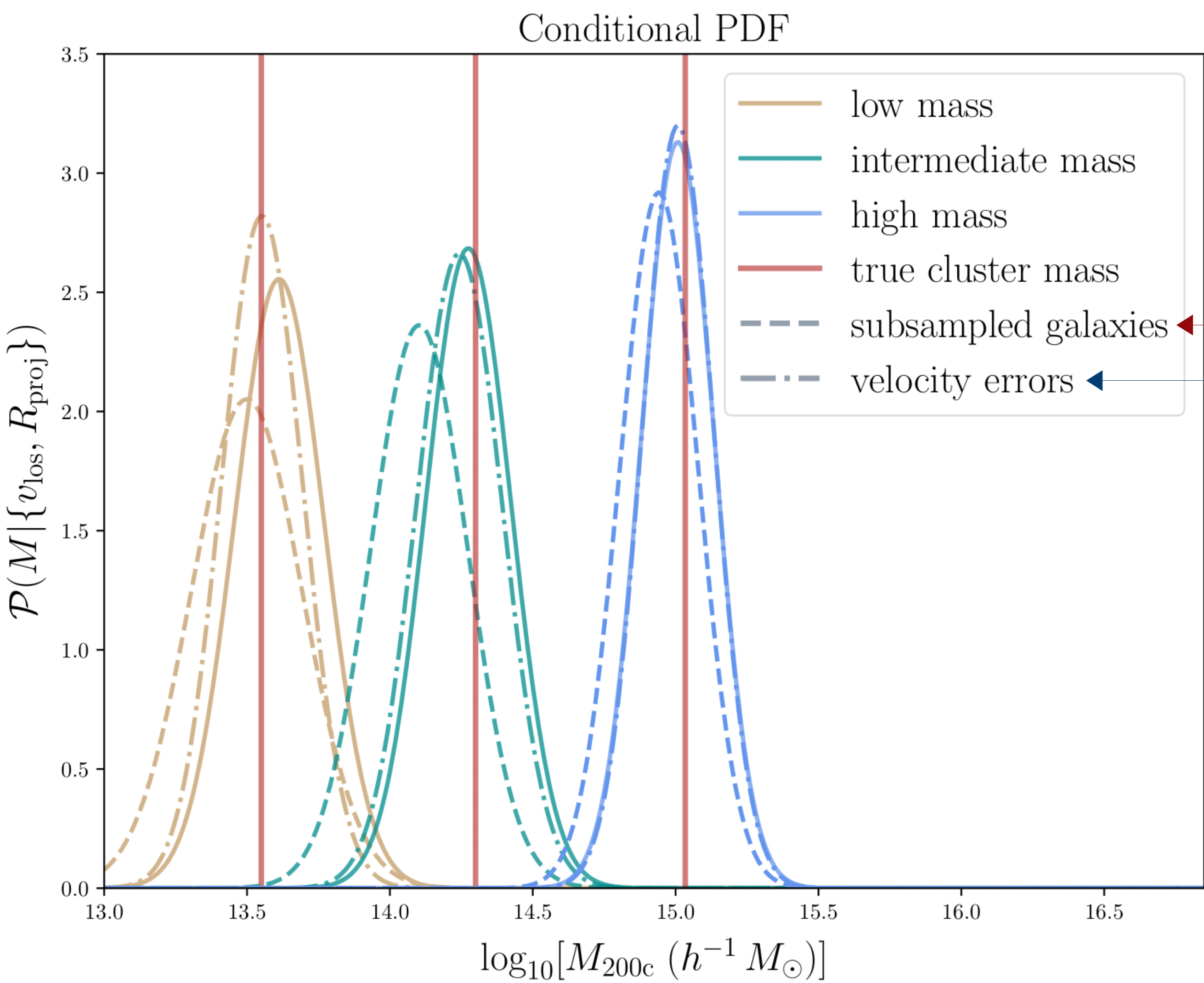
- NF mass estimator outperforms classical methods

- Systematic error due to interlopers scales with cluster richness



Robustness tests

- Verify robustness to galaxy selection effects & typical velocity errors

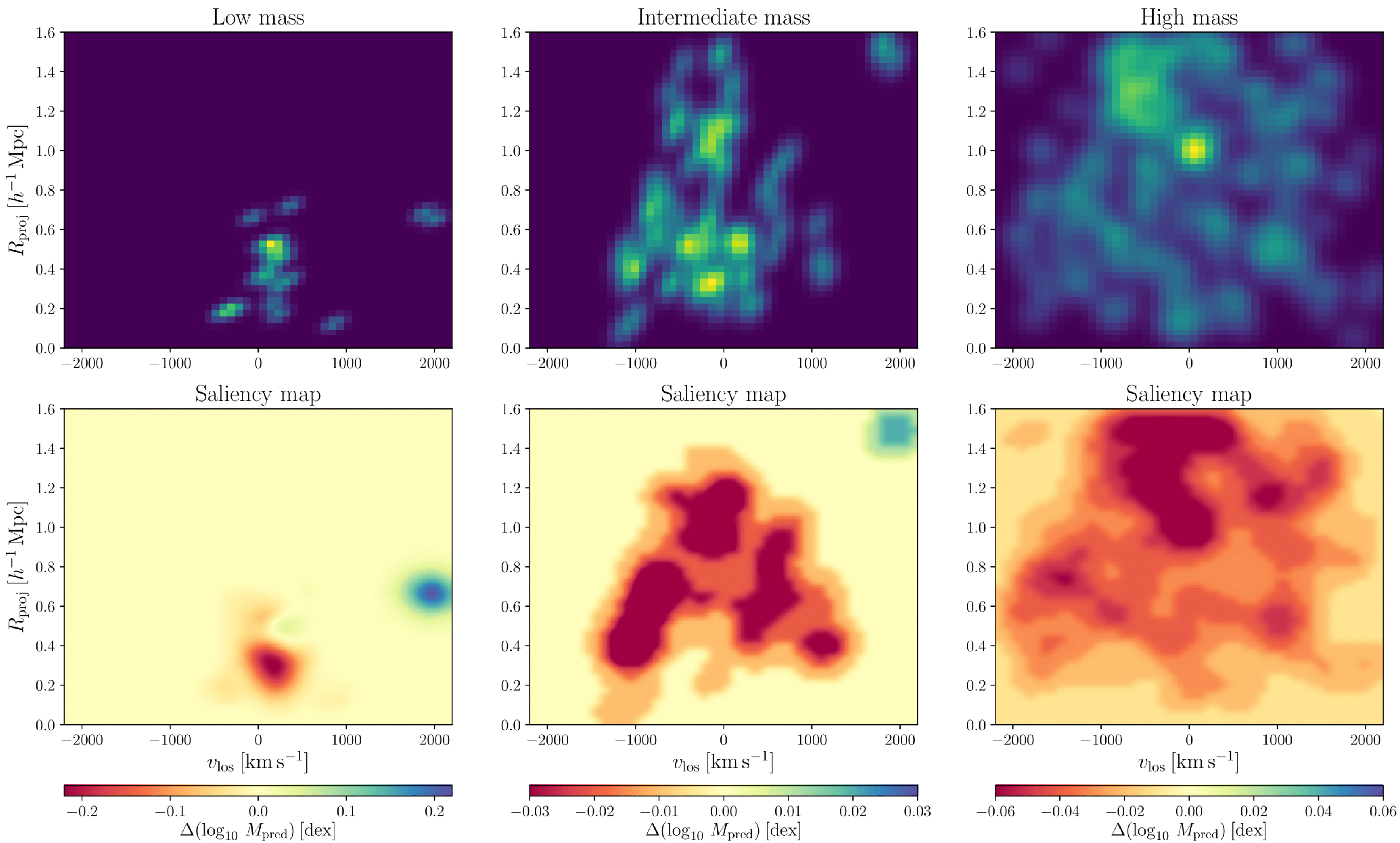


33% fewer galaxies

Gaussian scatter ~100 km/s

Saliency maps

- Topographical representation of the informative structures in input 2D phase space

