Interpretable Probabilistic Modelling Approach to Correct Hubble Flow around Galaxy Clusters

Dario Barone
d.barone@bham.ac.uk
University of Birmingham
School of Computer Science
Birmingham, UK

Marco Canducci
m.canducci@bham.ac.uk
University of Birmingham
School of Computer Science
Birmingham, UK
INAF
Osservatorio Astronomico di Padova
Padova, Italy

Rory Smith
rory.smith@usm.cl
Universidad Técnica Federico Santa María
Departamento de Física
Valparaíso, Chile

Peter Tino
p.tino@bham.ac.uk
University of Birmingham
School of Computer Science
Birmingham, UK

Abstract

Cosmological theories rely on understanding the three-dimensional structure of the Universe on large scale. Astrophysical observations of galaxy clusters are limited to measuring a two-dimensional position in the sky and a one-dimensional velocity along the line of sight, inherently losing part of the relevant information content. Commonly, the third spatial dimension is reconstructed by applying a deterministic scaling of the line-of-sight velocity that accounts for the expansion of the Universe (Hubble flow). However, due to gravitational effects, this transformation does not hold in the vicinity of massive structures, such as galaxy clusters. In this work, we propose an interpretable probabilistic model aiming to reconstruct the three-dimensional radial distance of an observed galaxy from its nearest galaxy cluster. Our model is completely transparent, and it explicitly embeds astrophysical prior knowledge to account for the non-trivial dynamics that are superimposed on the Hubble flow. The three-dimensional radial distance is reconstructed in the form of a posterior distribution, conditioned on the positional and kinematical observations, which enables coherent quantification of the predicted uncertainties. We train our model on data from cosmological simulations, and then use it to reconstruct the actual cluster neighbourhoods. We empirically show that our model is more accurate in recovering the radial profiles of several simulated clusters, for which the ground truth is known, compared to the standard techniques. Crucially, our model-driven approach also uncovers features of simulated cosmological data that are not yet fully understood from a theoretical perspective, suggesting new directions of research. This tool is designed for use by astronomical observers, offering improvements over existing methods and opening up possibilities for further scientific discoveries.

This work is licensed under a Creative Commons Attribution 4.0 International License.
KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3819009

CCS Concepts

• **Mathematics of computing** → **Density estimation**; • **Computing methodologies** → **Modeling methodologies**; • **Applied computing** → **Astronomy**.

Keywords

probabilistic modelling; astrophysics; galaxy clusters; Bayesian model; Hubble flow; observational correction

ACM Reference Format:
Dario Barone, Marco Canducci, Rory Smith, and Peter Tino. 2026. Interpretable Probabilistic Modelling Approach to Correct Hubble Flow around Galaxy Clusters. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3819009

1 Introduction

According to the standard cosmological model, we live in an expanding Universe where, on large cosmological scales, every point expands away from every other point with a velocity v that depends only on the distance d separating them. The expansion follows the so-called "Hubble-Lemaitre law" $v = H_0\ d$, where H_0 is Hubble's constant. Using the observed redshift of a galaxy's spectra, we can measure its line-of-sight velocity v with respect to us, and use the Hubble-Lemaitre law to calculate the distance d to the galaxy along the line of sight. In this way, it is possible to build large three-dimensional distributions of galaxy locations within cosmological volumes and map out the large-scale structure of the Universe.

However, in certain circumstances, the Hubble-Lemaitre law is not fully valid. This is particularly the case in the vicinity of galaxy clusters, structures with vast masses (10^{14-15}M$_\odot$, where M$_\odot$ denotes the unit of solar mass), often containing thousands of individual galaxies. The deep gravitational potential of the cluster pulls on neighbouring galaxies and slows their expansion below the velocities predicted by the Hubble flow. This effect is significant out to large distances from a cluster (up to tens of Megaparsecs

[1]) as shown in Figure 1. If only the Hubble flow is assumed, the reconstruction of the three-dimensional distribution of galaxies from observation gets distorted. For example, a hypothetical galaxy located at $r_{\text{true}} = 6$ Mpc from a cluster of mass $\approx 5 \times 10^{14}$ M$_\odot$ would be de-projected at an average distance $r_{\text{HF}} \approx 0.32$ Mpc, incurring in a relative error of $|r_{\text{HF}}-r_{\text{true}}|/r_{\text{true}} \approx 95\,\%$.

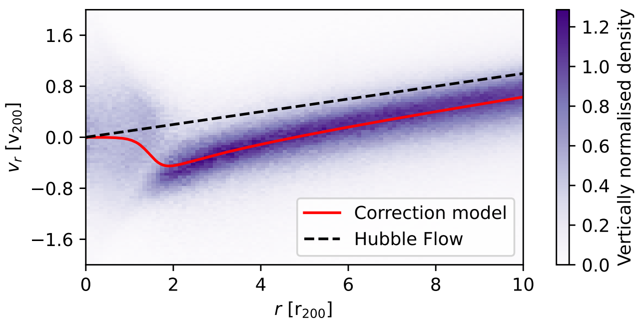

Figure 1: Radial velocity v_r from the cluster centre as a function of the distance r from the cluster centre of galaxies in the simulated data. The dashed line represents the Hubble flow. The red solid line is the mean radial velocity predicted by our model, accounting for the gravitational pull exerted by the galaxy cluster.

Moreover, the motion of a galaxy is not only caused by the expansion of the Universe or the gravitational attraction towards the major cluster structure. Indeed, it is also gravitationally influenced by smaller, nearer sub-structures (such as galaxy groups), or by local interactions with nearby galaxies, or by the conservation of the momentum that the galaxy has had since the birth of the Universe, or by the drag exerted by intergalactic gas. All these effects are usually interplaying, making the deterministic modelling of the motion of a specific galaxy infeasible (note the dispersion of the velocity in Figure 1). If these elements are not taken into account when calculating the distance along the line of sight, significant distortions may be introduced [14].

A further impediment to the three-dimensional reconstruction is the impossibility of observing all six dynamical quantities (three positions and three velocities) that describe a galaxy's motion. In fact, only two positions (in the sky plane) and one velocity (along the line of sight) can be measured, causing a loss of information that could never be physically avoided.

Our goal is to create a tool able to reproduce the radial distribution of galaxies, up to a few tens of Megaparsecs around massive galaxy clusters, by correcting for the distortion effects due to both the gravitational pull of the large-scale structures and the shorter-scale dynamical interactions affecting the peculiar motion. A user-friendly tool of this type would be of enormous value to the astronomical community. Indeed, there is currently a surge of interest in studying the surrounding of clusters, with dedicated deep spectroscopic surveys actively measuring galaxy spectra around multiple nearby clusters, both in the northern [17] and southern hemisphere [26]. The growth and evolution of galaxy clusters and their member galaxies are known to be directly influenced by their surrounding large-scale structure [7, 9, 10, 13, 16]. The ability to create more accurate three-dimensional representations of the vicinities of observed galaxy clusters will provide crucial insights also into the fields of cluster evolution [30], large-scale structure formation [3, 8, 22], and cosmology [4, 29], currently intensively studied in the astrophysics community.

The two sources of distortion are known phenomena affecting cosmological observations. However, common correction techniques focus either on the large-scale distortion [18] or on the peculiar motion inside clusters and groups [27]. Only a few works account for both effects and apply the correction on the single-galaxy basis [25]. However, the statistical dispersion is never taken into consideration.

Related works in the field of galaxy cluster dynamics have adopted a probabilistic modelling approach. In [11], assuming spherical symmetry, they estimate the mass of a galaxy cluster by analytically modelling the average deviation from the Hubble flow that galaxies show in the outer regions. Other approaches, with the same goal, discriminate between the dynamics along the radial and tangential directions [32], and take into account differences in the galaxies orbital behaviour (infalling vs back-splash) [12]. The orbital classification itself is a subject of study that has been addressed with probabilistic models built in the observation space [2]. All these approaches are characterised by addressing the forward problem, i.e. by building the model in the six-dimensional space and by varying the target parameters until the projected quantity of interest matches the measurements.

Fewer works aim to solve the inverse problem, namely reconstructing the full space from the projected quantities. In [24], the Richardson-Lucy algorithm is used to recover the three-dimensional gravitational potential of the cluster from the line-of-sight velocity. Similarly, the two-point correlation function of line-of-sight velocity has been used to correct dynamical distortions for the galaxies within the cluster [28]. The three-dimensional shape of clusters can also be reconstructed using different observables than the aforementioned dynamical quantities [19].

We approach the problem by constructing a *transparent probabilistic model* that estimates the distribution of the three-dimensional distance of a galaxy from the cluster centre, given its projected distance in the sky plane from the cluster centre, its line-of-sight velocity relative to the cluster and the mass of the cluster of reference. We leverage Bayes' theorem to solve the inverse problem, so that we can define the likelihood in the full six-dimensional space and fit its parameters on detailed cosmological simulations. This enables us to embed the astrophysical knowledge about both the Hubble flow correction and the dynamical perturbations in our model, through explicit functional forms that depend on interpretable and physically-meaningful free parameters. Our probabilistic formulation allows to preserve all the possible information, encoded in the posterior distribution, in a context where a deterministic de-projection of the observables is impossible. In addition, by providing a full distribution, we provide a coherent treatment of the uncertainties along the reconstructed dimension.

Our methodology is the first, to the best of our knowledge, which (a) addresses the de-projection of the three-dimensional structure of (b) both the galaxy cluster and its surrounding, (c) with a continuous

[1]The *parsec* is a distance unit of measure. The Milky Way's diameter is ≈ 27 kpc.

model, (d) that is physically informed and transparent, (e) naturally incorporates the treatment of uncertainties and (f) relies jointly on both the positional and the velocity observables.

