THE PAST RECENT MERGER HISTORY OF GALAXY CLUSTERS FROM X-RAY AND RADIO OBSERVATIONS, WITH SIMULATION-BASED INFERENCE AND CONTRASTIVE LEARNING

SHERA JAFARITABAR

Heidelberg University

September 2025

Referees: Dr. Annalisa Pillepich, Prof. Ralf Klessen

ABSTRACT

Galaxy clusters assemble hierarchically through mergers whose shocks, turbulence, and core disruption leave rich thermal (X-ray) and non-thermal (radio) imprints. Inferring the underlying merger parameters, e.g., time since/to pericenter, collision velocity, mass ratio, pericenter distance, and component masses is scientifically valuable but much of the relevant information is lost or distorted in observations. Moreover, observation provide only single-epoch snapshots rather than time-resolved tracks over Gyr, making reliable ground-truth labels scarce. We therefore adopt a simulation-based inference approach, using the TNG–Cluster simulation and its 352 zoom-in halos in three orthogonal projections across redshifts $0 \leqslant z \leqslant 1$ to create intrinsic maps and their merger parameters. Building on these data, this thesis develops a framework that tests whether merger properties can be reliably inferred solely from images. First, we train SimCLR, a contrastive learning framework, separately on intrinsic X-ray, radio, and joint X-ray+radio maps to learn morphology-aware representations that distill high-resolution maps into feature vectors that capture the relevant structure. A physics-aware augmentation suite (rotations, flips, Gaussian blur, additive noise, and affine zoom/shift) was also designed and used for training the SimCLR to promote invariance to observing nuisances without erasing merger signatures. Second, a conditional invertible neural network (cINN) with rational–quadratic spline couplings and a mixture-of-experts partition in representation space, enables inference of the $p(\mathbf{x} \mid \mathbf{c})$, where x is the merger parameter and c is the conditioner. Two conditioning modes are evaluated: (i) representation-conditioned, where c is the learned embedding of (X-ray, radio, or joint maps); and (ii) scalar-conditioned, where c is a vector of scalar observable parameters that are fed directly to the cINN, without the preceding contrastive learning step. Across four conditioners schemes; scalars, X-ray, radio, and joint X-ray+radio maps, the method is evaluated on both last- and next-merger properties. Conditioning on the radio maps delivers the most precise and accurate posteriors for last merger, with typical maximum a posteriori (MAP) relative error ranges of: collision time $\sim [-5, 10]\%$, velocity $\leqslant \pm 1\%$, main cluster's M_{500c}, and subcluster mass $\sim \pm 0.5\%$, and pericenter distance $\sim \pm 3\%$, reflecting the close coupling of radio emission and its geometry to recent cluster dynamics. Conditioning on X-ray maps remains informative, but produces broader posteriors and larger scatter in MAP estimates. Joint conditioning is consistently intermediate; it narrows and stabilizes inferences relative to X-ray alone, but it does not surpass radio conditioning. Conditioning on scalar observable properties on the other hand, performs competitively only for collision time and the main cluster mass, while elsewhere it is neither accurate nor precise, and exhibits heteroscedastic dispersions, pointing to the information bottle neck that could be introduced by only using scalar observable parameters. In all cases, the cINN not only recovers ΛCDM-consistent cross-parameter correlations, but also could use the learned correlations to propagate information from well constrained merger properties to weaker ones. Forecasts for the next merger mirror the last-merger performance with physically expected weaker precision (most visible for timing). Methodologically, this thesis demonstrates that label-free contrastive representation of maps can be used directly as

cINN condition enabling survey-scale, uncertainty-aware inference of galaxy cluster's assembly from imaging data alone. Practically, the results advocate prioritizing high-fidelity radio imaging. In future research, this framework can be extended toward instrument-aware mock observables for simulation-to-real transfer, multi-resolution fusion including SZ and weak-lensing, and scalable pipelines suitable for eROSITA, Chandra, XMM paired with LOFAR, MeerKAT, or VLA.

ZUSAMMENFASSUNG

Galaxienhaufen entstehen hierarchisch durch Verschmelzungen, deren Stoßwellen, Turbulenzen und Kernstörungen reiche thermische (Röntgen) und nicht-thermische (Radio) Signaturen hinterlassen. Die zugrunde liegenden Parameter solcher Verschmelzungen, z. B. die Zeit seit/zur Perizentrenpassage, Kollisionsgeschwindigkeit, Massenverhältnis, Perizentrenabstand und Komponentenmassen, zu bestimmen, ist wissenschaftlich wertvoll. Doch ein Großteil der relevanten Informationen geht in Beobachtungen verloren oder wird verzerrt. Zudem liefern Beobachtungen nur Momentaufnahmen zu einzelnen Zeitpunkten statt zeitaufgelöster Entwicklungen über Gigajahre hinweg, wodurch verlässliche Ground-Truth-Labels selten sind. Wir verfolgen daher einen simulationsbasierten Inferenzansatz, indem wir TNG-Cluster Simulationen und deren 352 Zoom-in-Halos in drei orthogonalen Projektionen über Rotverschiebungen $0 \leqslant z \leqslant 1$ verwenden, um intrinsische Karten und die zugehörigen Verschmelzungsparameter zu erzeugen. Aufbauend auf diesen Daten entwickelt diese Arbeit einen durchgängigen, simulationsbasierten Ansatz, um aus Bildern kalibrierte Posteriorverteilungen über die Fusionsphysik abzuleiten. Zunächst trainieren wir SimCLR, ein Framework für kontrastives Lernen, getrennt auf Röntgen-, Radio- und kombinierte Röntgen+Radio-Karten, um morphologiebewusste Repräsentationen zu erlernen, die hochaufgelöste Karten in Merkmalsvektoren verdichten, welche die relevante Struktur erfassen. Ein physikbewusstes Augmentationspaket (Rotationen, Spiegelungen, Gaußsche Unschärfe, additives Rauschen sowie affine Zoom-/Shift-Operationen; bei kombinierten Karten kanalkonsistent angewendet) fördert Invarianz gegenüber beobachtungsbedingten Störeinflüssen, ohne Fusionssignaturen zu verwischen. Anschließend ermöglicht ein bedingtes invertierbares neuronales Netzwerk (cINN) mit rational–quadratischen Spline-Kopplungen und einer Mixture-of-Experts-Aufteilung im Repräsentationsraum die Inferenz von $p(\mathbf{x} \mid \mathbf{c})$, wobei x den Fusionsparameter und c den Conditioner bezeichnet. Wir betrachten zwei Modi der Konditionierung: (i) Repräsentationsbasiert, bei dem c die erlernte Einbettung der (Röntgen-, Radio- oder kombinierten Karten) ist; und (ii) Skalarbasiert, bei dem c ein Vektor skalierbarer beobachtbarer Parameter ist, der direkt in das cINN eingespeist wird, ohne kontrastives Lernen (den ersten Schritt). Über vier Conditioner-Varianten Skalare, Röntgen, Radio und kombinierte Röntgen+Radio-Karten, wird die Methode sowohl für vergangene als auch für zukünftige Verschmelzungseigenschaften evaluiert. Radio-Konditionierung liefert die präzisesten und genauesten Posteriorverteilungen, mit typischen relativen Maximum a Posteriori (MAP)-Fehlerbereichen von Kollisionszeit $\sim [-5, 10]\%$, Geschwindigkeit $\leqslant \pm 1\%$, Hauptcluster-M_{500c}, Subcluster-Masse $\sim \pm 0{,}5\%$ und Perizentrenabstand $\sim \pm 3\%$. Dies spiegelt die enge Kopplung von Radioemission und ihrer Geometrie an die jüngste Dynamik von Clustern wider. Röntgen-Konditionierung bleibt informativ, führt jedoch zu breiteren Posterioren

und größerer Streuung der MAP-Schätzungen. Kombinierte Konditionierung liegt durchgehend dazwischen: Sie verengt und stabilisiert die Inferenz im Vergleich zu Röntgen allein, übertrifft aber Radio nicht. Skalarkonditionierung hingegen ist nur bei Kollisionszeit und Hauptcluster-Masse konkurrenzfähig; ansonsten ist sie weder genau noch präzise und zeigt heteroskedastische Streuungen, was auf den Informationsverlust hinweist, der durch die ausschließliche Nutzung skalarer Beobachtungsparameter entstehen kann. In allen Fällen rekonstruiert das cINN nicht nur ΛCDM-konsistente Kreuzparameterkorrelationen, sondern kann die gelernten Zusammenhänge auch nutzen, um Informationen von gut bestimmbaren zu schwächer bestimmbaren Fusionsparametern zu propagieren. Vorhersagen für die nächste Verschmelzung spiegeln die Ergebnisse für vergangene Fusionen wider, jedoch mit der erwarteten, physikalisch begründeten geringeren Präzision (am deutlichsten bei der Zeitabschätzung). Methodisch zeigt diese Arbeit, dass kontrastive, labelfreie Repräsentationen, direkt als cINN-Conditioner genutzt, unsicherheitsbewusste Inferenz der Galaxienhaufen-Entstehung aus reinen Bilddaten im Survey-Maßstab ermöglichen. Praktisch sprechen die Ergebnisse dafür, hochqualitative Radioabbildungen zu priorisieren. Zukünftige Forschung kann diesen Rahmen auf instrumentenbewusste Mock-Beobachtungen für Sim-to-Real-Transfer, Multi-Resolution-Fusion inklusive SZ- und Weak-Lensing-Daten sowie skalierbare Pipelines für eROSITA, Chandra, XMM in Kombination mit LOFAR, MeerKAT oder VLA erweitern.

CONTENTS

LIST OF TABLES

Part I

INTRODUCTION

1

GALAXY CLUSTERS AS COSMOLOGY AND ASTROPHYSICS LABORATORIES

1.1 GROUPS, GALAXY CLUSTERS AND SUPER CLUSTERS

Definitions and Their Difference in Mass and Size

In observational and extragalactic astrophysics, galaxy groups, galaxy clusters, and superclusters serve as distinct yet connected systems within the hierarchy of large-scale cosmic structures. Their definitions primarily depends on the number of galaxies within, gravitational binding, equilibrium states, mass, and other physical scales.

Galaxy Groups: Galaxy groups are typically the smallest gravitationally bound systems of galaxies, often comprising tens of galaxies. The Local Group, containing prominent galaxies such as the Milky Way, Andromeda (M31), and approximately 50 other galaxies[96]. Galaxy groups generally exhibit total masses ranging from approximately $10^{12.5}$ to $10^{14} M_{\odot}$, with radii spanning about 0.5 to 1 megaparsecs (Mpc) [109]. Given their relatively shallow gravitational potential wells, groups often display considerable internal galaxy interactions, frequent mergers, and tidal effects are common. These sytems are often dynamically complex states rather than perfect equilibrium [34].

Galaxy Clusters: Galaxy clusters are the largest gravitationally bound and virialized structures in the Universe, hosting hundreds to thousands of galaxies embedded in an extensive, hot, diffuse intracluster medium (ICM). Their gravitationally binding mass typically ranges between 10^{14} and 10^{15} $M_{\odot}$, with radii commonly extending from 1 to 5 Mpc. Clusters are deep gravitational potential wells capable of heating their intracluster gas to temperatures often exceeding 10^{7} Kelvin, which causes them to be luminous in X-ray wavelengths. Galaxy clusters generally approach hydrostatic equilibrium, though disturbances such as mergers and accretion events frequently perturb their internal structure [84].

Superclusters: At an even larger scale, superclusters constitute collections of multiple galaxy groups and clusters, spanning immense spatial extents; tens to hundreds of megaparsecs. Unlike virialized galaxy clusters, superclusters typically are not gravitationally bound structures and are instead in the process of continuous expansion, influenced significantly by the Universe's overall Hubble flow. As such, they exhibit complex dynamical states, characterized by relatively weaker gravitational binding and ongoing structure formation processes [45]. The Local Supercluster, also known as the Virgo Supercluster, exemplifies this category, comprising numerous galaxy groups and clusters, including our Local Group and the Virgo Cluster. Superclusters are vital for investigating the large-scale web-like structure of the Universe, which includes cosmic filaments and expansive voids [156].

The transitions between these categories are not always sharp, as intermediate system exist (e.g, poor clusters or rich groups), and environmental processes can

blur the boundaries. However, the above distinctions hint us toward a hierarchical structure of the universe.

Hierarchical Structure of the Universe and Their Place in the Cosmic Web

The Universe exhibits a hierarchical structure, characterized by complex networks of matter distributions that evolve through gravitational interactions. At large scales, this structure is collectively known as the cosmic web, a network composed primarily of galaxies, galaxy groups, galaxy clusters, filaments, sheets, and voids, intricately connected by gravitational interactions and cosmological evolution processes.

Hierarchical Formation: According to the prevailing Λ Cold Dark Matter (ΛCDM) cosmological model, structure formation proceeds in a hierarchical manner, meaning smaller objects form first and subsequently merge to form larger and more massive structures [171]. Initial density fluctuations in the early Universe, imprinted during inflation, serve as seeds for this hierarchical buildup. Over cosmic time, these fluctuations grow via gravitational instability, collapsing first into dark matter halos and subsequently attracting baryonic matter to form galaxies and larger cosmic structures [171].

The large-scale structure can be visualized as a web-like arrangement of matter, the cosmic web, containing several distinct features:

- *Filaments*: These elongated structures connect galaxy groups and clusters, containing the majority of baryonic matter outside clusters. Filaments are characterized by their dense environments, facilitating galaxy growth and migration towards more massive gravitational potentials.
- *Galaxy Clusters and Groups*: These are nodes within the cosmic web where filaments intersect, forming the densest and most gravitationally bound regions. Clusters and groups represent the largest structures in dynamical equilibrium.
- *Voids*: These are expansive, low-density regions encompassing the majority of cosmic volume, characterized by significantly fewer galaxies. Voids expand faster due to lower gravitational attraction, shaping the surrounding filaments and sheets.

Galaxy clusters and groups hold crucial positions as nodes within the cosmic web, marking the peaks in the cosmic density field. They act as gravitational attractors, driving the flow of matter along filaments, which contributes directly to their growth. Understanding galaxy clusters and groups is critical because they encapsulate the processes of hierarchical merging, gravitational interactions, and baryonic physics in a relatively confined yet massive environment [17].

Galaxy clusters, therefore, represent the culmination of hierarchical structure formation. Their positions within the cosmic web, at the center of cosmic filaments, make them invaluable laboratories for testing cosmological models, probing dark matter properties, and studying baryonic physics processes such as galaxy formation, gas dynamics, and feedback mechanisms. They also provide observational benchmarks to refine cosmological simulations, ultimately enhancing our understanding of the Universe's evolution [171].

Why Clusters are Important: Largest Gravitationally Bound Systems, Key for Structure Formation, and Cosmology

Galaxy clusters are exceptionally significant astrophysical structures due to their unique physical characteristics and their essential role in cosmological studies. Their importance spans multiple dimensions, including astrophysics, cosmology, and galaxy evolution, and is primarily due to their status as the largest gravitationally bound systems in the Universe.

Largest Gravitationally Bound Systems: Galaxy clusters represent the most massive and extensive gravitationally bound objects known in the cosmos, with masses ranging typically from 10^{14} to 10^{15} solar masses $M_\odot$. This immense gravitational binding power creates a deep gravitational potential well, enabling clusters to retain vast amounts of hot intracluster medium (ICM) gas and thousands of galaxies within their volume. Such gravitational dominance enables clusters to be stable, quasi-equilibrium systems that offer valuable insights into gravitational dynamics and dark matter properties [142].

Key Role in Structure Formation: Clusters serve as critical probes in studying hierarchical structure formation. According to the Λ Cold Dark Matter (ΛCDM) cosmological framework, large-scale structures form hierarchically from initial small density perturbations in the early Universe. Galaxy clusters are thus endpoints of this hierarchical process, formed via successive mergers and accretion events involving smaller halos and galaxy groups. Analyzing clusters helps astronomers reconstruct how cosmic structures evolved over time, offering direct tests of theoretical predictions [159, 171].

Cosmological Laboratories: Galaxy clusters are potent cosmological tools because their number density, spatial distribution, and internal structures strongly depend on cosmological parameters such as matter density (Ω_m), dark energy density (Ω_Λ), and the Hubble constant (H_0). By comparing observed cluster abundances, mass distributions, and scaling relations (e.g., between mass, temperature, and luminosity) with theoretical predictions, astronomers can place constraints on cosmological models and parameters. This makes clusters vital in addressing outstanding cosmological puzzles, including the nature of dark matter, dark energy, and the Universe's expansion history [3].

Laboratories for Galaxy Evolution and Baryonic Physics: The cluster environment profoundly influences galaxy evolution processes. Interactions such as ram-pressure stripping, galaxy harassment, and tidal interactions are intensified in clusters due to the dense environment and high relative velocities. Clusters thus offer valuable observational settings to study how environment-driven processes regulate star formation rates, morphological transformations, and AGN activities in galaxies [18, 41].

Tests of Fundamental Physics: Galaxy clusters also facilitate tests of fundamental physics, including gravity theories and neutrino physics. Gravitational lensing observations, galaxy motions, and the thermal state of the ICM can probe deviations from general relativity, constrain neutrino masses, and test potential modifications to standard particle physics [35].

In summary, galaxy clusters are pivotal across astrophysics and cosmology, serving as laboratories to understand gravitational dynamics, cosmological structure

formation, galaxy evolution processes, and fundamental physical laws governing the Universe.

1.2 ICM AND ITS OBSERVATIONAL PROBES

What is Intracluster Medium?

The Intracluster Medium (ICM) constitutes the diffuse hot plasma that pervades the space between galaxies within galaxy clusters. Composed primarily of ionized hydrogen and helium, with trace amounts of heavier elements, this plasma represents the majority of baryonic mass in clusters and plays a fundamental role in their evolution and observable properties. Typical temperatures of the ICM range from approximately 10^7 to 10^8 Kelvin. Such high temperatures result from the virialization of gravitational potential energy released during cluster formation and subsequent merger events [142].

The temperature profile of the ICM typically shows variations from the cluster core to the outskirts, influenced by heating and cooling processes, including shocks induced by mergers, AGN feedback, and radiative cooling in the densest regions. Spatially resolved temperature measurements can be used for understanding cluster dynamics and energy distribution [165].

Mass fractions within galaxy clusters reveal that dark matter dominates, accounting for approximately 85% of the total cluster mass. However, the baryonic component, primarily the ICM, constitutes roughly 12-15% of the total mass, significantly exceeding the mass contained within galaxies. Precise measurements of ICM mass fractions provide cosmological constraints, particularly regarding the matter density parameter, Ω_m, and the baryon fraction, impacting cosmological models [32].

Within galaxy clusters, the ICM overwhelmingly dominates the baryonic mass budget, representing around 80-90% of the total baryonic content, while galaxies themselves typically account for merely 10-20%. This distribution underscores the ICM's significance for understanding cluster formation and evolution, as well as broader cosmological processes [32].

The mass fraction of the ICM is typically derived using X-ray observations of its diffuse emission, combined with gravitational mass estimates from lensing or dynamical methods. Consistent with hydrodynamical cosmological simulations, these measurements provide essential validation of cosmological models, particularly the predictions of baryon-to-dark matter ratios and the thermal state of baryonic matter on large scales [32].

The spatial variation of the ICM mass fraction within clusters also carries vital information about processes such as gas cooling, star formation, and feedback mechanisms. Cluster cores generally exhibit lower ICM fractions due to enhanced cooling and galaxy formation, whereas outer regions better reflect cosmic baryonic abundances, making them ideal for cosmological studies [133].

Main Observational Probes of the ICM

The ICM is studied through a diverse range of observational techniques spanning the electromagnetic spectrum, each providing complementary insights into its properties and dynamics.

X-ray Observations: The dominant probe of the ICM is its diffuse X-ray emission, which arises primarily from thermal bremsstrahlung (free-free radiation) produced as electrons are deflected by the Coulomb fields of ions in the hot plasma. For a fully ionized, optically thin plasma in collisional ionization equilibrium, the volume emissivity of thermal bremsstrahlung is approximately

$$\epsilon_\nu^{\mathrm{ff}} \propto n_e n_i T^{-1/2} \exp\left(-\frac{h\nu}{k_B T}\right), \tag{1}$$

where n_e and n_i are the electron and ion number densities, respectively, and T is the plasma temperature [141]. The exponential cutoff at photon energies $\sim k_B T$ renders the bremsstrahlung spectrum a sensitive diagnostic of the ICM temperature, while the normalization of the continuum emission scales with n_e^2, making X-ray brightness a powerful tracer of gas density.

At temperatures typical of clusters (10^7–10^8 K), the X-ray emission extends over the 0.1–10 keV band, well matched to the sensitivity of current-generation satellites such as *Chandra*, *XMM-Newton*, and *eROSITA*. These instruments provide high-resolution imaging and spectroscopy that enable the reconstruction of radial profiles of gas density, temperature, entropy, and pressure, which are essential for constraining hydrostatic mass estimates and baryon fractions [133, 165].

Superimposed on the continuum emission are prominent emission lines from highly ionized heavy elements, most notably the Fe XXV and Fe XXVI Kα complexes near 6.7–6.9 keV, as well as lines from lighter elements such as O, Ne, Mg, Si, and S. The equivalent widths and ratios of these lines provide direct measurements of the ICM metallicity and relative abundance patterns [16]. Such data reveal that the ICM is enriched to $\sim$0.3 solar metallicity, implying a long history of star formation and feedback processes (including both core-collapse and Type Ia supernovae) that distributed metals from galaxies into the surrounding medium [170]. The spatial distribution of metals further encodes the efficiency of feedback, turbulent mixing, and cluster assembly history.

Beyond static thermodynamic diagnostics, X-ray observations can also probe dynamical processes. Sharp surface-brightness discontinuities reveal shock fronts and cold fronts generated during cluster mergers, allowing direct measurements of shock Mach numbers and insights into plasma transport processes such as viscosity, conduction, and magnetic field draping [98]. Measurements of line broadening and centroid shifts with high-resolution spectroscopy (e.g., with *Hitomi* and its planned successor *XRISM*) open a new window on ICM turbulence and bulk motions, quantifying the non-thermal pressure support that affects hydrostatic mass determinations [62].

X-ray observations provide the most direct and comprehensive means of characterizing the thermodynamic state, chemical enrichment, and dynamical activity of the ICM. They form the backbone of cluster astrophysics, enabling determination of cluster mass profiles, baryon fractions, and feedback histories, while also serving as indispensable inputs to cosmological applications based on cluster scaling relations and abundance studies.

Radio Observations: Radio observations provide a complementary window into the intracluster medium (ICM), probing its non-thermal components (e.g., relativistic particles and magnetic fields) that are otherwise invisible in thermal X-ray studies. The dominant mechanism is synchrotron radiation, produced by relativistic

electrons (with Lorentz factors $\gamma \sim 10^3$–10^5) spiraling around cluster-scale magnetic fields of order μG. The synchrotron emissivity at frequency ν is approximately

$$j_\nu \propto N(\gamma) B^{1+\alpha} \nu^{-\alpha}, \tag{2}$$

where $N(\gamma)$ is the electron energy distribution, B is the magnetic field strength, and α is the synchrotron spectral index [141]. The observed radio continuum spectrum thus directly encodes the relativistic electron population and the magnetic field properties.