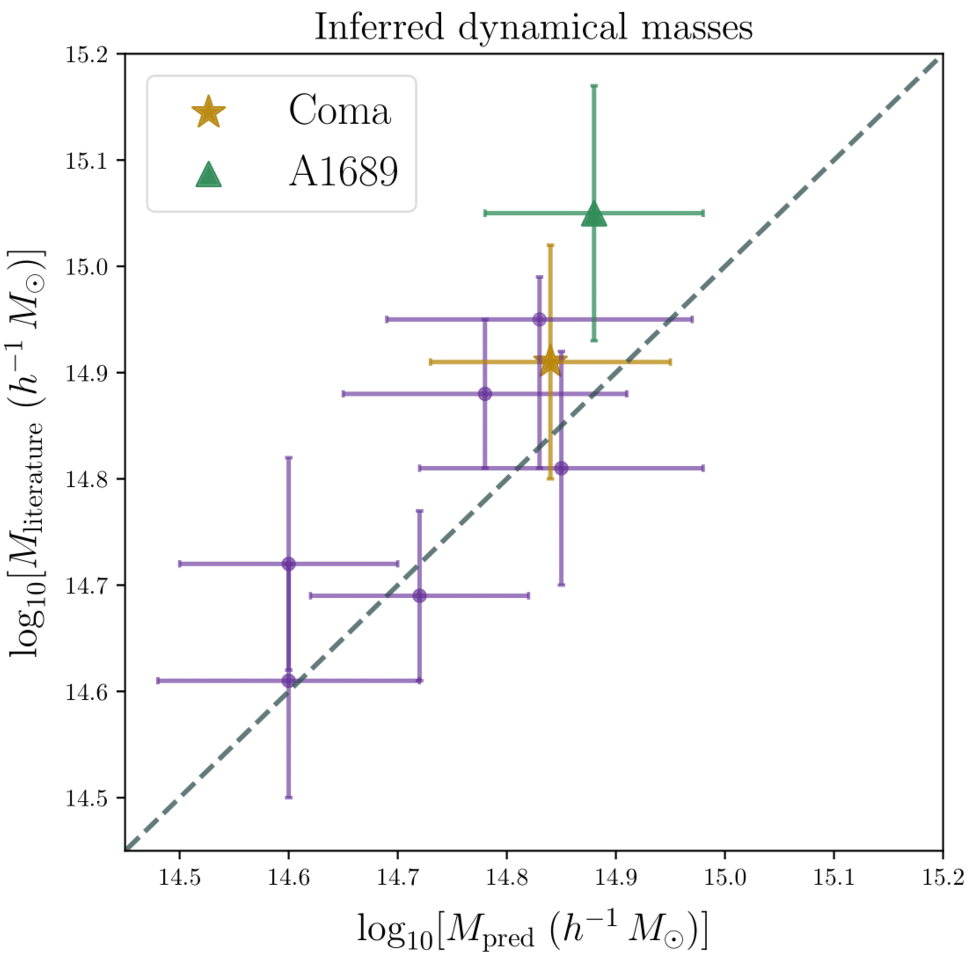


Real world applications

Infer masses of some well-known clusters

Galaxy cluster	NF dynamical mass	Literature value
Coma	14.84 ± 0.11	14.91 ± 0.11[1]
A1689	14.88 ± 0.10	15.05 ± 0.12[2]
A85	14.78 ± 0.13	14.88 ± 0.07[3]
A119	14.60 ± 0.12	14.61 ± 0.11[4]
A576	14.72 ± 0.10	14.69 ± 0.08[3]
A1651	14.85 ± 0.13	14.81 ± 0.11[3]
A2142	14.83 ± 0.14	$14.95^{+0.04}_{-0.14}$[5]
A2670	14.60 ± 0.10	14.72 ± 0.10[4]

[1] Łokas & Mamon (2003)
[2] Lemze et al. (2009)
[3] Wojtak & Łokas (2007)
[4] Abdullah et al. (2020)
[5] Munari et al. (2014)

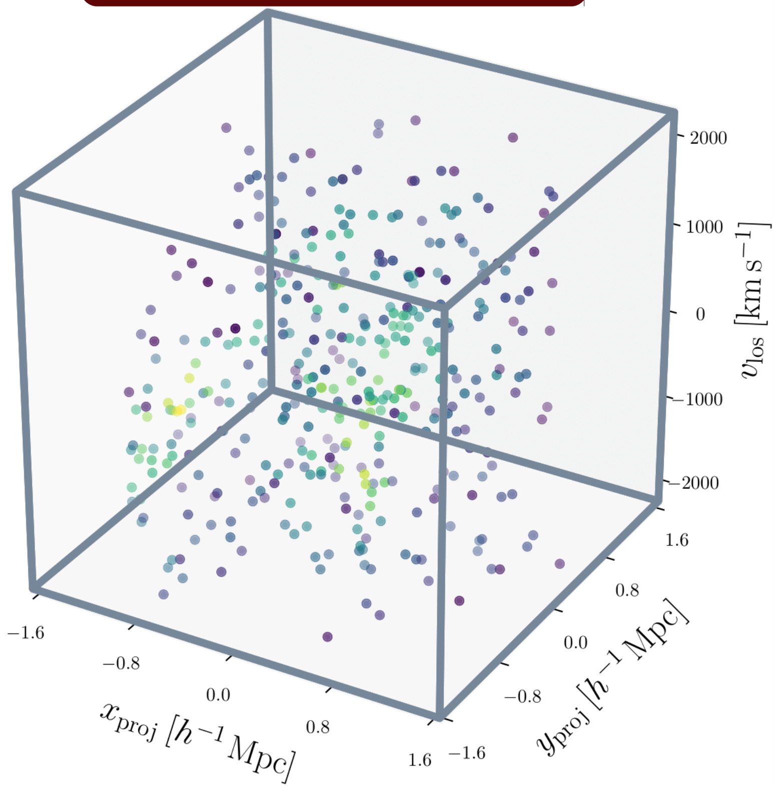


Simulation-based inference

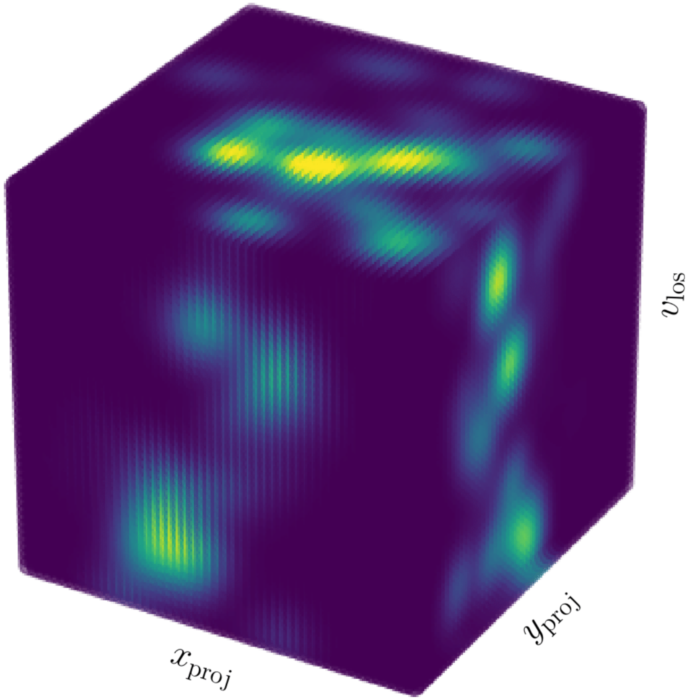
3D phase-space distribution

- **Galaxy cluster observables:**
  Galaxy positions projected on to plane of sky → $(x_{\rm proj}, y_{\rm proj})$
  Line-of-sight velocities of galaxy members → $v_{\rm los}$
- **Motivation for 3D** → render model more sensitive to interlopers
- **Mock SDSS cluster catalogue:** MDPL2 + semi-analytical model of galaxy formation (SAG)

3D galaxy distribution



Normalized 3D Gaussian KDE



Simulation-based inference

Why neural networks don't work and how to use them

Neural networks as universal model approximators

We can think of a neural network, $\mathbb{NN}(\boldsymbol{w}, \boldsymbol{\alpha}) : \mathbf{d} \to \boldsymbol{\tau}$, as an approximation of a model, $\mathcal{M} : \mathbf{d} \to \mathbf{t}$, where $\mathbf{d}$ is some input data to the network and the output of the network is $\boldsymbol{\tau}$ which is an estimate of some target, $\mathbf{t}$, associated with the data. The neural network itself is a function of some trainable parameters called weights, $\boldsymbol{w}$, and some hyperparameters, $\boldsymbol{\alpha}$, which encompass the architecture of the network, the initial values of the weights, the form of activation functions, the choice of cost function, etc.

Likelihood of obtaining targets given a network

In a traditional sense, the training of a neural network is equivalent to minimising a *cost* or *loss* function, $\Lambda(\mathbf{t}, \boldsymbol{\tau})$, with respect to the weights of the network, $\boldsymbol{w}$ (and hyperparameters, $\boldsymbol{\alpha}$) given a set of pairs of data and targets for training and validation, $\{\mathbf{d}_i^{\rm train}, \mathbf{t}_i^{\rm train} | i \in [1, n_{\rm train}]\}$ and $\{\mathbf{d}_i^{\rm val}, \mathbf{t}_i^{\rm val} | i \in [1, n_{\rm val}]\}$. The cost function, $\Lambda(\mathbf{t}, \boldsymbol{\tau})$, measures how close the outputs of a fixed network, $\mathbb{NN}(\boldsymbol{w}^*, \boldsymbol{\alpha}^*) : \mathbf{d} \to \boldsymbol{\tau}$, are to some target, $\mathbf{t}$, given a data-target pair, $\{\mathbf{d}, \mathbf{t}\}$, at some fixed network parameters and hyperparameters, $\boldsymbol{w} = \boldsymbol{w}^*$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$. That is, how likely is it that the output of the network provides the true target for the input data given a chosen set of weights and fixed network hyperparameters, i.e. the cost function is equivalent to the (negative logarithm of the) likelihood function

$$\Lambda(\mathbf{t}, \boldsymbol{t}) \simeq -\ln\mathcal{L}(\mathbf{t}|\mathbf{d}, \boldsymbol{w}^*, \boldsymbol{\alpha}^*).$$

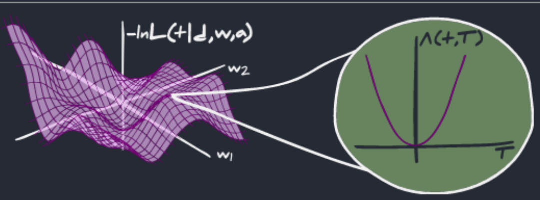


Figure 1 : The likelihood surface, although regular for a given set of network parameters and hyperparameters, is extremely complex, degenerate, and even discrete and non-convex in the directions of the network parameters and hyperparameters.