2 Methodology

Given the goal of the project, we fix a comoving reference frame with origin at the galaxy cluster centre, with the axis orientated such that x, y describe the projection plane on the sky and z is orientated along the line of sight. The vectors $\vec{r}$ and $\vec{v}$ represent, respectively, the position and velocity of a galaxy in this reference frame. Notation-wise, for a vector $\vec{a} \in \mathbb{R}^3$, we denote its i-th component by the subscript a_i, the corresponding unit vector by $\hat{a}$ and its norm induced by the dot product $a = \|\vec{a}\| = \sqrt{\vec{a} \cdot \vec{a}}$.

Consider a galaxy observed at the position (x, y) in the sky with the line-of-sight velocity v_z. The output of our methodology is the posterior distribution over the actual three-dimensional distance r of the galaxy from the cluster centre, given the velocity v_z, the observed projected distance from the cluster centre $R \coloneqq \sqrt{x^2 + y^2}$ and the mass of the galaxy cluster in which the reference frame is centred M:

$$P(r \mid R, v_z, M) = \frac{1}{\mathcal{N}} \, P(R, v_z \mid r, M) \, P(r) \tag{1}$$

where $\mathcal{N}$ is the normalisation factor. The factor $P(r)$ is the empirical prior distribution obtained from the training data distribution, while $P(R, v_z \mid r, M)$ is the likelihood term which encodes the information about gravitational and kinematical effects due to the presence of the cluster.

Given our notation, geometrical considerations lead to

$$R = \|\vec{r} - [\vec{r} \cdot \hat{z}] \, \hat{z}\| \tag{2}$$

$$v_z = \vec{v} \cdot \hat{z} \tag{3}$$

2.1 The Physics Involved

We model the galaxy's motion within and around a galaxy cluster as

$$\vec{v}(r, M) = \vec{u}(r, M) + [H_0 \, r + v_{\mathrm{corr}}(r, M)] \, \hat{r} \tag{4}$$

where H_0 is the Hubble constant, $\vec{u}$ a the noise term and v_{corr} is the velocity correction. The noise $\vec{u}$ is due to local perturbations and can be considered as a zero-mean isotropic random variable. Inside the cluster, $\vec{u}$ is dominated by the orbital motion around the centre, while, on the outskirts, the dominant contribution is due to the reciprocal attraction exerted by nearby galaxies and the momentum conserved since the birth of the Universe. The second term, aligned with the radial direction $\hat{r}$, accounts for the expansion of the Universe (through $H_0\vec{r}$)[2], and for the gravitational pull exerted by the galaxy cluster (through $v_{\mathrm{corr}}\hat{r}$)[3].

The magnitude of the velocity dispersion of galaxies inside the cluster, namely $\langle u^2(r, M) \rangle$ for small r, depends on the cluster mass, and so does the magnitude of the velocity correction $v_{\mathrm{corr}}(r, M)$ due to gravitational attraction. These two dependencies follow approximately[4] the same scaling law, which can be absorbed by a wise choice of coordinates. Therefore, we express positions and velocities in terms of Virial units (explained in Appendix D), commonly used in astrophysics [5], as they provide a scaling relation for distances and velocities that has the same dependency on the cluster mass. The definition of the Virial units depends on an over-density parameter Δ (see Appendix D for details), which in this work is set to $\Delta = 200$, following the common practice in observational cosmology. The Virial radius r_{200} sets the scale for the cluster size, and the Virial velocity v_{200} sets the scale for kinematics inside it (as discussed in Subsection D.2). Conveniently, the Hubble constant H_0, which is a constant in physical units, remains a constant also when expressed in Virial units (proof in Subsection D.1). This is particularly important because it allows us to stack together data related to clusters with different masses and generalise our model more easily. The only residual dependency on mass lies in the noise term outside the cluster, namely $\vec{u}(r, M)$ for $r \gtrsim \mathrm{r}_{200}$.

We point out that stacking together data from different clusters implies destroying the information about the three-dimensional morphology, and inevitably implies the spherical symmetry assumption. Even though the neighbourhoods of galaxy clusters are known not to be spherically symmetric, the shape of a generic observed environment is not known a-priori. Therefore, assuming spherical symmetry allows us to generalise our model at the expense of some accuracy loss, due to the non-spherical nature of the clusters.

2.2 The Likelihood

Given eqs. (2),(3), the two random variables R and v_z are not independent, thus the likelihood $P(R, v_z \mid r, M)$ cannot be factorised. However, they are *conditionally independent* if conditioned over both r and $\hat{z}$. The likelihood term can be written as

$$P(R, v_z \mid r, M) = P(v_z \mid R, r, M) \, P(R \mid r) \tag{5}$$

so factorised in a geometrical component $P(R \mid r)$ and a dynamical component $P(v_z \mid R, r, M)$.

The geometrical factor can be computed considering that it represents the probability of finding, at projected distance R, a point that is uniformly distributed on a sphere of radius r. The cumulative probability distribution $P(R' \leq R \mid r)$ represents the fraction of area of the sphere of radius r that is covered by the two spherical caps of radius R (one cap for each z-semi-axis). The area of one of these spherical caps is $A = 2\pi r \left[r - \sqrt{r^2 - R^2} \right]$, leading to $P(R' \leq R \mid r) = 1 - \sqrt{r^2 - R^2}/r$. Taking the derivative with respect to R, we obtain

$$P(R \mid r) = \frac{R}{r \, \sqrt{r^2 - R^2}} \tag{6}$$

Another, more formal, derivation of this distribution is addressed in Appendix C.

The geometrical component encodes the information of the orientation of the reference frame $\hat{z}$, since eq. (2) relates $\hat{z}$ implicitly to r and R. However, eq. (2) is not immediately invertible, as $R(\vec{r}, \hat{z}) = R(\vec{r}, -\hat{z})$. By defining $\varsigma \coloneqq \mathrm{sgn}(\vec{r} \cdot \hat{z})$, the implicit equation $\hat{z}(r, R, \varsigma)$ is well defined. When projected along the radial direction, this implicit equation assumes the explicit form

[2] The expansion happens in all directions equally. Since our reference frame is centred in the cluster, our coordinates are affected along the radial direction.

[3] The gravitational attraction of the cluster acts on the radial direction only, as gravity is a central force.

[4] If the density profile of different galaxy clusters is consistent, the scaling laws are exactly the same. Despite, in reality, different clusters show a slightly different density profile, the Navarro-Frenk-White model [20, 21] is commonly considered a solid generalisation within few Megaparsecs from the cluster centre.

$\hat{r} \cdot \hat{z}(r, R, \varsigma) = \varsigma\sqrt{1 - R^2/r^2}$. Thus, substituting eq. (4) in eq. (3) we obtain

$$v_z(r, M) = u_z(r, M) + \varsigma \sqrt{1 - \frac{R^2}{r^2}} \left[H_0\, r + v_{\mathrm{corr}}(r)\right] \tag{7}$$

It is then easy to deduce that the probability distribution $P(v_z \mid u_z, v_{\mathrm{corr}}, \varsigma, R, r, M)$ is a Dirac's delta distribution (denoted by δ_{D}) with the implicit form of eq. (7) as argument. Among the conditioning variables, u_z, v_{corr} and ς are dummy variables, so they can be marginalised out to obtain the dynamical component of the likelihood

$$P(v_z \mid R, r, M) = \int \mathrm{d}v_{\mathrm{corr}}\, P(v_{\mathrm{corr}} \mid R, r, M) \int \mathrm{d}u_z\, P(u_z \mid R, r, M) \cdot \cdot \int \mathrm{d}\varsigma\, P(\varsigma)\, P(v_z \mid u_z, v_{\mathrm{corr}}, \varsigma, R, r, M) \tag{8}$$

where $P(\varsigma) = \delta_{\mathrm{D}}(\varsigma-1)/2 + \delta_{\mathrm{D}}(\varsigma+1)/2$ because of the reflectional symmetry of the problem with respect to the plane $z = 0$.

The last two elements to be determined in the likelihood terms are the two probability distributions over u_z and v_{corr} in eq. (8). To properly model them, we project eq. (4) on the radial direction $\vec{v} \cdot \hat{r}$ to obtain

$$v_r(r, M) = u_r(r, M) + H_0\, r + v_{\mathrm{corr}}(r) \tag{9}$$

Given that $\vec{u}$ is isotropically distributed, we have that $P(u_z) = P(u_r)$. Therefore, modelling the radial component is sufficient to fully determine the likelihood distribution.

Similarly to [32], the probability distribution $P(v_r)$ is modelled as the skewed t (ST) distribution[5]

$$P(v_r) = \mathrm{ST}(v_r; \mu, \sigma, \nu, \lambda) \tag{10}$$

where μ is the average, σ is the standard deviation, ν is the degrees of freedom and λ is the skewness parameter. The explicit form of the ST distribution can be found in Appendix E.

As a sum of the two random variables u_r and v_{corr}, the probability distribution of v_r is obtained by the convolution

$$P(v_r) = \int_{\infty}^{\infty} \mathrm{d}u_r\, P(u_r)\, P(v_{\mathrm{corr}} = v_r - H_0\, r - u_r) \tag{11}$$

Given the assumption of isotropy and zero mean of the stochastic noise $\vec{u}$, all the information content about the offset and asymmetry of $P(v_r)$ must be contained in $P(v_{\mathrm{corr}})$. In particular, $P(v_{\mathrm{corr}})$ can be reconstructed by deconvolving eq. (11) once the explicit form of $P(u_r)$ is determined.