CLUSTER-SCALE RADIO PHENOMENA. Large diffuse radio sources, unassociated with individual galaxies, are now well established as signatures of merger-driven activity in clusters. Three principal classes are identified:

- *Radio halos:* Mpc-scale, centrally located diffuse emissions with steep spectra ($\alpha \sim 1.1$–1.5), typically tracing the distribution of the thermal ICM. Radio halos are correlated with dynamically disturbed, merging clusters and are thought to be powered by turbulence reaccelerating relativistic electrons throughout the cluster volume [19, 27].
- *Radio relics:* Elongated, peripheral sources aligned with merger shock fronts. Their morphology and polarization patterns support an origin in diffusive shock acceleration of cosmic-ray electrons at large-scale merger shocks [74, 180]. Relics can reach sizes of several Mpc and provide direct constraints on shock Mach numbers and magnetic field amplification processes.
- *Mini-halos:* Smaller (~100–500 kpc) diffuse sources surrounding brightest cluster galaxies (BCGs) in cool-core clusters. These are likely sustained by turbulence generated by AGN feedback and gas sloshing, offering insights into the interplay between cooling flows and non-thermal processes [54].

Faraday rotation measurements of background and embedded radio sources demonstrate that intracluster magnetic fields typically have strengths of order a few μG, with coherence lengths of 10–100 kpc [22]. The presence of large-scale synchrotron emission implies efficient acceleration or reacceleration of relativistic electrons, since their radiative lifetimes ($\sim 10^8$ yr) are much shorter than cluster dynamical times. This necessitates in-situ acceleration, achieved either through merger-driven turbulence (turbulent reacceleration models) or through shock acceleration (diffusive shock acceleration, DSA). These processes also tie radio emission directly to the cluster dynamical state.

Recent advances in low-frequency radio interferometry, particularly with the Giant Metrewave Radio Telescope (GMRT), the Low Frequency Array (LOFAR), and the upgraded Very Large Array (VLA), have enabled detailed studies of diffuse cluster emission at high sensitivity and resolution [179]. Upcoming surveys with the Square Kilometre Array (SKA) promise to revolutionize the field, providing orders-of-magnitude improvements in sensitivity and sky coverage. These facilities will enable statistical studies of halos and relics across cosmic time, constraining the non-thermal energy budget of clusters and the microphysics of cosmic-ray acceleration and magnetic field amplification.

Radio observations uniquely probe the relativistic particle populations and magnetic fields of the ICM, offering important insights into merger dynamics, feedback processes, and plasma microphysics. Together with thermal diagnostics from X-rays and SZ measurements, they are indispensable for constructing a complete picture of the thermal and non-thermal energy balance in galaxy clusters.

Gravitational Lensing: Gravitational lensing provides an independent, mass-based observational probe of galaxy clusters. By measuring the deflection of light from distant background galaxies, lensing enables precise determinations of total cluster mass profiles without relying on assumptions about dynamical equilibrium or the state of the gas. Strong lensing phenomena, such as giant arcs and Einstein rings, occur in dense cluster cores, while weak lensing provides mass measurements extending to larger radii. Lensing studies are essential for validating hydrostatic mass estimates derived from X-ray and SZ measurements and for testing cosmological models [104].

Optical Observations: Optical observations complement other techniques by studying cluster galaxies and their interactions with the ICM. Spectroscopic and photometric surveys provide crucial insights into galaxy evolution processes such as star formation quenching, morphological transformations, and AGN feedback driven by interactions with the ICM. Optical studies also identify cluster memberships, measure galaxy velocities, and contribute significantly to cluster mass estimations via dynamical methods. Furthermore, optical imaging facilitates the identification of gravitational lensing signatures and the study of the galaxy luminosity function and stellar populations within clusters, shedding light on galaxy formation histories and their connection to ICM processes [41, 139].

Sunyaev–Zeldovich Effect: The Sunyaev–Zeldovich (SZ) effect describes the distortion of the cosmic microwave background (CMB) radiation spectrum caused by inverse Compton scattering of CMB photons by high-energy electrons in the ICM. The SZ effect provides a unique, redshift-independent probe of cluster properties, particularly valuable for studying distant clusters. Observatories such as the Atacama Cosmology Telescope (ACT), the South Pole Telescope (SPT), and Planck have systematically surveyed clusters via SZ observations, significantly expanding our cluster catalogs and enhancing cosmological constraints. SZ measurements allow for precise determination of cluster pressures, masses, and even peculiar velocities through kinematic SZ effects, offering a powerful method for cosmological and astrophysical studies [23, 129].

Figure 1 highlights the complementarity of observational probes: galaxies and dark matter from lensing, thermal gas from X-ray emission, and relativistic components from radio synchrotron emission. Together, they establish galaxy clusters as laboratories where gravitational, thermal, and non-thermal physics can be jointly constrained.

This multiwavelength view (Fig 1) highlights the complementarity of observational probes: galaxies and dark matter from lensing, thermal gas from X-ray emission, and relativistic components from radio synchrotron emission. Together, they establish galaxy clusters as laboratories where gravitational, thermal, and non-thermal physics can be jointly constrained.

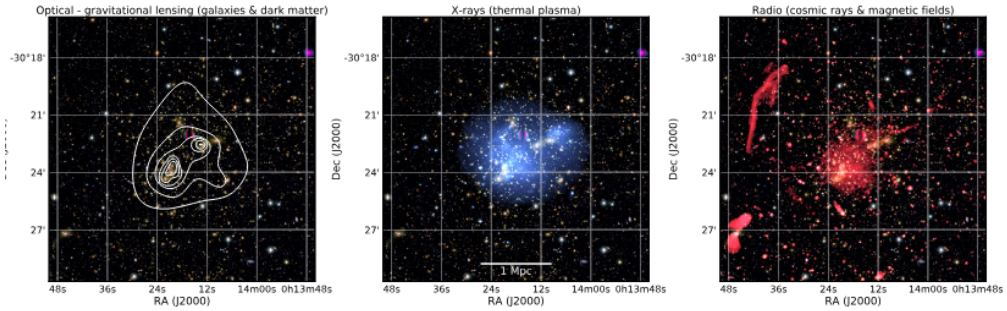

Figure 1: The galaxy cluster Abell 2744. **Left:** optical (Subaru BRz; Medezinski et al. 2016) view of the cluster. White linearly spaced contours represent the mass surface density ($\kappa = \Sigma/\Sigma_{cr}$) from weak-lensing studies [95, 105]. **Middle:** *Chandra* X-ray emission (0.5–2.0 keV band) from the hot thermal ICM (blue). **Right:** 1–4 GHz VLA radio image (red) tracing cosmic rays and magnetic fields. Figure adapted from van Weeren et al. [179].

2

GALAXY CLUSTERS AS THE END RESULTS OF HIERARCHICAL GROWTH IN LAMBDA CDM

2.1 LAMBDA CDM MODEL AND COSMOLOGICAL SIGNIFICANCE OF GALAXY CLUSTERS

Components of ΛCDM: Cosmological Constant, Cold Dark Matter, Ordinary Matter

The cosmological model, Λ Cold Dark Matter (ΛCDM), specifies the homogeneous expansion through the Friedmann equations and the inhomogeneous growth of structure through gravitational instability. Its matter-energy content is summarized by dimensionless density parameters $\Omega_i = \rho_i/\rho_{crit}$ with

$$\rho_{crit}(z) \equiv \frac{3H^2(z)}{8\pi G}, \qquad H^2(a) = H_0^2\left[\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda\right], \tag{3}$$

where $a = (1+z)^{-1}$ is the scale factor, $\Omega_m = \Omega_c + \Omega_b$ is the total non-relativistic matter (cold dark matter + baryons), Ω_r includes photons and (effectively massless) neutrinos, $\Omega_k \equiv 1 - \sum_i \Omega_i$ encodes spatial curvature, and Ω_Λ represents the cosmological constant [24].

COSMOLOGICAL CONSTANT (Λ) Originally introduced to permit a static solution to Einstein's field equations [46], the cosmological constant acts as a uniform energy density with stress-energy tensor $T^{(\Lambda)}_{\mu\nu} = -\rho_\Lambda g_{\mu\nu}$, corresponding to an equation of state $p_\Lambda = w_\Lambda \rho_\Lambda$ with $w_\Lambda = -1$. As emphasized by Zel'dovich [174], Λ can be interpreted as the energy density of the quantum vacuum. A positive Λ accelerates the expansion at late times and provides a simple explanation of the supernova Hubble diagram, in which Type Ia supernovae appear dimmer than expected in a decelerating universe [124, 137]. In the background dynamics, Ω_Λ is spatially homogeneous and does not cluster on sub-horizon scales; its principal role in the present context is to set the recent expansion history against which cluster abundances and growth are calibrated.

COLD DARK MATTER (CDM). Evidence for a non-luminous gravitating component dates back to galaxy clusters [177] and galaxy rotation curves [140]. In the CDM hypothesis, this component is non-relativistic by the epoch of matter-radiation equality and effectively collisionless on astrophysical scales. Its pressure is negligible ($w_c = 0$), so its background density scales as $\rho_c \propto a^{-3}$ and it clusters on all scales above its tiny free-streaming length. The CDM paradigm provides a successful framework for hierarchical structure formation: small halos collapse first, later merging into progressively larger systems [14]. In linear theory, CDM dominates the gravitational potential wells that seed the growth of baryonic structure; in the non-linear regime, it sets the halo mass function and clustering against which groups and clusters are identified and modeled throughout this thesis.

Cold dark matter (CDM) is effectively pressureless and non-relativistic by matter–radiation equality, with a negligible free-streaming length; it preserves small-

scale power and yields bottom-up (hierarchical) assembly in which small halos form first and merge into larger systems. Warm dark matter (WDM) consists of particles with keV-scale masses and residual thermal velocities that erase primordial fluctuations below a characteristic *half-mode* scale (free-streaming lengths $\sim$ 0.1–1 Mpc), suppressing low-mass halos and substructure relative to CDM [15]. Hot dark matter (HDM), exemplified by eV-scale neutrinos, free-streams over tens of Mpc when non-relativistic, wiping out small-scale perturbations and producing a top-down ("pancake") sequence of structure formation, incompatible with the observed galaxy distribution and clustering [38]. Current large-scale structure data, particularly the Lyman-α forest, strongly limit WDM-like suppression, implying lower bounds on the WDM particle mass of a few keV (model dependent) [164].

ORDINARY (BARYONIC) MATTER. Baryons (with electrons) constitute the ordinary matter content, sharing $w_b \approx 0$ at late times and thus $\rho_b \propto a^{-3}$. Unlike CDM, baryons experience pressure forces and, prior to recombination, are tightly coupled to photons. Acoustic oscillations in this photon-baryon fluid imprint a characteristic scale in the matter distribution [122] and, due to diffusion (Silk) damping, erase small-scale fluctuations in the baryons relative to CDM [152]. After recombination and reionization, baryons fall into CDM potential wells and undergo complex hydrodynamics, cooling, star formation, and feedback. In the high-mass halos of interest here, the bulk of baryons resides not in stars but in the hot, diffuse ICM, whose thermodynamics and observable tracers (X rays, and Radio effect) we will use as probes of growth and assembly in later chapters.

ENERGY-BUDGET PERCENTAGES AT $z \approx 0$. In the concordance ΛCDM cosmology, the present-day energy density is dominated by dark energy with a subdominant matter component split into cold dark matter and baryons. Using the Planck Collaboration et al. [130] base ΛCDM solution, the inferred parameters are $\Omega_m = 0.315 \pm 0.007$ with $\Omega_b \simeq 0.049$ and $\Omega_c \simeq 0.264$, and $\Omega_\Lambda = 0.685 \pm 0.007$; radiation today is negligible ($\Omega_r \sim 10^{-4}$).[1] Expressed as percentages, the cosmic inventory is therefore $\sim$ 68.5% dark energy, $\sim$ 26.4% cold dark matter, and $\sim$ 4.9% baryons, implying a cosmic baryon fraction $f_b = \Omega_b/\Omega_m \approx 0.156$ [130].

In this framework, galaxy clusters emerge as rare, high-mass peaks in the cosmic density field, making them exceptionally sensitive tracer of both the matter content and the expansion history of the universe.

Success of the Theory

The ΛCDM model gives a coherent, quantitative picture of the Universe from early times to the present. It links the initial conditions seen in the cosmic microwave background (CMB) to the web of galaxies we map today, and it connects the growth of structure to the measured expansion history. Here we highlight three pillars of this success.

1 Values vary at the percent level with data combinations and neutrino assumptions; curvature is consistent with zero, $\Omega_k \approx 0$.

LARGE-SCALE STRUCTURE. Wide-area galaxy surveys measure the clustering of galaxies over hundreds of megaparsecs. Their two-point statistics and power spectra agree with ΛCDM predictions when the model is seeded with nearly scale-invariant, Gaussian fluctuations. The baryon acoustic oscillation (BAO) feature appears at the expected comoving scale and serves as a standard ruler in the late-time Universe [48]. On the theory side, large N-body and hydrodynamical simulations within ΛCDM reproduce the observed filament–node pattern and its statistics, providing a bridge from initial conditions to the present-day galaxy distribution [159]. Independently, CMB temperature and polarization maps pin down the initial fluctuation spectrum and key background parameters, and are well fit by the same model [130].

GALAXY AND CLUSTER ABUNDANCES. In ΛCDM, the number of dark-matter halos as a function of mass and redshift (the halo mass function) can be predicted and calibrated on simulations. These predictions match observations over many decades in mass when selection effects and mass proxies are handled carefully [163]. In particular, counts of massive galaxy clusters-identified in X-ray, SZ, or optical data-and their evolution constrain Ω_m and σ_8 in a way that is consistent with CMB-inferred values, within current systematic uncertainties [166]. The overall picture supports hierarchical growth: small halos form first and later assemble into groups and clusters.

EXPANSION HISTORY. Distance-redshift measurements from Type Ia supernovae show that the expansion of the Universe is accelerating at late times [124, 137]. When combined with the CMB acoustic scale and the BAO ruler, these data select a spatially flat model with matter plus a cosmological constant that also accounts for structure growth and halo abundances [48, 130]. In this sense, the background expansion and the growth of structure are described by a single, self-consistent parameter set-one of the hallmark successes of ΛCDM.

Because galaxy clusters sit on the high-mass tail of the halo population and form at the intersections of filaments, their statistics and internal properties provide sensitive tests of all three pillars above. We will make use of this in the next sections.

Open Issues and Challenges

Despite its broad successes, ΛCDM faces several active challenges:

SMALL-SCALE TENSIONS. On galaxy scales, three long-standing discrepancies are often discussed. The *cusp–core* problem refers to central density profiles of dwarfs that appear shallower than the cusps seen in dark-matter–only simulations; the *missing satellites* problem is the apparent shortfall of observed dwarf satellites compared with the large number of predicted low-mass subhalos; and *too big to fail* highlights simulated subhalos that are too dense to host the brightest observed satellites. Baryonic processes (bursty feedback, tides, reionization) can reduce some of these gaps, but a full, joint solution across systems is still being tested [21].

THE HUBBLE TENSION. Independent measurements of the present-day Hubble constant disagree at the several-sigma level. Inference from the early Universe

using CMB data within base ΛCDM gives $H_0 \simeq 67.4 \pm 0.5\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$, while local distance-ladder determinations using Cepheids and Type Ia supernovae prefer $H_0 \simeq 73\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$ with $\sim 1\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$ uncertainty. This persistent offset may point to unrecognized systematics or to extensions of the model (e.g., early dark energy), but no consensus solution has emerged [130, 138].

EARLY MASSIVE GALAXIES. Deep near-infrared imaging has revealed candidates for very massive galaxies at $z \geqslant 7$ with stellar masses $\geqslant 10^{10}\ M_\odot$ and high number densities. If confirmed, such systems require rapid assembly shortly after reionization and put pressure on simple forms of star-formation histories and feedback at early times. Improved spectroscopic redshifts and mass estimates are refining these samples, but the overall picture continues to motivate tests of early galaxy formation within ΛCDM [87].

How Clusters Serve as Cosmological Probes?

Galaxy clusters are rare, high-mass halos that sit on the exponential tail of the halo mass function. Their *number density and spatial distribution* are therefore very sensitive to the amplitude and growth of matter fluctuations and to the expansion history. In parallel, clusters host a hot, X-ray–bright intracluster medium (ICM) whose thermodynamics encodes the depth of the potential well. Together, these facts allow cluster surveys to constrain key cosmological parameters when mass–observable relations and selection effects are under control.

ABUNDANCE AND DISTRIBUTION. The basic idea is simple: very massive halos are rare, and their rarity depends strongly on how fast structure grows. This is captured by the halo mass function

$$\frac{dn}{dM}(M,z) = f(\sigma)\,\frac{\bar{\rho}_m}{M}\,\frac{d\ln\sigma^{-1}(M,z)}{dM},$$

where $\sigma(M,z) = D(z)\,\sigma(M,0)$ is the rms fluctuation of the density field on the mass scale M, and $D(z)$ is the linear growth factor. The function $f(\sigma)$ summarizes the collapse statistics and is calibrated from theory and simulations [134, 149, 163]. Because clusters live on the high-mass (low-σ) tail, even small changes in the growth (through $D(z)$), in the fluctuation amplitude (σ_8), or in the matter density (Ω_m) lead to large changes in the predicted number of clusters.

To compare with data, one maps mass to an observable O (e.g., X-ray flux, SZ signal, optical richness) and folds in the survey *selection function* $S(O,z)$ that tells which objects are detected at each redshift. Counts as a function of redshift then test the combination of growth and expansion encoded by the cosmological model.

Beyond the total counts, the *spatial distribution* of clusters also carries information. Massive halos are more strongly clustered than the matter field; this "bias" $b(M,z)$ grows with mass and can be measured from the two-point clustering of the cluster sample [71]. Fitting counts and clustering together helps break degeneracies (e.g., between σ_8 and Ω_m) and provides a stronger, self-consistent cosmological test.

INTERNAL STRUCTURE AND SELF-SIMILAR SCALINGS. In a gravity-only, self-similar picture, halo structure and thermodynamics follow simple scalings with mass and redshift: $T \propto M^{2/3}E^{2/3}(z)$ and (for bremsstrahlung-dominated emission)

$L_X \propto E^{7/3}(z)\, M^{4/3}$, where $E(z) \equiv H(z)/H_0$ [72]. The ICM thermal energy scales as $E_{th} \propto M^{5/3} E^{2/3}(z)$, motivating low-scatter mass proxies that trace integrated pressure or gas mass. Departures from exact self-similarity-driven by cooling, star formation, and feedback-are informative systematics that can be modeled and, increasingly, calibrated with simulations and multiwavelength data.

X-RAY AND SZ OBSERVABLES AS MASS PROXIES. The X-ray emissivity of hot, optically thin plasma is $\epsilon_X \propto n_e^2 \Lambda(T, Z)$, allowing deprojection of gas density profiles; with spatially resolved temperatures, the hydrostatic equation gives

$$M(< r) = -\frac{k_B T(r)\, r}{G\, \mu m_p} \left(\frac{d \ln n_e}{d \ln r} + \frac{d \ln T}{d \ln r} \right),$$

yielding M_Δ^{HSE} profiles under the assumption of equilibrium [142]. Non-thermal pressure support (bulk motions, turbulence) biases M^{HSE} low at the ~10% level in relaxed systems and more in disturbed ones, motivating external calibration. Two robust, low-scatter X-ray mass proxies widely used in counts analyses are the gas mass M_{gas} and $Y_X \equiv M_{gas} T$, the latter tracing total thermal energy and scaling nearly self-similarly with mass ($Y_X \propto M^{5/3} E^{2/3}$) [85].

The thermal Sunyaev-Zel'dovich (tSZ) effect measures the line-of-sight integral of electron pressure,

$$y = \frac{\sigma_T}{m_e c^2} \int n_e k_B T \, dl, \qquad Y \equiv \int y \, d\Omega \propto \frac{E_{th}}{D_A^2},$$

and thus provides an (approximately) redshift-independent selection for massive clusters [160, 161]. The integrated SZ signal Y (or Y_{500}) exhibits low intrinsic scatter at fixed mass and follows the same $M^{5/3} E^{2/3}$ trend expected from self-similarity; its normalization and scatter can be calibrated with X-ray data and weak-lensing masses [7].

MASS CALIBRATION AND SYSTEMATICS. Cosmological inference from cluster counts hinges on the calibration of the mean relation $\langle \ln O | M, z \rangle$ and its intrinsic scatter, along with the survey selection $S(O, z)$ and observable noise. Malmquist and Eddington biases must be modeled when mapping between O and M. *Weak gravitational lensing* offers a direct, nearly unbiased mass calibration for ensemble averages [73], anchoring the mass scale used with X-ray and SZ proxies. Additional cross-checks come from the *gas-mass fraction* method, which leverages the near-universality of f_{gas} in massive, relaxed clusters to constrain Ω_m and the distance-redshift relation [2]. With these elements in hand, cluster counts and their redshift evolution provide competitive constraints on σ_8 and Ω_m and can probe the dark-energy equation of state when combined with BAO and CMB information [130, 166].

Cluster *abundances and clustering* test the growth of structure; *scaling relations* connect observables to mass; *X-ray and SZ measurements* supply low-scatter proxies and selection. When tied together with careful mass calibration, clusters act as precise cosmological probes while simultaneously informing baryonic physics in massive halos.

2.2 FORMATION OF CLUSTERS

Initial Conditions and Early Structure Formation

This subsection covers the *linear* stage only; how the initial perturbations were set and processed up to the time when baryons could fall into dark-matter wells. The non-linear build–up of protoclusters and clusters follows in the next subsection.

INFLATION AND THE ORIGIN OF PERTURBATIONS ($t \sim 10^{-36}$–$10^{-32}\,s$). A short period of accelerated expansion stretches quantum fluctuations to super-horizon scales, turning them into classical, nearly Gaussian curvature perturbations with a close-to scale-invariant spectrum [11, 59]. These perturbations are the seeds of all later structure.

REHEATING AND THE RADIATION ERA ($t \leqslant 1\,s$ TO $t \sim 50{,}000\,yr$; $z \geqslant 3400$). When inflation ends, the inflation decays and fills the Universe with a hot plasma (*reheating*). The Universe is then radiation dominated. During this phase, sub-horizon modes in the photon–baryon fluid undergo acoustic oscillations, while cold dark matter (CDM) does not feel pressure and begins to set up the potential wells that will later guide baryons.

MATTER–RADIATION EQUALITY ($t \approx 50{,}000\,yr$; $z \approx 3400$). Once matter dominates the energy density, CDM perturbations grow roughly as the scale factor in linear theory. The linear matter power spectrum can be written as

$$P(k, z) = A_s\, k^{n_s}\, T^2(k)\, D^2(z), \tag{4}$$

with primordial amplitude and tilt (A_s, n_s) set by inflation, transfer function $T(k)$ encoding horizon-entry physics, and growth factor $D(z)$ describing linear growth [10]. Modes that entered the horizon during radiation domination were suppressed relative to large scales, imprinting the CDM "turnover" in $P(k)$.

RECOMBINATION AND DECOUPLING ($t \approx 380{,}000\,yr$; $z \approx 1100$). Electrons and protons combine, the photon mean free path jumps, and the CMB is released. After decoupling, baryons fall into existing CDM potential wells. The earlier sound waves in the photon–baryon fluid leave baryon acoustic oscillations (BAO) as a standard comoving ruler in the matter distribution [122, 152].