https://www.aquila-consortium.org/method/machine%20learning/nn.html

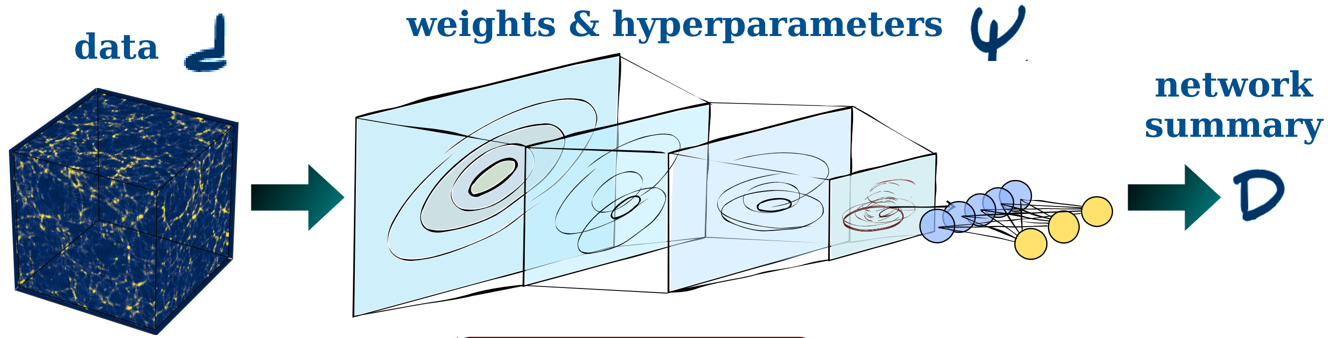
Tom Charnock

Charnock, Lavaux & Wandelt 2018 (PRD)
arXiv: 1802.03537

Data compression via the Information Maximizing Neural Network (IMNN)

Simulation-based inference

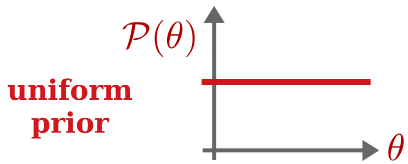
1 Train a network to compress input data to desired summary



$$\boxed{\mathcal{D} = \mathbb{NN}(d, \psi)}$$

Generate (training) data

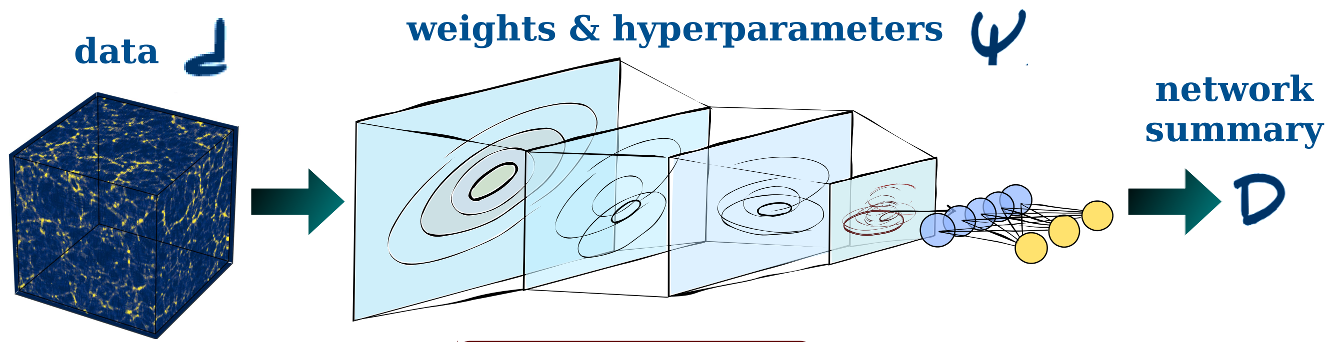
$$d_{\rm train} = \mathcal{F}(\theta), \quad \theta \sim \mathcal{P}(\theta)$$



Credit: Schematics adapted from Tom’s lectures

Simulation-based inference

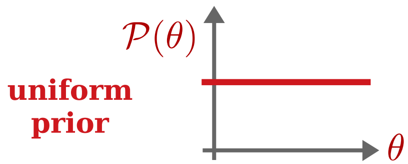
1 Train a network to compress input data to desired summary



$$\mathcal{D} = \mathbb{NN}(d, \psi)$$

Generate (training) data

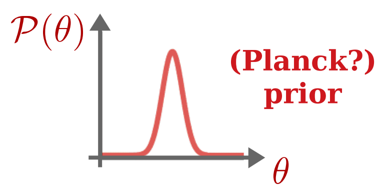
$$d_{\rm train} = \mathcal{F}(\theta), \quad \theta \sim \mathcal{P}(\theta)$$

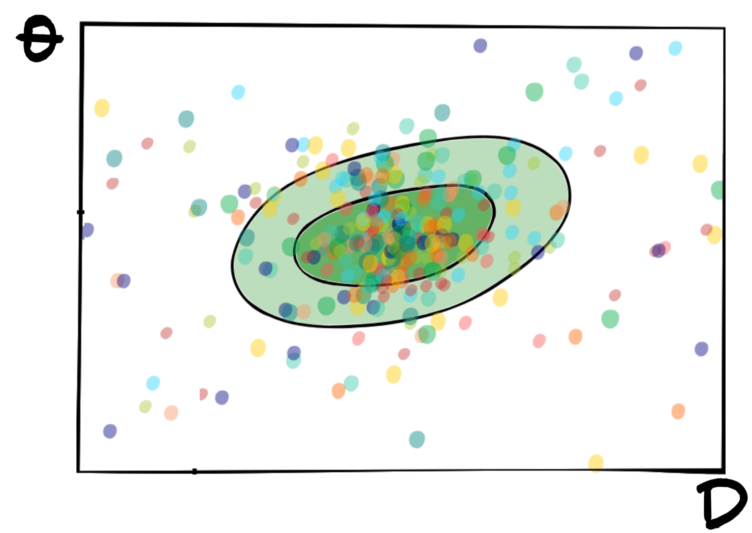


2 Feed a separate test set to obtain $\{\theta, \mathcal{D}\}$ & compute a density estimate of joint PDF

Generate (test) data

$$d_{\rm test} = \mathcal{F}(\theta), \quad \theta \sim \mathcal{P}(\theta)$$

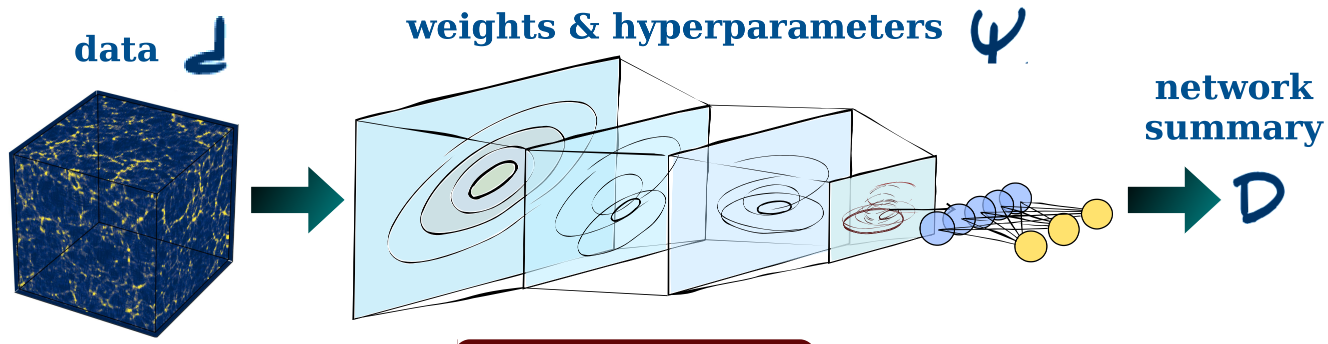




Credit: Schematics adapted from Tom's lectures

Simulation-based inference

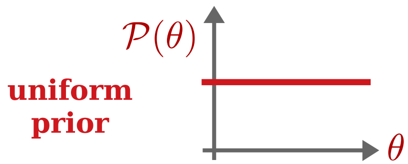
1 Train a network to compress input data to desired summary



$$\boxed{\mathcal{D} = \mathbb{NN}(d, \psi)}$$

Generate (training) data

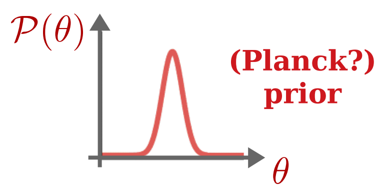
$$d_{\rm train} = \mathcal{F}(\theta), \quad \theta \sim \mathcal{P}(\theta)$$