This can be attained by again leveraging the isotropy assumption. According to the model in eq. (4), the only contribution to the velocity that is not along the radial direction is due to $\vec{u}$. Placing a Cartesian reference frame on a galaxy with velocity $\vec{v}$, orientated so that one axis is $\hat{r}$ and the other two $(\hat{e}_1, \hat{e}_2)$ define the tangential plane, we have $\vec{v} \cdot \hat{e}_i = v_i = u_i$, with $i = 1, 2$. Since $\vec{u}$ is isotropic, $P(u_r) = P(u_1) = P(u_2)$. Defining the tangential speed as $u_\Omega := \sqrt{u_1^2 + u_2^2}$, the probability distribution $P(u_\Omega)$ is related to $P(u_r)$ as

$$P(u_\Omega) = P(\|(u_1, u_2)\|) \quad \text{with } u_{1,2} \sim P(u_r) \tag{12}$$

This relation is particularly important because it allows us to unravel the convoluted distributions in the radial direction and to determine both the appropriate functional form and the parameter values for $P(u_r)$. We also used this relation to regularise the parameter fitting procedure, as explained in the following section. The data shows that the choice of a t distribution for $P(u_r)$ with $\mu = 0$ and $\lambda = 0$ leads to a distribution $P(u_\Omega)$ that mimics well the empirical distribution of the tangential speed, as shown in Figure 2.

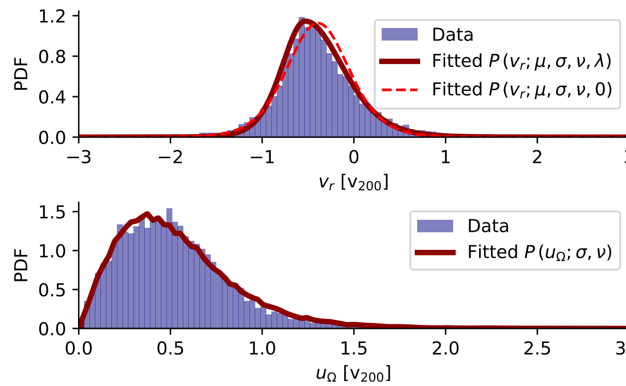

Figure 2: Comparison between the fitted distributions in eqs. (10)–(12) **and the respective data, in the bin with** $M \in [2.1, 3.71] \times 10^{14} \mathrm{M}_\odot$ **and** $r \in [2, 2.5]\mathbf{r_{200}}$**. The dashed red line in the top panel shows the distribution when the skewness parameter is forced to be** 0**.**

2.3 Model Fitting

The dataset used to fit the model parameters is obtained from the Uchuu simulation[6], a large high-resolution cosmological simulation [1, 15] that creates a complete mock Universe within a $2\,\mathrm{Gpc}\,\mathrm{h}^{-1}$ volume. First, from the simulation we select 1000 random clusters with a uniform probability, to preserve the statistical representation that each cluster has in real observational data (more massive clusters are rarer). We also ensure these clusters are central halos, meaning they are the dominant structure in their local environment, and that they are not enclosed within a more massive halo. For each cluster, we associate all the galaxies with distance up to $20\,\mathrm{r}_{200}$ from the cluster centre, and we augment the dataset by applying 64 random rotations of the Cartesian reference frame to generalise with respect to the point of view. At the end, we combine the data for all the clusters and their associated galaxies to obtain ≈ 8.5 million galaxies. The dataset is composed of one entry per galaxy, with the relative positions and velocities in the respective cluster-centred reference frame, together with the information about the cluster mass.

Considering eq. (10), a unique set of parameters for modelling the radial velocity cannot be assumed for all r and M, because of the dependencies in eq. (9). It is natural to look for estimates that are only locally (in the (r, M) space) valid, as is also done in [12, 32]. Therefore, the galaxy data is first binned into 40 linear bins in r and 5 logarithmic bins in M[7]. The binning scheme is chosen to

[5] A different probabilistic model for v_r was used in [12], but our formulation contains physically interpretable parameters.

[6] For data availability, see Section 6.

[7] Since massive clusters are rarer, there is an imbalance in the number of clusters in each mass bin. However, the number of galaxies associated with massive clusters is much higher, and it compensates for the bin imbalance.

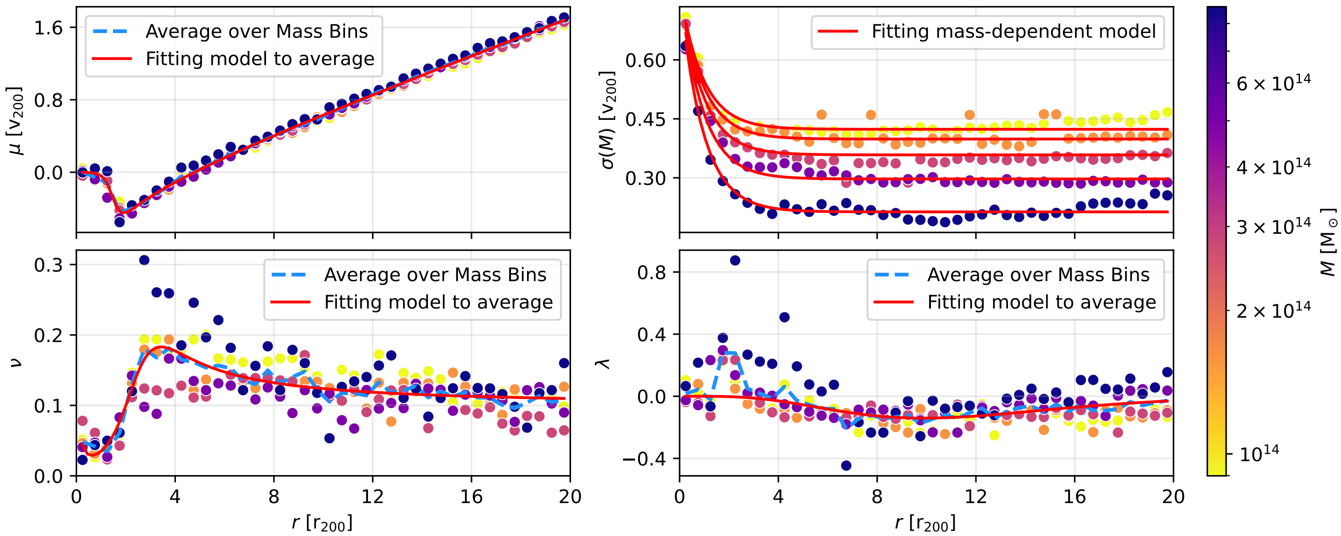

Figure 3: Fitting results for the parameter of the radial velocity distribution (eq. (10)). The dots represent the parameter values obtained by the independent fits in each (r, M)-bin. The blue dashed line represents the average for fixed r across different M. The red solid line represents the functional forms in eqs. (13)–(16) fitted to the average trend.

be fine enough to capture the variation of the data features and coarse enough to ensure a statistically reliable sample in each bin. The model (10) is fitted in each bin independently (the upper plot in Figure 2 shows the result within one bin), and then smoothed out by finding closed-form dependencies of the parameters of the ST distribution on (r, M). Thus, we obtain a smooth conditional density $P(v_r \mid r, M)$.

In Figure 3, each point represents the independent fit in the respective (r, M) bin. While the dependency of the various parameters on the distance from the cluster is evident, the dependency on the mass of the cluster is weak in all cases but for σ. This effect is expected and coherent with the reason behind the use of the Virial units, as discussed in Subsection 2.1.

The fitting of the ST distribution (eq. (10)) on the radial component of the training data was found to be unstable for σ and ν over consecutive bins in r. In fact, although the shape of the ST distribution is affected by ν alone, its width is determined both by σ and ν, leading to identifiability problems. To mitigate this, we opted for a simultaneous fit of both $P(v_r; \mu, \sigma, \nu, \lambda)$ and $P(u_\Omega; \sigma, \nu)$, as shown in Figure 2 (note that the values of σ, ν are the same in both distributions). Since $P(u_\Omega)$ has a higher sensitivity to the two parameters, the simultaneous fit allows to break the identifiability issue and stabilise the results of the independent fits.

The functional forms used to smooth the binned structure are

$$\mu^*(r) = \frac{1}{1+\exp\left(-\tau_\mu\left[r-\rho_\mu\right]\right)}\left[H_0\, r + \frac{\alpha_\mu + \kappa_\mu\, r^{\gamma_\mu}}{r^{\gamma_\mu}-\beta_\mu}\right] \tag{13}$$

$$\sigma^*(r, M) = \kappa_\sigma(M) + \left[\alpha_\sigma - \kappa_\sigma(M)\right]\exp\left(-\tau_\sigma\, r\right) \tag{14}$$

$$\nu^*(r) = \left[\frac{1}{1+\exp\left(-\tau_\nu\left[r-\rho_\nu\right]\right)}\,\frac{\alpha_\nu + \kappa_\nu\, r^{\gamma_\nu}}{r^{\gamma_\nu}-\beta_\nu}\right]^{-1} \tag{15}$$

$$\lambda^*(r) = \alpha_\lambda\, r^{\gamma_\lambda}\exp\left(-\tau_\lambda\, r\right) \tag{16}$$

Equation (10) together with eqs. (13)–(16), specify a transparent and interpretable radial velocity model $P(v_r \mid r, M)$.

