The Dynamical Mass of the Coma Cluster from Deep Learning

Matthew Ho * [1,2], Michelle Ntampaka[3,4], Markus Michael Rau[1,2], Minghan Chen[5], Alexa Lansberry[1], Faith Ruehle[1], and Hy Trac[1,2]

[1] *McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

[2] *NSF AI Planning Institute for Physics of the Future, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

[3] *Space Telescope Science Institute, Baltimore, MD 21218, USA*

[4] *Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA*

[5] *University of California, Santa Barbara, Department of Physics, Santa Barbara, California, United States*

July 1, 2022

In 1933, Fritz Zwicky's famous investigations of the mass of the Coma cluster led him to infer the existence of dark matter [1]. His fundamental discoveries have proven to be foundational to modern cosmology; as we now know such dark matter makes up 85% of the matter and 25% of the mass-energy content in the universe. Galaxy clusters like Coma are massive, complex systems of dark matter in addition to hot ionized gas and thousands of galaxies, and serve as excellent probes of the dark matter distribution. However, empirical studies show that the total mass of such systems remains elusive and difficult to precisely constrain. Here, we present new estimates for the dynamical mass of the Coma cluster based on Bayesian deep learning methodologies developed in recent years. Using our novel data-driven approach, we predict Coma's M_{200c} mass to be $10^{15.10\pm0.15}$ $h^{-1}\mathrm{M}_\odot$ within a radius of 1.78 ± 0.03 $h^{-1}\mathrm{Mpc}$ of its center. We show that our predictions are rigorous across multiple training datasets and statistically consistent with historical estimates of Coma's mass. This measurement reinforces our understanding of the dynamical state of the Coma cluster and advances rigorous analyses and verification methods for empirical applications of machine learning in astronomy.

Due to its close proximity, high sample richness, and historical significance, the Coma system is one of the most well studied galaxy clusters in the sky and has served as a hotbed for applications of both new and established astronomical survey techniques for almost a century [2]. Over the years, astronomers have produced multiple mass estimates of Coma, each improving upon the previous with higher quality survey data, better control of systematics or new physically-motivated inference methods. Existing analyses have inferred Coma's mass from a variety of known observables, including the gravitational lensing of background light [3, 4], the X-ray emission of its hot intracluster gas [5], and the spectra of its galaxies [6, 7, 8], but have generally exhibited large predictive scatter and wide modeling uncertainties. In addition to distinct phenomenological studies of the Coma cluster, the introduction of novel cluster mass inference methods is also a crucial step towards progressing modern cosmological analyses. For example, robust inference of cluster masses allows for accurate

*Corresponding author, mho1@andrew.cmu.edu

determination of universal observables such as the halo mass function, which in turn enables us to constrain cosmological models [9]. In addition, cluster mass and accretion history is needed to understand galaxy formation and evolution.

With the upcoming data releases of large-scale spectroscopic surveys from the Dark Energy Spectroscopic Instrument (DESI), the Vera C. Rubin Observatory, and Euclid [10], the question of how to accurately and efficiently measure cluster masses from galaxy spectra is a popular topic in the cosmology research. Spectroscopic redshifts of galaxies are an effective probe of the dynamical state of a cluster, from which the depth of the system's gravitational well and thereby its total mass can be analytically derived [11]. The resulting power-law relationship between cluster mass M and the dispersion of its line-of-sight galaxy velocities σ, as measured by spectroscopic surveys, is traditionally known as the M-σ relation. While fundamentally sound and historically significant, the M-σ is notably susceptible to a variety of systematic biases, both physical [12] and selection-based [13]. For example, gravitational instabilities such as cluster mergers can distort or reshape the velocity profile of constituent galaxies, violating any assumptions of dynamical equilibrium. Similarly, unbound interloping galaxies along the line-of-sight can be mistaken for cluster members, contaminating our galaxy sample. Recent research has made progress towards quantifying and accounting for the effect of these systematics, with new machine learning based methods introducing innovative data-driven approaches to the field [14, 15, 16].

The deep learning methods applied in this analysis extend the body of literature on dynamical mass estimates and have been shown to effectively mitigate the impact of their systematics. Previous publications have demonstrated that these deep-learning-based mass reconstruction methods can reduce the scatter of mass measurements by a factor of three [14] and produce well-calibrated estimates of predictive uncertainty [15] when evaluated on realistic simulated observations of galaxy clusters. These improvements can be attributed to the ability of deep learning to learn and model highly non-linear relationships in data. The distortions and sample contamination implicit in dynamical observables parallel those studied in classic deep learning problems like image-recognition. Also, whereas traditional dynamical mass measurements seek to analytically characterize the complex relationship between cluster masses and dynamical observables, deep learning models can be trained to learn these intricacies automatically through experiencing thousands of simulated mock examples. In the context of these complex systematic features and rich, high-fidelity training data, deep learning models excel, making them an apt candidate for modeling dynamical masses of clusters.

Figure 1 details the inference pipeline behind our deep learning analysis of dynamical observables. From either mock simulation or real measurements, we calculate a set of dynamical observables from all cluster-galaxy pairs in our sample, namely the set of relative line-of-sight velocities, $v_{\rm los}$, and projected radial distances $R_{\rm proj}$. The space spanned by these two dimensions $\{v_{\rm los}, R_{\rm proj}\}$ is canonically referred to as *dynamical phase space*. We assume each cluster's mass dictates a unique distribution in this space and our galaxy population is a representative, but limited sample from this distribution. We then design neural network architectures to recover masses from each cluster's galaxy sample in this space. To study the improvements of including $R_{\rm proj}$ as an additional input dimension, we design two neural architectures which utilize either the univariate $\{v_{\rm los}\}$ distribution or the joint $\{v_{\rm los}, R_{\rm proj}\}$ distribution to estimate masses, subsequently referred to as 1D or 2D models, respectively. For either input type, we estimate each cluster's galaxy dynamical distribution using a Kernel Density Estimator [17] and then sample it at regular intervals across a pre-defined range. This has the effect of both normalizing our data to guard against sample richness dependency; a proxy which is generally difficult to constrain in observation, and creating an image of our galaxy distributions, in a fixed shape which is acceptable to our deep learning models.