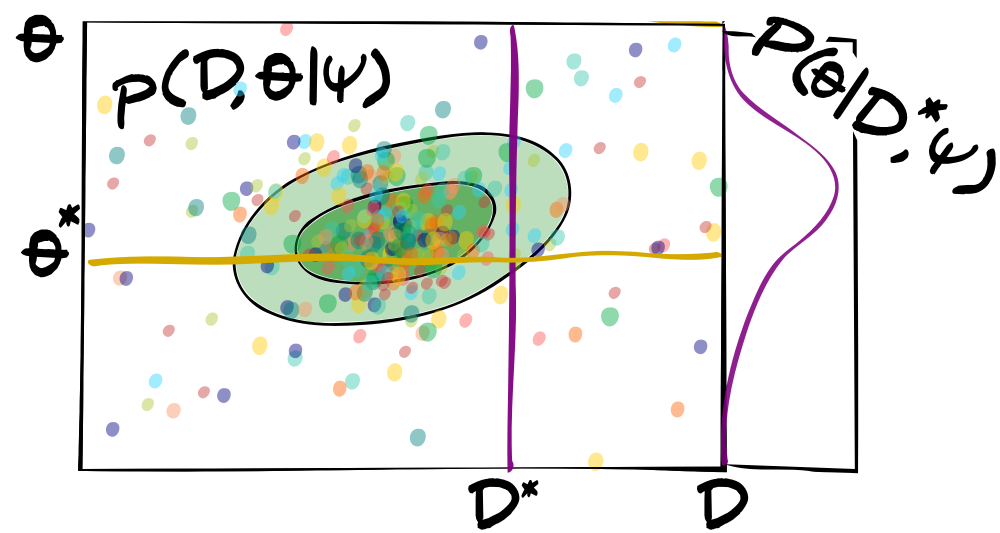


2 Feed a separate test set to obtain $\{\theta, \mathcal{D}\}$ & compute a density estimate of joint PDF

Generate (test) data

$$d_{\rm test} = \mathcal{F}(\theta), \quad \theta \sim \mathcal{P}(\theta)$$





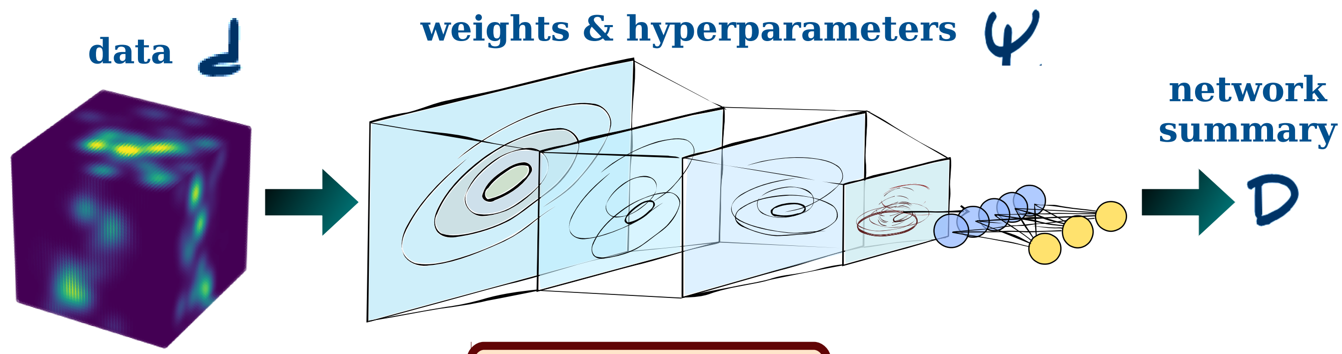
3 Slice through joint PDF $\mathcal{P}(\mathcal{D}, \theta)$, for a given observation d^*, at $\mathcal{D}^* = \mathbb{NN}(d^*)$

Approximate posterior $\boxed{\mathcal{P}(\theta|\mathcal{D}^*) \approx \mathcal{P}(\theta|d^*, \psi)}$

Credit: Schematics adapted from Tom's lectures

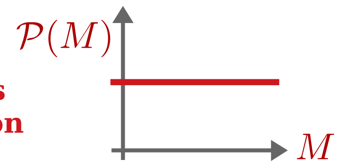
Simulation-based inference

1 Train a network to compress input data to desired summary



Generate (training) data

$$d_{\text{train}} = \mathcal{F}(M), \quad M \sim \mathcal{P}(M)$$

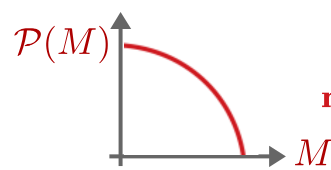


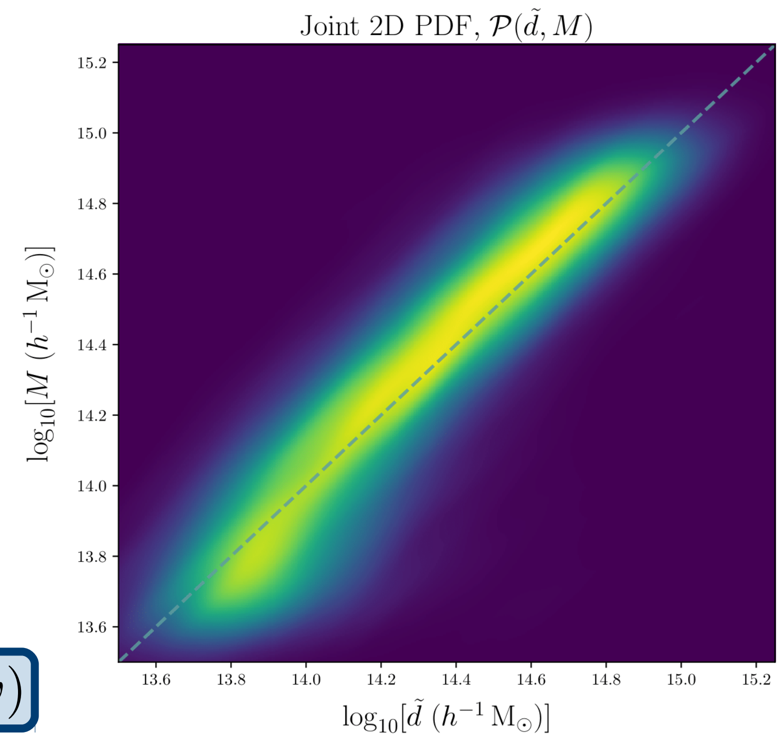
$$\boxed{\mathcal{D} = \mathbb{NN}(d, \psi)}$$

2 Feed a separate test set to obtain $\{M, \mathcal{D}\}$ & compute a density estimate of joint PDF

Generate (test) data

$$d_{\text{test}} = \mathcal{F}(M), \quad M \sim \mathcal{P}(M)$$



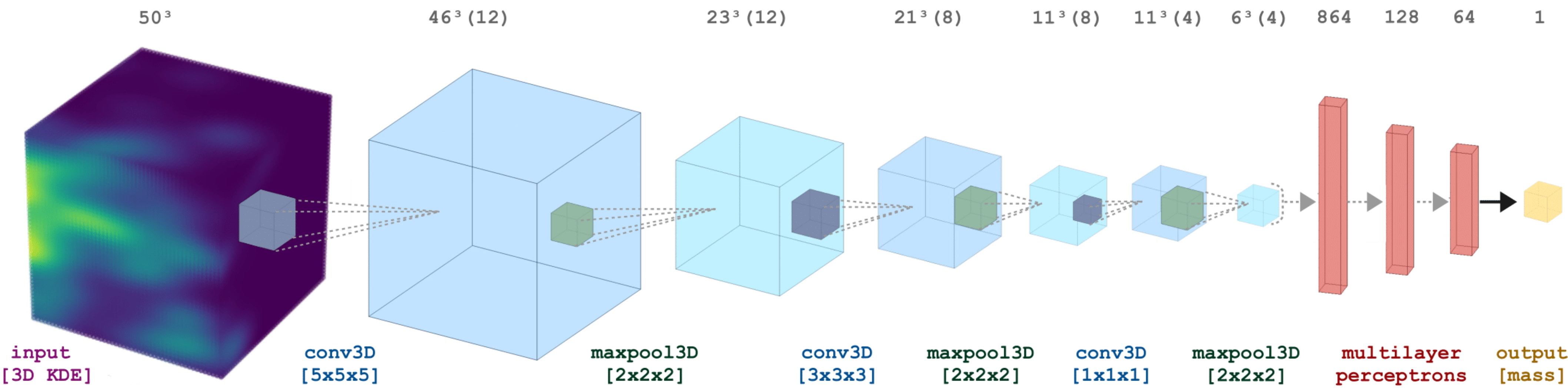


3 Slice through joint PDF $\mathcal{P}(\mathcal{D}, M)$, for a given observation d^*, at $\mathcal{D}^* = \mathbb{NN}(d^*)$

Approximate posterior $\boxed{\mathcal{P}(M|\mathcal{D}^*) \approx \mathcal{P}(M|d^*, \psi)}$

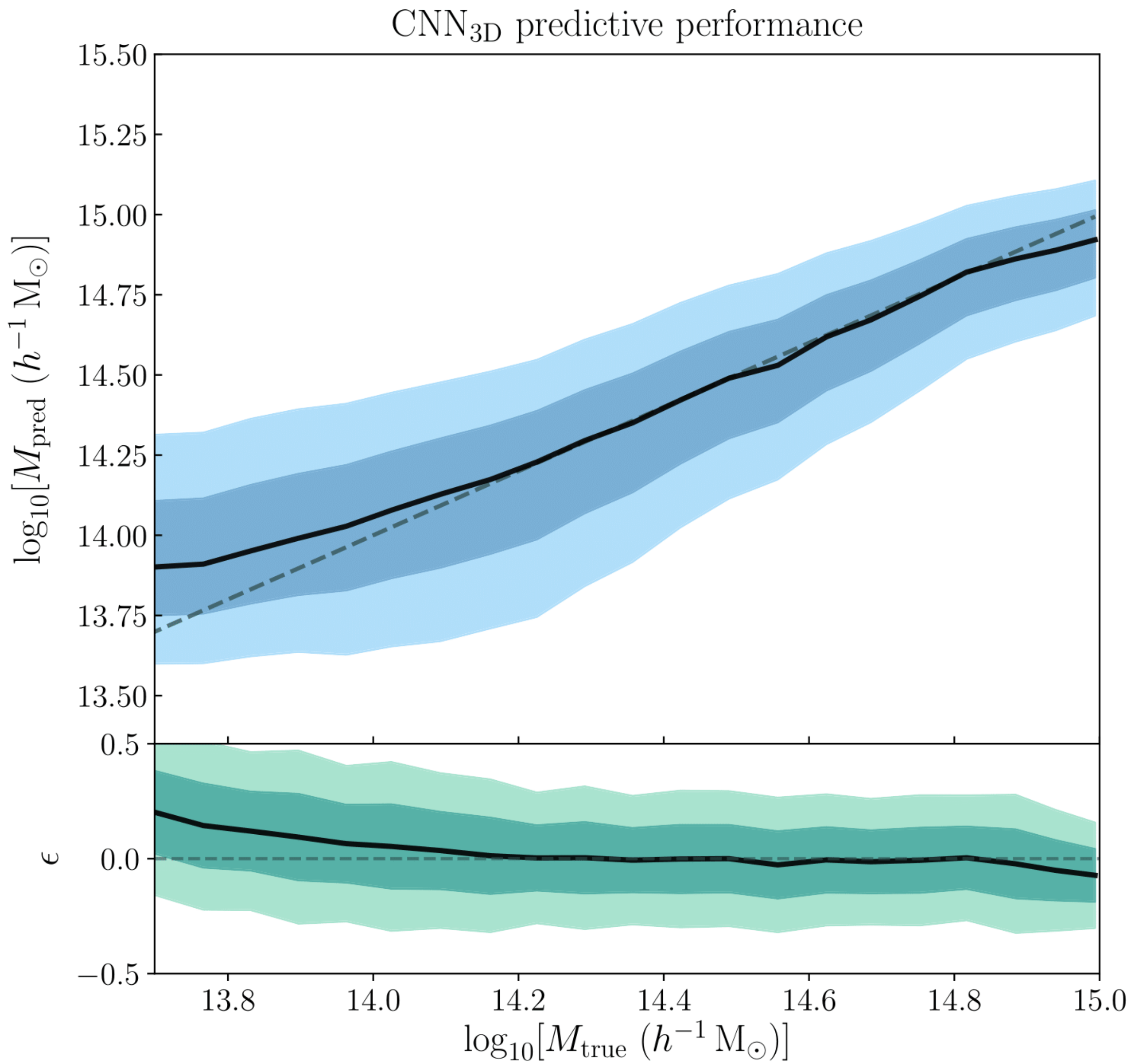
CNN architecture

- Nothing fancy – just a standard $\mathrm{CNN}_{\mathrm{3D}}$ model
- Feature extraction → compress to a single scalar (dynamical cluster mass)
- Relatively simple network with ~100k trainable parameters