Most of the parameters in these equations have an actual physical meaning, and the functional forms themselves are inspired by physical considerations. Eq. (16) is the only equation that is completely data-driven, because the physical nature of the skewness in the radial velocity distribution is uncertain. Eq. (13) is composed of a transition term that smoothly interpolates between the average dynamics within the cluster, given by $\mu^*(r) = 0$ since orbit-dominated, and outside the cluster, given by the term in square brackets. This transition happens at $r = \rho_\mu$, and τ_μ dictates how sharply the average galaxy changes dynamical behaviour. The second term in the square brackets represents the average velocity correction due to the gravitational pull of the cluster. Its shape is inspired to the form $v_{\text{corr}}(r) = \alpha\, [r/r_0]^{-\beta}$ used in [11], but generalised so that it does not diverge in $r = 0$ and it converges to an offset κ_μ for $r \to \infty$. The reason for the first is purely the numerical necessity to have a unique function that fulfils $\lim_{r\to 0}\mu(r) = 0$. The reason for the second is that, moving farther away from the cluster, the mass density reaches the average value of the Universe, leading to an expansion that has the same gradient predicted by the Hubble flow, but with an offset that depends on the accumulated mass. Eq. (15) has a transition term that is semantically equivalent to the one in eq. (13). Indeed, the dynamical behaviour inside the cluster is more Gaussian-like, while on the outskirts the distribution is more fat-tailed, even though the weight in tails decreases for large r. The hyperbola term is data-derived. The functional form of eq. (14) is data-driven as well, but its parameters have a physical interpretation: α_σ is the velocity dispersion inside the cluster, which is known to have a specific relation to the mass of the cluster, and it should assume approximately a constant value in Virial units; $\kappa_\sigma(M)$ is the velocity dispersion outside the cluster, which is the only parameter to have a strong dependency on the mass of the cluster;

τ_σ represents the sharpness of transition between the dynamical behaviour inside and outside the cluster.

The parameter κ_σ shows an interesting behaviour: if one assumes that the stochastic noise affecting the velocity outside of the cluster is independent of cluster mass when expressed in physical units, then expressing it in Virial units should introduce the dependency $\kappa_\sigma \propto M^{-\frac{1}{3}}$ (the reciprocal of the scaling in eq. (32)). However, we find the following empirical relation in the data (see Figure 4):

$$\kappa_\sigma(M) = \alpha_\kappa \exp\left(-\frac{M}{\tau_\kappa}\right) \tag{17}$$

The empirical measures of κ_σ in each mass bin M are obtained by computing the average $1/30 \sum_{5\le r\le 20} \sigma(r, M)$, where $5 \le r \le 20$ is discretised into 30 bins. $r = 5$ has been chosen conservatively such that all the curves $\sigma^*(r, M)$ are in their saturated regime (see Figure 3).

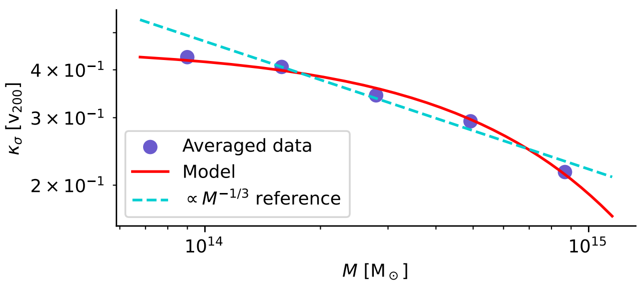

Figure 4: The red solid line represents eq. (4) **fitted on the data (purple dots). Each point is obtained as the average standard deviation of the radial velocity outside the cluster. The dashed blue line is the expected $\kappa_\sigma(M)$ relation in Virial units if κ_σ had no dependency on M in physical units.**

Given the small skewness λ in Figure 3, we opt for an approximation of the model that discards the asymmetry in the distribution of v_r. The top panel in Figure 2 shows the comparison between the asymmetric distribution and the approximated one. Such an approximation has only second-order effects on the accuracy of the posterior distribution, but facilitates a significantly simpler implementation of the model. Even though the non-Gaussian effects in the radial velocity have already been discussed in [31], a linear theory that discards them is commonly adopted in cosmology. Thus, after fitting all the model parameters, we force $\alpha_\lambda = 0$ in eq. (16)[8]. Doing so, all the information content about the velocity noise can be included in $\vec{u}$, and all the information content about the offset must be included in v_{corr}, since $\vec{u}$ has zero mean. This leads to the following model formulation:

$$P(u_r \mid r, M) = \text{ST}(0, \sigma^*(r, M), \nu^*(r), 0) \tag{18}$$

$$P(v_{\text{corr}} \mid r) = \delta_{\text{D}}(\mu^*(r) - H_0\, r) \tag{19}$$

Given the isotropy of $\vec{u}$, eq. (18) sets also $P(u_z \mid r, M)$. Together with eq. (19), they can be inserted in eq. (8) to obtain the dynamical component of the likelihood. This one, together with the geometrical component in eq. (6) constitute the full likelihood. Finally, considering the empirical prior $P(r)$, we obtain the posterior distribution in eq. (1). A schema of the logic of our probabilistic framework is sketched in Appendix A, Figure 9.

[8]Alternatively, one could constrain the model to $\lambda = 0$ and then fit the other parameters. However, we found that, in such case, the tails of the distributions get over-emphasised in the model, with poorer fitting of the parameter values.

3 Results

Once the training is completed, we apply our model to the out-of-sample (test) data. The test dataset is obtained by sampling, with the same strategy, a new set of 1000 clusters from the Uchuu simulation.

We contrast the performance of our model with that of a simple baseline, which assumes only Hubble flow and neither gravitational pull nor complex dynamical effects. According to this baseline model, the reconstructed radial distance from the cluster is:

$$r_{\text{bl}} = \sqrt{R^2 + \left[\frac{v_z}{H_0}\right]^2} \tag{20}$$

We first check how accurately our model can reconstruct the radial profile of the neighbourhood of an observed cluster. To do so, we group galaxies that are associated with the same cluster. Given the i-th galaxy, we define C^c as the set of indices corresponding to the galaxies associated with the c-th cluster. The true number-density radial profile of the c-th cluster is then given by aggregating r^i_{true} for $i \in C^c$, and getting the empirical distribution $n^c_{\text{true}}(r) = P^c_{\text{true}}(r)$. We then use our model to obtain the posterior $P^i_{\text{mod}}(r \mid R^i, v^i_z)$ from eq. (1) for each galaxy. To obtain the model-reconstructed number-density profile, we average over the posteriors $n^c_{\text{mod}}(r) = 1/|C^c| \sum_{i\in C^c} P^i_{\text{mod}}(r \mid R^i, v^i_z)$, where $|C^c|$ is the cardinality of the set. We benchmark against the baseline by computing $n^c_{\text{bl}}(r) = P^c_{\text{bl}}(r)$, where $P^c_{\text{bl}}(r)$ is obtained by aggregating r^i_{bl}, computed according to eq. (20), $\forall i \in C^c$. In Figure 5, we compare the true and reconstructed radial profiles for three random clusters in the test set. Our model is capable of accurately reconstructing the galaxy cluster body (region within $r \lesssim 1$) and following the radial profile in the outer regions, especially for higher masses. More massive clusters have more galaxies associated with them, and the wider statistical sample helps the convergence of our probabilistic model.

The various modes in the data distribution in Figure 5 are due to the presence of sub-structures around the main galaxy cluster. We notice that our model often captures the presence of these sub-structures, but blurs their edges. The motion of a sub-structure as a whole, together with the motion of field galaxies, is well captured, since it is mostly driven by the presence of the main cluster. On the contrary, those galaxies that are members of sub-structures undergo a stronger (than the local average) dynamical perturbation, which misleads our model about de-projecting the observed velocity v_z and, eventually, creates the blurring effect.

A more important benchmark concerns the performance of our model in different regions of the observation space (R, v_z) (often addressed as *phase space* in the literature), shown in Figure 6. In other words, we address the question "what is the accuracy of the model in a given region of the phase space?". To answer, we set up a regular grid in the phase space, where each point has coordinates (R^j, v^k_z). We set two windows of amplitude Δ_R and Δ_v centred in each point of the grid, and we aggregate the galaxies,

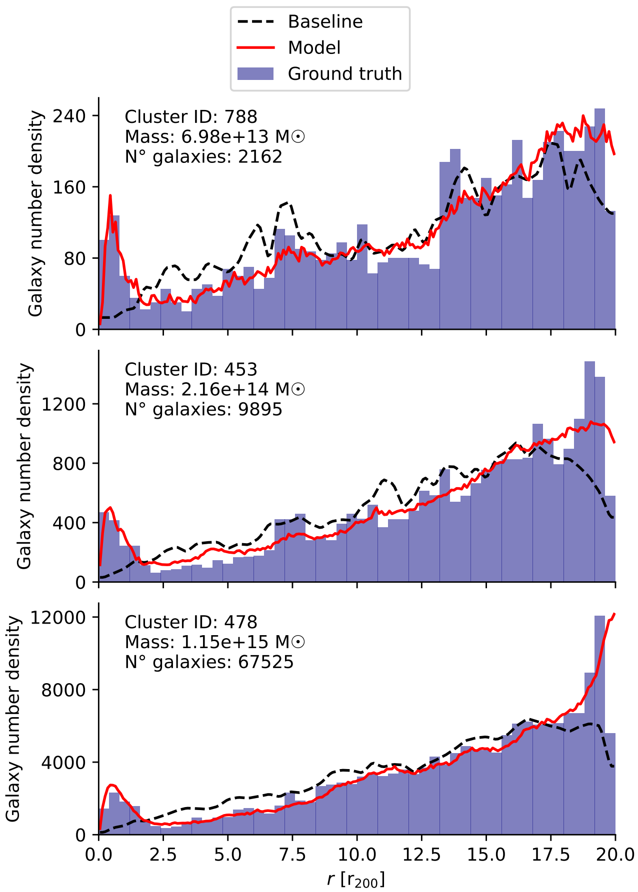

Figure 5: Comparison between the true (purple histogram) and the reconstructed (red solid line for our model, black dashed line for the baseline) three-dimensional radial profile for three clusters in the test set selected at random. The galaxies in these plots are selected such that $r \leq 20\,\mathrm{r}_{200}$.