EARLY LINEAR GROWTH INTO THE COSMIC WEB SKELETON ($t \sim 10^7$–$10^9\,yr$; $z \sim 30$–6). As $D(z)$ increases, density contrasts on progressively larger mass scales approach unity. The initial field is well modeled as Gaussian, so its peaks and tidal field set the preferred directions of collapse that will later become sheets and filaments; the skeleton onto which non-linear structures will assemble [10, 175].

Inflation sets the initial spectrum; the radiation and matter eras shape $T(k)$; recombination frees baryons to follow CDM. By $z \sim 6$ ($t \sim 1\,Gyr$), the stage is set for non-linear assembly along the emerging web.

Proto-Clusters and High-Redshift Cluster Formation

We now move to the *non-linear* regime: how rare peaks turn into protoclusters and, later, massive, bound clusters.

ONSET OF NON-LINEARITY AND ANISOTROPIC COLLAPSE ($t \sim 0.1$–$1\,\mathrm{Gyr}$; $z \sim 30$–6). When the variance on a given mass scale reaches unity, linear theory breaks down. In the Zel'dovich picture, matter collapses first along one axis to form sheets (walls), then along a second to form filaments, and finally along the third to form dense nodes [175]. This produces the filamentary *cosmic web* seen in simulations and surveys [17]. Rare, high peaks (large $\nu = \delta/\sigma$) sit at filament intersections, i.e. preferred sites for future clusters.

SPHERICAL COLLAPSE AND HALO FORMATION TIMES. A region of mass M virializes when its *linearly extrapolated* overdensity reaches $\delta_c \simeq 1.686$ (weakly cosmology dependent), setting a mapping between the initial field and collapse redshift [58]. This framework underlies the halo mass function and the hierarchical picture in which low-mass halos form earlier and later merge into larger systems [88, 134].

PROTOCLUSTERS ($t \sim 1$–$3\,\mathrm{Gyr}$; $z \sim 6$–2). Before full virialization, overdense regions traced by multiple converging filaments appear as *protoclusters*. They are extended (tens of comoving Mpc), highly anisotropic, and actively accreting. Star formation and black-hole growth are vigorous, and early intra-halo gas is being shock-heated. These structures mark the future nodes of the web but are not yet relaxed clusters [117].

ASSEMBLY INTO BOUND CLUSTERS ($t \sim 4$–$8\,\mathrm{Gyr}$; $z \sim 1$–0.5) AND MATURATION TO TODAY ($t \approx 13.8\,\mathrm{Gyr}$; $z = 0$). Through a mix of smooth accretion and mergers (major and minor) guided by the surrounding filaments, many protoclusters reach $M_{200c} \geqslant 10^{14}\,M_\odot$ and become fully bound clusters by $z \sim 1$. Their intracluster medium (ICM) records this assembly via shocks, cold fronts, and turbulence; substructure and ellipticity reflect ongoing anisotropic infall. By $z \leqslant 0.5$, a large fraction of the most massive systems ($M_{200c} \geqslant 10^{15}\,M_\odot$) have assembled, though accretion continues along preferred directions.

OBSERVATIONAL EVIDENCE FOR PROTO-CLUSTERS. Deep surveys have identified high-redshift proto-clusters ($z > 2$) as overdensities of galaxies, star-forming regions, and quasars. They exhibit vigorous star formation, intense AGN activity, and significant gas inflows, consistent with rapid growth in a dense environment [26, 31]. These observational results complement numerical simulations, which link proto-clusters to the mature clusters observed today. Understanding their properties is crucial for constraining early structure formation and feedback processes.

In conclusion, galaxy clusters trace their origin to rare peaks in the primordial density field. Their growth from anisotropic collapse to proto-clusters, and finally into massive bound clusters, encapsulates the full history of hierarchical structure formation.

3

THE INTERACTIONS BETWEEN GALAXY CLUSTERS AND THEIR SIGNATURES

3.1 DYNAMICS AND OBSERVATIONAL EFFECTS

Classification Between Relaxed and Disturbed

Galaxy clusters span a continuum of dynamical states, but are often classified into two broad categories: *relaxed* and *disturbed* systems. Relaxed clusters exhibit regular, approximately spherical morphologies, centrally peaked X-ray surface brightness, and smooth mass distributions that are consistent with hydrostatic equilibrium. They frequently host cool cores; dense, low-entropy gas in the center and their thermodynamic profiles (density, temperature, entropy, pressure) follow simple, nearly self-similar forms [133].

Disturbed clusters, in contrast, display irregular or multimodal morphologies, asymmetric X-ray isophotes, offsets between the brightest cluster galaxy (BCG) and the X-ray peak, and substantial substructures in both galaxies and dark matter. These features typically indicate recent or ongoing mergers. The distinction is not merely morphological: relaxed and disturbed clusters exhibit systematically different scaling relations, baryon distributions, and lensing–X-ray mass offsets, making dynamical classification essential for cosmological applications [136].

Quantitative metrics have been developed to distinguish between these states, including: (i) *centroid shifts* (variance of X-ray centroid positions with aperture), (ii) *power ratios* of X-ray surface brightness multipoles, and (iii) *concentration parameters* (ratio of central to global surface brightness). These provide robust, reproducible measures of relaxation across large samples.

Substructure and Merger Signatures

Galaxy cluster mergers, the most energetic events since the Big Bang, leave distinct multiwavelength imprints on the intracluster medium (ICM), galaxy population, and dark matter distribution.

X-RAY SIGNATURES. Mergers drive shocks and turbulence into the ICM. Shock fronts appear as sharp surface-brightness and temperature jumps, heating gas and increasing X-ray luminosity locally. The Bullet Cluster is the canonical example: a supersonic subcluster has generated a bow shock visible in *Chandra* images, with Mach number $\mathcal{M} \sim 3$ [98]. Cold fronts, contact discontinuities between gas phases of different entropies, arise from gas sloshing or core–core encounters, appearing as sharp edges in surface brightness but without shock heating [98]. Disturbed clusters often show highly asymmetric X-ray morphologies, multiple peaks, and deviations from hydrostatic equilibrium.

RADIO SIGNATURES. Radio observations reveal the non-thermal consequences of mergers. *Radio relics*, such as the Toothbrush Relic, trace peripheral shocks through elongated, polarized synchrotron emission aligned with the shock surface. *Radio halos*, in contrast, are diffuse, centrally located sources extending over ~Mpc scales, powered by turbulence re-accelerating relativistic electrons throughout the cluster volume [19, 179]. *Mini-halos* in cool-core clusters trace turbulence generated by AGN feedback and gas sloshing. The presence, morphology, and spectrum of these diffuse sources provide a direct window on the efficiency of particle acceleration and magnetic field amplification during mergers.

GRAVITATIONAL LENSING SIGNATURES. Mergers can decouple collisionless dark matter from collisional baryons, producing offsets between galaxy, gas, and dark matter distributions. In the Bullet Cluster, weak and strong lensing maps reveal dark matter peaks displaced from the hot gas, offering compelling evidence for the collisionless nature of dark matter [35]. Lensing also uncovers mass substructures otherwise invisible in X-rays, enabling direct tests of structure formation and dark matter physics.

Together, these signatures establish a multiwavelength framework for identifying and characterizing mergers. Each observational window highlights different aspects of the dynamical state: X-rays trace the thermal plasma, radio emission traces relativistic particles and magnetic fields, and lensing reveals the total (dark + baryonic) mass distribution. As illustrative examples of merger diagnostics, Figure 2a shows the Bullet Cluster where a Mach ~2–3 bow shock and lensing–X-ray offsets are evident [35], while Figure 2b shows the "Toothbrush" relic whose elongated ridge and polarized bristles trace a large-scale merger shock [135].

Mass and Velocity Estimates

Accurate estimation of galaxy cluster masses and internal velocities is pivotal for cosmological analyses. Here we will explain primary methods used for this purpose:

DYNAMICAL METHODS. Galaxy redshift surveys provide velocity dispersions, which, under the virial theorem, yield dynamical mass estimates. However, these assume isotropy and equilibrium, which are violated in merging clusters. Substructures and velocity caustics can inflate dispersions in disturbed systems [67].

HYDROSTATIC METHODS. X-ray and SZ observations probe gas density and temperature (or pressure), from which hydrostatic equilibrium (HSE) masses can be derived:

$$M(<r) = -\frac{k_B T(r)\, r}{G\mu m_p}\left(\frac{d\ln n_e}{d\ln r} + \frac{d\ln T}{d\ln r}\right).$$

Non-thermal pressure from turbulence, bulk motions, and cosmic rays leads to a systematic *hydrostatic bias*, typically ~10–20% but larger in disturbed clusters [85].

GRAVITATIONAL LENSING. Weak and strong lensing provide mass estimates independent of dynamical assumptions, mapping the projected mass distribution di-

rectly. However, mergers and triaxiality can complicate the inversion and increase scatter in mass estimates [104].

SYSTEMATIC BIASES. Mergers bias scaling relations (e.g., M-T, M-L_X, M-Y_{SZ}) by temporarily boosting luminosities and temperatures, leading to mass overestimates if equilibrium is assumed. Combining X-ray, SZ, lensing, and dynamical methods is thus key for mass calibration, particularly in the disturbed regime [104, 133].

In summary, mergers imprint themselves across all cluster observables, complicating mass estimation but also providing powerful diagnostics of cluster dynamics, plasma physics, and the nature of dark matter.

3.2 GALAXY-GALAXY AND GALAXY-ICM INTERACTIONS

Galaxy-Galaxy Encounters in Clusters

Interactions among galaxies within clusters are fundamental in shaping galaxy properties. These encounters vary broadly, classified as either fast or slow. Fast encounters involve high-velocity, brief interactions predominantly in the dense cluster core regions, affecting galactic morphology and internal gas reservoirs minimally but potentially inducing transient star formation events. Slow encounters, on the other hand, predominantly occur within galaxy groups falling into clusters, allowing prolonged gravitational interactions that trigger significant tidal forces and gas stripping, profoundly influencing galaxy evolution [108].

Ram Pressure Stripping and Bow Shocks

As galaxies traverse the hot, dense intracluster medium at high velocities, they experience ram pressure, stripping gas from their disks. This phenomenon, known as ram-pressure stripping, significantly influences galaxy evolution, removing gas necessary for sustained star formation. Galaxies undergoing strong ram-pressure stripping exhibit characteristic observational signatures such as truncated gas disks, trailing gas tails, and enhanced star formation in the leading edges. Additionally, in extreme cases, supersonic motion can create bow shocks in the intracluster medium, visible in X-ray [58, 131].

Tidal Interactions and Stripping

Tidal interactions within clusters arise from gravitational influences exerted by the cluster potential and close galaxy encounters. These interactions distort galaxies' stellar and gaseous components, leading to tidal stripping; i.e. removal of material from galaxies outer regions. Tidal stripping profoundly influences galaxy morphology and mass, often producing tidal tails and bridges observable in optical and radio wavelengths. Particularly in dense cluster cores, tidal interactions significantly reshape galaxy structures, affecting subsequent evolutionary trajectories [18].

Impact on Galaxy Evolution: Morphological Transformation, Quenching, and AGN Activity

Galaxy interactions within clusters drastically alter galaxy evolution pathways, manifesting as morphological transformations, star formation quenching, and active galactic nucleus (AGN) activity modifications. Environmental processes, such as ram-pressure and tidal stripping, drive morphological transitions from late-type spiral galaxies to early-type ellipticals or lenticulars (S0). Additionally, removal of gas reservoirs efficiently quenches star formation, observable as a pronounced galaxy color-morphology-density relation. Environmental influences also modulate AGN activities, potentially either suppressing or enhancing nuclear activities through gas removal or funneling toward galactic centers, significantly affecting cluster galaxy properties [41, 123].

3.3 FEEDBACK PROCESSES

AGN Feedback in Clusters

Feedback from active galactic nuclei (AGN) plays a crucial role in regulating cluster thermal dynamics. Central cluster galaxies commonly host supermassive black holes actively accreting matter, releasing vast energies into the intracluster medium. AGN feedback manifests observationally through X-ray cavities, indicative of evacuated regions by relativistic jets, and radio bubbles visible at lower frequencies. Such feedback regulates intracluster gas cooling, maintaining the delicate balance necessary to explain observed cluster cooling flows and preventing catastrophic starburst scenarios [51].

Star Formation and Supernova Feedback

Star formation and associated supernova (SN) explosions significantly influence the intracluster medium's thermodynamic state. Star formation triggered by galaxy interactions and mergers releases energetic stellar winds and supernova-driven outflows into the ICM, redistributing metals and heating gas. These processes profoundly influence cluster gas dynamics, metal enrichment patterns, and temperature distributions, essential for interpreting cluster scaling relations and their cosmological implications [157].

3.4 OPEN QUESTIONS

Uncertainties in Cluster Mass and Dynamical State Estimates

Cluster mass estimation remains challenging due to systematic uncertainties introduced by projection effects, substructure contamination, and departures from hydrostatic equilibrium. Projection effects complicate distinguishing cluster-bound structures from background and foreground galaxies. Substructures, prevalent in dynamically active clusters, bias mass estimates, leading to discrepancies between observational and theoretical predictions, challenging cosmological constraints derived from cluster abundances [136].

Role of Preprocessing and Environmental Effects

Understanding galaxy transformations necessitates distinguishing environmental effects within clusters from preprocessing in galaxy groups. Preprocessing significantly affects galaxy properties before cluster infall, complicating interpretations of cluster-specific influences. Quantifying the relative roles of preprocessing versus direct cluster environment interactions remains essential for comprehensive galaxy evolution models [101].

Non-Thermal Processes and Cosmic Rays

Non-thermal processes, particularly cosmic ray acceleration and magnetic fields in clusters, pose intriguing unresolved issues. Cosmic rays contribute significantly to cluster energetics, evident through radio halos and relics. Understanding their origin, acceleration mechanisms, and influence on cluster dynamics remains challenging, demanding integrated observational and theoretical investigations [19].

Future Observations and Their Purposes

Upcoming observational facilities such as the Square Kilometer Array (SKA), Athena X-ray observatory, and Euclid satellite promise revolutionary advancements in understanding cluster dynamics. These instruments will provide unprecedented resolution and sensitivity, clarifying unresolved cluster mass measurement issues, environmental effects on galaxy evolution, and cosmic ray processes, ultimately refining cosmological models [110].

This comprehensive exploration highlights the profound astrophysical and cosmological significance of galaxy clusters, necessitating ongoing investigation to unravel their complexities and contributions to cosmic evolution.

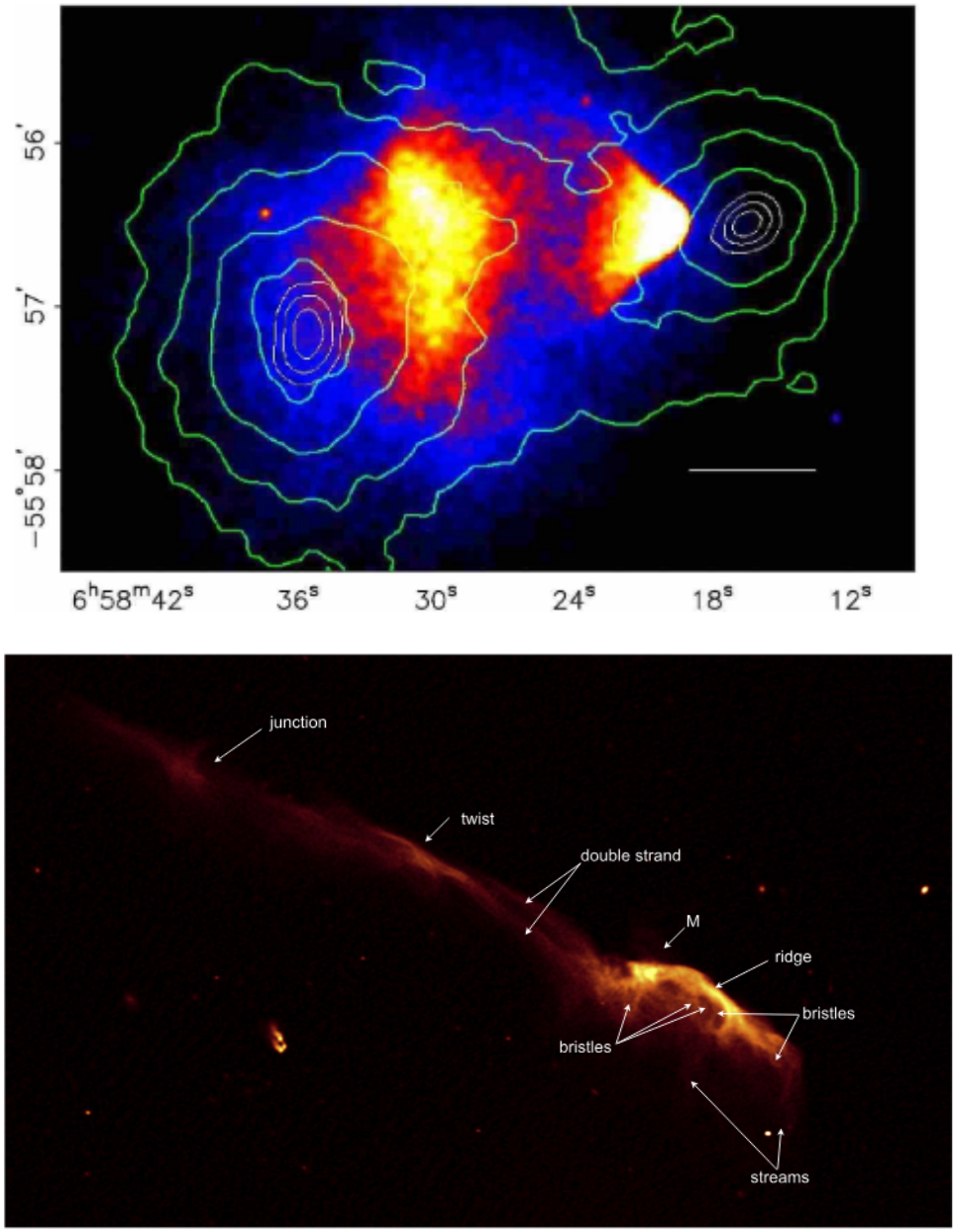

Figure 2: **Top:** *Bullet Cluster* (1E 0657−56): *Chandra* X-ray surface brightness with a bow shock (right) and a cool "bullet" core, with weak-lensing mass contours offset from the X-ray gas, illustrating the collisionless nature of dark matter. Image is from Clowe et al. [35]. **Bottom:** The "*Toothbrush*" relic (1RXS J0603.3+4214): deep GMRT/LOFAR/VLA radio imaging of the linear Mpc-scale relic; the bright ridge and fainter bristles/streams trace a merger shock and ordered magnetic fields. Image taken from Rajpurohit et al. [135].

4

GALAXY CLUSTERS IN SIMULATIONS

4.1 INTRODUCTION TO COSMOLOGICAL SIMULATIONS

The study of galaxy clusters requires bridging an enormous dynamic range: from the large-scale cosmic web that channels matter into cluster nodes, down to the internal processes of star formation, black-hole growth, and plasma physics that govern the observable properties of galaxies and the ICM. Observations provide snapshots of this complexity, but they cannot, by themselves, reveal the full three-dimensional and temporal evolution of clusters. Cosmological simulations have therefore become indispensable tools. They act as numerical laboratories, enabling us to follow the coupled evolution of dark matter, baryons, stars, and black holes under well-defined physical prescriptions, and to test theoretical models against observational data.

At their core, cosmological simulations solve the coupled system of gravity, hydrodynamics, and microphysical models for processes such as cooling, star formation, and feedback, within a cosmologically representative volume. Early simulations in the 1980s and 1990s focused primarily on collisionless N-body dynamics, successfully reproducing the hierarchical growth of dark-matter halos and the large-scale filamentary pattern of the Universe [38]. However, they could not address the baryonic physics that determine cluster observables such as X-ray luminosities, temperature profiles, or radio synchrotron structures. With the advent of hydrodynamical methods (e.g. smoothed-particle hydrodynamics, adaptive mesh refinement, moving-mesh schemes), simulations began to incorporate the gaseous component, tracking the heating, cooling, enrichment, and dynamical state of the ICM.

Today, cosmological simulations have matured into a central pillar of cluster astrophysics. They are not only capable of reproducing global statistical measures, such as the halo mass function, galaxy luminosity function, or scaling relations between cluster mass and X-ray or SZ observables, but also of generating synthetic observations that can be compared directly with real data. This makes them critical for both interpreting current surveys and planning future ones (e.g. *eROSITA*, *Euclid*, the Vera Rubin Observatory). In parallel, "zoom-in" techniques allow us to study individual clusters with high resolution, capturing the interplay of mergers, AGN feedback, turbulence, and non-thermal components in exquisite detail.

In this chapter, we will review the different classes of cosmological simulations relevant to cluster science. We begin by discussing dark-matter-only (DMO) simulations and their role in establishing the backbone of structure formation. We then move to full hydrodynamical simulations, where baryonic physics and feedback processes are modeled explicitly, before considering zoom-in simulations that target individual clusters at high resolution. Finally, we highlight the current state of the art, focusing on the IllustrisTNG and TNG-Cluster projects, which combine large cosmological volumes with sophisticated baryonic models and resolution sufficient to study cluster physics in depth.

4.2 TYPES OF GALAXY CLUSTER SIMULATIONS

Dark Matter Only (DMO) Simulations

The earliest generation of cosmological simulations focused exclusively on the collisionless dynamics of dark matter, motivated by the ΛCDM framework in which dark matter dominates the matter density of the Universe. In these simulations, the cosmic mass distribution is represented by a large number of particles interacting solely via gravity, evolved forward in time using N-body techniques. This approach is computationally efficient and isolates the role of dark matter in structure formation, without the additional complexities of baryonic physics.

Historically, N-body experiments in the 1980s and 1990s demonstrated the viability of hierarchical growth under cold dark matter. Davis et al. [38] performed one of the first large-scale simulations to show that the CDM paradigm could reproduce the observed clustering of galaxies. Later, increasingly sophisticated calculations, such as the Millennium Simulation [159], provided high-resolution predictions of the cosmic web, halo mass functions, and subhalo abundance in cosmological volumes. More recent efforts, such as the Bolshoi [80] and Bolshoi-Planck [82] simulations, refined these predictions with updated cosmological parameters, enabling direct comparisons to galaxy surveys such as SDSS and DES. These studies have firmly established the DMO framework as the backbone of modern structure formation theory.

DMO simulations yield robust predictions for a variety of key observables:

- *Halo mass function:* The abundance of dark matter halos as a function of mass and redshift can be calibrated with high precision, providing a cornerstone for cosmological tests using galaxy and cluster counts [163].
- *Clustering statistics:* Two-point correlation functions and power spectra of halos in DMO simulations match the observed large-scale distribution of galaxies when coupled with halo occupation models [173].
- *Substructure:* DMO simulations predict a wealth of subhalos within massive halos, providing the theoretical framework for galaxy–halo connection studies and dark matter annihilation searches [81].

Despite these successes, DMO simulations have inherent limitations when applied to galaxy clusters. Without baryonic processes, they cannot predict directly observable properties such as X-ray luminosities, temperature profiles, or radio emission from the intracluster medium. Moreover, the absence of baryonic physics leads to discrepancies in the internal structure of halos: for example, pure dark matter simulations predict overly cuspy inner density profiles compared to observations, a tension known as the cusp–core problem [107]. Similarly, the overabundance of predicted subhalos compared to satellite galaxies in the Local Group, the missing satellites problem, arises in the DMO framework and highlights the necessity of baryonic processes to suppress or transform low-mass halos

Hydrodynamical Simulations with Baryonic Physics

While dark-matter-only simulations successfully describe the hierarchical growth of halos, they neglect the baryonic component that gives rise to observable galax-

ies, stars, and the ICM. To capture these phenomena, cosmological simulations must incorporate gas dynamics, radiative cooling, star formation, chemical enrichment, and feedback processes. These hydrodynamical simulations thus provide a bridge between the underlying dark matter framework and the multi-wavelength observations of galaxy clusters.