Performance validation

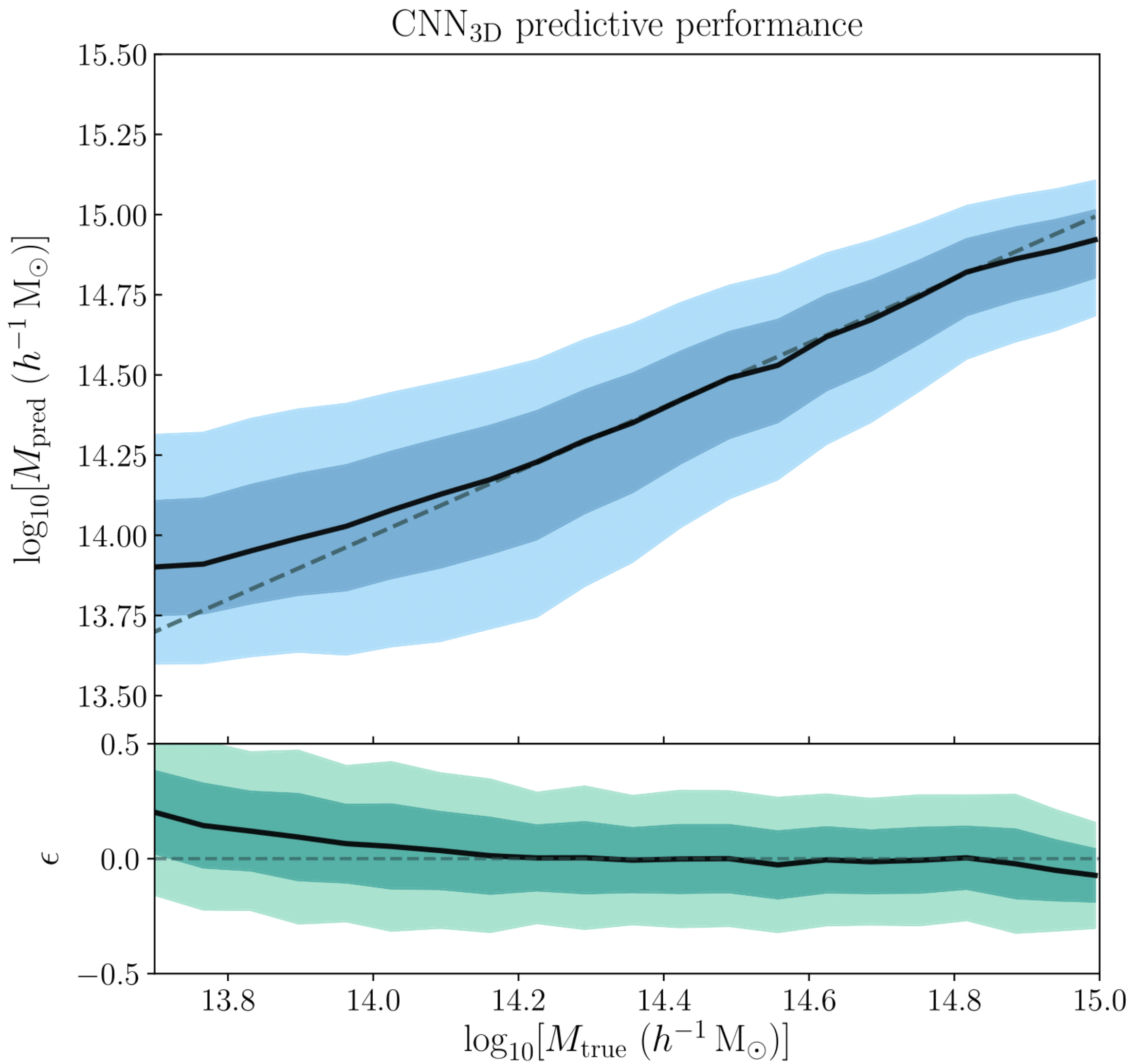
$$\epsilon \equiv \log_{10}(M_{\mathrm{true}}/M_{\mathrm{pred}})$$



Performance validation

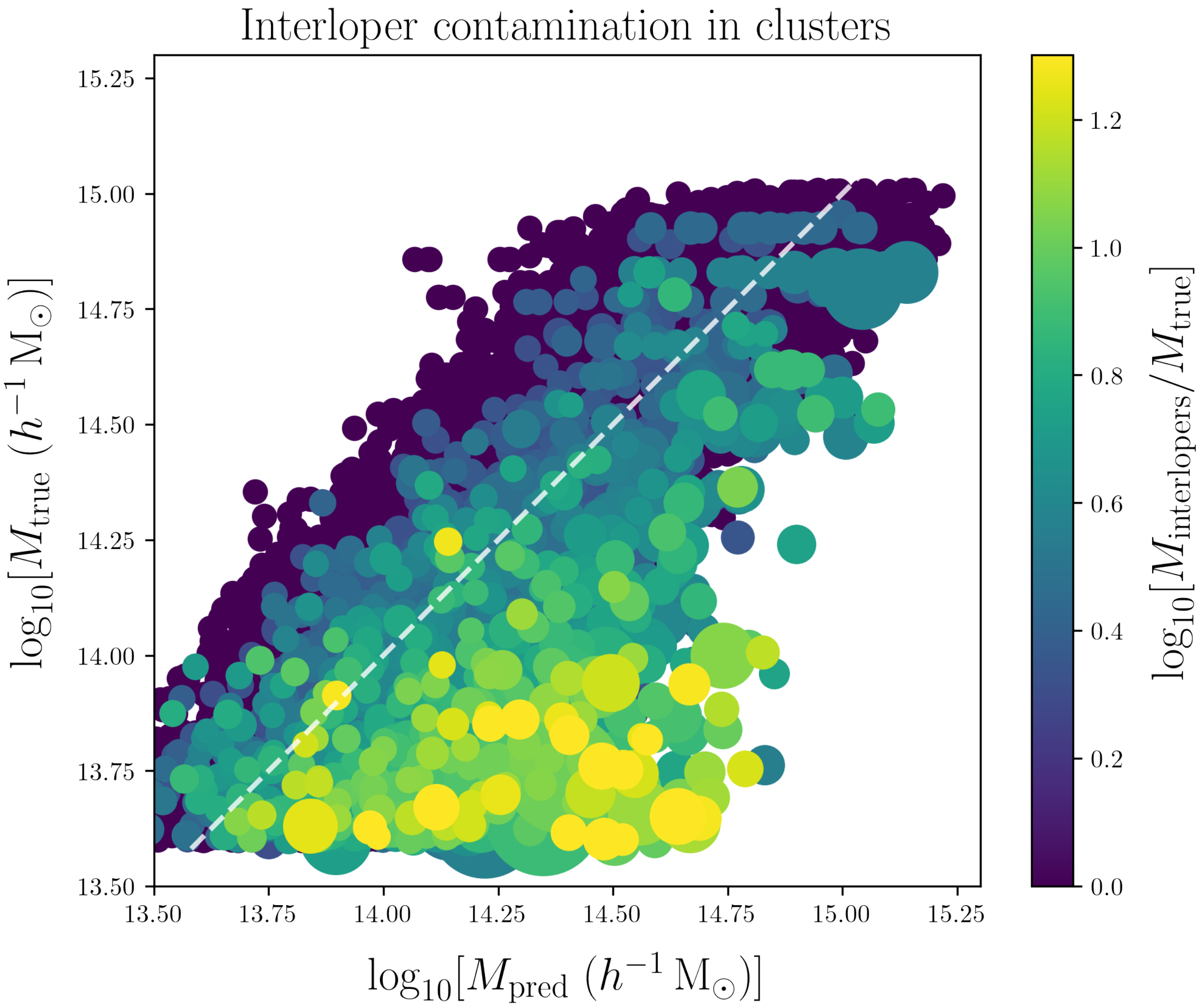
- $\mathrm{CNN_{3D}}$ tends to overestimate masses below ~14.1 dex
- Larger uncertainties for low-mass clusters
- Robust uncertainties with simulation-based inference
- Posterior is unbiased:
  Sub-optimal network → inflated posterior (but not incorrectly biased)