To map our cluster phase space distributions to masses, we use deep learning architectures based

on Approximate Bayesian Convolutional Neural Networks [18]. Convolutional Neural Networks (CNNs) [19, 20] are a class of neural networks which utilize *convolutional filters* to encourage learning localized patterns in sub-regions of high-dimensional, spatially-distributed datasets. This behavior is widely considered the gold standard in applications of computer vision and is well-suited for our task of dynamical mass estimation, as the aforementioned physical and selection systematics appear as distortions or artifacts in each cluster's dynamical phase distribution. To then recover estimates of predictive uncertainty or confidence intervals, we utilize Dropout marginalization to approximate our model as a Bayesian Neural Network [21]. Dropout layers [22] are a popular tool used to regularize learning in deep networks by randomly eliminating neural pathways during each epoch of training and, under certain conditions, can be used to emulate a Bernoulli variational distribution on free parameters. Here, we also allow Dropout layers to activate during predictive inference and marginalize over 100 realizations of their predictions to approximate marginalization over all parameter uncertainties. This method and architecture produces a Gaussian predictive posterior over cluster mass and was empirically validated to high precision on independent mock observations of simulated clusters [15]. Alternative methods for uncertainty reconstruction of deep learning cluster mass estimates are also investigated in [16, 23].

To demonstrate the robustness of our architecture, we train each model under one of two catalogs of realistic mock cluster observations derived from independent N-body simulations, namely the DR1 Uchuu simulation [24, 25] and the MultiDark Planck 2 simulation [26]. In each catalog, the mock cluster observations were designed to faithfully reproduce real systematics that affect dynamical mass estimates in practice. In each simulation, we assign galaxies to subhalos using the UniverseMachine [27, 25] labeling procedure and restrict our galaxy sample to a stellar mass cut of $M_{\rm stellar} \geq 10^{9.5}\ h^{-1}{\rm M}_{\odot}$. We then 'observe' clusters in each simulation from singular lines-of-sight and make observational cuts on the dynamical observables that we recover. Namely, we restrict our galaxy sample to a large cylinder in dynamical phase space defined by a maximum relative line-of-sight velocity of $v_{\rm los} \leq 3800$ km s^{-1} and a maximum radial projected distance of $R_{\rm proj} \leq 2.3\ h^{-1}$Mpc. This cylinder preserves the physical systematics of each cluster within our observable data and is sufficiently large and simple enough to allow interloping galaxies to contaminate the sample. After performing these selection cuts on clusters in each simulation, we produce the mock observation catalogs hereafter referred to as Uchuu-UM and MDPL2-UM. Due to its larger simulation volume, the Uchuu-UM catalog has more training examples than MDPL2-UM, with $10,000$ samples per training fold vs. MDPL2-UM's $7,000$ samples, though a flat mass prior is assumed in both training catalogs.

Using labeled data from these mock catalogs, we train neural network models to relate dynamical observables to cluster masses and extend their learned behavior to the observational Coma system. Figure 2 shows the catalog of galaxy sky positions and spectroscopic redshifts in the vicinity of the Coma cluster [28] drawn from the 12th Data Release of the Sloan Digital Sky Survey (SDSS) [29] around the center identified from the Abell catalogue [30]. It also demonstrates the dynamical observables $\{v_{\rm los}, R_{\rm proj}\}$ derived from these measurements as well as the corresponding KDE-processed distribution images that serve as our deep learning model inputs. We perform the same observational cuts on SDSS's Coma data as those imposed on our simulated catalogs, namely the cylinder cuts in dynamical phase space centered and the stellar mass cut. Stellar masses are assigned to Coma's galaxy spectra using the Portsmouth passive stellar galaxy model [31].

Table 1 details the architecture specifications, validation performance, and predictive inference of the models presented in this analysis. From the percentile statistics of each model's predictive residuals, we find that the deep learning models predict masses of our mock clusters consistently across the MDPL2-UM and Uchuu-UM simulations. Each model's mean predictions are statistically consistent with zero bias, regardless of whether it is validated on its own or alternative training

catalog. The average variances predicted by our models are also nearly identical across simulations, further assuring generalization between the mock catalogs. We also find that predictions of 2D models have lower scatter than those of 1D models, as in [15]. This increase in constraining power is similarly reflected in the recovered uncertainties for the Coma mass estimates. Lastly, we can construct traditional M-σ estimates of each simulated cluster's mass and compare these with our predictions. This analysis shows that the 1D and 2D CNN model 1-σ uncertainties are on average 56% and 69% smaller than those of the simplistic M-σ, respectively.

Figure 3 shows the final predictive mass posteriors we estimate for the Coma cluster, as well as how they compare to previous estimates of Coma's M_{200c} using a variety of other mass measurement methods [6, 5, 7, 32, 3, 4, 8, 16]. We show that our mass predictions of the Coma cluster are statistically consistent with most historical estimates. The only notable exceptions is the weak lensing estimates produced by [4]. In this case, the M_{200c} estimates are at a $\sim 2.3\sigma$ tension with our predictions, but were also analyzed by [3]with a shallower observational sample and were found to be consistent to within $\sim 1.6\sigma$ with our measurements. All other historical estimates are consistent with our results to within $\sim 2\sigma$.

As an experiment in observational selection, we investigate the impact of radial galaxy selection on our Coma mass estimates. We retrain both 1D and 2D models using newly-generated mock catalogs wherein all inputs are re-weighted to match the projected radial distribution of the Coma galaxy sample (details in Methods: Preprocessing of Observables). The validation performance and predictive inference on Coma are displayed in Table 1 and Figure 3. We see a slight increase in mean mass prediction for 1D models, and a universal decrease in predictive variance for both model. We suggest that this is the result of the ML models optimizing their predictions by learning the average masses for general $R_{\rm proj}$ cluster distributions, whereas enforcing the appropriate radial distribution results in lower uncertainties for Coma-like systems. However, the impact of re-weighted training sets on our Coma mass predictions does not affect our validation performance or Coma predictions to a statistically significant degree.