NUMERICAL METHODS.

- *Smoothed Particle Hydrodynamics (SPH):* a Lagrangian, particle-based method in which gas is represented by discrete particles with kernel-smoothed properties [55, 97]. Classic formulations can suppress fluid instabilities and mixing, although modern variants (e.g. pressure–entropy SPH) mitigate these issues.

- *Adaptive Mesh Refinement (AMR):* an Eulerian approach that solves the hydrodynamics on a grid with local, on-the-fly refinement in regions of high density or strong gradients; widely used in cluster and cosmic-web studies by codes such as ENZO [20] and RAMSES [162].

- *Moving-mesh schemes:* hybrid finite-volume approaches, exemplified by AREPO [155], which combine Lagrangian adaptivity with grid-based accuracy and robust shock capturing. This method underpins large cosmological volumes such as *Illustris* and *IllustrisTNG* [114, 127, 167].

BARYONIC PHYSICS. Hydrodynamical simulations must include a variety of sub-grid models to represent processes occurring below the resolution limit:

- *Radiative cooling and heating:* primordial and metal-line cooling, plus photoionization from a UV background, regulate the thermal state of the gas [77].

- *Star formation and stellar feedback:* cold, dense gas forms stars following empirical laws (e.g. Kennicutt-Schmidt). Supernovae inject thermal and kinetic energy, driving galactic winds that redistribute metals and regulate star formation [157].

- *Chemical enrichment:* stellar evolution models track the production and distribution of heavy elements, enriching the ICM and enabling comparison with X-ray abundance measurements [172].

- *Active galactic nuclei (AGN) feedback:* accretion onto supermassive black holes injects large amounts of energy into the surrounding medium, preventing runaway cooling flows in cluster cores and shaping galaxy evolution. Thermal and kinetic AGN feedback modes are implemented in state-of-the-art simulations [151].

- *Additional physics:* modern simulations increasingly incorporate magnetic fields [119], cosmic rays [126], and anisotropic transport processes (e.g. conduction, viscosity), which influence the structure and thermodynamics of the ICM.

KEY PROJECTS AND SUCCESSES. Several landmark hydrodynamical simulation campaigns have transformed our understanding of cluster physics:

- The *Illustris* project [167] was among the first to model a cosmological volume with detailed baryonic physics, successfully reproducing galaxy stellar mass functions, morphologies, and enrichment patterns, but overproducing stellar masses in massive galaxies.
- The *EAGLE* simulations [36, 144] calibrated feedback models to match the observed galaxy stellar mass function and sizes, achieving excellent agreement across a range of galaxy properties.
- *Horizon-AGN* [43] provided an alternative AGN feedback implementation and explored black-hole driven galaxy quenching and morphological transformations.

In the cluster regime, these simulations have reproduced many observed properties, including X-ray scaling relations, gas fractions, metallicity distributions, and the suppression of cooling flows, establishing feedback as a fundamental ingredient of structure formation.

LIMITATIONS AND CHALLENGES. Despite their successes, hydrodynamical simulations remain subject to uncertainties. Sub-grid feedback prescriptions are calibrated to reproduce specific observables, introducing model dependence and degeneracies. Different codes implementing similar physical processes often yield divergent predictions, as demonstrated by the "nIFTy" cluster comparison project [146]. Resolution limitations remain severe, particularly in large volumes, where the internal structure of galaxies and the multiphase nature of the ICM cannot be fully resolved. Moreover, processes such as plasma instabilities, cosmic-ray acceleration, and anisotropic conduction are only beginning to be incorporated.

In summary, hydrodynamical simulations provide the crucial link between dark matter structure formation and the observable Universe. By modeling baryonic physics, they reproduce the thermodynamics and scaling relations of galaxy clusters, while also highlighting the uncertainties and open questions that motivate the next generation of simulation efforts.

Zoom-in Simulations

While large-volume hydrodynamical simulations capture statistically representative samples of clusters, their finite resolution limits the fidelity with which internal structures and small-scale processes can be studied. To overcome this, zoom-in techniques selectively increase resolution within a chosen region of interest; typically around a single massive halo or a set of progenitors, while retaining the correct cosmological environment on large scales. This approach dynamically allocates computational resources, allowing the same simulation to capture both the megaparsec-scale accretion flows and the kiloparsec-scale details of intracluster physics.

NUMERICAL TECHNIQUE. Zoom-in simulations begin with a large-volume, low-resolution cosmological run in which halos of interest are identified. The initial conditions are then re-generated, embedding the target halo within a region of much higher particle or cell resolution, while surrounding large-scale structures

are modeled with lower resolution. This ensures that the growth and environment of the cluster are modeled consistently with the cosmic web, but with sufficient resolution to study its internal dynamics in detail [78, 112].

APPLICATIONS TO CLUSTERS. Zoom-in simulations have been especially powerful for galaxy cluster studies. They enable:

- *Merger dynamics:* By resolving substructure and shock fronts at high spatial resolution, zoom-ins allow detailed modeling of merger-induced phenomena such as cold fronts, turbulence, and radio relic shocks.
- *Gas physics:* The fine resolution permits more accurate treatment of the intracluster medium (ICM), including entropy profiles, cooling flows, and the interaction of AGN jets with the surrounding gas.
- *Galaxy populations:* Higher resolution also allows galaxies within clusters to be resolved, making it possible to study star formation quenching, morphological transformation, and satellite stripping in dense environments.

KEY PROJECTS. Several landmark zoom-in campaigns have targeted massive halos to study clusters in detail:

- The *Phoenix Project* [53] extended the Aquarius approach to the cluster regime, producing ultra-high-resolution simulations of massive halos and their substructures, enabling direct comparison with strong lensing and subhalo statistics.
- The *MUSIC* simulations (Marenostrum-MultiDark SImulation of galaxy Clusters; [145]) combined zoom-in techniques with baryonic physics to investigate gas fractions, scaling relations, and the impact of feedback in large cluster samples.
- The *Cluster-EAGLE* simulations [12] employed zoom-ins of clusters within the EAGLE framework, resolving both cluster-scale gas profiles and galaxy populations.
- The *TNG-Cluster* simulations [113] represent a next-generation zoom-in campaign within the IllustrisTNG framework, targeting more than 300 galaxy clusters with masses $M_{200c} \geqslant 10^{14}\,M_{\odot}$. With resolution comparable to the TNG300-1 simulation but applied to the cluster regime, TNG-Cluster captures both large-scale environmental effects and the fine-grained baryonic physics of the ICM, including AGN feedback, metal enrichment, and magnetic fields. This enables statistically robust predictions for observable properties such as X-ray scaling relations, SZ signatures, and radio synchrotron emission, while preserving the detailed physics usually only available in small zoom-in samples.

STRENGTHS AND LIMITATIONS. The chief strength of zoom-in simulations lies in their ability to combine cosmological context with exquisite internal resolution. They allow detailed, case-by-case analyses of complex processes such as mergers, AGN feedback, or turbulence, which cannot be resolved in large uniform-volume

runs. However, their computational expense restricts the number of clusters that can be studied in this way, limiting their statistical power. As such, zoom-in studies are best viewed as complementary to large-volume hydrodynamical simulations: the former provides physical insight into cluster microphysics, while the latter supplies the statistical framework for cosmological applications.

In summary, zoom-in simulations act as high-resolution laboratories for studying the internal structure and evolution of individual galaxy clusters. They are especially valuable for understanding the signatures of mergers and feedback processes, which strongly shape the thermodynamic and non-thermal properties of the ICM.

4.3 COMPARATIVE ANALYSIS OF SIMULATION TYPES

The three classes of cosmological simulations: dark-matter-only (DMO), full hydrodynamical volumes, and zoom-in cluster simulations, offer complementary insights into structure formation. Each comes with characteristic strengths and limitations in terms of resolution, scale, and the range of physical processes included. Here we provide a comparative overview, focusing on aspects most relevant to galaxy cluster studies.

Resolution and Scale

DMO simulations are computationally the most efficient, allowing the largest cosmological volumes and the highest statistical precision for halo mass functions and large-scale clustering. For example, the Millennium Simulation evolved more than 10^{10} particles in a $(500\,h^{-1}\,\mathrm{Mpc})^3$ box [159], enabling precise predictions of halo abundances and clustering. However, since baryons are neglected, these runs cannot resolve internal cluster structure or predict observable gas and galaxy properties.

Hydrodynamical simulations add baryons but at the cost of computational expense. To maintain large volumes, such as the $(300\,\mathrm{Mpc})^3$ box of TNG300 [114], the spatial and mass resolution must be coarser than in DMO runs. This resolution is adequate for statistical studies of cluster populations and scaling relations, but insufficient to fully resolve multiphase gas, AGN jets, or galaxy morphologies.

Zoom-in simulations invert this trade-off: by focusing on a single halo (or a limited set), they achieve spatial and mass resolutions an order of magnitude higher than uniform-volume runs. For example, the Phoenix Project resolved subhalos down to dwarf-galaxy scales within clusters [53], while the TNG-Cluster project [113] reaches baryonic resolution comparable to TNG300-1 but for hundreds of clusters. The drawback is limited sample size and cosmic variance, since only a small fraction of the cosmic volume is resimulated at high resolution.

Physical Processes Included

DMO simulations track only gravitational dynamics, and thus their predictions are limited to quantities such as halo mass functions, merger trees, and dark matter

density profiles. They form the gravitational backbone of structure formation but cannot directly connect to observables such as X-ray luminosities or SZ signals.

Hydrodynamical simulations incorporate baryonic physics via sub-grid models for radiative cooling, star formation, stellar and AGN feedback, and chemical enrichment [43, 144, 167]. These processes enable them to reproduce cluster gas fractions, metallicities, and scaling relations between mass, temperature, and luminosity. However, differences in feedback implementations lead to systematic variations across codes, as highlighted by the nIFTy cluster comparison project [146]. Moreover, additional physics; magnetic fields, cosmic rays, plasma transport, are only beginning to be included in a systematic way [119, 126].

Zoom-in simulations push the frontier by applying high resolution to these baryonic processes within cluster environments. They allow detailed studies of turbulence, shocks, cold fronts, AGN jet–ICM coupling, and galaxy transformations within dense environments. TNG-Cluster, for example, can simultaneously resolve galaxy populations and the thermodynamic state of the ICM across hundreds of massive clusters, while retaining cosmological context [113]. Nevertheless, zoom-ins still rely on sub-grid prescriptions and cannot capture cluster-to-cluster statistical variations at the level of full-volume simulations.

In summary, DMO, hydrodynamical, and zoom-in simulations are best viewed as complementary approaches: DMO establishes the dark matter scaffolding; hydrodynamical simulations add baryonic realism over large statistical samples; and zoom-ins provide high-resolution laboratories for detailed cluster physics. Together, they form a hierarchy of tools for connecting theoretical models of structure formation to the rich observational data available for galaxy clusters.

4.4 THE NEXT GENERATION: ILLUSTRISTNG AND TNG-CLUSTER SIMULATIONS

The past decade has seen remarkable progress in cosmological simulations, moving from dark-matter-only frameworks and early hydrodynamical runs toward comprehensive models that simultaneously capture large-scale structure, galaxy formation, and intracluster medium physics. Among these, the IllustrisTNG project and its cluster-focused extension, TNG-Cluster, represent the state of the art. These simulations combine cosmological volumes with sophisticated baryonic physics implementations, providing unprecedented opportunities to confront theory with observations of galaxy clusters across the electromagnetic spectrum.

Overview of IllustrisTNG Simulations

The IllustrisTNG (The Next Generation) project is the successor to the original Illustris simulation [167]. It consists of three large cosmological volumes; TNG50, TNG100, and TNG300, evolved with the moving-mesh code Arepo [155], each balancing resolution against volume [114, 127]. Compared to its predecessor, TNG includes several major advances in baryonic modeling:

- *AGN feedback:* a dual-mode model that injects thermal energy at high accretion rates and kinetic winds at low accretion rates, stabilizing cluster cores and preventing catastrophic cooling flows [169].

- *Stellar feedback:* improved prescriptions for galactic winds driven by supernovae and stellar feedback, regulating star formation in galaxies across a wide mass range [127].
- *Magnetic fields:* full magnetohydrodynamics (MHD) via the ideal MHD equations, allowing self-consistent amplification of primordial seed fields and their impact on galaxy and ICM evolution [119].
- *Chemical enrichment:* tracking of nine metal species from Type Ia/II supernovae and AGB stars, enabling direct comparison with observed metallicity patterns in the ICM and galaxies.

These improvements allow IllustrisTNG to reproduce a broad range of observables: galaxy stellar mass functions, morphologies, colors, star-formation histories, gas fractions, and cluster scaling relations. The three volumes are complementary: TNG50 achieves exquisite resolution within a 50 Mpc box, TNG100 balances resolution and statistics, and TNG300 covers a $(300\,\mathrm{Mpc})^3$ volume, providing hundreds of massive clusters for statistical analyses. In this sense, IllustrisTNG offers both a cosmologically representative framework and the physical fidelity required for cluster astrophysics.

Advantages of TNG-Cluster Simulations

While TNG300 contains hundreds of massive clusters, their resolution is limited compared to the smaller-volume TNG50 and TNG100 runs. To overcome this, the *TNG-Cluster* project [113] applies a zoom-in approach to 352 galaxy clusters with $M_{200c} \geqslant 10^{14}\,M_\odot$, achieving baryonic resolution comparable to TNG300-1. This enables detailed modeling of the intracluster medium, merger-driven shocks, and feedback-regulated cores, while maintaining a statistically significant sample size.

Key advantages of TNG-Cluster include:

- *Resolution:* sufficient to resolve both cluster galaxies and ICM substructure, including shocks, turbulence, and magnetic-field amplification.
- *Physics:* incorporates the full IllustrisTNG baryonic model (stellar and AGN feedback, MHD, metal enrichment) tuned for the cluster regime.
- *Statistics:* unlike previous zoom-in studies of individual clusters, TNG-Cluster provides hundreds of high-resolution clusters, enabling population-level comparisons to X-ray, SZ, and radio surveys.

Together, IllustrisTNG and TNG-Cluster represent the cutting edge of cosmological simulation capabilities, offering both the breadth of cosmological volumes and the depth of targeted zoom-in studies. They serve as essential theoretical laboratories for interpreting observations of clusters in the current era of large surveys. In this thesis, we will return to these simulations in detail in Part iii, where both IllustrisTNG and TNG-Cluster will be discussed comprehensively.

4.5 CONCLUDING REMARKS

The study of galaxy clusters through cosmological simulations has progressed enormously over the past four decades. The earliest dark-matter-only (DMO) N-

body simulations demonstrated the viability of the hierarchical ΛCDM framework and provided precise predictions for halo abundances and clustering. However, by neglecting baryons they could not connect directly to observed cluster properties such as X-ray luminosities, gas fractions, or radio emission.

The next stage, large-volume hydrodynamical simulations, incorporated baryonic processes; e.g. cooling, star formation, stellar and AGN feedback, and chemical enrichment, into cosmological contexts. Projects such as Illustris, EAGLE, and Horizon-AGN significantly advanced our understanding of the intracluster medium and galaxy populations, reproducing many observed scaling relations and enrichment patterns. At the same time, systematic uncertainties in sub-grid physics highlighted the need for careful model calibration and inter-code comparisons.

Zoom-in simulations provided a complementary avenue, achieving far higher spatial and mass resolution within selected cluster environments. These runs allowed detailed study of merger-driven shocks, turbulence, cold fronts, and feedback-regulated cores. Yet, their limited sample sizes restricted statistical applications. As a result, they have been most effective as physical laboratories, deepening our understanding of the microphysics of the ICM.

The most recent generation, exemplified by IllustrisTNG and its cluster-focused extension TNG-Cluster, combines the breadth of large cosmological volumes with the depth of high-resolution zoom-ins. These simulations implement state-of-the-art baryonic physics, including AGN feedback, galactic winds, chemical enrichment, and magnetohydrodynamics, and provide both statistically representative cluster samples and detailed internal structure. TNG-Cluster, in particular, delivers unprecedented resolution across hundreds of clusters, enabling direct comparison with X-ray, SZ, and radio surveys.

In summary, cosmological simulations now constitute indispensable laboratories for studying galaxy clusters. They reveal how the gravitational backbone provided by dark matter interacts with baryonic physics to shape the thermodynamic, chemical, and non-thermal state of the ICM. This dual perspective, statistical and physical, makes simulations essential tools for interpreting observations of cluster populations and their mergers. In the following parts of this thesis, we will leverage these simulations, in particular IllustrisTNG and TNG-Cluster, to investigate the merger histories of clusters and their observable signatures in the X-ray and radio regimes.

5 MACHINE LEARNING

5.1 DEEP LEARNING INTRODUCTION

Deep learning has emerged as one of the most transformative paradigms in modern machine learning, enabling computers to autonomously discover and extract complex, hierarchical features from raw data. It is a subfield of machine learning that leverages *artificial neural networks* (ANNs) composed of multiple processing layers, loosely inspired by the interconnected structure of biological neurons in the human brain [56, 90]. Unlike traditional machine learning methods, which rely heavily on hand-crafted features, deep learning models learn representations directly from data by optimizing millions (and sometimes billions) of parameters across deep architectures. This capacity to model highly non-linear and abstract relationships has led to breakthroughs across disciplines ranging from computer vision and natural language processing to medical imaging, astrophysics, and cosmology.

BASIC STRUCTURE OF NEURAL NETWORKS. At the core of deep learning lies the artificial neuron, which computes a weighted sum of its inputs and passes it through a non-linear activation function such as the rectified linear unit (ReLU), sigmoid, or hyperbolic tangent. Neurons are organized into layers: an *input layer*, multiple *hidden layers*, and an *output layer*. Stacking many hidden layers allows the network to learn increasingly abstract and complex features, a property often referred to as *representation learning*. The universal approximation theorem guarantees that even shallow neural networks can approximate any continuous function under certain conditions, but in practice deep networks achieve far more efficient and scalable representations for high-dimensional data [66].

ARCHITECTURAL INNOVATIONS. Several specialized architectures have been developed to address different data modalities:

- *Convolutional Neural Networks (CNNs)* are designed for grid-like data (e.g., images, maps, or 3D cubes). Convolutional layers apply learnable filters that exploit spatial locality and parameter sharing, enabling efficient feature extraction at multiple scales. CNNs have revolutionized computer vision, with landmark architectures such as AlexNet [86], VGGNet [153], and ResNet [61] achieving human-level or better performance on image classification benchmarks. In astrophysics, CNNs have been used to classify galaxy morphologies [39], detect gravitational lenses [125], and identify transient events in time-domain surveys [147].
- *Recurrent Neural Networks (RNNs)* and their gated variants such as Long Short-Term Memory (LSTM) units [63] and Gated Recurrent Units (GRUs; [33]) are tailored for sequential and temporal data. By maintaining hidden states that evolve with input sequences, RNNs capture temporal dependencies crucial

for tasks such as language modeling, speech recognition, and time-series prediction. In astrophysics, RNNs have been used for analyzing variable stars, and exoplanet transit detection.

OPTIMIZATION AND TRAINING. Training deep networks requires minimizing a loss function that quantifies the discrepancy between predictions and ground truth. This is achieved through gradient-based optimization, typically using stochastic gradient descent (SGD) and its adaptive extensions such as Adam [79], RMSProp [103], and Adagrad [168]. The backpropagation algorithm efficiently computes gradients of the loss with respect to millions of network parameters, enabling scalable learning. To prevent overfitting and improve generalization, techniques such as dropout, batch normalization, weight decay, and data augmentation are commonly employed.

IMPACT ACROSS DISCIPLINES. The impact of deep learning extends far beyond traditional computer science. In natural sciences, it is rapidly becoming indispensable: in astronomy, deep learning has been applied to galaxy classification, gravitational wave detection, cosmological parameter inference, and simulation-based surrogate modeling [9, 116]. Its ability to process high-dimensional, noisy, and incomplete data makes it especially well-suited for astrophysical applications, where observational data are often sparse or uncertain. Moreover, deep learning is increasingly coupled with physical models and simulations, forming the foundation of hybrid approaches that combine data-driven learning with theoretical priors.

In summary, deep learning provides a flexible and powerful framework for extracting complex patterns from data. Its architectures ranging from convolutional to recurrent models, enable the analysis of images, sequences, and other structured inputs at unprecedented levels of accuracy. In astrophysics, these methods are now central to tackling some of the most challenging problems, from understanding galaxy morphologies to constraining cosmological models. In the following sections, we will introduce extensions of deep learning that are particularly relevant to this thesis, including self-supervised learning methods and conditional invertible neural networks.

5.2 SELF-SUPERVISED LEARNING

One of the main limitations of supervised deep learning is its reliance on large labeled datasets. In many scientific domains, including astrophysics, labels are expensive, uncertain, or altogether absent: for instance, galaxy morphologies may require expert visual classification, and merger histories of galaxy clusters are not directly observable. This motivates *self-supervised learning* (SSL), a paradigm that leverages the intrinsic structure of data itself to provide supervisory signals without human annotation [70, 89].

GENERAL PRINCIPLES. Self-supervised learning constructs *pretext tasks*; auxiliary prediction tasks whose solutions require the network to learn meaningful representations of the data. Examples in computer vision include predicting the relative positions of image patches, colorizing grayscale images, or solving jigsaw

puzzles [40, 115, 176]. The key idea is that, by learning to solve these surrogate tasks, the network develops generalizable feature representations that can then be transferred to downstream tasks such as classification, regression, or clustering, even when only limited labeled data are available.

CONTRASTIVE LEARNING. Among the various SSL approaches, *contrastive learning* has emerged as the most powerful and widely adopted framework in computer vision and beyond [29, 57, 60]. Contrastive methods are built on the idea of learning representations that maximize agreement between different views of the same data point (positive pairs) while minimizing agreement with representations of other data points (negative pairs). Formally, given an encoder network f_θ that maps an input x to a representation $z = f_\theta(x)$, the objective is to learn a representation space where

$$\mathrm{sim}(z_i, z_j) \gg \mathrm{sim}(z_i, z_k),$$

for positive pairs (i, j) of the same underlying sample under different augmentations, and negatives (i, k) drawn from other samples. The similarity function is often the normalized dot product (cosine similarity).

CORE ELEMENTS OF CONTRASTIVE LEARNING. Several components determine the success of contrastive learning:

- *Data augmentations:* Strong, stochastic augmentations (e.g., random cropping, flipping, color distortion, blurring) create different "views" of the same image, forcing the encoder to learn invariant features. The choice and diversity of augmentations are critical for performance [29].
- *Contrastive loss:* The most common formulation is the InfoNCE loss, which normalizes similarity scores across all positives and negatives within a batch. This encourages clustering of positive pairs while repelling negatives in the representation space.
- *Negative sampling and memory banks:* Methods such as MoCo (Momentum Contrast) [60] address the need for large and diverse negatives by maintaining a dynamic memory bank or queue of past representations, allowing contrastive learning with manageable batch sizes.
- *Non-contrastive extensions:* More recent methods such as BYOL (Bootstrap Your Own Latent) [57] and SimSiam [30] achieve strong performance without explicit negatives, instead relying on asymmetric architectures and stop-gradient operations to avoid representational collapse.