$$\epsilon \equiv \log_{10}(M_{\mathrm{true}}/M_{\mathrm{pred}})$$



Interloper contamination

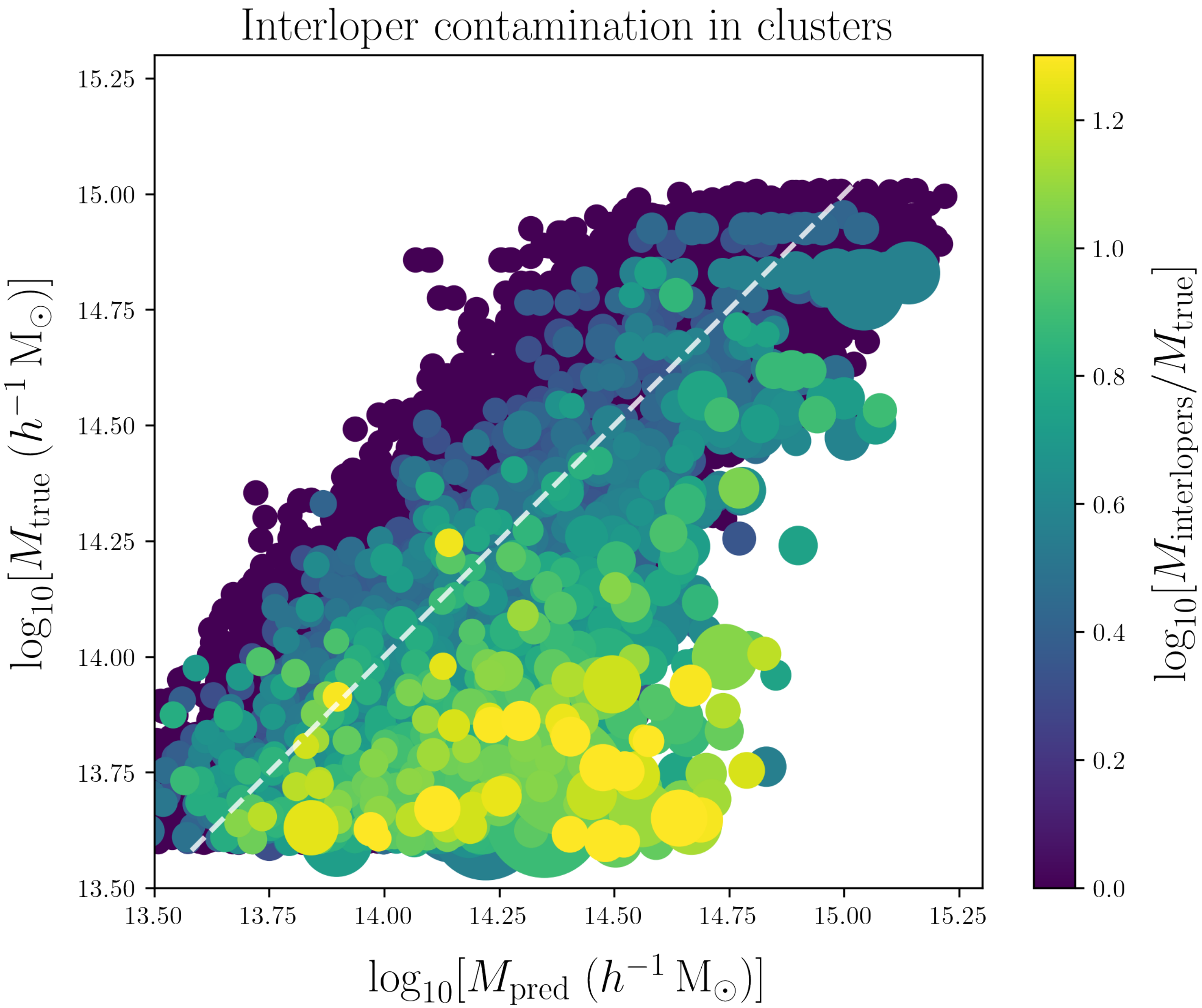
Colours – mass ratio of interloper cluster(s) to original cluster
Size – inverse distance between the clusters (i.e. closer = larger)



Interloper contamination

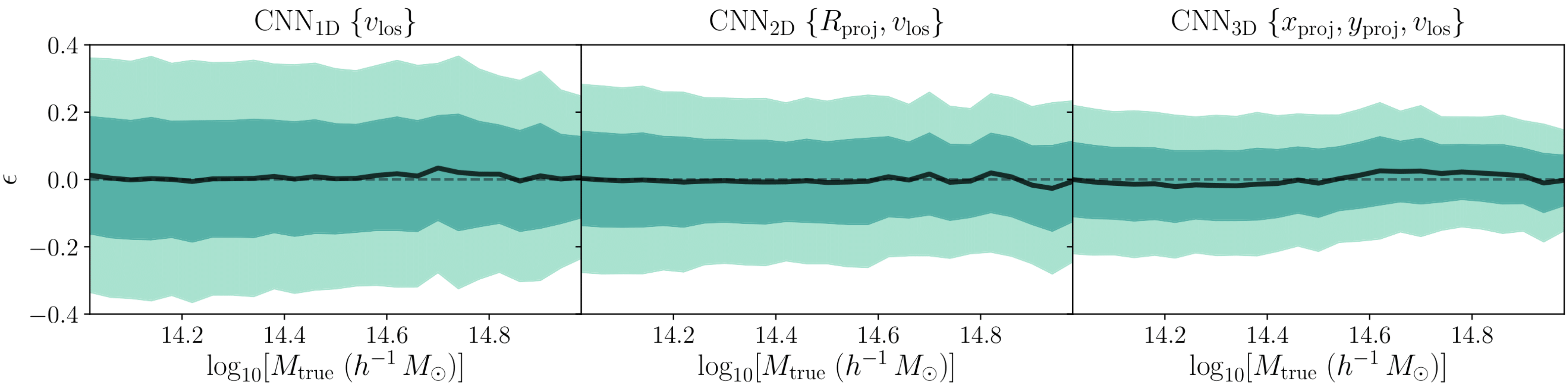
Colours – mass ratio of interloper cluster(s) to original cluster
Size – inverse distance between the clusters (i.e. closer = larger)

- Some interloper clusters up to 20 times more massive than original cluster
- Induces primarily a bias (overestimation)
- This bias is properly accounted via simulation-based inference



Information gain with higher dimensionality

- $\mathrm{CNN_{1D}}$ & $\mathrm{CNN_{2D}} \rightarrow$ results reproduced from **Ho+ 2019 (ApJ)** - **arXiv: 1902.05950**
- Performance quantified in terms of residual scatter
- Same mock catalogue for comparison
- Gain in constraining power when exploiting full 3D phase-space distribution



$$\boxed{\epsilon \equiv \log_{10}(M_{\mathrm{true}}/M_{\mathrm{pred}})}$$

Precision of ML cluster mass estimators

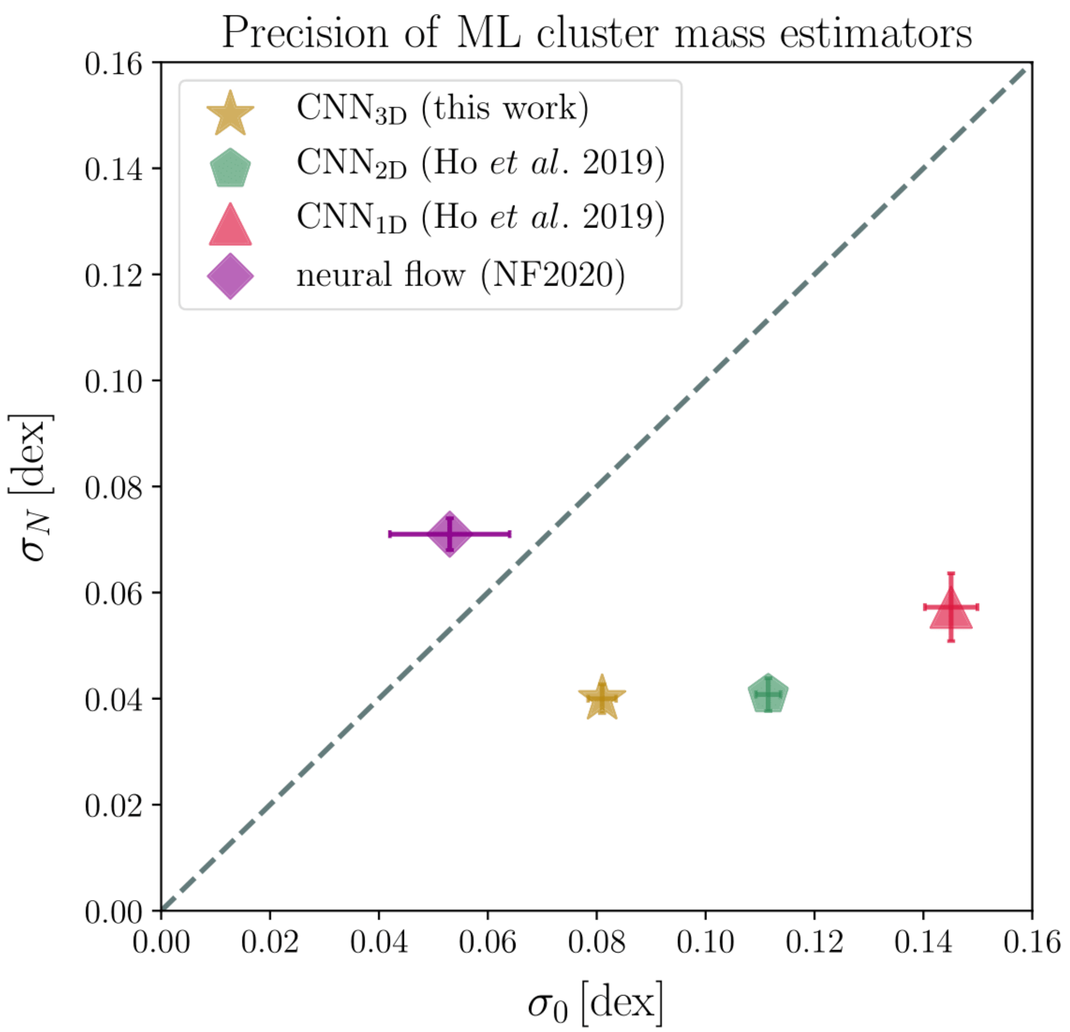
- As before, express total scatter as:

$$\boxed{\sigma^2 = \sigma_N^2 (N_{\text{members}}/100)^{-1} + \sigma_0^2}$$

Richness-dependent component **("statistical error")**

Richness-independent component **("systematic error")**

- Progressive improvement in precision of CNN models with higher dimensionality
- $\text{CNN}_{3\text{D}}$ is less sensitive to cluster richness than neural flow mass estimator