Identifying the standard Uchuu-UM 2D model as our best mass estimator for its robust training set, low validation scatter, and generality, we present a M_{200c} mass estimate of $10^{15.10\pm0.15}$ $h^{-1}\mathrm{M}_{\odot}$ for the Coma system. This corresponds to a $R_{200c} = 1.78 \pm 0.03$ h^{-1}Mpc estimate according to the training simulation's original cosmology [33]. The predictions are self-consistent and fall within the range of reliability for our model predictions. We find that our Coma mass predictions are statistically consistent with historical estimates, including strong agreement with estimates from the past decade. This empirical validation is a strong indicator of the applicability of deep learning models to observational study of galaxy clusters, and the methodology applied here serves as a template for future empirical extensions of the increasingly popular data-driven applications of ML in cosmology. Verification on well-studied systems such as Coma lay important groundwork for the eventual extension of ML methods to ensemble prediction of many clusters, resulting in constraints on cosmological models. In addition, future work will include studying the observational effects of 3D dynamical information as in [23] and adding robust predictive marginalization over astrophysical priors as in [34].

Table 1: Table of Cross-Validation Metrics and Coma Mass Posteriors for all Investigated Models

Training Sim	Input	Reweighted?	$\hat{\epsilon}_{\rm val}$[1,2]	$\hat{\epsilon}_{\rm x-val}$[1,3]	$\sqrt{\bar{\rm Var}_{\rm val}}$[4]	$\sqrt{\bar{\rm Var}_{\rm x-val}}$[5]	$\hat{m}_{\rm Coma}$[6]
Uchuu-UM	1D		$0.06^{+0.12}_{-0.12}$	$0.12^{+0.13}_{-0.12}$	0.14	0.16	14.99 ± 0.19
Uchuu-UM	2D		$0.04^{+0.08}_{-0.09}$	$0.06^{+0.09}_{-0.08}$	0.11	0.11	15.10 ± 0.15
MDPL2-UM	1D		$0.11^{+0.13}_{-0.12}$	$0.04^{+0.12}_{-0.15}$	0.15	0.14	14.85 ± 0.16
MDPL2-UM	2D		$0.03^{+0.08}_{-0.09}$	$-0.03^{+0.09}_{-0.11}$	0.11	0.11	15.01 ± 0.11
Uchuu-UM	1D	✓	$0.07^{+0.11}_{-0.12}$	$0.10^{+0.12}_{-0.12}$	0.14	0.15	14.90 ± 0.14
Uchuu-UM	2D	✓	$0.05^{+0.08}_{-0.09}$	$0.06^{+0.10}_{-0.09}$	0.12	0.11	14.86 ± 0.11
MDPL2-UM	1D	✓	$0.10^{+0.13}_{-0.13}$	$0.04^{+0.12}_{-0.15}$	0.15	0.14	14.83 ± 0.15
MDPL2-UM	2D	✓	$0.06^{+0.10}_{-0.08}$	$0.02^{+0.10}_{-0.11}$	0.11	0.12	14.87 ± 0.13

[1] Predictive residual $\epsilon \equiv m_{\rm pred} - m_{\rm true}$, where $m \equiv \log_{10}\left[M_{\rm 200c}\ h^{-1}{\rm M}_{\odot}\right]$

[2] Predictive residual median and 16-84 percentile range (dex) on independent test set of training simulation catalog

[3] Predictive residual median and 16-84 percentile range (dex) on test set of alternate simulation catalog

[4] Square root of average predictive variance on independent test set of training simulation catalog

[5] Square root of average predictive variance on test set of alternate simulation catalog

[6] Coma mass prediction with $\pm 1\sigma$ error bounds

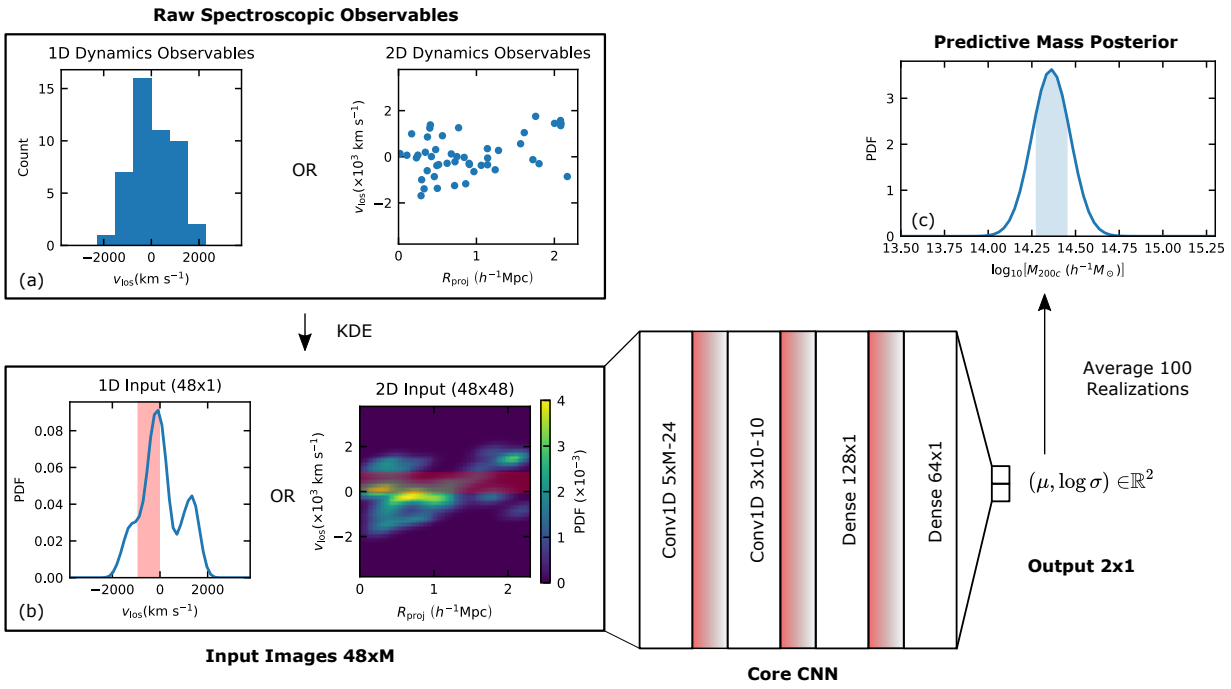

Figure 1: **Machine learning workflow for dynamical cluster mass inference**. **(a):** Raw 1D and 2D dynamical observables calculated from the galaxy sample. We generalize our input images to have shapes $48 \times M$, where M is equal to 1 or 48 for 1D or 2D models, respectively. **(b):** Our chosen CNN design and neural architecture. In the neural architecture, we show an example convolutional filter highlighted in red over the input distributions. Dropout connections exist in between all layers and are activated for both model training and inference . All layers utilize a rectified linear activation function (ReLU). In the diagram, convolutional layers are described using their filter shape and number of filters, respectively. **(c):** shows an example output predictive mass posterior for a single input.