indexed by i, in sets $S^{j,k} = \{i : |R^i - R^j| \leq \Delta_R \;\wedge\; |v_z^i - v_z^j| \leq \Delta_v\}$. The cardinality $|S^{i,j}|$ of all these sets is shown in Figure 8. Then, we compute the probability distribution of finding a galaxy at distance r in each of these windows. With our model, we obtain $P^{j,k}_{\mathrm{mod}}(r) = {}^1/_{|S^{j,k}|} \sum_{i \in S^{j,k}} P_{\mathrm{mod}}(r \mid R^i, v_z^i)$. To compare with the true and baseline analogous distribution, we compute both $P^{j,k}_{\mathrm{true}}(r)$ and $P^{j,k}_{\mathrm{bl}}(r)$ as the empirical distributions of, respectively, r^i_{true} and r^i_{bl}, with $i \in S^{j,k}$. For some windows, the three probability distributions are plotted in the panels (a)–(g) of Figure 6. The baseline is a bell-shaped distribution centred in $r^{j,k} = \sqrt{(R^j)^2 + (v_z^k/H_0)^2}$, and does not match the true distribution. On the contrary, the posterior distribution accurately recovers the shape of the true distribution in every region of the phase space. The bimodality of the distributions can also be physically explained: there is one mode at $r^{j,k} = R^j$ due to the geometrical component of the likelihood (eq. (6)) and another mode due to the dynamical component (eq. (8)). The two modes have a relative weight that gets naturally adjusted by the model across the phase space. Both these relative weights and the shape of the posterior only derive from physics-constrained transformations and data-driven functional forms of $P(u_r)$ and $P(v_{\mathrm{corr}})$.

In the top-left panel of Figure 6, we show the different performance of our model in different regions of the phase space. We compute the error in each window as $\varepsilon^{j,k}_{\mathrm{mod}} = \int \mathrm{d}r\, |P^{j,k}_{\mathrm{model}}(r) - P^{j,k}_{\mathrm{true}}(r)|$. In the bottom-left panel we show the ratio $\varepsilon^{j,k}_{\mathrm{bl}}/\varepsilon^{j,k}_{\mathrm{mod}}$, where $\varepsilon^{j,k}_{\mathrm{bl}}$ is computed analogously. The worst performance of our model is found in the annular region around the central galaxy cluster, i.e. in the region $R^2 + v_z^2 \approx 2$. The reason for this effect can be found in Figure 1, by observing the density plot corresponding to the edge of the cluster (for $1 < r < 2.5$). Here, two different populations of galaxies are present: those infalling into the cluster (having $\langle v_r \rangle < 0$) and those orbiting around the cluster (having $\langle v_r \rangle = 0$). Although we do model such a smooth change of dynamical behaviour through the sigmoid terms in eqs. (13),(15) and the exponential term in eq. (14), our probabilistic model for v_r (eq. (10)) is unimodal. This leads our model to misinterpret those galaxies that lie in the transition between the two dynamical behaviours, hence lowering the accuracy.

The unmodelled bimodality of $P(v_r)$ on the edge of the cluster is also the cause for the kinks in the fitted values of the parameters ν and λ in Figure 3: the fitting procedure tends to increase the shape and skewness of the ST distribution to account for the presence of the second mode.

Considering the poor performance of the baseline methodology proven in Figure 6, the reconstruction of the radial number-density profile in Figure 5 can look surprisingly good. We explain this effect by noticing that, in the panels (a)–(g) of Figure 6, r_{bl} is inaccurately recovered, sometimes underestimating r_{true}, sometimes overestimating it. When aggregating over different galaxies around the same cluster, these inaccuracies tend to cancel out, building up a strong flat noise. Moreover, the baseline reconstructs r_{bl} discarding any dynamical perturbation to the Hubble flow. For all the galaxies on which these dynamical perturbations act in the xy-plane, the distortion is zero, and the baseline reconstruction is exact (but for an offset due to v_{corr}).

3.1 Application to SDSS Data

Our methodology is developed with the aim to be applied to observational data. Since the ground truth is not available on non-simulated datasets, we resort to compare our methodology with an alternative reconstruction technique. In [25], the authors develop a correction method for the redshift and apply it to the galaxies observed by SDSS[9]. We randomly choose a massive cluster in the catalogue and select all the observed galaxies with measurements contained in a square window in the phase space (R, v_z). On these galaxies, we apply our methodology and obtain one posterior for each galaxy. In Figure 7, similarly to what is already shown in Figure 5, we show the reconstructed radial profile in three dimensions of one cluster of the SDSS dataset. The number density distribution of the baseline model is obtained as the empirical distribution of r^i_{bl} computed as in eq. (20) on the selected galaxies; the distribution of the alternative correction method is obtained as the empirical distribution of r^i_{alt} computed by [25]; the distribution of our model is obtained as the average (flat mixture) of the posteriors on each selected galaxy.

[9] For data availability, see Section 6.

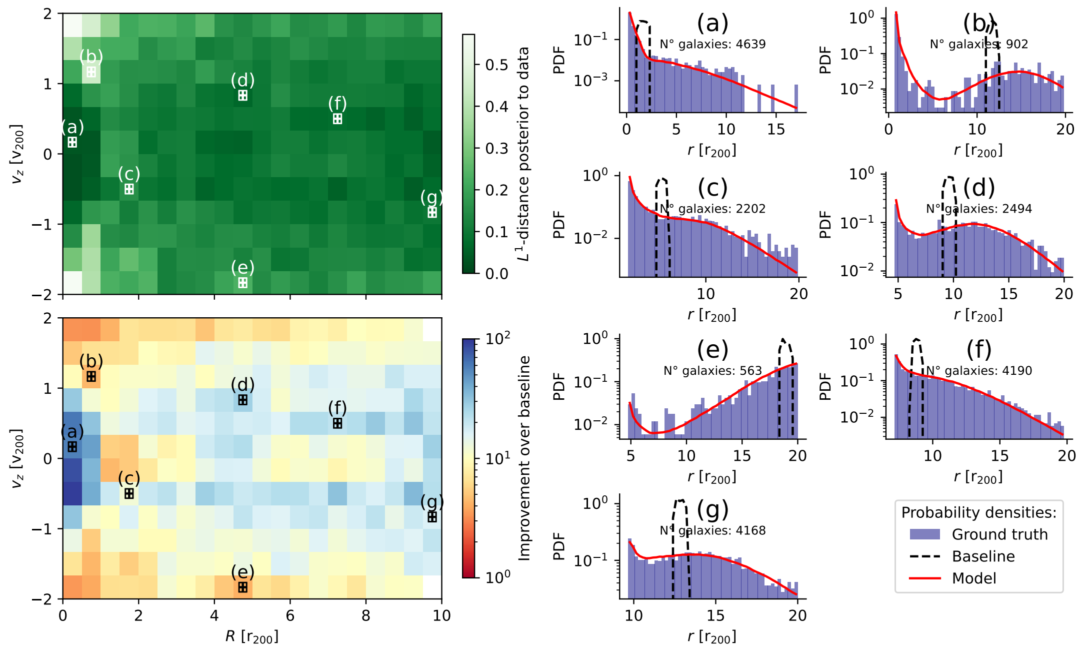

Figure 6: Model performance across the phase space. Panels (a)–(g) compare the distribution of the recovered r from our model (red solid line) and the baseline (black dashed line), contrasted with the empirical true distribution. The top-left panel shows the L^1-distance between the true and the model-reconstructed distribution. The bottom-left panel shows the reduction factor of the error computed on our reconstruction compared to the baseline.

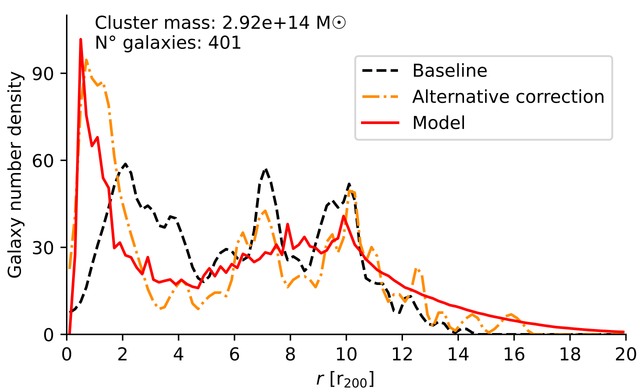

Figure 7: Reconstruction of the three-dimensional radial profile, in terms of galaxy number density, of a cluster observed by SDSS with coordinates RA = 223.28° **and** DEC = 16.79° **in the sky, and redshift** $z = 0.04556$. **Our methodology (red solid line) is compared to the baseline model (black dashed line) and to an alternative methodology proposed in [25] (orange dash-dotted line). The galaxies in these plots are selected such that $R \leq 10\,\mathrm{r}_{200}$ and $|v_z| \leq 1\,\mathrm{v}_{200}$.**

In Figure 7, we notice that our model is able to de-project real data and find a radial density profile similar to the one found on simulated data (true distribution in Figure 5) at small scales ($r \leq 2\,\mathrm{r}_{200}$). As already observed, the baseline model fails completely at recovering the inner cluster. The alternative correction correctly recovers the core of the cluster, but the drop in the density is found too far out from the centre. A sub-structure is coherently found by all three reconstruction techniques around $r = 10\,\mathrm{r}_{200}$. Another sub-structure is found around $r = 7\,\mathrm{r}_{200}$ by the baseline and alternative methods, while our model locates it at $r = 8\,\mathrm{r}_{200}$. In between the previous two sub-structure, a smaller third one is found coherently by our and the alternative methodologies, while it is absent in the baseline reconstruction. Due to the lack of a ground truth to benchmark against, it is not possible to determine which reconstruction method provides the most accurate result.