Precision of ML cluster mass estimators

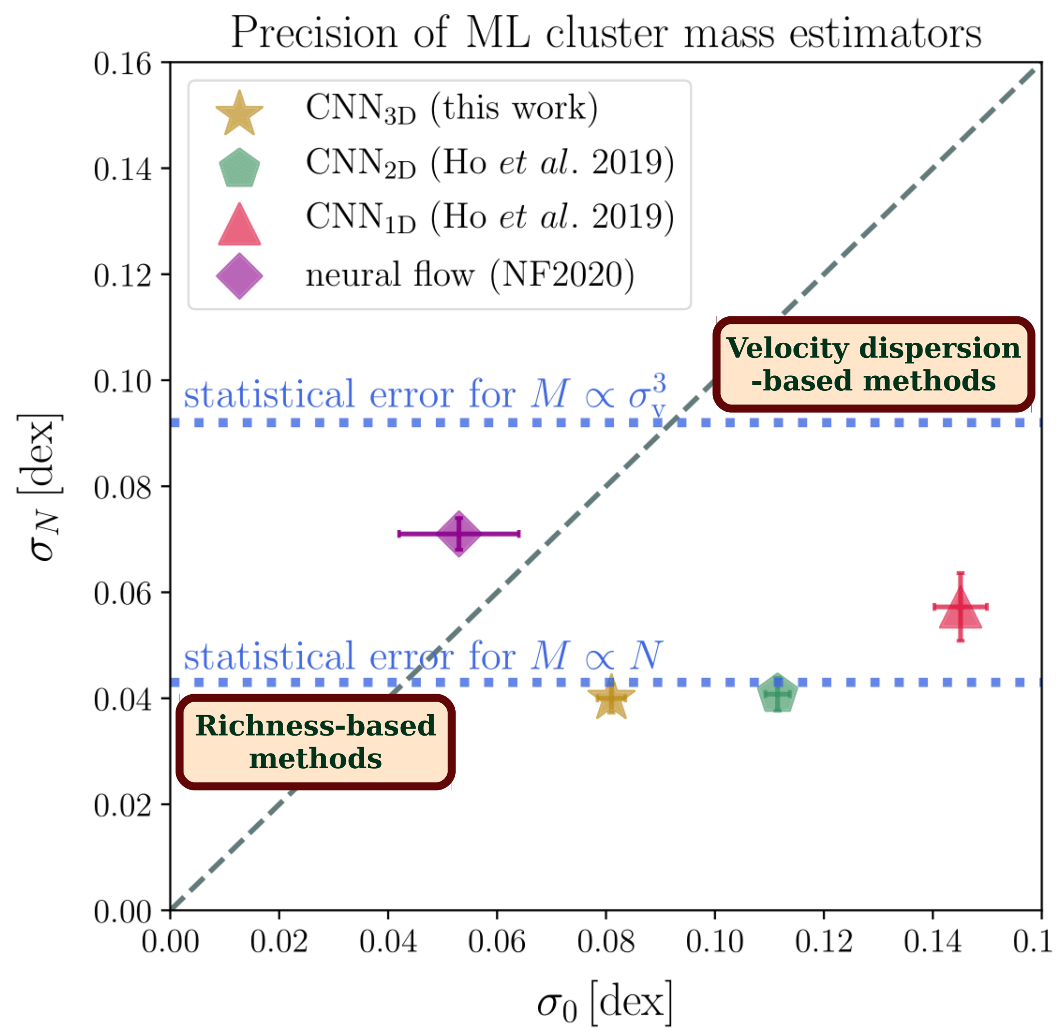
- As before, express total scatter as:

$$\boxed{\sigma^2 = \sigma_N^2 (N_{\mathrm{members}}/100)^{-1} + \sigma_0^2}$$

Richness-dependent component ("statistical error")

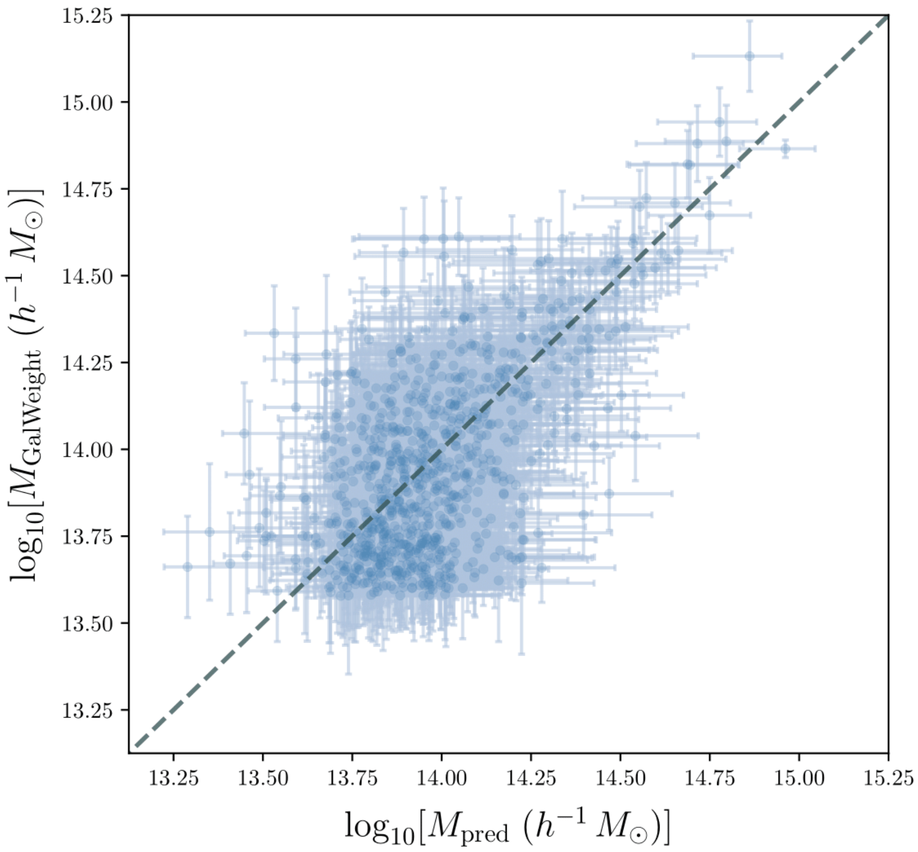
Richness-independent component ("systematic error")

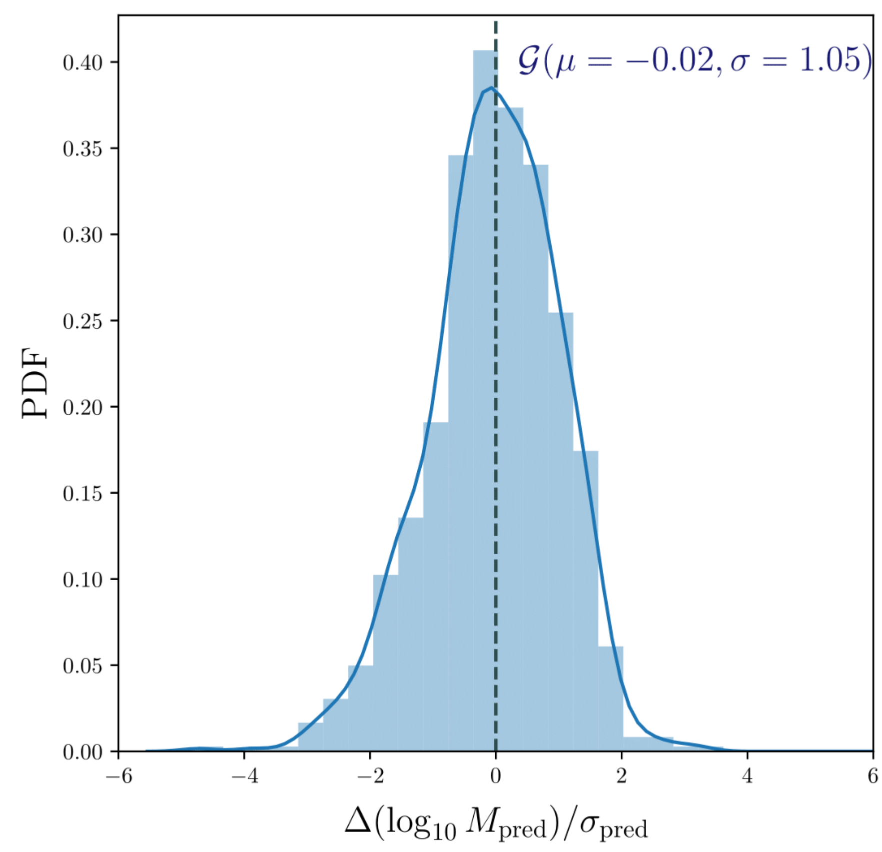
- Progressive improvement in precision of CNN models with higher dimensionality
- $\mathrm{CNN_{3D}}$ is less sensitive to cluster richness than neural flow mass estimator