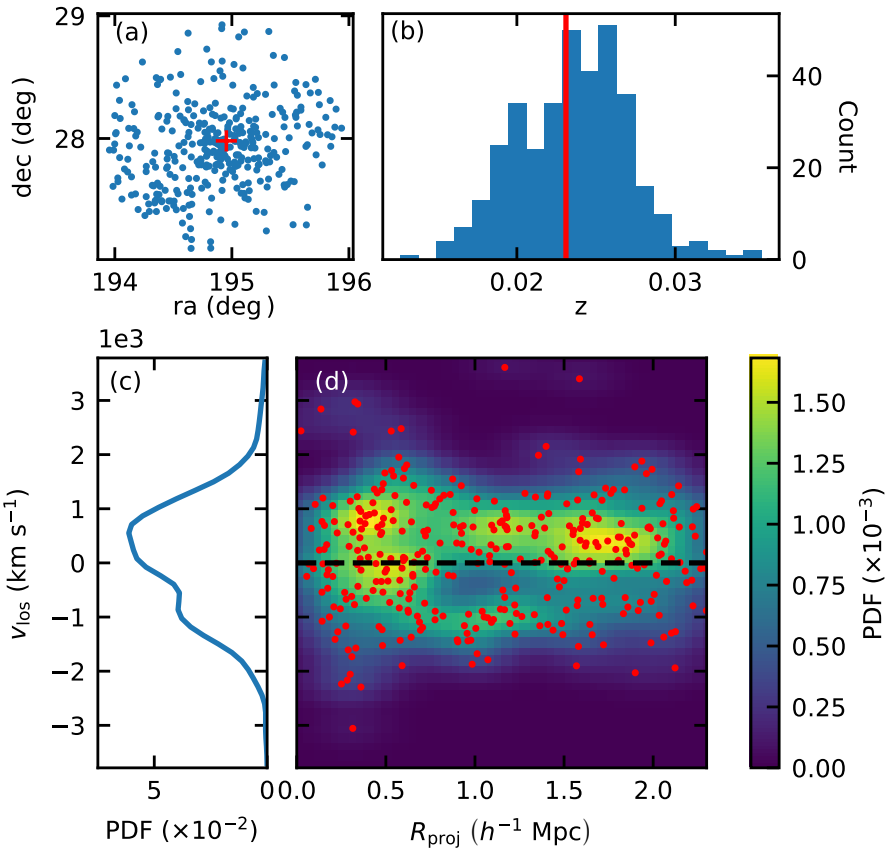

Figure 2: **Observational Sample of Coma Cluster Galaxies. (a):** Projected sky distribution of Coma galaxies in our sample. **(b):** Redshift distribution of Coma galaxies in our sample. **(c,d):** 1D and 2D KDE-processed images derived from the Coma data to be used as input to our machine learning model.