SCIENTIFIC RELEVANCE. Contrastive learning is particularly attractive for astrophysics and cosmology, where unlabeled data are abundant but labeled sets are scarce or subjective. By leveraging augmentations appropriate to scientific data (e.g., rotations, noise injection, smoothing), one can train encoders that learn representations of galaxy images, cluster X-ray/radio maps, or cosmological simulations without requiring explicit labels. These representations can then be used for downstream tasks such as cluster classification, merger state identification, or

simulation-based inference of physical parameters. Recent studies have demonstrated that contrastive learning can rival or even outperform supervised training in low-label regimes, while also producing embeddings that transfer robustly across datasets and observational conditions [1, 28, 49].

In summary, self-supervised learning, and contrastive learning in particular, provides a principled and scalable way to extract scientific knowledge from large unlabeled datasets. Its ability to learn invariant, semantically meaningful features makes it a natural fit for applications in astrophysics, where the cost of labeling is high but the potential for representation learning is enormous. In this thesis, we will build upon these methods to analyze cluster simulations and develop models capable of connecting raw observables to the physical histories of galaxy clusters.

5.3 CONDITIONAL INVERTIBLE NEURAL NETWORKS

In Bayesian statistics, the goal of inference is to obtain the posterior distribution of latent variables x (here unobservable cluster properties such as merger mass ratios, timescales, or pericenter distances) given observed data c, such as X-ray or radio maps. Bayes' theorem provides the formal relation:

$$p(x|c) = \frac{p(c|x)\,p(x)}{p(c)}, \tag{5}$$

where $p(x)$ denotes the prior distribution over latent variables, $p(c|x)$ is the likelihood of observing c given x, and $p(c)$ is the evidence or marginal distribution of observables [13].

In principle, one could compute $p(c|x)$ for a given physical model and thus directly evaluate $p(x|c)$. In practice, however, for complex astrophysical systems such as galaxy clusters, the likelihood is intractable. The mapping from merger histories to observables is highly non-linear, and depends sensitively on projection effects [84]. For this reason, we must turn to the framework of *likelihood-free inference*, where cosmological simulations provide paired samples (x, c) of galaxy cluster merger properties properties and their observables properties or radio/X-ray maps. A machine learning model is then trained to approximate the posterior $p(x|c)$ directly from these samples. It is important to emphasize that such learned posteriors are always constructed relative to the prior and marginal distribution encoded in the training data [6].

Traditional regression models such as multilayer perceptrons (MLPs), trained with a mean-squared error (MSE) loss, can only approximate point estimates corresponding to the posterior mean [56]. This approach implicitly assumes a bijective mapping from observables to latent variables, or at most Gaussian-like uncertainties around them. However, in reality, clusters with nearly identical X-ray morphologies may have experienced very different merger histories, while projection effects can make the same system appear very different depending on orientation. Thus, the full posterior distribution $p(x|c)$; including multimodality and correlations between parameters, is required for a meaningful characterization of cluster assembly histories [6].

NORMALIZING FLOWS. Normalizing flows (NFs) provide a flexible way to model complex distributions while retaining exact likelihood evaluation [69]. The idea is

to represent a complicated target distribution by applying a sequence of invertible and differentiable transformations to a simple base distribution. Let $z \sim p_Z(z)$ denote a latent variable with tractable density (usually Gaussian), and let $x = f_\theta(z)$ be the output of an invertible neural network f_θ parameterized by θ. By the change-of-variables formula, the density of x is

$$p_X(x) = p_Z(f_\theta^{-1}(x)) \left| \det \frac{\partial f_\theta^{-1}(x)}{\partial x} \right|. \tag{6}$$

Because f_θ is bijective and differentiable by construction, both density evaluation and sampling are efficient: one can map from $z \to x$ (generation) and from $x \to z$ (inference) with exact likelihoods.

CONDITIONAL FLOWS AND CINNS. While standard flows model the unconditional distribution $p(x)$, many scientific problems require conditional distributions $p(x|c)$, where c denotes conditioning variables (e.g., observables or measurements). From a Bayesian perspective, the posterior is written as Equaiton 5; direct evaluation is generally impossible for high-dimensional astrophysical problems, but conditional flows can learn this distribution directly.

Conditional Invertible Neural Networks (cINNs) [6] extend normalizing flows by explicitly conditioning the transformations on c. They define an invertible mapping

$$f : (x, c) \mapsto z, \qquad z \sim \mathcal{N}(0, I),$$

where x are latent cluster properties and c are the observables (e.g., X-ray or radio images). Because f is invertible, posterior samples can be drawn as

$$x = f^{-1}(z, c), \qquad z \sim \mathcal{N}(0, I).$$

Thus, the complex posterior $p(x|c)$ is mapped to a tractable Gaussian prior in latent space, and the inverse transformation recovers x samples conditioned on new observables c.

TRAINING OBJECTIVE. CINNs are trained by minimizing the negative log-likelihood (NLL) of the latent variables $z = f(x, c)$ under the Gaussian prior:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{2}\|z\|^2 - \log|\det J|, \tag{7}$$

where $J = \partial f(x, c)/\partial x$ is the Jacobian of the transformation. This objective ensures that z follows $\mathcal{N}(0, I)$ and that the network correctly models the posterior $p(x|c)$ [44].

IMPLEMENTATION AND APPLICATIONS. For practical implementations we employ the `FrEIA` framework [4], a PyTorch-based library for invertible neural networks. CINNs and related conditional flows have already been applied in astrophysics and cosmology, for example to reconstruct initial conditions of large-scale structure [106], to perform cosmological parameter inference from weak lensing [68], to infer the assembly history of simulated galaxies in IllustrisTNG [50], and to model galaxy–halo connections [83]. Their strength lies in combining Bayesian

rigor (explicit posteriors and likelihoods) with the flexibility of deep learning, enabling principled uncertainty quantification in regimes where traditional regression is inadequate.

In this thesis, we employ CINNs to model the posterior distributions of galaxy cluster merger parameters x conditioned on multi-wavelength observables (radio or X-ray maps) or scalar observable propertiesc. This approach enables not only point predictions of merger properties, but also a full characterization of their uncertainties and degeneracies; providing a more complete and scientifically robust description of cluster assembly histories.

Part II

RATIONAL OF THE THESIS

6 RATIONALE

6.1 SCIENTIFIC MOTIVATION

Why infer cluster merger histories from images?

Galaxy clusters assemble hierarchically through mergers. These events inject gravitational energy into the ICM, driving shocks, turbulence, and bulk flows that reshape both the thermal X–ray emitting plasma and the non–thermal, synchrotron–bright relativistic component [84, 98]. The timing, geometry, and energetics of a cluster's most recent mergers imprint themselves on its projected morphology: peaked or disturbed X–ray surface brightness, ellipticity and substructure in the core and outskirts, and the presence and layout of radio relics/halos tied to shock acceleration and ageing [19, 98, 179]. Recovering *merger parameters*, (e.g. time since (or to) pericenter, collision velocity, mass ratio, and pericenter distance) from these images therefore offers a direct window into the dynamical state and recent assembly of the most massive bound structures in the Universe.

This inference is scientifically valuable for at least four reasons:

1. **Dynamical state as a hidden variable.** Merger phase and geometry strongly influence widely used mass–observable relations (e.g. scatter and bias in X–rays). Identifying where a system sits in its merger timeline helps disentangle astrophysical variance from cosmological signal [132].

2. **Baryonic physics and feedback.** Shocks and turbulence influence cooling, metal transport, magnetic field amplification, and cosmic–ray acceleration. Linking observed structures to merger parameters sharpens tests of ICM microphysics and AGN feedback models [19, 98].

3. **Multi–wavelength synergy.** Thermal X–ray and non–thermal radio morphologies respond differently to the same event (e.g. cores versus shock–traced relics). Joint inference exploits their complementarities to constrain the same underlying dynamics [19, 179].

4. **Survey–scale forecasting and follow–up.** Fast, image–based posteriors on merger stage and geometry can prioritize systems for deeper, expensive observations (e.g. high–resolution spectroscopy, weak lensing) at phases of maximal diagnostic power [132, 150].

From a practical standpoint, imaging is the most abundant, homogeneous, and cost–effective mode of observation across surveys. Deep spectroscopy or tailored hydrodynamical modeling can constrain individual systems, but do not scale to thousands of clusters. In contrast, modern X-ray and radio surveys deliver large archival image sets with uniform processing, enabling statistical studies, if we can translate 2D morphology into 3D merger physics with quantified uncertainty [132].

This thesis takes that route. Namely, we want to see whether it is possible to learn the merger parameters of galaxy clusters by only using maps (e.g. X-ray, radio

or joint). We use cosmological hydrodynamical simulations for constructing the intrinsic maps, and then design a self–supervised (contrastive learning) pipeline to learn morphology–aware embeddings from the X–ray and radio maps (separately and jointly). In the next step, we then condition a *conditional invertible neural network* (cINN) on the learned embeddings to infer posterior distributions over merger parameters. In doing so, we turn abundant imaging into probabilistic constraints on cluster assembly, at scale and with quantified uncertainty [6, 29].

Challenges: projection effects, sparsity of labels, degeneracies

Inferring a cluster's 3D merger history from 2D images is intrinsically ill–posed. Three intertwined difficulties dominate:

PROJECTION EFFECTS.

- **Random orientation and triaxiality.** The same merger, viewed at different angles, can produce dramatically different morphologies. In X–ray, line–of–sight integration ($\propto \int n_e^2 \, d\ell$) smears substructure and boosts dense cores; in radio, thin shock sheets brighten when viewed edge–on and can look as halo–like features when seen face–on [136, 179].
- **Superposition and confusion.** Multiple subhaloes, filaments, or foreground/background groups along the sight line can mimic or hide merger signatures (e.g. double–relic counterparts hidden by projection or limited FOV) [136, 179].
- **Instrumental realism.** PSF/beam convolution, mosaicking, depth variations, and redshift–dependent dimming (plus $B_{CMB} \propto (1+z)^2$ losses for radio) reshape the apparent morphology, further entangling physics with observing conditions [132, 179].

SPARSITY (AND NOISINESS) OF LABELS.

- **No ground truth in observations.** Key dynamical quantities (e.g., time since pericenter, collision velocity, pericenter distance, mass ratio), are not directly observable. Proxy labels (centroid shifts, concentration, relic curvature) are informative but incomplete and survey–dependent [27, 100].
- **Small, heterogeneous samples.** Well–imaged X–ray+radio cluster samples remain modest and selection–biased (e.g. relic detectability depends on depth and orientation), limiting supervised training and risking overfitting to survey idiosyncrasies [132, 179].
- **Sim-to-real domain shift.** Simulations furnish merger "truths" and abundant images, but subgrid physics, resolution, and emissivity modeling can mismatch reality. Bridging this gap requires representations that are robust to such shifts and a probabilistic mapping rather than hard classification [37, 120].

PHYSICAL AND GEOMETRIC DEGENERACIES.

- **Non–uniqueness of the inverse map.** Different parameter combinations can yield visually similar images: e.g. (higher mass ratio, larger pericenter) vs.

(lower mass ratio, smaller pericenter) at another viewing angle; recent pericenter with weak B vs. older shock with stronger B in radio; core sloshing vs. minor merger in X–ray [98, 179].

- **Coupled nuisance physics.** Magnetic-field strength/topology, electron acceleration efficiency, and ageing shape radio brightness independently of dynamics; cooling, AGN feedback, and multiphase structure modulate X–ray cores. These nuisance factors widen posteriors if unaccounted for [19].
- **Bounded/fractional targets.** Parameters like mass ratio are intrinsically bounded and highly skewed, making absolute errors appear small while relative errors blow up at the extremes; discrete snapshot timing likewise inflates percentage errors for small intervals [132].

IMPLICATIONS FOR METHODOLOGY. These challenges motivate (i) self–supervised representation learning to exploit abundant unlabeled images and become robust to projection/augmentation; (ii) multi wavelength study (thermal X–ray + non–thermal radio) to break geometry/physics degeneracies; and (iii) simulation–based, probabilistic inference (cINNs) that return full, possibly multi–modal posteriors rather than point estimates, explicitly acknowledging non–uniqueness [6, 29, 37]. In practice, at the first step, for getting the embedding, our contrastive learning pipeline, uses a set of augmentations that mimic observational variance, treat orthogonal projections as independent views to expose orientation variability and returns a representation space for X-ray, radio, and joint maps. Next, a mixture–of–experts trains local cINNs on different regions of the representation space, to leverage on clustered diversity. The final outcome, will be posteriors which will give a probabilistic distribution of the merger parameters, and maximum-a-posteriori (MAP) estimates which can be used to measure the relative error.

6.2 METHODOLOGICAL GAP

Limits of hand-crafted indicators

Classical morphology metrics (e.g., X–ray concentration/cuspiness, centroid/COM shifts, power ratios, and radio relic length/curvature/polarization), compress $10^5 - 10^6$ pixels into a few scalars. Similar approaches have been employed by, for example, Lee et al. [92], who estimate the time since collision from the observed separation of a pair of radio relics. While compact, these summaries discard multi-scale spatial information that is important for disentangling merger timing, three-dimensional geometry, and mass partition. Their values are also sensitive to analysis choices (e.g., aperture definitions, background modeling, PSF/beam convolution, exposure depth, and redshift), and are further confounded by projection and line-of-sight superposition effects [100, 132, 136, 179]. Consequently, methods that operate directly on image-level data (e.g., radio and X-ray maps) and preserve morphological information promises to provide stronger and more informative constraints.

Why self-supervised representation learning?

We use contrastive self-supervision (e.g., SimCLR) to learn an image-level embedding that serves as the *conditioning input* to the cINN. This addresses two needs:

1. *Dimensionality reduction for the cINN*: mapping $10^5 - 10^6$ input pixels into a compact vector of $\mathcal{O}(10^2 - 10^3)$ descriptors makes the conditioning signal tractable, stabilizes training, and reduces overfitting that would arise from feeding raw images into the flow.

2. *Morphology-preserving organization*: the learned representation groups images by physically similar structure (e.g. cores, shocks, tails, asymmetries), while suppressing nuisance variation (absolute flux, arbitrary orientation), effectively distilling the information most relevant for downstream inference [29, 132]. Physics-aware augmentations encode the desired invariances, and encoders integrate thermal (X-ray) and non-thermal (radio) channels without handcrafted features [49, 50].

Why simulation-based inference with cINNs?

Simulations are indispensable because the physical variables that define a galaxy cluster's merger parameters (e.g., time of collision, pericenter distance, relative velocity, and mass components) are not directly observable, and each real cluster is seen only as a single, projected snapshot of a Gyr-scale process [84]. Cosmological simulation of galaxies such as TNG-Cluster provide us with time–resolved merger trees and self–consistent 3D thermodynamic and magnetic fields from which one can synthesize X–ray and radio maps across snapshots and viewing angles. This enables (i) large, controllable training dataset with causal labels; (ii) systematic coverage of parameter space (mass ratios, impact parameters, redshifts); and (iii) explicit study of projection and selection effects by rendering multiple lines of sight for the same event [132, 136]. Simulation–anchored datasets permit principled *simulation–based inference*: flows such as cINNs can be trained on paired merger parameters and conditions (such as observable or maps) samples $((\mathbf{x}, \mathbf{c}))$ and be validated, and further applied on survey images [37].

Merger inversion is many–to–one and inherently degenerate; many different interaction histories can yield similar morphologies, so point estimates are misleading. Conditional invertible neural networks (cINNs) model the full posterior $p(\mathbf{x} \mid \mathbf{c})$ with exact likelihoods via invertible flows, capturing non–Gaussianity and multi–modality while amortizing inference over large datasets [6, 37]. Relative to alternatives: deterministic regressors (no uncertainty), ABC/rejection SBI (sample–inefficient), or VAEs/GANs (no tractable likelihood), cINNs yield precise and accurate posteriors suited to survey–scale image→physics inference. In short: contrastive learning helps us by compressing high–dimensional maps into informative representation space; and cINN will further *lift* those features into full posteriors over merger parameters.

6.3 THESIS TOOLS, OBJECTIVES AND RESEARCH QUESTIONS

Tools

The primary prerequisite is a cosmological hydrodynamical simulation. In this work, we use TNG-Cluster, and its 352 zoom-in halos [113], sampled over $0 \leqslant z \leqslant 1$ and projected along three orthogonal lines of sight. The X-ray dataset consists of intrinsic emission maps produced for TNG-Cluster by Nelson et al. [113], with a field of view of $\pm 2\,R_{200c}$ and a line-of-sight depth of $2\,R_{200c}$. The radio dataset is the intrinsic synchrotron maps at $\nu_{obs} = 1.4\,\mathrm{GHz}$, obtained by post-processing TNG-Cluster shock surfaces as produced by Lee et al. [91], using the same $\pm 2\,R_{200c}$ field of view. For each (halo, snapshot) pair, observable quantities are extracted directly from the simulation, while next/last merger parameters are taken from the Lee et al. [91], and defined with respect to the time of the first pericenter passage.

Objectives

1. **Learn morphology–aware representations from images.**
   a) Train SimCLR on *intrinsic* TNG-Cluster maps for (i) X–ray, (ii) radio, and (iii) paired X–ray+radio inputs with channel–consistent augmentations.
   b) Quantitatively assess the representations via kNN retrieval, UMAP organization, and label–aware hexbin overlays for halo, ICM, and merger properties.
2. **Perform simulation–based inference of merger physics.**
   a) Develop a conditional invertible neural network (cINN) with *rational–quadratic spline* couplings to infer posteriors $p(\mathbf{x} \mid \mathbf{c})$ for last/next–merger parameters: collision time, collision velocity, mass ratio, pericenter distance, and component masses.
   b) Compare conditioning on embeddings learnt from X–ray, radio, and joint maps.
3. **Improve performance with a Mixture–of–Experts (MoE).**
   a) Partition the learned embedding space with k–means (fit on training data only) and train expert–local cINNs.
4. **Establish evaluation and error quantification.**
   a) Use posterior–vs–truth heatmaps, prior–posterior contraction, and MAP error statistics (medians, 10-90% envelopes).
   b) Test cross–target correlations with corner plots (posterior samples, MAPs, ground truths) for physical consistency.

Primary research questions

1. **Main Question:** Fundamentally, can merger parameters be reliably inferred from imaging data alone? Furthermore, is it more effective to condition the

cINN directly on raw maps, or to first extract and use learned representations derived from those maps?

2. **Representation quality:** Do self–supervised SimCLR encoders trained on cluster maps learn morphology–aware embeddings that (i) form smooth neighborhoods and (ii) exhibit coherent, physically meaningful gradients when colored by halo/ICM/merger labels?

3. **Modality comparison:** How does conditioning on *radio* embeddings compare to *X–ray* embeddings for inferring merger parameters? Is joint (X–ray+radio) conditioning strictly better than either modality alone, or does one modality dominate?

4. **Inference fidelity:** When conditioning the cINN on learned embeddings, are posteriors $p(\mathbf{x} \mid \mathbf{c})$ *accurate* and *precise* across targets (collision time/velocity, mass ratio, pericenter, component masses)? How do posterior–vs–truth histograms, and prior–posterior contraction behave?

5. **Last vs. next merger:** Does inference performance differ between *past* (last) and *future* (next) merger parameters, and if so, which targets degrade most when forecasting?

6. **Against scalar baselines:** How does representation–conditioned inference compare to conditioning on scalar observables (core entropy, concentration, offsets, etc.) in terms of precision and accuracy?

Relation to Prior Work

Two recent studies Eisert et al. [49, 50] apply contrastive learning and conditional invertible networks to *galaxy* images to infer aspects of their assembly histories. They pretrain on large optical/near-IR cutouts of galaxies to learn representation spaces correlated with physical observable properties, and then, at inference time, condition their cINN on scalar summaries distilled from the images. In addition, their cINN flow uses GLOW-style affine coupling layers.

In this thesis we tackle a different regime, galaxy clusters, and a different data design (thermal X-ray and non-thermal radio, including paired inputs). Moreover, our cINN is conditioned directly on the learned representation vectors from contrastive pretraining, not on scalar observable properties. Moreover, the cINN pipeline employs rational–quadratic spline coupling blocks. The goals are related (recovering merger/assembly histories), but the methodology and implementation are different: the entire pipeline (e.g., preprocessing and the contrastive learning code, conditioning strategy, and the cINN code) was developed independently for this work, without reusing code from Eisert et al. [49, 50].

Another related study, is Chadayammuri et al. [28] which is a cluster-centric study that uses contrastive learning to connect simulated cluster X-ray maps to merger–related properties. Chadayammuri et al. [28] train contrastive learning on these maps and show that the learned representation space correlate with merger related quantities without needing to rely on morphology scalars. Downstream, they attach deterministic heads (linear probes or shallow regressors/classifiers) to the embedding to predict merger proxies and to retrieve semantically similar systems. The method does not model full posteriors: there is no conditional

flow/cINN and no uncertainty. It also remains single-modality (X-ray only), without radio maps or paired multi-wavelength training.

In spirit it is close to our aims: learn feature spaces directly from simulated cluster maps and use those features to probe dynamical state or merger history. However in this thesis, we condition a cINN directly on learned representation space of radio maps, and paired (radio + X-ray) maps besides X-ray. Moreover, this thesis applies the cINN on the representation space, returning posteriors and a distribution for merger parameters instead of only point estimations.

6.4 THESIS ROADMAP

The thesis proceeds as multiple, end–to–end pipelines whose intermediate artifacts (maps, embeddings, and trained experts) flow forward from one chapter to the next:

1. **Data & physical setup (Chs. 13, 17).** We define the simulation inputs and map–making for the two observables: intrinsic X–ray surface–brightness maps from TNG-Cluster (made by Nelson et al. [113]), and intrinsic synchrotron radio maps constructed with the shock–based emissivity model (made by Lee et al. [91]). Each chapter specifies the detail description on how the maps are built, their fields of view, projection axes, redshift snapshots, pixelization, and per–modality normalizations.

2. **Self-supervised representations (Chapters 14, 18, 20).** We train SimCLR separately on X–ray and radio maps, then on paired two–channel inputs. For each setting we extract 512-D embeddings (or fused codes), probe their structure (UMAP grids, kNN), and show astrophysical gradients when colored by labels. *Output:* fixed, morphology–centric embeddings for every map in single- and joint-modality variants.

3. **Inference model (Chapters 11.1, 15.1-15.3).** We introduce the conditional invertible neural network (cINN), detail conditioning on embeddings, and construct a Mixture–of–Experts (MoE) partition in representation space via k-means. We describe training (NLL objective), sampling, and postprocessing (MAP, priors, and posteriors). *Output:* an accurate, expert–conditioned $p(\mathbf{x} \mid \mathbf{c})$ for merger parameters.

4. **X–ray–conditioned results (Part v, Chapter 16).** Using the X–ray embeddings as conditions, we evaluate posterior calibration, MAP accuracy, and cross–target correlations for last– and next–merger parameters. We analyze strengths/weaknesses versus scalar baselines and visualize representative posteriors. *Output:* validated inferences from thermal morphology alone.

5. **Radio–conditioned results (Part vi, Chapter 19).** We repeat the full evaluation with radio embeddings, comparing against X–ray. We show systematically tighter posteriors for several targets and assess robustness across the test population. *Output:* validated inferences from non–thermal morphology and a quantitative modality comparison.