3.2 Computational Cost and Scalability

While our methodology is inherently more complex than the baseline approach, its low computational cost makes it practical even for large datasets. Training on approximately 8 million simulated galaxies, as described in this paper, takes around 0.7 CPUh on a single core of an Intel® Xeon® Platinum 8360Y. In the evaluation stage, on the same processor, the posterior of each galaxy can be generated in approximately 10 CPUμs. In both stages, RAM usage is dominated by dataset loading. This computational efficiency results from a lean methodological design with fewer than twenty free parameters. Our implementation ensures, both in the training and in the evaluation stage, a linear scaling of the computational requirements with respect to the number of galaxies considered.

4 Discussion and Conclusion

Our study shows how probabilistic modelling can be used to recover the three-dimensional distance of a galaxy from a nearby galaxy cluster given its astrophysical observations. In a context where the information content is inevitably lost due to physical observational constraints, our approach is able to reconstruct the true distance with good accuracy, at least in a statistical sense (see Figure 5–6). Since no ground truth can be known on real data, we resort to detailed cosmological simulations to both train and test our model. Our model design explicitly embeds cosmological theories and relies on free parameters that have an actual physical meaning (as discussed in Subsection 2.3), allowing for better control and understanding of the model behaviour.

The use of an explicit model highlights features in the data that merit further theoretical investigation. One of these features is the skewness of the radial velocity distribution (see Figure 2 and Subsection 2.3). Although a negative skewness has already been noted and discussed [31], we observe in the data (bottom-right panel in Figure 3) an unexpected correlation between the skewness and the mass of the cluster, both in magnitude and in sign. Positive skewness requires further investigations, especially related to big galaxy clusters [6, 23].

Another interesting feature found in the data is the exponential suppression of the radial velocity dispersion (in Virial units) with respect to the mass of the cluster (see Figure 4). Some works try to relate the velocity dispersion along the line of sight to the mass of the cluster [12, 32], but the exponential behaviour is not mentioned. This aspect should be further inquired as well.

The model presented in this work has two main limitations. The first is due to the modelling of dynamical perturbations caused by the galaxy cluster alone. Although this assumption holds for field galaxies and sub-structures as a whole, it is not valid for member galaxies of those sub-structures (as discussed in Section 3). This effect can be strongly mitigated by using our model in a hierarchical fashion: first, identifying all the sub-structures and the galaxies that are contained; next, de-projecting these galaxies with respect to the sub-structure and removing them from the galaxies to still be processed; then, de-projecting the sub-structure as a whole with respect to the main cluster; finally, aggregating the hierarchically reconstructed sub-structures in the bigger picture.

The second limitation is related to modelling the radial velocity distribution as unimodal (as discussed in Section 3). This will be addressed in future work by rewriting the likelihood as a mixture model for galaxies with different orbital classifications, as inspired by [2, 12, 32].

A minor, additional limitation is the assumption of spherical symmetry. While spherical symmetry is a necessary ingredient for generalisation outside of the galaxy cluster, the shape of the cluster can be recovered by incorporating a probabilistic model for triaxiality [19] in the likelihood. This would only marginally improve the accuracy of our model, and only for $r \lesssim 1\,\mathrm{r}_{200}$.

One important future development of this methodology regards the target variable of the posterior distribution, changing from the radial distance r to the line-of-sight distance z. It will allow users to directly obtain a probability distribution over the unobservable position, without needing to compute it from $z_{\mathrm{mod}} = \pm\sqrt{r_{\mathrm{mod}}^2 - R^2}$. The major improvement will lie in the coherent treatment of the sign of z, since dealing with the radial distance r implies reflective symmetry with respect to the plane $z = 0$.

A great feature of our methodology is the possibility to incorporate new dependencies from additional observable characteristics of galaxies. For instance, the stellar velocity dispersion within a galaxy [16], the star formation rate [10] and the colour of a galaxy [7]. All these observables can be independently measured from the kinematical quantities used in this work, and can be added as conditioning variables of the posterior to improve accuracy and, especially, reduce uncertainties.

Another, even more important, feature of our methodology is the possibility to naturally account for observational errors on individual galaxies. Suppose a galaxy is observed having velocity $v_{z\,\mathrm{obs}}$ at a distance R_{obs} in the sky from a cluster of mass M_{obs}. These measurements come with the associated uncertainties, which can be represented by an error model $P(R, v_z, M \mid R_{\mathrm{obs}}, v_{z\,\mathrm{obs}}, M_{\mathrm{obs}})$. By sampling this error model, one can obtain multiple triplets (R, v_z, M), and, for each, a posterior distribution. Then, averaging over these posteriors, one obtains the distribution $P(r \mid R_{\mathrm{obs}}, v_{z\,\mathrm{obs}}, M_{\mathrm{obs}})$, which blends the information about the dynamical model with the observational uncertainties. The fully-probabilistic modelling ensures extreme versatility as it imposes no constraints on the error model.

To conclude, in this work, we present the first attempt, to the best of our knowledge, towards the three-dimensional reconstruction of the surroundings of an observed galaxy cluster. We tackle the problem with a probabilistic formulation that is transparent (all the parameters are interpretable and the complexity is kept low), versatile (answers a general question from which more specific problems can be addressed), extensible (more observables can be included to improve the performance), conceptually-consistent (probabilities and uncertainties are propagated through physics-based transformations) and computationally cheap (applicable to massive datasets with reasonable computing resources).

5 Limitations and Ethical Considerations

The data we use are not sensitive and no concerns for privacy or vulnerability are to be raised.

Acknowledgments

DB and PT acknowledge financial support from the Engineering and Physical Sciences Research Council [EP/Y031032/1 (EDUCADO)]. DB training is co-funded by the Marie Sklodowska-Curie Action [grant agreement 101119830 (EDUCADO)]. MC and PT acknowledge financial support from the Engineering and Physical Sciences Research Council [EP/X025454/1 (ARCANE)]. RS acknowledges financial support from FONDECYT Regular 2023 [1230441]; and from ANID–MILENIO [NCN2024_112]. The Authors acknowledge the use of the Uchuu simulation [1, 15] [link to credits]; and the SDSS DR7 corrected dataset [25].

6 Data Availability and Reproducibility

The simulation dataset is publicly available at Uchuu's website [link]. The corrected SDSS dataset is publicly available online [link]. The code used to produce this paper can be found on GitHub [link].

7 GenAI Disclosure

Generative AI was used in the second stage of the literature research, to integrate possible missing references. Two of the works cited in the references were directly suggested by AI tools. Three others were found with a manual search that originated from those references.

References

[1] Han Aung, Daisuke Nagai, Anatoly Klypin, Peter Behroozi, Mohamed H Abdullah, Tomoaki Ishiyama, Francisco Prada, Enrique Pérez, Javier López Cacheiro, and José Ruedas. 2022. The Uchuu-universe machine data set: galaxies in and around clusters. *Monthly Notices of the Royal Astronomical Society* 519, 2 (12 2022), 1648–1656. arXiv:https://academic.oup.com/mnras/article-pdf/519/2/1648/48447434/stac3514.pdf doi:10.1093/mnras/stac3514

[2] Han Aung, Daisuke Nagai, Eduardo Rozo, Brandon Wolfe, and Susmita Adhikari. 2023. Accurate model of the projected velocity distribution of galaxies in dark matter haloes. *Monthly Notices of the Royal Astronomical Society* 521, 3 (03 2023), 3981–3990. arXiv:https://academic.oup.com/mnras/article-pdf/521/3/3981/49672134/stad601.pdf doi:10.1093/mnras/stad601

[3] Petra Awad, Reynier Peletier, Marco Canducci, Rory Smith, Abolfazl Taghribi, Mohammad Mohammadi, Jihye Shin, Peter Tiňo, and Kerstin Bunte. 2023. Swarm-intelligence-based extraction and manifold crawling along the Large-Scale Structure. *Monthly Notices of the Royal Astronomical Society* 520, 3 (02 2023), 4517–4539. arXiv:https://academic.oup.com/mnras/article-pdf/520/3/4517/49287036/stad428.pdf doi:10.1093/mnras/stad428

[4] Julien Bel, Julien Larena, Roy Maartens, Christian Marinoni, and Louis Perenon. 2022. Constraining spatial curvature with large-scale structure. *Journal of Cosmology and Astroparticle Physics* 2022, 09 (sep 2022), 076. doi:10.1088/1475-7516/2022/09/076

[5] James Binney and Scott Tremaine. 2008. *Galactic Dynamics: Second Edition*. Princeton university press.