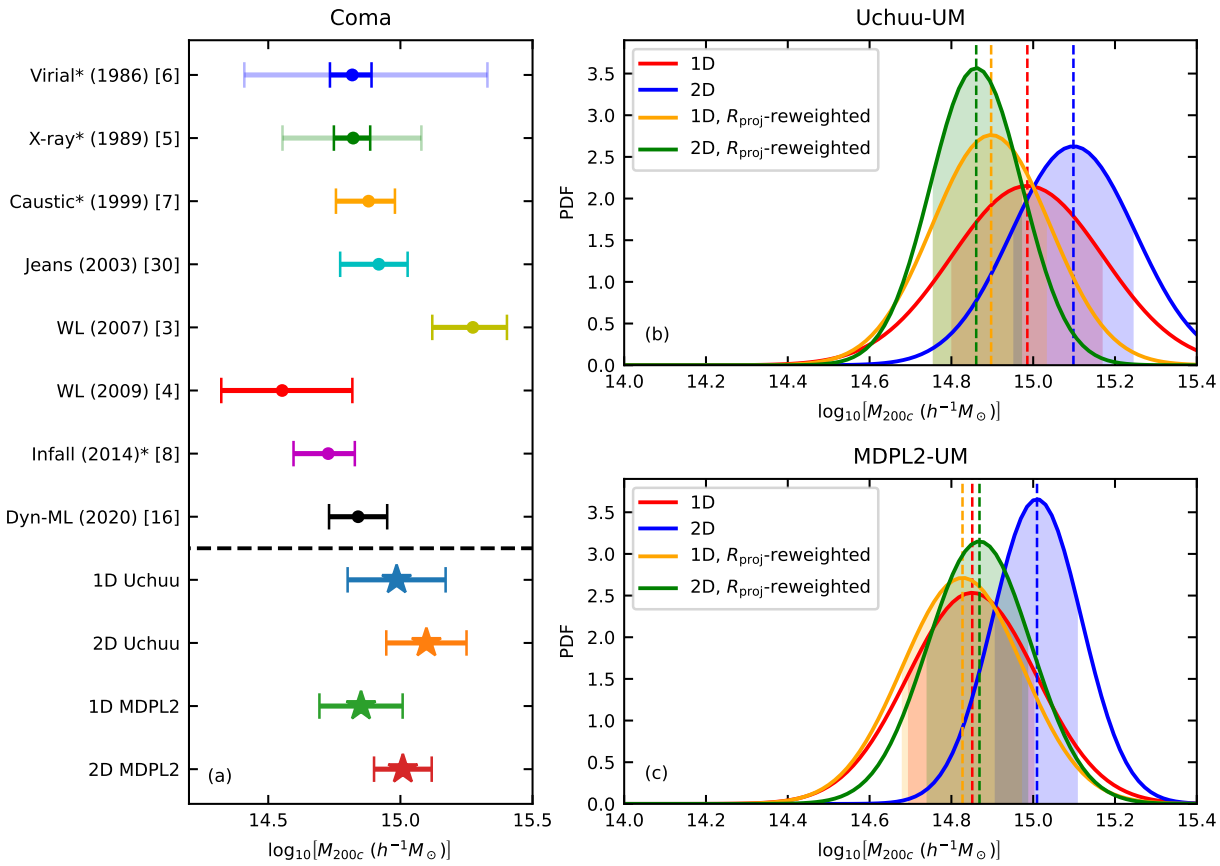

Figure 3: M_{200c} **Mass Estimates of the Coma cluster. Left:** Deep learning Coma mass estimates relative to historical predictions of M_{200c} derived from virial methods [6], X-ray profiles [5], caustics [7], Jeans analyses [32], weak lensing measurements [3, 4], infalling galaxy kinematics [8], and ML-based dynamical estimators [16]. All mass estimates are shown with their median and 16th to 84th percentile confidence intervals. The confidence intervals on the two oldest mass estimates [6, 5] are shown in a darker color for when one makes strict assumptions about the shape of the mass profile, and again in a lighter color when those assumptions are relaxed. An asterisk (*) identifies analyses in which M_{200c} estimates were not explicitly published, but where we have reconstructed them from reported masses and radii using an assumed NFW profile (See Methods: Historical Mass Estimates of the Coma Cluster for further details). **Right:** Mass posteriors of the Coma cluster estimated by our deep learning models. Predictive posteriors are shown for each model type across both MDPL2-UM and Uchuu training sets. Each mass posterior's mean and 1σ confidence interval are highlighted.

Methods

Approximate Bayesian Deep Neural Networks

Deep neural networks [20] are a class of parametric ML models which are commonly used for learning complex relationships in data-rich environments. Mathematically, a DNN can be viewed as a highly non-linear functional mapping $\mathbf{y} = f(\mathbf{x};\boldsymbol{\theta})$ between inputs $\mathbf{x}$ and outputs $\mathbf{y}$, which is characterized by some set of weight matrices $\boldsymbol{\theta}$. Classically, training a DNN involves attempting to find the optimal weights $\boldsymbol{\theta}^*$ which produce the best mapping of inputs to outputs according to minimization of some error metric called the loss function $\mathcal{L}(\mathbf{y}, f(\mathbf{x};\boldsymbol{\theta}))$ averaged over a given, labeled training dataset $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$. This optimization of weight parameters is numerically tractable, even for DNNs with million of parameters, which makes these models excellent candidates for data-driven discovery. For a more detailed explanation of DNNs and their evaluation, see [14].

Whereas classical DNNs are effective at tasks relating to point regression and classification, modern advancements in machine learning have described how to characterize the outputs of these models in a Bayesian setting. In our application, we use the functional output of a DNN to dictate a distribution of outputs $p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta},\boldsymbol{\eta})$ [35], namely the means and variances of a univariate Gaussian over cluster mass, $f(\mathbf{x};\boldsymbol{\theta}) = (\mu, \log\sigma) \in \mathbb{R}^2$. This framework is a method of modeling intrinsic (or aleatoric) uncertainties in the data and allows the DNN to express not only what output predictions it can make, but also the statistical confidence that it has in those predictions for a given input. Under realistic modeling conditions, even with an idealized training procedure, the recovered setting $\hat{\boldsymbol{\theta}}$ is often highly degenerate over the parameter space $\boldsymbol{\Theta}$. When training data is limited, it is possible to recover parameter settings which minimize loss over the training set but are not representative of the data at large. To model epistemic uncertainties, we marginalize predictive distributions over the conditional probability of all possible weight parameters given the training data.

$$p(\mathbf{y}|\mathbf{x},\boldsymbol{\eta},\mathcal{D}) = \int p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta},\boldsymbol{\eta})\, p(\boldsymbol{\theta}|\boldsymbol{\eta},\mathcal{D})\, d\boldsymbol{\theta}, \tag{1}$$

where $p(\mathbf{y}|\mathbf{x},\boldsymbol{\eta},\mathcal{D})$ is the weight-marginalized posterior distribution, $p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta},\boldsymbol{\eta})$ is the chosen predictive distribution, and $p(\boldsymbol{\theta}|\boldsymbol{\eta},\mathcal{D})$ is the distribution of weight parameters informed by training data. Unfortunately, the full calculation of Eqn. 1 is numerically intractable for large DNNs. The integration over the space of hundreds of thousands of DNN weights is not feasible, even with highly efficient Monte Carlo methods.

Instead, we can use a technique known as variational inference to approximate the true weight distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta},\mathcal{D})$ with a variational distribution $q(\boldsymbol{\theta}|\hat{\boldsymbol{\phi}})$ whose form is chosen to simplify the integration in Eqn. 1. In our application, we use a multivariate Bernoulli distribution to model our varational distribution $q(\boldsymbol{\theta}|\hat{\boldsymbol{\phi}})$, a technique pioneered by [21]. In their implementation, they utilized the popular regularization technique, Dropout, to perform stochastic integration (Eqn. 1). In both the training and inference stages, Dropout layers are allowed to randomly set some fraction, $p_d \in [0,1]$, of the weight parameters equal to 0. The Dropout layers are stochastic, causing each functional evaluation of the model to use a different weight configuration. During training, this acts to regularize the iterative updates of stochastic gradient descent [22]. During inference, one can average many realizations of the Dropout layers to effectively produce a Monte Carlo estimate of the model output. [21] showed that such a training and evaluation procedure approximates a Gaussian Process and is able to accurately recover uncertainties for both in- and out-of-sample data.

Mock catalog generation

The catalogs used in this analysis are generated from a $z = 0.022$ snapshot of the MDPL2 simulation [26] and a $z = 0$ snapshot of the Uchuu-UM simulation [24], each of which assumes a ΛCDM cosmology consistent with the 2013 [33] and 2015 [36] Planck data, respectively. MDPL2 simulates 3840^3 particles at a mass resolution of $1.51 \times 10^9\ h^{-1}\mathrm{M}_\odot$ within a $\left(1000\ h^{-1}\mathrm{Mpc}\right)^3$ volume box. Uchuu has a higher spatial and mass resolution, simulating 12800^3 particles at a mass resolution of 3.27×10^8 within a box of volume $\left(2000\ h^{-1}\mathrm{Mpc}\right)^3$. Host halos and subhalos are identified in each simulation using the ROCKSTAR halo finder [37]. We model clusters as host halos in each catalog with spherical overdensity masses of $M_{200c} \geq 10^{14}\ h^{-1}\mathrm{M}_\odot$. Galaxies are assigned to subhalos in both simulations using the UniverseMachine [27, 25] labeling procedure. Clusters and galaxies in our sample inherit mass, position, and velocity from their respective halos in the MDPL2 and Uchuu Rockstar catalogs.