6. **Joint X–ray+Radio conditioning (Part vii, Chapter 21.5).** We join modalities (using joint contrastive learning code) and assess whether joint condition-

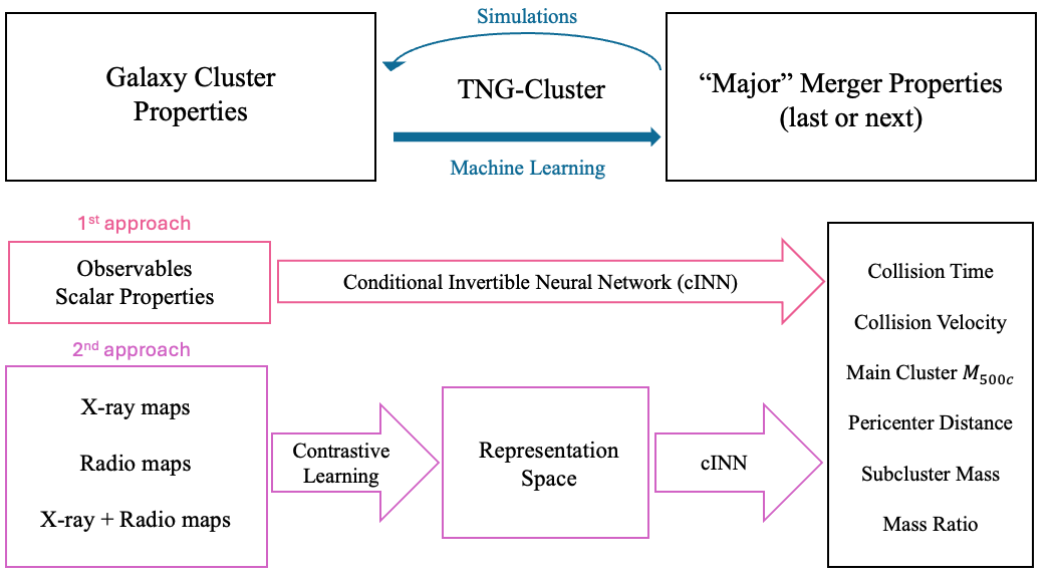

Figure 3: Schematic overview of the thesis workflow. Starting from simulated galaxy cluster properties derived from *TNG-Cluster*, two approaches are pursued to infer merger properties: (i) a direct route using scalar observables as cINN inputs, and (ii) an image-driven route where contrastive learning compresses X-ray, radio, or combined X-ray+radio maps into morphology-aware representations for conditioning the cINN.

ing improves over single–modality results. We report posterior calibration, MAP accuracy, and correlation recovery for last/next mergers, highlighting intermediate performance relative to radio–only. *Output:* a multi–wavelength inference pathway.

7. **Appendices (Appendix .1 and .2).** We extend the cINN to predict a wide set of *observables* from embeddings (halo/BCG/ICM/dynamical), provide similar diagnostic plots as for unobservable.

The overall workflow of this thesis is summarized in Figure 3. The diagram shows how simulation inputs are used to infer galaxy cluster merger properties via two different approaches: (i) a scalar–observable route, where a cINN is applied directly to the scalars; and (ii) an imaging route, where X-ray/radio maps are first encoded into a morphology-aware representation by contrastive learning, and the cINN is then conditioned on these embeddings.

Part III

GALAXY CLUSTERS IN THE TNG-CLUSTER SIMULATIONS

7

TNG-CLUSTER SIMULATION

7.1 ILLUSTRIS TNG AND TNG-CLUSTER PROJECT

Numerical Framework: AREPO *and Ideal MHD*

The IllustrisTNG simulations are built on the moving-mesh code AREPO, which solves the coupled equations of self-gravity and (magneto-)hydrodynamics on an unstructured Voronoi tessellation [127, 155]. Hydrodynamics is advanced with a finite-volume, Godunov scheme using directionally unsplit second-order time integration and Riemann solvers at cell interfaces; the mesh moves approximately with the local flow (arbitrary Lagrangian-Eulerian), providing adaptive spatial and temporal resolution without preferred directions [155]. Gravity is computed with a Tree-PM scheme, while the Friedmann-Lemaître background sets cosmic expansion [127]. IllustrisTNG evolves ideal magneto-hydrodynamics (MHD), advecting the cell-averaged magnetic field and allowing self-consistent amplification of a weak seed field by structure formation and turbulence [127]. This combination is central for cluster applications because it resolves shocks and mixing with low advection errors, while MHD controls magnetic pressure support, anisotropic-transport proxies, and the magnetization level of the ICM [113].

Cosmology

TNG adopts a flat ΛCDM cosmology consistent with Planck Collaboration et al. [129], with parameter values $\Omega_m = 0.3089$, $\Omega_b = 0.0486$, $\Omega_\Lambda = 0.6911$, $\sigma_8 = 0.8159$, $n_s = 0.9667$, and $H_0 = 67.74\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ [114, 129]. These parameters set the halo mass function and growth histories against which the baryonic model operates; for clusters, they fix the expected abundance of $M_{200c} \geqslant 10^{14}\,M_\odot$ systems and the timing and other properties of mergers that we will try to infer in this thesis. In particular, Ω_m and Ω_Λ govern the overall growth rate of cosmic structure, determining how rapidly massive clusters assemble at different epochs [25]. The power spectrum normalization σ_8 controls the amplitude of density fluctuations, directly affecting the number of massive progenitors available for merging [163]. The spectral index n_s influences the scale dependence of these fluctuations, shaping the relative contribution of mergers across mass scales [47]. Finally, H_0 sets the relationship between redshift and cosmic time, which fixes the temporal spacing between merger events and the physical scales over which clusters can interact [65]. Together, these parameters establish the statistical backdrop for the cluster merger rate, mass ratio distribution, and redshift evolution that our analysis will probe.

Galaxy Formation Physics

The IllustrisTNG model couples hydrodynamics and gravity to a calibrated, physically motivated set of unresolved (subgrid) processes [127, 169].

- *Radiative processes and star formation.* Gas cooling and heating account for both primordial and metal-line emission, in the presence of a time-dependent UV/X-ray background. At high densities, gas is treated with the Springel–Hernquist two-phase ISM model, where stars form stochastically according to an effective pressure law [157]. Even though the cores of massive clusters remain hot, pockets of gas with short cooling times can still form, leading to multiphase condensation. The star-formation model therefore regulates the residual star formation in brightest cluster galaxies (BCGs) and the consumption of condensed cold gas.

- *Chemical enrichment.* The simulation follows nine elemental species (H, He, C, N, O, Ne, Mg, Si, Fe), with time-resolved mass and metal return from core-collapse supernovae (SNe II), Type Ia supernovae (SNe Ia), and asymptotic giant branch (AGB) stars, using updated stellar yields and lifetimes [127]. These metals regulate gas cooling (and hence the cooling time, t_{cool}) and determine the strengths of X-ray line emission.

- *Stellar (galactic wind) feedback.* Relative to the original Illustris model, TNG launches winds *isotropically*, injects a controlled thermal fraction, and uses updated velocity/energy scalings (including metallicity dependence and a floor), improving regulation of stellar mass and the metal budget in infalling groups and satellites [127]. For clusters, the primary impact is indirect but important: by setting the stellar mass of satellites and the BCG, winds determine the overall stellar-to-gas partition, the amount and distribution of metals pre-processed in infalling halos, and the fuel available for central cooling flows between AGN episodes. These also influence satellite gas removal and the ICM's enrichment/entropy budget during group–group and group–cluster mergers [113, 127].

- *SMBHs and AGN feedback.* Black holes are seeded in massive halos, grow by gas accretion and mergers, and inject energy in two modes. At high accretion rates the model uses thermal (quasar-mode) feedback; at low Eddington ratios it switches to a kinetic (wind-mode) that directly imparts momentum/energy to surrounding gas. This kinetic low-accretion mode, introduced specifically in TNG, efficiently couples to hot halo atmospheres at late times, keeping massive galaxies quenched and restructuring the central ICM entropy/density [127, 169]. For galaxy clusters this ingredient is pivotal: it competes with radiative cooling to set cool-core versus non-cool-core states, can reheat or displace low-entropy gas after mergers, and shapes the observables (e.g. central entropy, cooling time) which will be later discussed in Part iv [113].

- *Magnetic fields.* Regarding the magnetic fields, IllustrisTNG evolves a weak, uniform primordial seed field that is amplified by compression, shear, and small-scale dynamos [127]. While the exact seed value is unimportant for galaxy statistics, evolving ideal MHD yields cluster-scale magnetic fields and morphologies that affect buoyancy, mixing, and the confinement of AGN-driven structures. For merger studies, MHD controls how sloshing/turbulence grow and decay, how cold fronts persist, and how quickly metal/entropy inhomogeneities are erased, and therefor influencing the longevity of cool cores post-merger [113].

From IllustrisTNG to TNG-Cluster

Galaxy clusters lie on the exponentially suppressed tail of the halo mass function, so any uniform-volume simulation must trade numerical resolution against statistical power at the highest masses. Within the base IllustrisTNG boxes, even the largest (TNG300; $\sim$ 300 Mpc per side) contains only a limited number of Coma-mass systems at $z = 0$, constraining population inferences at $M_{200c} \geqslant 10^{15}\,M_\odot$ [113, 127]. The *TNG-Cluster* project addresses this by constructing a large, mass-representative sample of cluster halos while keeping the AREPO numerics and the galaxy-formation physics *unchanged* from TNG, thereby offering population-level statements at the top end of the mass function and clean, like-for-like comparisons across mass and redshift [113]. Detailed aspects of the TNG-Cluster are presented in the next section.

7.2 TNG-CLUSTER: SETUP, DATA PRODUCTS

Setup

TNG-Cluster draws its targets from a dark-matter-only parent simulation of side length 1 Gpc (TNG-Cluster-Dark). Halos are identified at $z = 0$ and selected purely by mass in narrow 0.1 dex bins to yield an approximately flat sampling in $\log M_{200c}$ over $\log(M_{200c}/M_\odot) \simeq 14.3$–$15.4$. Above $M_{200c} \geqslant 10^{15} M_\odot$ all halos in the parent box are included, providing volume-limited statistics at the rarest masses. In total, 352 regions are re-simulated; the 1 Gpc parent (roughly 36$\times$ the TNG300 volume) yields of order ninety halos above $10^{15}\,M_\odot$ [113].

For each selected cluster, all FoF-member DM particles at $z = 0$ are traced back to the initial redshift to define the Lagrangian region. An adaptive oct-tree marks occupied cells, which are then expanded to enclose $\sim 3\times$ the original Lagrangian volume; this conservative padding suppresses late-time incursion of low-resolution particles into the virial region. The refined patch is built at $4\times$ higher linear resolution than the parent (i.e. $64\times$ better mass resolution), and embedded in a progressively coarser buffer comprising eight discrete mass levels.

The zoom-in simulations start from a large dark-matter-only (DMO) box, 1 Gpc on a side (TNG-Cluster-Dark). From this box, candidate halos at redshift $z = 0$ are chosen based only on their mass. For each selected halo, all the dark matter particles in its FoF group at $z = 0$ are traced back to the initial conditions to identify the halo's Lagrangian region (the patch of the early universe that collapses to form the halo). This region is built using an adaptive oct-tree grid. First, all cells containing the traced particles are marked. Then the marked volume is expanded until it is about three times larger than the original. This padding ensures that no low-resolution (i.e. massive) particles drift into the halo's virial radius at late times. Within this high-resolution patch, the grid is further refined: the linear resolution is increased by a factor of four, so each cell is $4^3 = 64$ times smaller in volume (and mass) compared to the parent box. The surrounding space is then filled with particles of progressively lower resolution, arranged in eight discrete mass levels that smoothly increase with distance from the target halo [113]

The initial conditions are generated with the Zel'dovich approximation using N-GenIC. Starting from a uniform particle grid, small displacements are applied

to imprint the desired density fluctuations. Outside the zoom-in region, the large-scale phases of the parent simulation are preserved so that the tidal environment remains identical. Within the refined region, however, additional small-scale modes are introduced, providing the extra power needed to resolve structure at higher resolution [113, 127]. This setup improves the effective spatial (mass) resolution inside the zoom-in volume by about a factor of 4 (64) relative to the parent DMO run, intentionally matching the resolution of TNG300-1. Afterwards, possible contamination from low-resolution particles is measured (Appendix A in [113]) and found to be negligible.

Baryons are added at the universal fraction in the high-resolution region and the unchanged IllustrisTNG galaxy-formation model is employed: radiative cooling and a time-dependent UV/X-ray background, star formation in a pressurized two-phase ISM, time-resolved chemical enrichment from SNe II/Ia and AGB stars (nine tracked elements), an updated galactic-wind model, and the dual-mode SMBH feedback scheme [127, 169]. Cosmological parameters follow the Planck 2015/2016 ΛCDM set used throughout TNG (see Section 7.1). Mass and force resolutions are held to strict parity with TNG300-1 (identical mean gas-cell and DM particle masses; collisionless softenings comoving at early times transitioning to fixed physical at late times; gas with adaptive softenings tied to cell size), ensuring like-for-like comparisons of ICM structure and merger-driven transients across the joint TNG300+TNG-Cluster ensemble [113, 127].

The simulations include hydrodynamics, gravity, and the full galaxy-formation physics model of IllustrisTNG (see Section 7.1). The zoom-in regions are run at exactly the same mass resolution as TNG300-1: a mean baryonic cell mass of $m_b \simeq 1.2 \times 10^7, M_\odot$ and a dark matter particle mass of $m_{DM} \simeq 6.1 \times 10^7, M_\odot$ (quoted for $h = 0.6774$). Gravitational softenings follow the same prescriptions: for collisionless species, softenings are comoving at high redshift and switch to fixed physical values at late times, reaching $\epsilon_{DM,\star} = 1.5$ kpc at $z = 0$; for gas, softenings adapt to the instantaneous cell size with a comoving minimum of ~ 0.25 ckpc/h [113, 114]. By construction, this strict consistency ensures that any differences between halos of the same mass in TNG300 and the TNG-Cluster ensemble reflect variations in their assembly histories, not numerical artifacts.

Data Products

TNG-Cluster follows the IllustrisTNG snapshot strategy with 100 outputs from high redshift to $z = 0$, comprising 20 *full* and 80 *mini* snapshots [113, 114]. Full snapshots store the complete set of particle/cell fields for all types, while mini snapshots provide a reduced, analysis-focused subset (per-field availability is documented in the public data specifications) [114]. The 20 full snapshots correspond to the snapshot numbers, redshifts, and cosmic times listed in table 1 scanning $z \simeq 12 \rightarrow 0$.

The halo and galaxy identifications used throughout this thesis follow the standard two-stage procedure of a friends-of-friends (FoF) group finder applied to the dark matter (DM) field, followed by the SUBFIND algorithm to decompose each FoF group into a bound central object and its substructures [38, 114, 158]. In all TNG and TNG-Cluster runs, FoF is executed with a linking length $b = 0.2$ times the mean inter-particle separation on DM particles; baryonic resolution elements

Snapshot	Redshift	Time (Gyr)
2	11.9802	0.3702
3	10.9756	0.4177
4	9.9966	0.4747
6	9.0023	0.5471
8	8.0122	0.6396
11	7.0054	0.7639
13	6.0108	0.9317
17	4.9959	1.1772
21	4.0079	1.5404
25	3.0081	2.1454
33	2.0020	3.2845
40	1.4955	4.2929
50	0.9973	5.8780
59	0.7001	7.3141
67	0.5030	8.5866
72	0.3999	9.3891
78	0.2977	10.2986
84	0.1973	11.3224
91	0.0994	12.4664
99	0.0000	13.8027

Table 1: Full snapshot numbers and their corresponding redshifts and cosmic times in TNG-Cluster.

(gas, stars, black holes) are then attached to the FoF group of their nearest DM particle. Within each FoF group, SUBFIND identifies locally overdense, gravitationally self-bound subhalos and computes their properties including masses, centers, and kinematics [114].

For each FoF halo, we define the center as the position of the particle with the minimum gravitational potential energy within the group (i.e., the center-of-potential). We identify the brightest cluster galaxy (BCG) with this central subhalo by default, while noting that it is usually, but not always, the most massive subhalo [114]. Throughout this work, the primary aperture is the critical overdensity sphere at $\Delta = 200$ or $\Delta = 500$, defined by the radius $R_{\Delta c}$ (R_{200c} or R_{500c})for which the mean enclosed density equals $\Delta\rho_c(z)$, and the corresponding enclosed mass is $M_{\Delta c}$(M_{200c} or M_{500c}). Within each FoF halo, SUBFIND ranks self-bound substructures by their number of bound elements, with the first entry corresponding to the central or primary subhalo.

8

TNG-CLUSTER SAMPLE

8.1 MASS-REDSHIFT DEMOGRAPHICS

Having established our data products and mass/centering conventions in the previous section, we begin by exploring how many halos we have at which masses and redshifts. This provides the statistical context for all subsequent results and makes explicit the consequences of the $z = 0$ mass-targeted selection. Figure 4 and Tables 2-3 (binned counts) summarize the demographics. At $z = 0$, *TNG-Cluster* contributes 95 halos with $\log_{10}(M_{200c}/M_\odot) \geqslant 15$ (versus 3 in TNG300-1), and 204 vs. 38 in $14.5 \leqslant \log_{10} M_{200c} < 15$. The lowest bin is dominated by TNG300-1 (239 vs. 53), reflecting the design goal of a flat mass $z = 0$ TNG-Cluster selection. By $z = 0.5$ the number of $> 10^{15}\,M_\odot$ halos drops to 16 (TNG-Cluster) + 1 (TNG300-1), and by z=1 there are none, consistent with hierarchical growth.

Mass range (TNG300)	$z = 0$	$z = 0.5$	$z = 1$	$z = 2$
$14 \leqslant \log_{10}(M_{200c}/M_\odot) < 14.5$	239	134	48	3
$14.5 \leqslant \log_{10}(M_{200c}/M_\odot) < 15$	38	14	2	0
$\log_{10}(M_{200c}/M_\odot) \geqslant 15$	3	1	0	0

Table 2: Number of halos in different $\log_{10}(M_{200c}/M_\odot)$ ranges for TNG300 at selected redshifts.

Mass range (TNG-Cluster)	$z = 0$	$z = 0.5$	$z = 1$	$z = 2$
$14 \leqslant \log_{10}(M_{200c}/M_\odot) < 14.5$	53	159	198	31
$14.5 \leqslant \log_{10}(M_{200c}/M_\odot) < 15$	204	156	48	0
$\log_{10}(M_{200c}/M_\odot) \geqslant 15$	95	16	0	0

Table 3: Number of primary zoom halos in different $\log(M_{200c}/M_\odot)$ ranges for TNG-Cluster at selected redshifts.

8.2 ICM AND ITS PROPERTIES

Baryon fraction

We begin by quantifying how baryons are partitioned among hot gas, cold gas, and stars as a function of halo mass. For each halo at z=0 we record the total baryon fraction, the total gas fraction, and their decomposition into a hot phase ($T > 10^6$k) and a cold phase ($T < 10^6$k), alongside the stellar fraction. The total baryonic mass is defined as the sum of gas, stellar, and black-hole masses (the

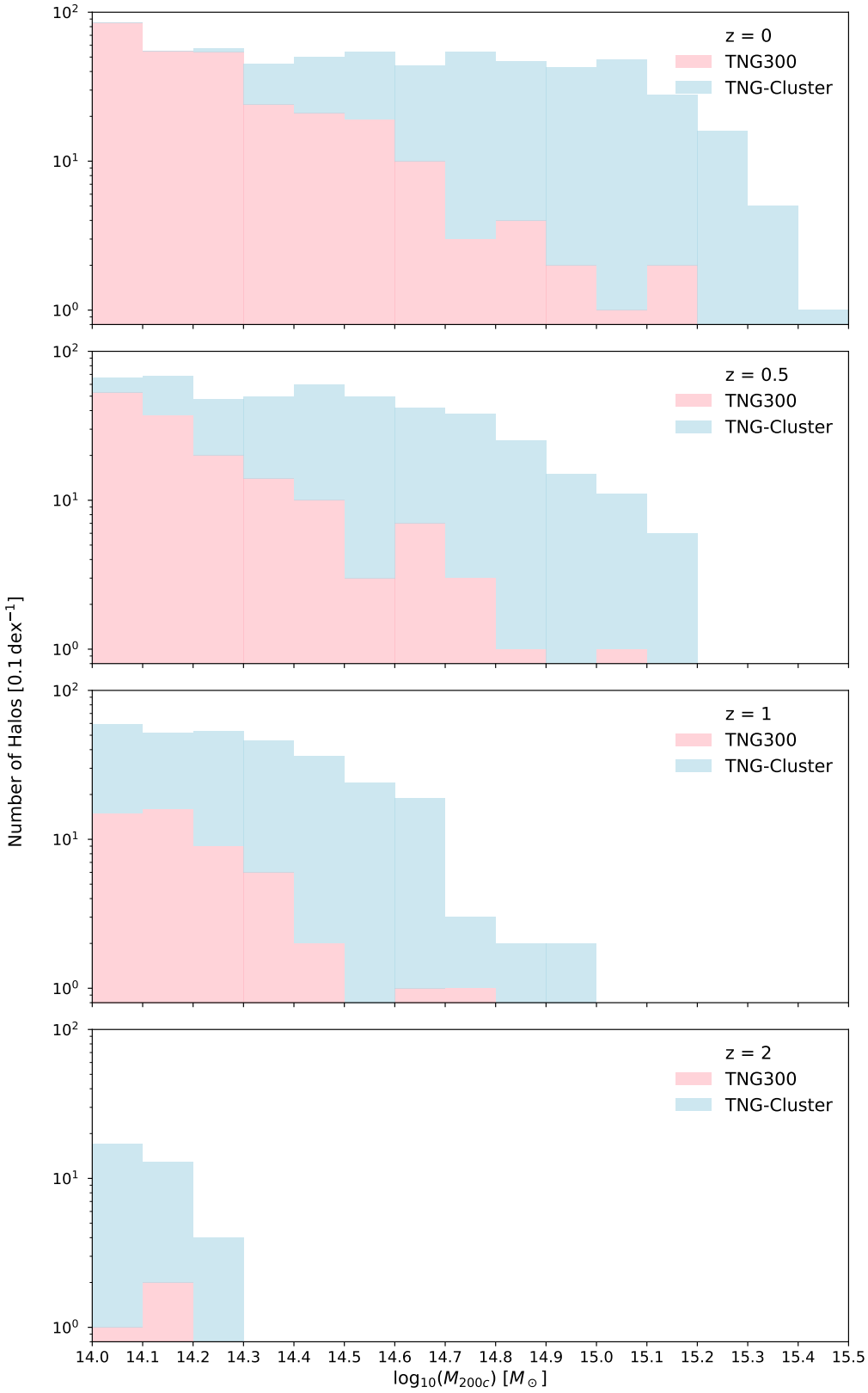

Figure 4: Cluster mass function for primary zoom targets of TNG-Cluster simulation (blue) stacked on top of the TNG300 across z=0, 0.5, 1, 2, with the bin width of 0.1 dex. *TNG-Cluster* supplies the vast majority of $M_{200c} \geqslant 10^{15}\,M_\odot$ systems, enabling ensemble analyses of rare mergers. Redshift panels visualize the progenitor-biased nature of the $z = 0$-selected sample at earlier times, a caveat we account for when presenting evolutionary trends.