[6] R. R. de Carvalho, A. L. B. Ribeiro, D. H. Stalder, R. R. Rosa, A. P. Costa, and T. C. Moura. 2017. Investigating the Relation between Galaxy Properties and the Gaussianity of the Velocity Distribution of Groups and Clusters. *The Astronomical Journal* 154, 3 (aug 2017), 96. doi:10.3847/1538-3881/aa7f2b

[7] Valeria Coenda, Martín de los Rios, Hernán Muriel, Sofía A. Cora, Héctor J. Martínez, Andrés N. Ruiz, and Cristian A. Vega-Martínez. 2021. Reconstructing orbits of galaxies in extreme regions (ROGER) – II: reliability of projected phase-space in our understanding of galaxy populations. *Monthly Notices of the Royal Astronomical Society* 510, 2 (12 2021), 1934–1944. arXiv:https://academic.oup.com/mnras/article-pdf/510/2/1934/41988066/stab3551.pdf doi:10.1093/mnras/stab3551

[8] Alison L. Coil. 2013. *The Large-Scale Structure of the Universe*. Springer Netherlands, Dordrecht, 387–421. doi:10.1007/978-94-007-5609-0_8

[9] K. de Vos, N. A. Hatch, and M. R. Merrifield. 2024. From outskirts to core: the suppression and activation of radio AGN around galaxy clusters. *Monthly Notices of the Royal Astronomical Society* 535, 1 (10 2024), 217–222. arXiv:https://academic.oup.com/mnras/article-pdf/535/1/217/60129039/stae2391.pdf doi:10.1093/mnras/stae2391

[10] K. de Vos, M. R. Merrifield, and N. A. Hatch. 2024. Local versus global environment: the suppression of star formation in the vicinity of galaxy clusters. *Monthly Notices of the Royal Astronomical Society* 531, 4 (06 2024), 4383–4390. arXiv:https://academic.oup.com/mnras/article-pdf/531/4/4383/58271076/stae1403.pdf doi:10.1093/mnras/stae1403

[11] Martina Falco, Steen H. Hansen, Radoslaw Wojtak, Thejs Brinckmann, Mikkel Lindholmer, and Stefania Pandolfi. 2014. A new method to measure the mass of galaxy clusters. *Monthly Notices of the Royal Astronomical Society* 442, 2 (Aug. 2014), 1887–1896. doi:10.1093/mnras/stu971

[12] Akinari Hamabata, Masamune Oguri, and Takahiro Nishimichi. 2019. Constraining cluster masses from the stacked phase space distribution at large radii. *Monthly Notices of the Royal Astronomical Society* 489, 1 (08 2019), 1344–1356. arXiv:https://academic.oup.com/mnras/article-pdf/489/1/1344/29229772/stz2227.pdf doi:10.1093/mnras/stz2227

[13] Akinari Hamabata, Taira Oogi, Masamune Oguri, Takahiro Nishimichi, and Masahiro Nagashima. 2019. New constraints on red-spiral galaxies from their kinematics in clusters of galaxies. *Monthly Notices of the Royal Astronomical Society* 488, 3 (07 2019), 4117–4125. arXiv:https://academic.oup.com/mnras/article-pdf/488/3/4117/29102209/stz1991.pdf doi:10.1093/mnras/stz1991

[14] A. J. S. Hamilton. 1998. *Linear Redshift Distortions: A Review*. Springer Netherlands, Dordrecht, 185–275. doi:10.1007/978-94-011-4960-0_17

[15] Tomoaki Ishiyama, Francisco Prada, Anatoly A. Klypin, Manodeep Sinha, R. Benton Metcalf, Eric Jullo, Bruno Altieri, Sofía A. Cora, Darren Croton, Sylvain de la Torre, David E. Millán-Calero, Taira Oogi, José Ruedas, and Cristian A. Vega-Martínez. 2021. The Uchuu simulations: Data Release 1 and dark matter halo concentrations. *Monthly Notices of the Royal Astronomical Society* 506, 3 (06 2021), 4210–4231. arXiv:https://academic.oup.com/mnras/article-pdf/506/3/4210/39554057/stab1755.pdf doi:10.1093/mnras/stab1755

[16] Yara L. Jaffé, Bianca M. Poggianti, Alessia Moretti, Marco Gullieuszik, Rory Smith, Benedetta Vulcani, Giovanni Fasano, Jacopo Fritz, Stephanie Tonnesen, Daniela Bettoni, George Hau, Andrea Biviano, Callum Bellhouse, and Sean McGee. 2018. GASP. IX. Jellyfish galaxies in phase-space: an orbital study of intense ram-pressure stripping in clusters. *Monthly Notices of the Royal Astronomical Society* 476, 4 (02 2018), 4753–4764. arXiv:https://academic.oup.com/mnras/article-pdf/476/4/4753/24565813/sty500.pdf doi:10.1093/mnras/sty500

[17] Shoko Jin et al. 2023. The wide-field, multiplexed, spectroscopic facility WEAVE: Survey design, overview, and simulated implementation. *Monthly Notices of the Royal Astronomical Society* 530, 3 (03 2023), 2688–2730. arXiv:https://academic.oup.com/mnras/article-pdf/530/3/2688/57347736/stad557.pdf doi:10.1093/mnras/stad557

[18] Francisco-Shu Kitaura, Raul E. Angulo, Yehuda Hoffman, and Stefan Gottlöber. 2012. Estimating cosmic velocity fields from density fields and tidal tensors. *Monthly Notices of the Royal Astronomical Society* 425, 4 (10 2012), 2422–2435. arXiv:https://academic.oup.com/mnras/article-pdf/425/4/2422/4876372/425-4-2422.pdf doi:10.1111/j.1365-2966.2012.21589.x

[19] Andrea Morandi, Kristian Pedersen, and Marceau Limousin. 2010. Unveiling the Three-Dimensional Structure of Galaxy Clusters: Resolving the Discrepancy between X-Ray and Lensing Masses. *The Astrophysical Journal* 713, 1 (mar 2010), 491. doi:10.1088/0004-637X/713/1/491

[20] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. 1996. The Structure of Cold Dark Matter Halos. *apj* 462 (May 1996), 563. arXiv:astro-ph/9508025 [astro-ph] doi:10.1086/177173

[21] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. 1997. A Universal Density Profile from Hierarchical Clustering. *The Astrophysical Journal* 490, 2 (dec 1997), 493. doi:10.1086/304888

[22] Maria Angela Raj, Petra Awad, Reynier F. Peletier, Rory Smith, Ulrike Kuchner, Rien van de Weygaert, Noam I. Libeskind, Marco Canducci, Peter Tiňo, and Kerstin Bunte. 2024. Large-scale structure around the Fornax-Eridanus complex. *Astronomy & Astrophysics* 690, Article A92 (Oct. 2024), A92 pages. arXiv:2407.03225 [astro-ph.GA] doi:10.1051/0004-6361/202450815

[23] A. L. B. Ribeiro, P. A. A. Lopes, and M. Trevisan. 2011. Non-Gaussian velocity distributions - the effect on virial mass estimates of galaxy groups. *mnras* 413, 1 (May 2011), L81–L85. arXiv:1103.0426 [astro-ph.CO] doi:10.1111/j.1745-3933.2011.01038.x

[24] Eleonora Sarli, Sven Meyer, Massimo Meneghetti, Sara Konrad, Charles L. Majer, and Matthias Bartelmann. 2014. Reconstructing the projected gravitational potential of galaxy clusters from galaxy kinematics. *A&A* 570 (2014), A9. doi:10.1051/0004-6361/201321748

[25] Feng Shi, Xiaohu Yang, Huiyuan Wang, Youcai Zhang, H. J. Mo, Frank C. van den Bosch, Shijie Li, Chengze Liu, Yi Lu, Dylan Tweed, and Lei Yang. 2016. Mapping the Real-Space Distributions of Galaxies in SDSS DR7. I. Two-Point Correlation Functions. *The Astrophysical Journal* 833, 2 (dec 2016), 241. doi:10.3847/1538-4357/833/2/241

[26] Cristóbal Sifón et al. 2025. CHANCES, the Chilean Cluster Galaxy Evolution Survey: Selection and initial characterisation of clusters and superclusters. *A&A* 697 (2025), A92. doi:10.1051/0004-6361/202452710

[27] Max Tegmark et al. 2004. The Three-Dimensional Power Spectrum of Galaxies from the Sloan Digital Sky Survey. *The Astrophysical Journal* 606, 2 (may 2004), 702. doi:10.1086/382125

[28] Huiyuan Wang, H. J. Mo, Xiaohu Yang, and Frank C. van den Bosch. 2012. Reconstructing the cosmic velocity and tidal fields with galaxy groups selected from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 420, 2 (01 2012), 1809–1824. arXiv:https://academic.oup.com/mnras/article-pdf/420/2/1809/3082715/mnras0420-1809.pdf doi:10.1111/j.1365-2966.2011.20174.x

[29] Di Wen, Athol J. Kemball, and William C. Saslaw. 2020. Halo Counts-in-cells for Cosmological Models with Different Dark Energy. *The Astrophysical Journal* 890, 2 (feb 2020), 160. doi:10.3847/1538-4357/ab6d6f

[30] Martin White, J. D. Cohn, and Renske Smit. 2010. Cluster galaxy dynamics and the effects of large-scale environment. *Monthly Notices of the Royal Astronomical Society* 408, 3 (10 2010), 1818–1834. arXiv:https://academic.oup.com/mnras/article-pdf/408/3/1818/18587684/mnras0408-1818.pdf doi:10.1111/j.1365-2966.2010.17248.x

[31] Ayako Yoshisato, Masahiro Morikawa, and Hideaki Mouri. 2003. Negative skewness of radial pairwise velocity in the quasi-non-linear regime: Zel'dovich approximation. *Monthly Notices of the Royal Astronomical Society* 343, 3 (08 2003), 1038–1044. arXiv:https://academic.oup.com/mnras/article-pdf/343/3/1038/3891824/343-3-1038.pdf doi:10.1046/j.1365-8711.2003.06760.x

[32] Ying Zu and David H. Weinberg. 2013. The redshift-space cluster–galaxy cross-correlation function – I. Modelling galaxy infall on to Millennium simulation clusters and SDSS groups. *Monthly Notices of the Royal Astronomical Society* 431, 4 (04 2013), 3319–3337. arXiv:https://academic.oup.com/mnras/article-pdf/431/4/3319/3953195/stt411.pdf doi:10.1093/mnras/stt411

A Additional Figures

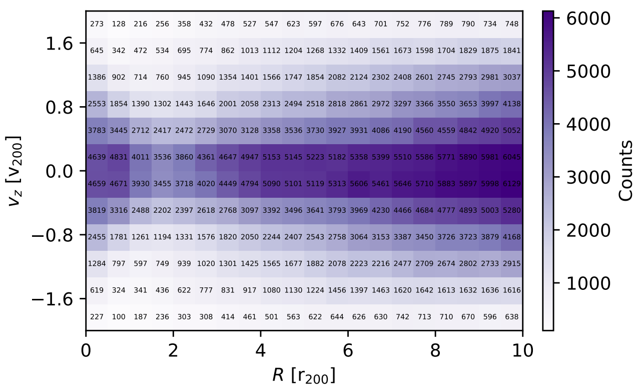

Figure 8: Number of galaxies contained in each square window of the phase space pictured in Figure 6.