As in [14, 15], the mock catalogs used to train our models are augmented to have a constant number density of $dn/d\log m = 10^{-5.2}\ h^3\mathrm{Mpc}^{-3}\mathrm{dex}^{-1}$ across all cluster masses $M_{200\mathrm{c}} \geq\ 10^{14}\ h^{-1}\mathrm{M}_\odot$. To achieve this evenly-distributed training set, abundant low-mass clusters are downsampled and scarce, high-mass clusters are upsampled. The upsampling procedure involves taking multiple observations of the same cluster from various, evenly-distributed LOSs, to fully capture orthogonal dynamical information. The test catalogs are not augmented and are distributed according to the simulation's halo mass function. For comprehensive details on the mock catalog generation code, see [14].

Preprocessing of Observables

Each cluster entry in our mock catalog contains dynamical information in the form of v_{los} and R_{proj} measurements of member galaxies. These variable-length data vectors are processed into fixed-length image representations using Kernel Density Estimators (KDEs) [17]. KDEs generate a non-parametric estimate of the probability density function (PDF) of an unknown given independent samples from its distribution (Eq. 2 in [14]). To turn dynamical observables into images, we first use KDEs to 'smooth' each cluster's list of discrete v_{los} and R_{proj} data points into a continuous estimated PDF. The nature and scale of this smoothing is determined by a chosen kernel function which, in our case, is a Gaussian kernel with a fixed bandwidth scaling factor of $h_0 = 0.25$. The fixed KDE bandwidth was chosen heuristically following Scott's rule [17] for an average cluster. The KDE smoothing allows our model inputs to be more robust to fluctuations in sample richness, a desirable property for galaxy-based cluster observations. Once smoothed, we create input images by evaluating each cluster's KDE-estimated PDF at regular intervals across the dynamical phase space. 1D inputs are generated querying v_{los} PDFs at 48 evenly-spaced points along the range $v_{\mathrm{los}} \leq v_{cut}$. 2D inputs are derived from joint $\{v_{\mathrm{los}}, R_{\mathrm{proj}}\}$ PDFs evaluated on a regular grid of 48×48 points spanning the area defined by $|v_{\mathrm{los}}| \leq v_{cut}$ and $0 \leq R_{\mathrm{proj}} \leq R_{\mathrm{aperture}}$.

To model projection-based selection effects, we include an alternative analysis of the Coma system using model training catalogs which are re-weighted to fit Coma's radial profile. To accomplish this, we first estimate the radial distribution of Coma galaxies from our observational sample using a KDE. Then, in the KDE estimation step for our mock clusters, we systematically upweight the importance of galaxies with the same R_{proj} as overdense regions of Coma's profile, and likewise downweight the importance of galaxies in scarce Coma regions. Following our Coma measurements, this tends to emphasize the impact of galaxies within $R_{\mathrm{proj}} \leq 1\ h^{-1}\mathrm{Mpc}$ and diminish the impact of outer galaxies on the dynamical phase space distribution. This method ensures that the resulting input images have approximately the same radial distribution as the Coma system.

Historical Mass Estimates of the Coma Cluster

The virial mass of the Coma system has been presented at various definitions throughout history. In order to make a reliable comparison of our methods with these estimates, we need first enforce a strict definition and unit convention of the halo virial mass. Throughout this work, we refer to the virial mass as M_{200c}, the mass enclosed within a spherical overdensity of 200 times the critical density of the universe, defined as $\rho_c = 3H^2(z)/8\pi G$. We present these virial masses in units of $h^{-1}\mathrm{M}_\odot$, where h is the dimensionless Hubble constant defined as $H_0 = 100h\ \mathrm{Mpc}^{-1}\,\mathrm{s}^{-1}\,\mathrm{km}$. The virial radius R_{200c} is defined as the comoving radius of the spherical overdensity enclosing M_{200c} and presented in units of h^{-1}Mpc. Due to its low redshift, mass determination of the Coma cluster should be relatively insensitive to cosmological parameters, but for completeness we assume a generic, flat ΛCDM cosmology ($h = 0.7$, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$).

When historical estimates of Coma's mass are not presented according to the M_{200c} definition used here, we convert them to our definition assuming an Navarro, Frenck, & White (NFW) profile [38]. Observational evidence suggests that this assumption is sound for the Coma system [7]. Below, we carefully cite each historical mass estimate used in our comparison (Figure 3) and detail how we convert each to the mass definition used in this work. Where explicit NFW fits are unavavailable [6, 5, 8], we assume a fiducial concentration for Coma of $c_{200} \simeq 7$ which is statistically consistent with all previous measurements [7, 32, 3, 4] to perform our conversions.

[6] derives the mass of the Coma cluster from the velocity dispersion of its galaxies using the virial theorem. Their primary results assume that Coma is spherically symmetric and that the radial distribution of galaxies exactly traces that of the dark matter. Under these assumptions, they present a mass estimate of $1.9 \times 10^{15}\ h_{50}^{-1}\,\mathrm{Mpc}$ with 30% uncertainty (15% error) within a radius of $5.4\ h_{50}^{-1}\,\mathrm{Mpc}$, where h_{50} is defined as $H_0 = 50h_{50}\ \mathrm{Mpc}^{-1}\,\mathrm{s}^{-1}\,\mathrm{km}$. However, when they relax their assumption of the matter radial distribution, they report that models with masses between $(0.6 - 5) \times 10^{15}\ h_{50}^{-1} M_\odot$ are consistent with available data. For completeness, we include both estimates in our analysis. Using our fiducial Coma concentration, we find that these mass estimates convert to $M_{200c} = (0.7 \pm 0.1) \times 10^{15}\ h^{-1}\mathrm{M}_\odot$ and $M_{200c} = (0.7^{+1.5}_{-0.4}) \times 10^{15}\ h^{-1}\mathrm{M}_\odot$ for strong and weak assumptions on the dark matter distribution, respectively.

[5] assumes that Coma is under hydrostatic equilibrium and that its dark matter mass distribution follows that of its optical light. Using X-ray imaging data from the Einstein Observatory, the author infers a Coma mass of $(1.84 \pm 0.24) \times 10^{15}\ h_{50}^{-1} M_\odot$ within a radius of $5\ h_{50}^{-1}\,\mathrm{Mpc}$. However, when a larger class of dark matter distributions is considered, the models suggest total mass can potentially be within $(1.1-3.0)\times 10^{15}\ h_{50}^{-1}\mathrm{M}_\odot$. Again, we consider both estimates in our analysis, resulting in converted masses of $M_{200c} = 0.66^{+0.11}_{-0.10} \times 10^{15}\ h^{-1}\mathrm{M}_\odot$ and $M_{200c} = 0.66^{+0.54}_{-0.31} \times 10^{15}\ h^{-1}\mathrm{M}_\odot$, respectively, using our fiducial concentration for Coma.