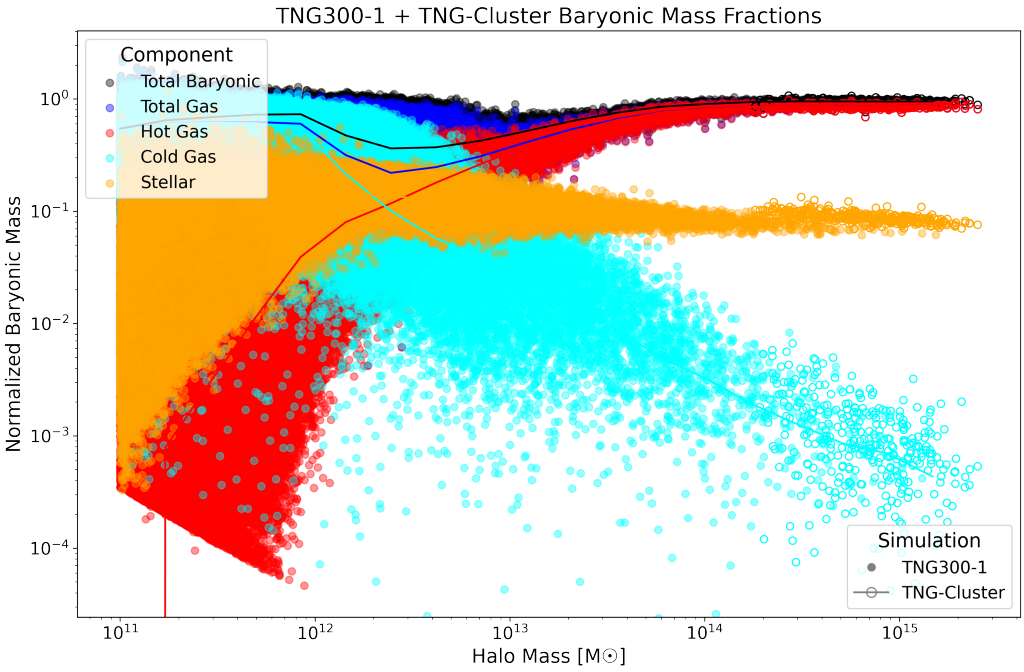

Figure 5: **Baryon and phase fractions vs. halo mass at** z=0. Component masses within R_{200c}, normalized by $f_b M_{200c}$, for TNG300-1 (filled points) and *TNG-Cluster* (open points). Lines show running medians in logarithmic mass bins. The gas fraction increases with halo mass and approaches the cosmic value at the top end, the stellar fraction declines, and the hot phase dominates the ICM budget across the cluster regime.

latter being negligible in clusters). All fractions are normalized by the product of the halo mass and the cosmic baryon fraction, i.e.

$$f_{comp} = \frac{M_{comp}}{f_b\, M_{200c}}, \qquad f_b = \Omega_b/\Omega_m \simeq 0.157,$$

with running medians shown for each component (Figure 5).

Figure 5 combines TNG300-1 (filled markers) and TNG-Cluster (open markers) to cover the full cluster mass range. Four trends emerge. (i) The gas fraction rises with halo mass and approaches the cosmic value at the top end, as expected, since deeper potentials better retain and thermalize baryons. (ii) The stellar fraction declines with mass, reflecting the reduced integrated star-formation efficiency in massive halos. (iii) Within the gas, the hot phase dominates across the cluster regime, while the cold component is dominant at all $M_{200c} \leqslant 10^{13} M_\odot$ and becomes negligible at the highest masses. (iv) In the cluster regime the total baryon fraction closely tracks the hot-gas fraction, indicating that most of the ICM mass budget resides in X-ray–emitting plasma rather than in stars or cold gas [3, 32, 133].

These mass trends and their scatter are the imprint of near self-similar gravitational physics modulated by feedback and assembly. As halos grow, baryon retention increases and the ICM becomes more thermalized, increasing the f_{gas} and decreasing the $f_\star$ [32].

Thermal structure and X-ray definitions

As discussed in figure 5, hot, diffuse plasma in the intracluster medium (ICM) contains the bulk of baryons in clusters and therefor, it radiates predominantly in X-rays. At the temperatures characteristic of rich clusters ($kT \sim 2$–10 keV), the emissivity is well approximated by thermal bremsstrahlung (free–free emission), with metal lines contributing increasingly toward group scales and in cool cores. X-ray measurements therefore encode the ICM's density, temperature, and metallicity.

For this chapter only and for simplicity, we compute X-ray luminosities with the bolometric bremsstrahlung estimator of Navarro, Frenk, and White [111]:

$$L_X = 1.2 \times 10^{-24} (\mu m_p)^{-2} m_g \sum_{i=1}^{N_{gas}} \rho_i T_i^{1/2} \text{ erg s}^{-1} \tag{8}$$

where m_p is the proton mass, $\mu \simeq 0.6$ for a fully ionized primordial plasma, and the sum runs over hot gas elements ($T_i \geqslant 10^6$ K). Here ρ_i and T_i denote the mass density and keV temperature of the i-th gas particle, and m_g is the gas mass. In our moving–mesh data the natural implementation replaces m_g by the individual cell mass m_i in the sum. It is important to note that the Equation 8 provides a bolometric, metal–independent estimate appropriate for hot clusters where bremsstrahlung dominates; it neglects line emission (important below ~ 3 keV) and bandpass/K-corrections.

Thermal structure: temperature and X-ray radial profiles

With the X-ray definition in Equation 8 established, we characterize the spherically averaged ICM at $z = 0$ for the primary zoom-in halos of TNG-Cluster. We build 3D radial profiles in shells about the potential minimum (halo's center), normalizing radii by R_{200c} and color–coding each halo by $\log M_{200c}$. Specifically in figure 6, we show (i) the mean shell temperature $\langle T \rangle(r)$ and (ii) the shell X-ray luminosity $L_X(r)$ from the bremsstrahlung estimator (Equation 8).

Temperature and X-ray profiles show strong outer self-similarity. They also display clear mass ordering. Inside $\leqslant 0.2\, R_{200c}$ the behavior diverges strongly, reflecting core state and recent dynamical activity. Because $L_X \propto \rho^2 T^{1/2}$ (free–free emission), X-ray profiles are very sensitive to gas clumping and to cool, dense cores. This sensitivity explains the larger scatter at small radii. The steep central rise is mainly a *cool–core* signature: high density and modest temperatures boost the emissivity. Sloshing on the other hand does not create the cusp, instead, it introduces azimuthal asymmetries, cold fronts, and small "wiggles" in the spherically averaged profile. It can also offset the X-ray peak from the potential center, which further broadens the inner scatter [98]. The mass coloring confirms a simple trend: at fixed r/R_{200c}, more massive systems are hotter and have brighter cores.

Temperature and X-ray maps

Having established the radial trends, we turn to two–dimensional ICM morphology. We use the most massive zoom-in halo of TNG-Cluster at z=0 (HaloID 0), and construct paired maps within a square of side $2R_{500c}$, centered on the potential minimum. Gas cells belonging to the FoF halo are projected along the line of sight onto an N×N grid (N=300).

For temperature map, for each pixel we compute the mean $\langle \log T \rangle$ (K). And for X-ray map, per pixel we sum the cell luminosities from the bolometric bremsstrahlung estimator (Equation 8) and divide by the pixel area to obtain SB_X in $\mathrm{erg\,s^{-1}\,kpc^{-2}}$.

The X-ray panel exhibits a bright central peak and a smooth global decline with radius, punctuated by faint substructures. The temperature field is comparatively smooth, with coherent gradients and localized hot/cool patches that trace recent stirring or minor accretion. Together, the maps reveal asymmetries and small features that spherical profiles can average out.

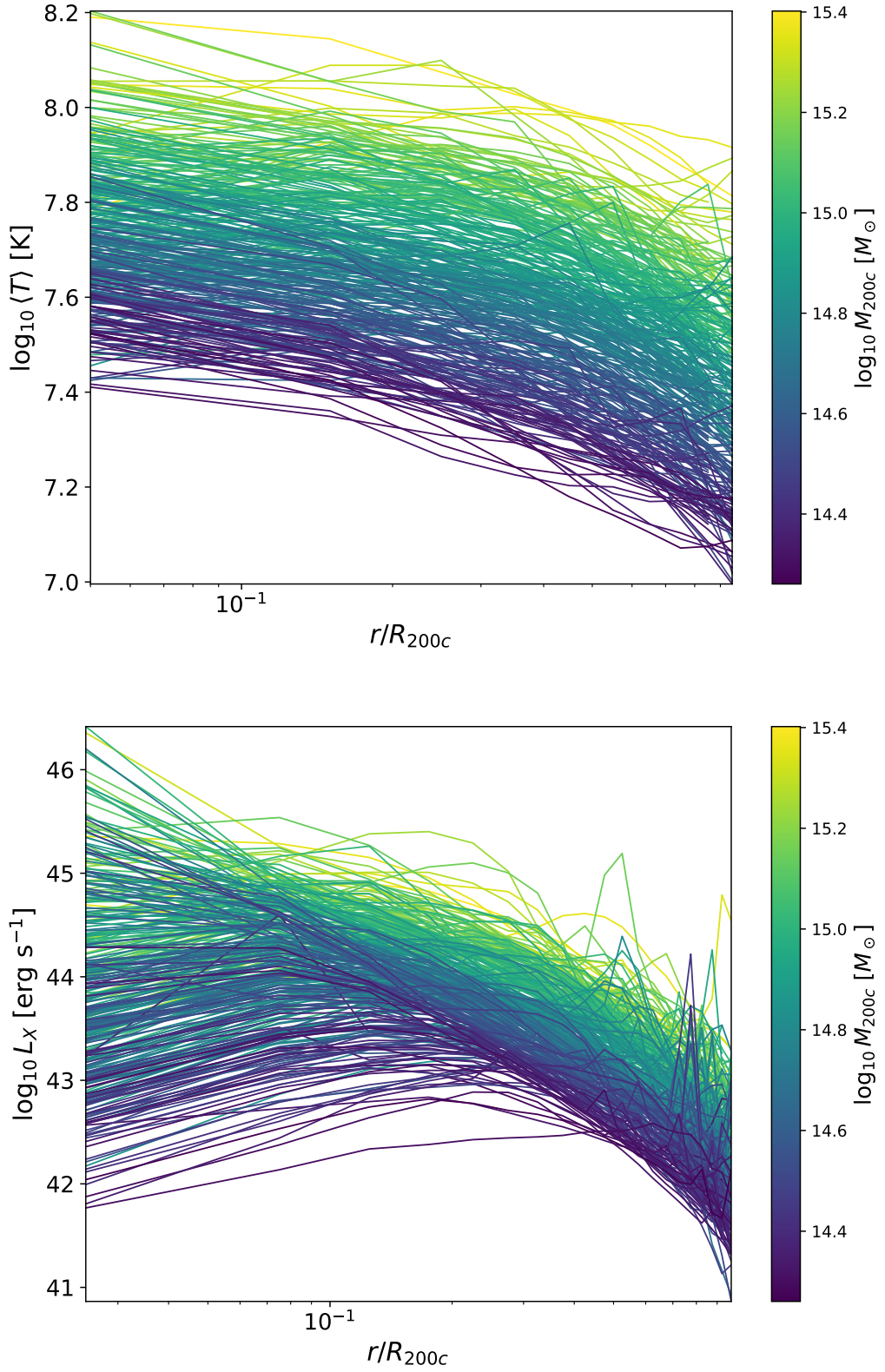

Figure 6: **ICM radial structure at $z = 0$.** (a) Normalized radial temperature profiles $\langle T\rangle(r)$ and (b) X-ray luminosity profiles $L_X(r)$ versus r/R_{200c} for TNG-Cluster zoom-in halos colored based on the M_{200c}. More massive halos are hotter and more X-ray luminous at fixed scaled radius; outside the core, profiles decline gently with clear mass ordering, while the inner $\leqslant 0.2R_{200c}$ shows substantial diversity indicative of cool-core vs. non–cool-core states. In panel (b), the steep central rise and enhanced small-scale fluctuations reflect the $L_X \propto \rho^2 T^{1/2}$ dependence and substructure/sloshing.

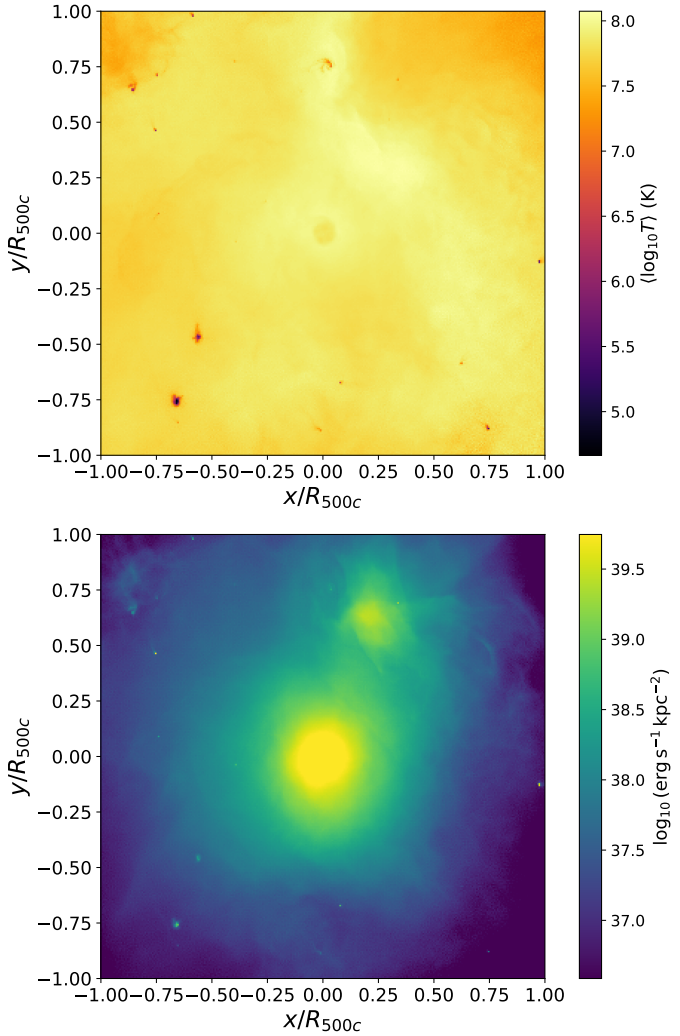

Figure 7: **Halo 0 at $z = 0$: emission structure within R_{500c}.** *Top:* mean $\langle \log T \rangle$ (K). *Bottom:* log X-ray surface brightness from the free–free Bremstrehlung (Eq. 8) in erg s^{-1} kpc^{-2}. The field spans $[-R_{500c}, +R_{500c}]$ in both directions and is centered on the potential minimum. The bright core and gentle outer gradient are evident in X-rays; with small asymmetries and substructures appearing in both panels.

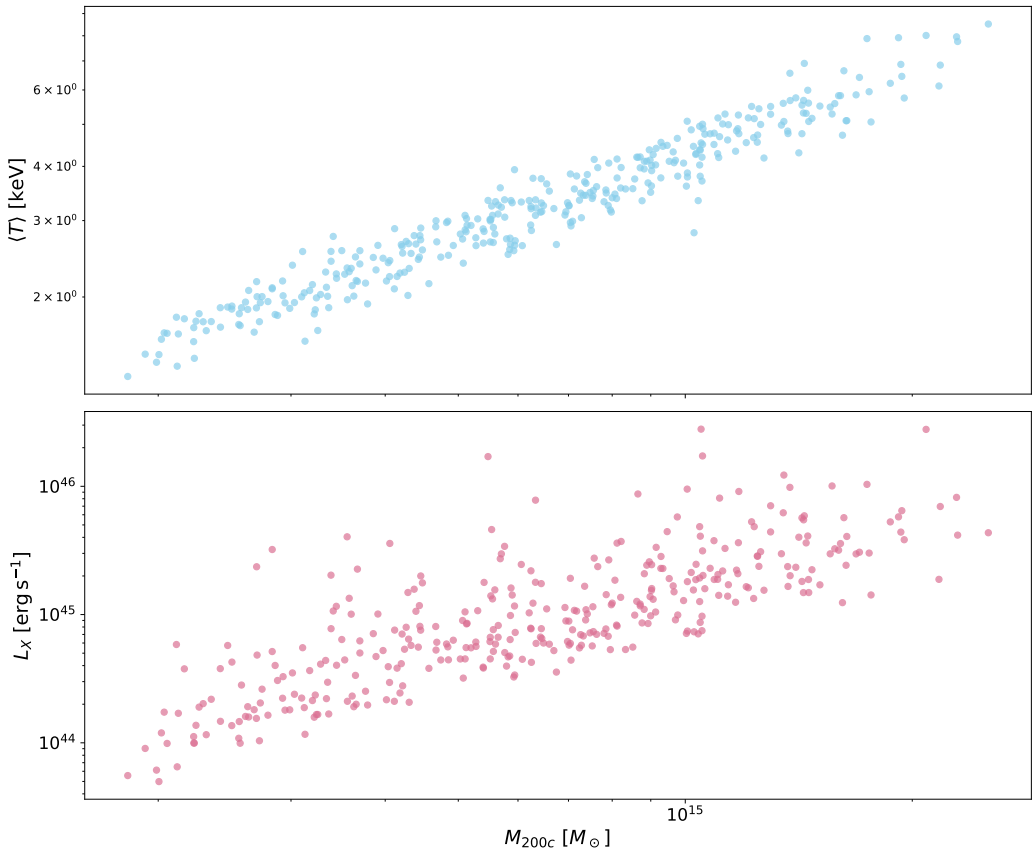

Figure 8: **Global ICM scaling at** z=0. *Top:* mass–weighted mean temperature within R_{200c} vs. M_{200c}. *Bottom:* bolometric X-ray luminosity (Eq. 8) within R_{200c} vs. M_{200c}. The T–M relation is tight, while L_X–M shows larger intrinsic scatter owing to the ρ^2 dependence that emphasizes core structure, clumping, and recent dynamical activity [84, 133].

Mass–observable scalings: T–M *and* L_X–M

We now turn from profiles to global ICM observables at z=0. For each primary zoom-in halo in TNG-Cluster we measure a mass–weighted mean temperature within R_{200c} and a bolometric X-ray luminosity using equation 8. Radii are centered on the potential minimum. Figure 8 shows the resulting T–M and L_X–M relations.

The T–M relation is tight and monotonic; hotter halos are more massive, which is consistent with virial self-similarity ($T \propto M^{2/3}$; Kaiser [72]) and relatively insensitive to core physics when temperatures are mass-weighted or core–excised [165]. By contrast, L_X–M also rises with mass but shows much larger intrinsic scatter because $L_X \propto \rho^2 T^{1/2}$: the density-squared weighting amplifies core structure, clumping, and substructure that vary with AGN feedback and recent mergers [84, 133]. It is important to not that our L_X is bolometric and metal–independent by construction (Equation 8). Band-limited observational comparisons (e.g. 0.5–2 keV) will differ depending on T and metallicity.

8.3 BRIGHTEST CLUSTER GALAXY

BCG mass components versus halo mass

We identify the BCG with the central galaxy in a cluster as explained in 7.1 and extract its: total BCG Mass, black hole, gas, dark matter, and stellar components. Figure 9 shows each component as a function of M_{200c} for the z=0 primary *TNG-Cluster* zoom-in halos; points are individual systems and thick lines indicate running medians in logarithmic mass bins.

- *DM:* dominates the BCG's bound mass and rises nearly in lockstep with M_{200c}, tracing the deepening inner potential.
- *Gas:* $\sim$ 0.5–1.0 dex below the total; including halo-to-halo scatter because of different cooling and AGN feedback feedback efficiency across clusters.
- *Stars:* increase sub–linearly with M_{200c}; typically $\sim$ 2 dex below the total and $\sim$ 1 dex below the gas curve, indicating the low in-situ efficiency and merger-driven growth.
- *BH:* correlated with BCG mass but gravitationally negligible; $\sim 10^{-2}$–10^{-3} of $M_{\star,BCG}$ (and $\sim 10^{-4} - 10^{-5}$ of the BCG's total mass).

These magnitudes provide a practical baseline for linking BCG growth modes to core thermodynamics and AGN energetics.

Relaxation Criteria and M_{12}

Following Ayromlou et al. [8], we use the (dimensionless) definition of *offset magnitude*:

$$x_{off} = \frac{|\mathbf{r}_{MBP} - \mathbf{r}_{CM}|}{R_{200c}}, \tag{9}$$

where $\mathbf{r}_{MBP}$ is the position of the most–bound particle (the halo center defined by the potential minimum), $\mathbf{r}_{CM}$ is the center of mass of the halo. As used in Ayromlou et al. [8] the systems are classified as relaxed for $x_{off} \leqslant 0.1$, and not-relaxed or disturbed for $x_{off} > 0.1$, indicating recent mergers [8].

On the other hand we define M_{12} within each halo as the ratio of the gravitationally bound masses of the two most massive subhalos at the same time, $M_{12} = M_1/M_2$, where M_1 and M_2 are the masses of the most– and second–most–massive subhalos. Where $M_{12} \gg 1$ indicates a dominant central with a much smaller secondary (minor accretion regime). Because a near equal mass pair ($M_{12} \sim 1 - 3$) moves the center of mass relative to the potential minimum, and as a result it makes the distance $|\mathbf{r}_{MBP} - \mathbf{r}_{CM}|$ larger because the center of mass, sits larger from the dominant subhalo's core (in most cases). As a result, lower M_{12} values are expected to correlate with larger offsets x_{off} and vice versa.

Figure 10 shows the subhalo mass ratio M_{12} versus the offset magnitude x_{off} on $\log_{10}$ axes. The purple band marks relaxed halos ($x_{off} \leqslant 0.1$) and the pink band non–relaxed halos ($x_{off} > 0.1$). We find a clear tendency for larger offsets at *low* M_{12} (two similarly massive subhalos) and smaller offsets at *high* M_{12} (one dominant subhalo).