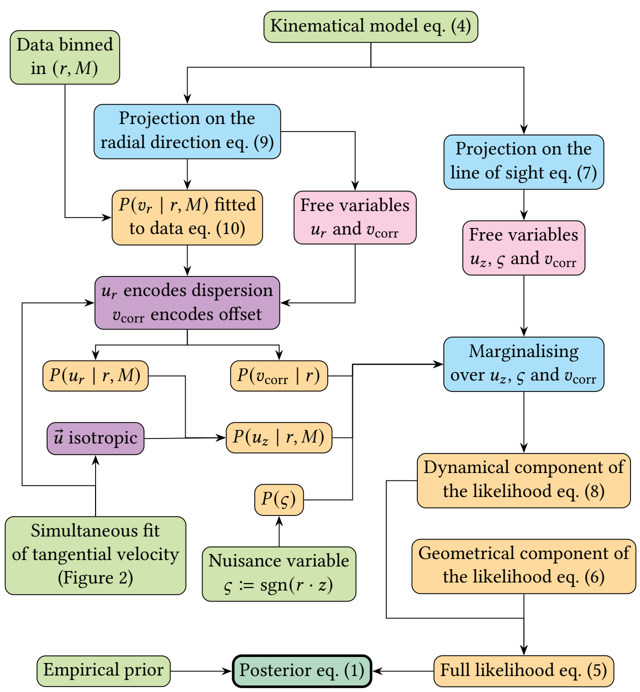

Figure 9: Probabilistic model schema. Light green boxes represent independent considerations that can be taken at the start by observing the data; cyan boxes represent mathematical operations; yellow boxes represent the key steps when the probability distributions are determined; pink boxes highlight important dependencies for the flow of the logic; purple boxes represent the assumptions made; the dark green box represent the output of our model.

B Additional Tables

Table 1: Symbols used in this paper and their meaning.

Variable	Meaning	Unit
$\vec{r}$	3-dimensional distance	r_{200}
R	distance projected on the sky-plane	r_{200}
$\vec{v}$	3-dimensional velocity	v_{200}
v_z	line-of-sight velocity	v_{200}
M	galaxy cluster mass	$M_\odot$
H_0	Hubble constant	$\mathrm{km\,s^{-1}\,Mpc^{-1}}$
Δ	Virial over-density parameter	

C Geometrical Component of the Likelihood

In this section, we analytically derive the probability distribution $P(R \mid r)$. We start by considering that, given a point at $\vec{r}$, this probability distribution is not a Dirac-delta because the orientation of the observation plane is unknown, which we can safely assume to be uniformly distributed in three-dimensional space. We can then invert the logic and consider the reference frame fixed and the orientation of $\vec{r}$ being uniformly distributed on the sphere. This leads to the probability distribution

$$P(x, y, z|r) = \frac{1}{4\,\pi\,[x^2 + y^2 + z^2]}\,\delta_{\rm D}\left(\sqrt{x^2 + y^2 + z^2} - r\right) \tag{21}$$

Transforming it in cylindrical coordinates $\{x, y, z\} \to \{R, \varphi, z\}$, given that the determinant of the Jacobian is $|J| = R$, we obtain

$$P(R, \varphi, z|r) = |J|\,P(x, y, z|r) = \frac{R}{4\,\pi\,[R^2 + z^2]}\,\delta_{\rm D}\left(\sqrt{R^2 + z^2} - r\right) \tag{22}$$

The probability distribution over R is then obtained by marginalizing over φ and z

$$P(R \mid r) = \int_0^{+\infty} \mathrm{d}z\,\frac{R}{R^2 + z^2}\,\delta_{\rm D}\left(\sqrt{R^2 + z^2} - r\right) \tag{23}$$

Using the property of the Dirac-delta

$$\delta_{\rm D}(f(x)) = \sum_{i=1}^{N} \frac{\delta_{\rm D}(x - x_i)}{|f'(x_i)|} \tag{24}$$

with $f(x) = \sqrt{R^2 + x^2} - r$, so $x_{1,2} = \pm\sqrt{r^2 - R^2}$ and $f'(x) = x/\sqrt{R^2+x^2}$, we can rewrite the delta term as

$$\delta_{\rm D}\left(\sqrt{R^2 + z^2} - r\right) = \frac{r}{\sqrt{r^2 - R^2}}\left[\delta_{\rm D}\left(z - \sqrt{r^2 - R^2}\right) + \delta_{\rm D}\left(z + \sqrt{r^2 - R^2}\right)\right] \tag{25}$$

In eq. (23), considering that $\delta_{\rm D}\left(z + \sqrt{r^2 - R^2}\right) = 0$ in the integration domain, the evaluation of the integral provides

$$P(R \mid r) = \frac{R}{r\,\sqrt{r^2 - R^2}} \tag{26}$$

D Virial Units

Virial units are defined from the Virial theorem, which relates the kinetic energy to the potential energy of a system of particles at equilibrium. Within a certain radius from the cluster centre, said *Virial radius*, the galaxies move as particles at equilibrium in a gravitational well. The Virial radius r_{vir} is defined implicitly as the distance at which the dark matter density of the galaxy cluster drops below a multiple Δ of the *critical density* of the Universe

$$\rho(r < r_{\text{vir}}) = \Delta\,\rho_c \tag{27}$$

The critical density of the Universe ρ_c is the density of matter and energy required for the universe to have a flat geometry, and it's given by

$$\rho_c = \frac{3\,H_0^2}{8\,\pi\,G} \tag{28}$$

An *over-density parameter* Δ can be defined to arbitrarily set the scale of the cluster size. Since r_{vir} depends on Δ, the Virial radius is commonly denoted as r_Δ to make this dependency explicit.

The Virial mass and Virial velocities are defined from the Virial radius as

$$M_\Delta = \frac{4}{3}\,\pi\,r_\Delta^3\,\Delta\,\rho_c \tag{29}$$

$$v_\Delta = \sqrt{\frac{G\,M_\Delta}{r_\Delta}} \tag{30}$$

The three Virial quantities scale with respect to each other as

$$v_\Delta \propto r_\Delta \tag{31}$$

$$v_\Delta \propto \sqrt[3]{M_\Delta} \tag{32}$$

D.1 Hubble Constant

Importantly, we need to compute the value of the Hubble constant expressed in Virial units. Isolating M_Δ from eq. (30) and substituting it in eq. (29), and then isolating ρ_c and substituting again in eq. (28), we obtain

$$\frac{3\,v_\Delta^2}{4\,\pi\,G\,\Delta\,r_\Delta^2} = \frac{3\,H_0^2}{8\,\pi\,G} \tag{33}$$

When expressing distances and velocities in Virial units, we have $r_\Delta = 1 = v_\Delta$, and simplifying all the coefficients, we obtain

$$H_0 = \sqrt{\frac{2}{\Delta}} \tag{34}$$

D.2 Velocity Dispersion

The velocity dispersion $\langle v^2 \rangle$ inside a galaxy cluster is approximately (discarding coefficients of order ≈ 1)

$$\langle v^2 \rangle \approx \frac{G\,M_\Delta}{r_\Delta} \tag{35}$$

Isolating M_Δ in eq. (30) and substituting it, we obtain

$$\langle v^2 \rangle \approx v_\Delta^2 \tag{36}$$

which further motivates the choice of Virial units for this work.

E Skewed *t* Distribution

The skewed t distribution is defined by the function

$$\mathrm{ST}(x;\mu,\sigma,\nu,\lambda) = \frac{1}{\mathcal{N}}\left[1 + \frac{2\,[x-\mu+b]^2}{\nu\;[a\,\sigma]^2\,[1+\lambda\,\mathrm{sgn}(x-\mu+b)]^2}\right]^{-\frac{1+\nu}{2}} \tag{37}$$

where $\mu, \sigma, \nu, \lambda$ are the parameters, and the normalisation factor $\mathcal{N}$ and the two placeholders a and b are

$$\mathcal{N} := \frac{\Gamma(\nu/2)\,a\,\sigma\,\sqrt{\pi\,\nu/2}}{\Gamma(1+\nu/2)}$$

$$b := 2\,\lambda\,a\,\sigma\sqrt{\frac{\nu}{2\,\pi}}\,\frac{\Gamma(\nu-1/2)}{\Gamma(\nu/2)}$$

$$a := \left[\left[1+3\,\lambda^2\right]\frac{\nu}{2\,\nu-4} - \frac{2\,\lambda^2\,\nu}{\pi}\left[\frac{\Gamma(\nu-1/2)}{\Gamma(\nu/2)}\right]^2\right]^{-\frac{1}{2}}$$