Using new spectroscopy of Coma galaxies from the FAST spectrograph at Whipple Observatory, [7] determines the mass profile of the Coma cluster from the shape of caustics of the galaxy distribution in redshift space. The mass profiles inferred in this work closely match an NFW profile with a scale radius of $r_s = 0.192 \pm 0.035\ h^{-1}$Mpc and a mass of $(1.44 \pm 0.29) \times 10^{15}\ h^{-1}\mathrm{M}_\odot$ within a radius of $5.5\ h^{-1}$Mpc. Using this scale radius estimate, we arrive at a Coma mass of $M_{200c} = 0.76^{+0.20}_{-0.19} \times 10^{15}\ h^{-1}\mathrm{M}_\odot$.

[32] infers the virial mass of the Coma cluster via arguments deriving from Jeans equations. Using a density contrast of $\Delta = 102$, the authors report a mass of $(1.4 \pm 0.4) \times 10^{15}\ h_{70}^{-1}\ \mathrm{M}_\odot$ within a radius of $2.9\ h_{70}^{-1}\,\mathrm{Mpc}$ and a concentration of $c_{102} = 9.4$. This converts to our mass definition as $M_{200c} = (0.83 \pm 0.24) \times 10^{15}\ h^{-1}\mathrm{M}_\odot$.

[3] and [4] reconstruct the Coma cluster mass through weak lensing signal using data from Data Release 5 of the Sloan Digital Sky Survey (SDSS) and deep exposures from the Canada France

Hawaii Telescope (CFHT), respectively. These two papers report estimates of Coma's virial mass at $M_{200c} = 1.88^{+0.65}_{-0.56} \times 10^{15}\ h^{-1}\mathrm{M}_\odot$ and $M_{200c} = 5.1^{+4.3}_{-2.1} \times 10^{14}\ h_{70}^{-1}\mathrm{M}_\odot$, respectively. We convert the latter estimate to our standard units as $0.36^{+0.30}_{-0.15} \times 10^{14}\ h^{-1}\mathrm{M}_\odot$. Despite both estimates arising from weak lensing measurements, we find that these measurements are at a minimum $\sim 2.7\sigma$ tension.

The most recent works, [8] and [16], introduce novel methods for determination cluster masses from galaxy kinematics, through explicit modeling of radial velocity profiles and implicit modeling of cluster dynamical distributions, respectively. After testing on simulated catalogs, each paper seeks to validate their novel methodology via predictions of the Coma cluster. [8] presents a spherical overdensity mass of $(9.2 \pm 2.4) \times 10^{14}\ \mathrm{M}_\odot$ at a density contrast of $\Delta = 93.8$. Using our fiducial concentration for Coma, this leads to a mass of $(0.53 \pm 0.14) \times 10^{15}\ h^{-1}\mathrm{M}_\odot$. [16] predicts $\log_{10}\left[M_{200c}\ (h^{-1}\mathrm{M}_\odot)\right] = 14.84 \pm 0.11$ for the Coma cluster.

Data Availability

The MDPL2 Rockstar catalog is made publicly available through the CosmoSim database at `https://www.cosmosim.org/`. The UniverseMachine catalogs of the MDPL2 simulation that support the findings of this study are available from Peter Behroozi and Andrew Hearin but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Peter Behroozi and Andrew Hearin. The Uchuu DR1 Rockstar halo catalog is available via the Skies and Universes website (`http://skiesanduniverses.iaa.es/`). The Uchuu UniverseMachine (Uchuu-UM) galaxy catalogs will be soon available via the Skies and Universes website (`http://skiesanduniverses.iaa.es/`). The sky positions, spectroscopic redshifts, and stellar masses are made available from SDSS DR12 (`https://www.sdss.org/dr12/spectro/galaxy_portsmouth/`). All mock cluster observation catalogs, trained ML models, and processed Coma observation catalogs generated during the current study are available from the corresponding author upon reasonable request.

Code Availability

All ML models are built in Python using the *Tensorflow* framework (`https://www.tensorflow.org/`). Code for generating the mock cluster observations, training the ML models, and running our inference pipeline is made available at `https://github.com/McWilliamsCenter/halo_cnn`. Jupyter notebooks detailing specific training and data analysis procedures are available from the corresponding author upon reasonable request.

Acknowledgements

We greatly appreciate the helpful insight, comments, and paper notes from Arya Farahi during the development of this research work. This work is supported by NSF AI Institute: Physics of the Future, NSF PHY-2020295, and the McWilliams-PSC Seed Grant Program. The computing resources necessary to complete this analysis were provided by the Pittsburgh Supercomputing Center. The CosmoSim database used in this paper is a service by the Leibniz-Institute for Astrophysics Potsdam (AIP). The MultiDark database was developed in cooperation with the Spanish MultiDark Consolider Project CSD2009-00064. We thank Instituto de Astrofísica de Andalucía CSIC, New Mexico State University, and the Spanish research and academic network (RedIRIS) for hosting the

Skies & Universes site for cosmological simulation products as well as Tomoaki Ishiyama, Francisco Prada, Anatoly Klypin, and Manodeep Sinha for contributing the Uchuu DR1 dataset.

Author Contributions Statement

M.H. coordinated the research, wrote the data analysis code, and prepared the manuscript. M.H., M.N., M.M.R, and H.T. designed the experiment and interpreted the results. M.N., M.M.R., and H.T. helped in presentation of the main findings and gave feedback on the manuscript. M.C., A.L., and F.R. gathered, parsed and analyzed observational measurements of the Coma system.

Competing Interests Statement

The authors declare no competing interests.

References

[1] Zwicky, F. Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta* **6**, 110–127 (1933).

[2] Biviano, A. Mazure, A., Casoli, F., Durret, F. & Gerbal, D. (eds) *Our best friend, the Coma cluster (a historical review).* (eds Mazure, A., Casoli, F., Durret, F. & Gerbal, D.) *Untangling Coma Berenices: A New Vision of an Old Cluster*, 1 (1998). `astro-ph/9711251`.