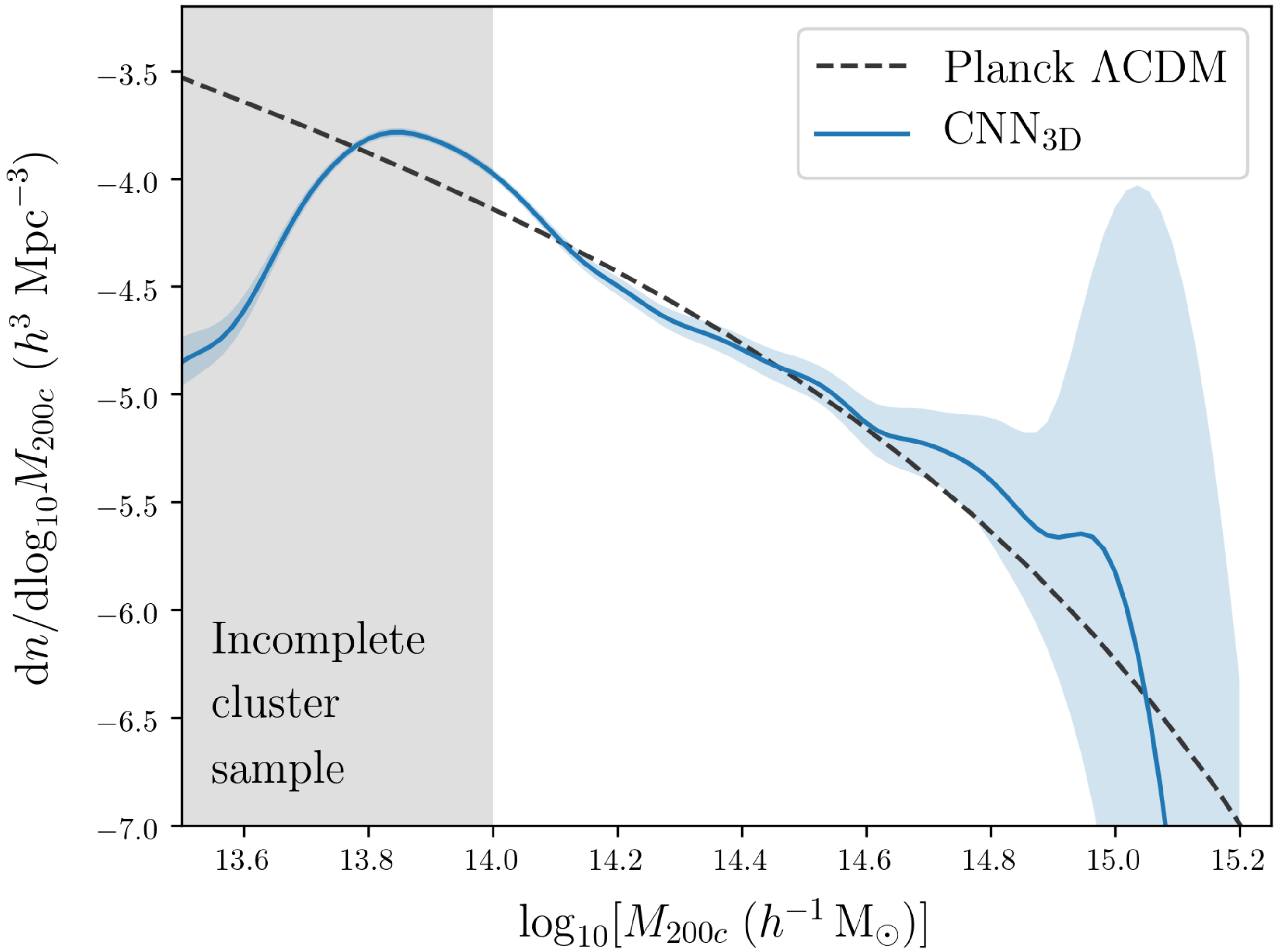
Application to SDSS clusters

- *GalWeight* catalogue – 910 galaxy clusters **(Abdullah+ 2020, ApJS)** - **arXiv: 1907.05061**
- Apply same phase-space cuts and preprocessing as for mock catalogue
- Overall consistency of our predictions with results from the **`GALWEIGHT`** mass estimator





Mass function from SDSS clusters



- Cluster mass function reconstructed from our SDSS dynamical mass estimates
- Recover mass function predicted by Planck ΛCDM down to mass completeness limit

Summary & future work

Neural flow mass estimator:

- Promising performance w.r.t classical & recent ML methods
- Robustness to velocity errors and galaxy subsampling (richness)
- Saliency maps to show informative regions
- Application to a set of real clusters

(DKR, Wojtak+ 2020, MNRAS)

Summary & future work

Neural flow mass estimator:

- Promising performance w.r.t classical & recent ML methods
- Robustness to velocity errors and galaxy subsampling (richness)
- Saliency maps to show informative regions
- Application to a set of real clusters

(DKR, Wojtak+ 2020, MNRAS)

Simulation-based inference ($\mathrm{CNN}_{\mathrm{3D}}$):

- Ensures uncertainties are not underestimated
- Optimally exploit information content of 3D dynamical phase-space distribution
- Application to SDSS catalogue – dynamical mass estimates with uncertainties
- Recover Planck ΛCDM mass function down to mass completeness limit

(DKR, Wojtak & Arendse 2020, submitted)

Summary & future work

Neural flow mass estimator:

- Promising performance w.r.t classical & recent ML methods
- Robustness to velocity errors and galaxy subsampling (richness)
- Saliency maps to show informative regions
- Application to a set of real clusters

(DKR, Wojtak+ 2020, MNRAS)

Simulation-based inference ($\mathrm{CNN_{3D}}$):

- Ensures uncertainties are not underestimated
- Optimally exploit information content of 3D dynamical phase-space distribution
- Application to SDSS catalogue – dynamical mass estimates with uncertainties
- Recover Planck ΛCDM mass function down to mass completeness limit

(DKR, Wojtak & Arendse 2020, submitted)

Future work:

- Cosmological inference from SDSS cluster abundance (normalizing flows, simulation-based inference & variational inference)

Collaboration with Matthew Ho

arXiv: 2006.13231

Back-up slides

Normalizing flows

Section 3.2 in paper

- Smooth **invertible** mapping with **tractable Jacobian** (2 fundamental requirements)

$$
\begin{aligned}
\boldsymbol{x} &= \mathcal{F}(\boldsymbol{u}) \\
\mathcal{P}(\boldsymbol{x}) &= \Psi\left[\mathcal{F}^{-1}(\boldsymbol{x})\right] \left|\frac{\partial \mathcal{F}^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|
\end{aligned}
$$

- **Composition** of series of relatively simple invertible transformations (key property) (flexible → characterize arbitrary complex distributions)

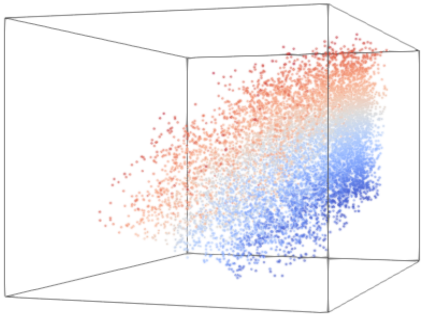
$$\mathcal{F} \equiv \mathcal{F}_1 \circ \mathcal{F}_2 \circ \ldots \mathcal{F}_k$$

- Invertibility allows both (1) sampling and (2) probability density evaluations (as long as it is possible to do so for the base distribution)

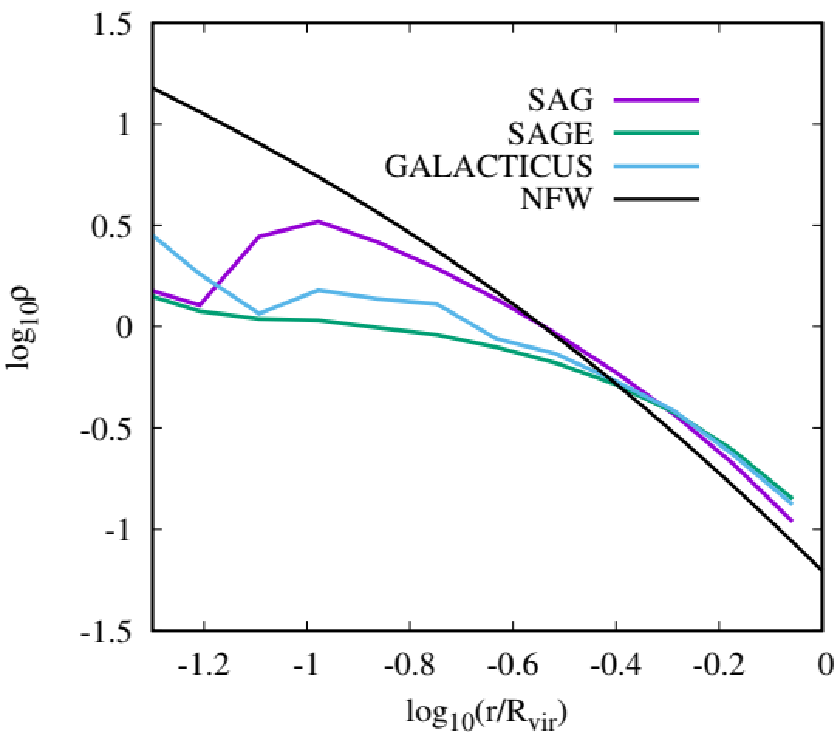
- In essence, neural network learns transformation from base distribution

Section 2.2 in paper

Mock SDSS cluster catalogue



- MDPL2 + SAG semi-analytical model (orphan galaxies)
- SAG - most complete implementation of modelling orphan galaxies
- Massive DM halos (**`ROCKSTAR`** catalogue) → galaxy clusters + central galaxy
- For every cluster → draw a LOS and compute phase-space diagram
- Phase-space coordinates computed relative to central galaxy
- Observed velocities → include Hubble flow w.r.t. cluster centre



- SAG → positions & absolute mags in SDSS filters
- Adopt SDSS-like flux limit
- Compute apparent mags by assigning an observer to each cluster
- Use maximum comoving distance of 250 h^{-1} Mpc (completeness ~10^{14} h^{-1} $M_{\odot}$)

- Account for distance-dependent completeness expected for flux limited selection of spectroscopic targets