[3] Kubo, J. M. *et al.* The Mass of the Coma Cluster from Weak Lensing in the Sloan Digital Sky Survey. *ApJ* **671**, 1466–1470 (2007).

[4] Gavazzi, R. *et al.* A weak lensing study of the Coma cluster. *Astron. Astrophys.* **498**, L33–L36 (2009).

[5] Hughes, J. P. The Mass of the Coma Cluster: Combined X-Ray and Optical Results. *ApJ* **337**, 21 (1989).

[6] The, L. S. & White, S. D. M. The mass of the Coma cluster. *Astron. J.* **92**, 1248–1253 (1986).

[7] Geller, M. J., Diaferio, A. & Kurtz, M. J. The Mass Profile of the Coma Galaxy Cluster. *Astrophys. J. Lett.* **517**, L23–L26 (1999).

[8] Falco, M. *et al.* A new method to measure the mass of galaxy clusters. *Mon. Not. R. Astron. Soc.* **442**, 1887–1896 (2014).

[9] Allen, S. W., Evrard, A. E. & Mantz, A. B. Cosmological Parameters from Observations of Galaxy Clusters. *Annu. Rev. Astron. Astrophys.* **49**, 409–470 (2011).

[10] Dodelson, S. *et al.* Cosmic Visions Dark Energy: Science. *arXiv e-prints* arXiv:1604.07626 (2016).

[11] Binney, J. & Tremaine, S. *Galactic dynamics* Vol. 13 (Princeton university press, 2011).

[12] Old, L. *et al.* Galaxy Cluster Mass Reconstruction Project - III. The impact of dynamical substructure on cluster mass estimates. *Mon. Not. R. Astron. Soc.* **475**, 853–866 (2018).

[13] Wojtak, R. *et al.* Galaxy Cluster Mass Reconstruction Project - IV. Understanding the effects of imperfect membership on cluster mass estimation. *Mon. Not. R. Astron. Soc.* **481**, 324–340 (2018).

[14] Ho, M. *et al.* A Robust and Efficient Deep Learning Method for Dynamical Mass Measurements of Galaxy Clusters. *ApJ* **887**, 25 (2019).

[15] Ho, M., Farahi, A., Rau, M. M. & Trac, H. Approximate Bayesian Uncertainties on Deep Learning Dynamical Mass Estimates of Galaxy Clusters. *ApJ* **908**, 204 (2021).

[16] Kodi Ramanah, D., Wojtak, R., Ansari, Z., Gall, C. & Hjorth, J. Dynamical mass inference of galaxy clusters with neural flows. *Mon. Not. R. Astron. Soc.* **499**, 1985–1997 (2020).

[17] Scott, D. W. *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons, 2015).

[18] Gal, Y. & Ghahramani, Z. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158* (2015).

[19] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).

[20] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).

[21] Gal, Y. & Ghahramani, Z. Balcan, M. F. & Weinberger, K. Q. (eds) *Dropout as a bayesian approximation: Representing model uncertainty in deep learning.* (eds Balcan, M. F. & Weinberger, K. Q.) *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48 of *Proceedings of Machine Learning Research*, 1050–1059 (PMLR, New York, New York, USA, 2016). URL `https://proceedings.mlr.press/v48/gal16.html`.

[22] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014). URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[23] Kodi Ramanah, D., Wojtak, R. & Arendse, N. Simulation-based inference of dynamical galaxy cluster masses with 3D convolutional neural networks. *Mon. Not. R. Astron. Soc.* **501**, 4080–4091 (2021).

[24] Ishiyama, T. *et al.* The Uchuu simulations: Data Release 1 and dark matter halo concentrations. *Mon. Not. R. Astron. Soc.* **506**, 4210–4231 (2021).

[25] et al., H. A. In prep (2022). Unpublished.

[26] Klypin, A., Yepes, G., Gottlöber, S., Prada, F. & Heß, S. MultiDark simulations: the story of dark matter halo concentrations and density profiles. *Mon. Not. R. Astron. Soc.* **457**, 4340–4359 (2016).

[27] Behroozi, P., Wechsler, R. H., Hearin, A. P. & Conroy, C. UNIVERSEMACHINE: The correlation between galaxy growth and dark matter halo assembly from z = 0-10. *Mon. Not. R. Astron. Soc.* **488**, 3143–3194 (2019).

[28] van Dokkum, P. G. & van der Marel, R. P. The Star Formation Epoch of the Most Massive Early-Type Galaxies. *ApJ* **655**, 30–50 (2007).

[29] Alam, S. *et al.* The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. *Astrophys. J. Suppl. Ser.* **219**, 12 (2015).

[30] Abell, G. O., Corwin, J., Harold G. & Olowin, R. P. A Catalog of Rich Clusters of Galaxies. *Astrophys. J. Suppl. Ser.* **70**, 1 (1989).

[31] Maraston, C. Evolutionary population synthesis: models, analysis of the ingredients and application to high-z galaxies. *Mon. Not. R. Astron. Soc.* **362**, 799–825 (2005).

[32] Łokas, E. L. & Mamon, G. A. Dark matter distribution in the Coma cluster from galaxy kinematics: breaking the mass-anisotropy degeneracy. *Mon. Not. R. Astron. Soc.* **343**, 401–412 (2003).

[33] Planck Collaboration *et al.* Planck 2013 results. XVI. Cosmological parameters. *Astron. Astrophys.* **571**, A16 (2014).

[34] Villaescusa-Navarro, F. *et al.* Robust marginalization of baryonic effects for cosmological inference at the field level. *arXiv e-prints* arXiv:2109.10360 (2021).

[35] Bishop, M. A. Mixture density networks. *Technical Report NCRG/94/004, Aston University* (1994). URL `https://publications.aston.ac.uk/id/eprint/373/1/NCRG_94_004.pdf`.

[36] Planck Collaboration *et al.* Planck 2015 results. XXIV. Cosmology from Sunyaev-Zeldovich cluster counts. *Astron. Astrophys.* **594**, A24 (2016).

[37] Behroozi, P. S., Wechsler, R. H. & Wu, H.-Y. The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. *ApJ* **762**, 109 (2013).

[38] Navarro, J. F., Frenk, C. S. & White, S. D. M. A Universal Density Profile from Hierarchical Clustering. *ApJ* **490**, 493–508 (1997).