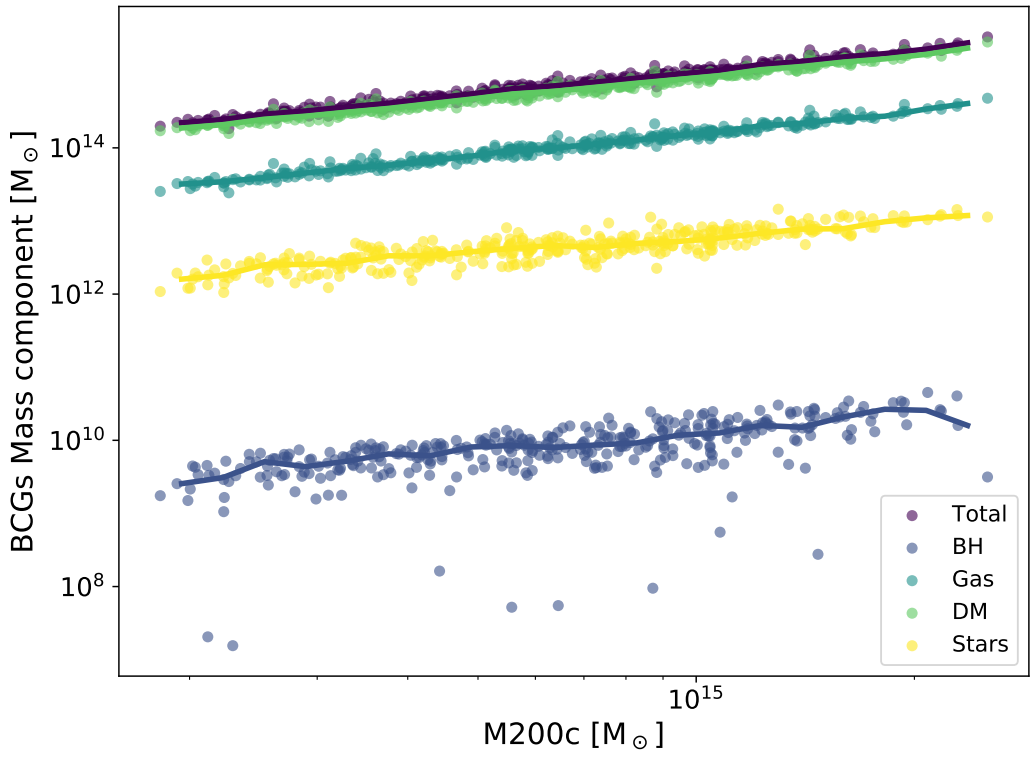

Figure 9: **BCG mass components vs. halo mass (M_{200c}) at z=0.** For each primary zoom in galaxy cluster in TNG-Cluster simulation at $z = 0$, we select the central galaxy and plot its bound total (purple), BH (dark blue), gas (light blue), dark matter (green), and stellar masses (yellow) against M_{200c}. Thick lines show running medians.

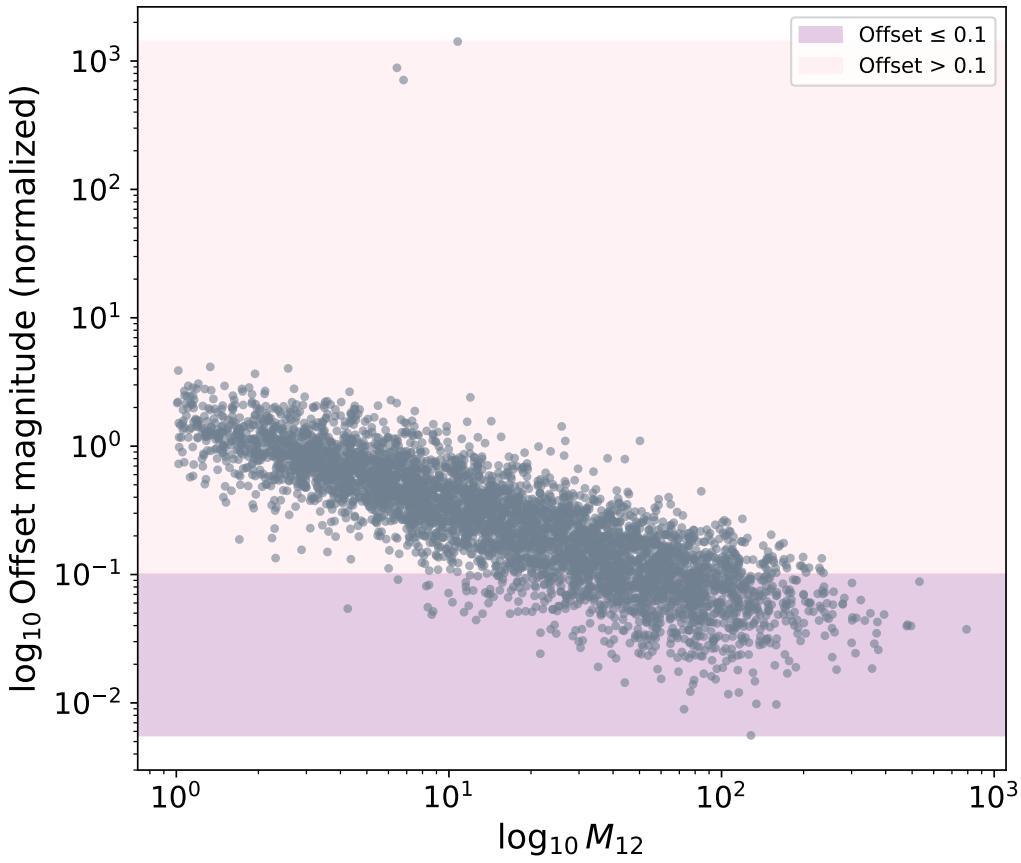

Figure 10: **Subhalo mass ratio versus offset magnitude.** Points show halos; M_{12} is the ratio of the most– to second–most–massive subhalo bound masses within the halo, and the offset magnitude is defined as in equation 9 following Ayromlou et al. [8]. Background shading indicates the relaxation cut at $x_{off} = 0.1$ (purple: relaxed; pink: non–relaxed). Systems with two comparably massive subhalos (low M_{12}) preferentially show large offsets, while halos with a dominant central (high M_{12}) concentrate at small offsets, with broad intrinsic scatter.

8.4 THERMODYNAMIC STATE OF THE ICM: COOL–CORE DIAGNOSTICS

To classify the thermodynamic state of clusters we adopt six standard cool–core (CC) diagnostics used in Lehle et al. [93]. Using this we can distinguish *strong* cool cores (SCC), *weak* cool cores (WCC), and *non*–cool–cores (NCC) using widely employed thresholds as will be discussed further.

To classify cluster thermodynamic states we adopt the six cool–core (CC) diagnostics and thresholds defined by Lehle et al. [93]. Central quantities (t_{cool}, K_0, n_e) are evaluated in 3D within a spherical aperture of radius $0.012\,R_{500c}$ about the gravitational potential minimum; the cuspiness uses the slope at $r = 0.04\,R_{500c}$; the X-ray concentrations use 0.5–5 keV luminosities in projected apertures of (40, 400) kpc (physical) or $(0.15\,R_{500c},\, R_{500c})$ (scaled). Following Lehle et al. [93], the gas selection includes only gravitationally bound, non–star-forming, actively cooling cells with $T > 10^6$ K. We briefly summarize the criteria used by Lehle et al. [93] to characterize cluster cool–core states.

(I) CENTRAL ENTROPY, K_0. The specific entropy proxy is defined as

$$K(r) = k_B\, T(r)\, n_e(r)^{-2/3} \quad [\mathrm{keV\,cm^2}],$$

with the central value $K_0 \equiv K(r_0)$ measured at $r_0 = 10$ kpc. Clusters are classified as SCC if $K_0 \leqslant 22$ keV cm², WCC if $22 < K_0 \leqslant 150$ keV cm², and NCC if $K_0 > 150$ keV cm².

(II) CENTRAL COOLING TIME, $t_{cool,o}$. The isobaric cooling time is defined as

$$t_{cool}(r) = \frac{3}{2}\,\frac{(n_e + n_i)\, k_B\, T}{n_e\, n_H\, \Lambda(T, Z)},$$

evaluated at r_0. The classification is SCC for $t_{cool,o} < 1$ Gyr, WCC for $1 \leqslant t_{cool,o} \leqslant 7.7$ Gyr, and NCC for $t_{cool,o} > 7.7$ Gyr.

(III) CENTRAL ELECTRON NUMBER DENSITY, n_e. The electron number density at r_0 is denoted $n_e = n_e(r_0)$ ($\mathrm{cm^{-3}}$). Clusters are classified as NCC if $n_e \leqslant 5.1 \times 10^{-3}$ $\mathrm{cm^{-3}}$, WCC if $5.1 \times 10^{-3} < n_{e,0} \leqslant 1.51 \times 10^{-2}$ $\mathrm{cm^{-3}}$, and SCC if $n_{e,0} > 1.51 \times 10^{-2}$ $\mathrm{cm^{-3}}$.

(IV) CUSPINESS OF THE DENSITY PROFILE, α_n. The cuspiness parameter is defined as

$$\alpha \equiv -\left.\frac{d \ln n_e}{d \ln r}\right|_{r = 0.04\,R_{500c}}.$$

The adopted classification is NCC for $\alpha_n \leqslant 0.5$, WCC for $0.5 < \alpha_n \leqslant 0.75$, and SCC for $\alpha_n > 0.75$.

(V) X-RAY CONCENTRATION (PHYSICAL APERTURES), c_{SB}^{phys}. The concentration parameter is defined as

$$c_{phys} = \frac{L_X^{0.5-5\,keV}(r_p < 40\ \mathrm{kpc})}{L_X^{0.5-5\,keV}(r_p < 400\ \mathrm{kpc})}$$

Clusters are classified as NCC if $c_{phys} \leqslant 0.075$, WCC if $0.075 < c_{phys} \leqslant 0.155$, and SCC if $c_{phys} > 0.155$.

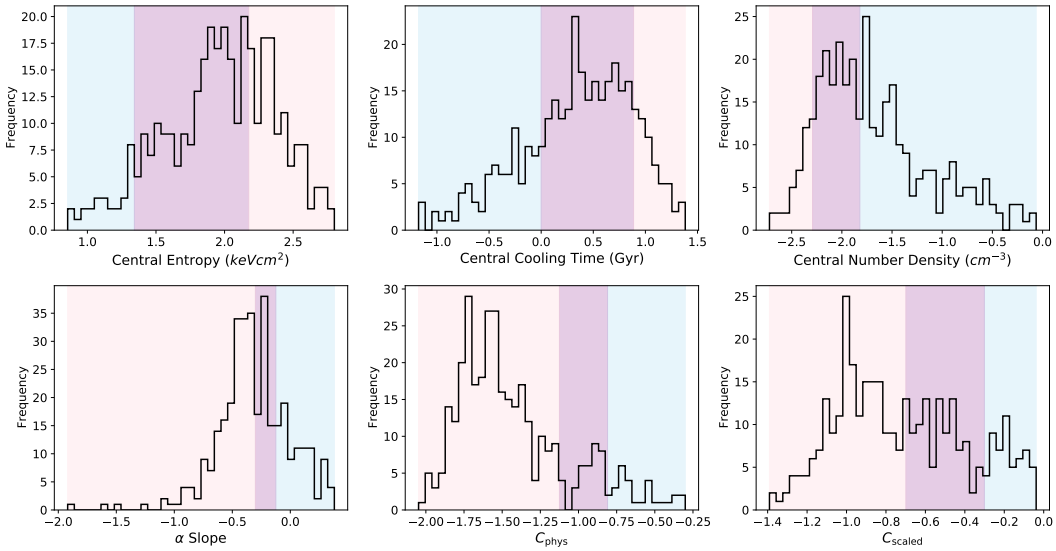

Figure 11: **Cool–core diagnostics Distribution.** Distributions of the six CC indicators with SCC (sky blue), WCC (purple), and NCC (pink) regions shaded using the thresholds listed in section 8.4 for the 352 primary zoom-in halos at $z = 0$. The diagnostics based on profiles (K_0, t_{cool}, n_e, α) and those based on imaging (c_{phys}, c_{scaled}) give a consistent partition of the sample into SCC/WCC/NCC, with differences reflecting sensitivity to core size and projection.

(VI) X-RAY CONCENTRATION (SCALED APERTURES), c_{SB}^{sca}. A redshift- and size-independent measure is given by

$$c_{scaled} = \frac{L_X^{0.5-5\,keV}(r_p < 0.15 r_{500c})}{L_X^{0.5-5\,keV}(r_p < r_{500c})}$$

Clusters are classified as NCC if $c_{scaled} \leqslant 0.2$, WCC if $0.2 < c_{scaled} \leqslant 0.5$, and SCC if $c_{scaled} > 0.5$.

The first four diagnostics rely on 3D thermodynamic profiles (T, n_e) and are closely related (low $K_0 \Leftrightarrow$ short $t_{cool} \Leftrightarrow$ high n_e and steep α_n). The latter two are purely imaging–based and robust to modest uncertainties in spectral modeling.

In figure 11, for all of the primary zoom-in halos of TNG-Cluster simulation at $z = 0$, we can see the histograms for each diagnostic with the SCC/WCC/NCC regions shaded in the background in blue/purple/pink.

8.5 MERGER IDENTIFICATION AND MEASUREMENT

In this work, for defining merger parameters, we use the data made by Lee et al. [91]. The mergers are recorded for the 352 main halos with $0 \leqslant z \leqslant 1$. Throughout, the main cluster is the most massive FoF host in the high-resolution region, and a subcluster (collider) is any subhalo that has undergone a *first pericenter passage* with respect to the main cluster. The mergers are recorded when the main halo has a $M_{500c} \geqslant 10^{14} M_\odot$ and the subcluster has a halo mass $M_{500c} \geqslant 10^{13} M_\odot$. In general there are $\sim$ 2000 merger events, with the distribution of main and subcluster masses that can be seen in figure 12.

For determining the time of collision (first pericenter passage), and pericenter distance, we use the time evolution of their orbit ($D(t)$) as seen in figure 13. Since

the time between snapshots in the TNG-Cluster simulations are large, for getting the correct time of pericenter passage, a quadratic function is fit to the first local minimum [91]:

$$D(t) \simeq at^2 + bt + c\,,$$

where the pericenter passage is at its minimum at its $t_{peri} = -b/2a$. The pericenter distance follows as [91]:

$$D_{peri} = D(t_{peri}) = c - \frac{b^2}{4a}\,.$$

The M_{500c} of the main cluster is also taken at the closest snapshot to t_{peri}. However, sometimes the group ID is not defined at this snapshot, but these values will be handled as explained in section 10.1. The collision velocity is also calculated as the time derivative of halo separation $(2at + b)$ at the t_{peri}. [91]

As a subhalo falls into the host potential, tidal stripping and ram–pressure reduce its bound mass, making the instantaneous mass at pericenter a poor proxy for the impacting mass. To correctly identify the subcluster mass that triggered the merger, the time where the subcluster reached its pick mass before the pericenter passage is chosen, as can be seen from figure 13. At this time, both the subcluster's mass, and the merger mass ratio $\mu = M_{sub,peak}/M_{main}$ is calculated [91].

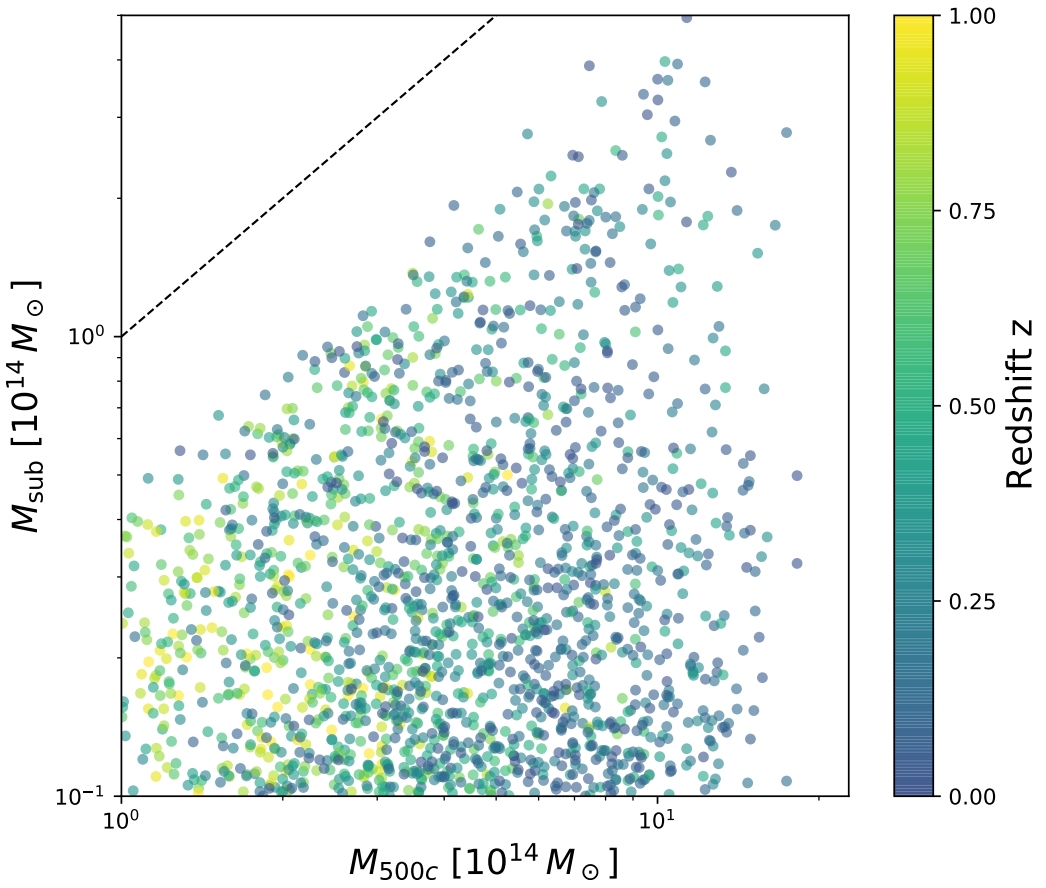

Figure 12: **Event demographics.** Host mass at collision, M_{500c}, versus the subcluster's pre-pericenter peak mass M_{sub} for all recorded mergers in the Lee et al. [91] catalog. Each point illustrates one merger with its color pointing to the redshift at which the collision has happened.

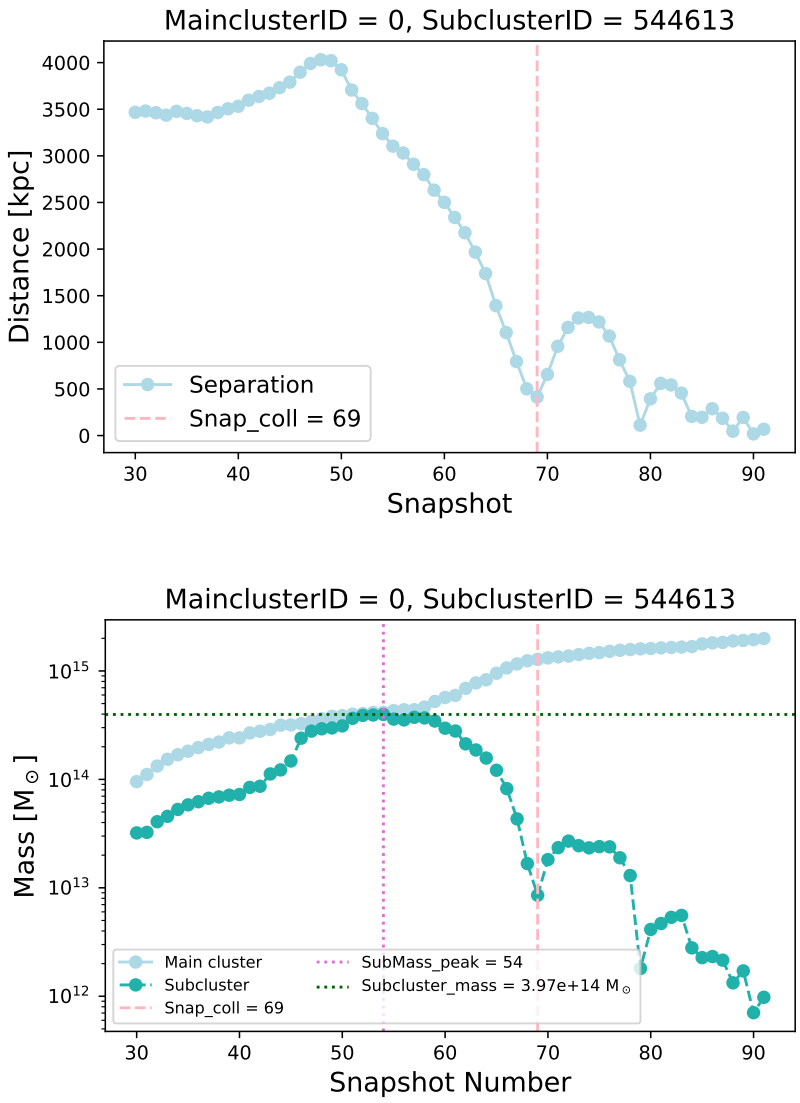

Figure 13: **Merger measurements.** *Top:* separation $D(t)$ for the main cluster and subcluster with the first curve marking the first pericenter where the quadratic function is fitted. The vertical pink dashed line marks the closest snapshot to t_{peri}, yielding sub-snapshot t_{peri} and D_{peri}. *Bottom:* bound masses of the main and sub-cluster versus TNG-Cluster snapshots; the collider's pre-pericenter maximum $M_{sub,peak}$ (purple dotted line) defines the impacting mass and sets the reference snapshot for the mass ratio (pink dashed line again marks t_{peri}).

Part IV

INFERRING THE MERGER HISTORIES OF GALAXY CLUSTERS FROM OBSERVABLES VIA CONDITIONAL INVERTIBLE NEURAL NETWORK

9

GALAXY CLUSTER PROPERTIES

9.1 OBSERVABLES

In this section, we describe the input observable properties extracted from the TNG-Cluster simulation. These properties can be grouped into four categories:

1- Halo-Scale Cluster Properties

All halo-scale observables are measured on Friends-of-Friends (FoF) halos, identified using the standard algorithm with a linking length of $b = 0.2$ applied to the dark matter particle distribution (see section 8). In table 4, these will be the properties: R_{500c}, M_{500c}, Gas Mass, Gas Metallicity, H Fraction, He Fraction, Stellar Metallicity, and Velocity.

2- Brightest Cluster Galaxy (BCG) Properties

The BCG properties are measured on the central Subfind subhalo within each FoF halo. By central we mean the subhalo containing the particle of minimum gravitational potential—it is not necessarily the most massive subhalo in the Fof halo (as explained in 8). In table 4, these will be the properties: BCG Total Mass, BCG Stellar Mass, BCG SFR, Central BH Mass, and Central BH Accretion Rate.

3- Intracluster Medium (ICM) Core Properties

We use the ICM properties derived by Lehle et al. [93], which were employed to characterize the cool-core state. Each property is measured on non–star-forming gas with $T > 10^6$ K and gravitationally bound to the central halo. In table 5, this will be the properties: Central Number Density, Central Cooling Time, Central Entropy, α Slope, C_{phys}, and C_{scaled}.

4- Global and Dynamical Diagnostics

In addition to structural and thermodynamic quantities, we include three global and dynamical diagnostics that capture the evolutionary state and relaxedness of each halo. In table 5, the properties are Cosmic Time, Offset Magnitude, and M_{12}.

Property	Definition
R_{500c}	Galaxy cluster's radius; centered on the halo's potential-minimum position, within which the mean density equals $500\,\rho_{crit}(z)$.
M_{500c}	Galaxy cluster's mass; total mass of all particles and cells enclosed within a radius, such that the average density inside this radius is $500 \times \rho_{crit}(z)$.
Gas Mass	Sum of the masses of all gas-type cells associated with the FoF halo (wind-phase cells included).
Gas Metallicity	Mass-weighted average metallicity of all gas cells in the FoF halo, defined as the ratio of metal mass (elements heavier than helium) to total gas mass.
H Fraction	Mass fraction of hydrogen, defined as the total mass of hydrogen divided by the gas mass.
He Fraction	Mass fraction of helium, defined as the total mass of helium divided by the gas mass.
Stellar Metallicity	Mass-weighted average metallicity of all star particles in the FoF halo, defined as the ratio of metal mass to total stellar mass.
Velocity	The peculiar velocity of the galaxy cluster, computed as the sum of the mass-weighted velocities of all particles and cells belonging to the FoF halo.
BCG Total Mass	Total mass of all particles and cells bound to the BCG (dark matter, gas—including wind phase—, stars and black holes).
BCG Stellar Mass	Total mass of all of the stars bound to the BCG.
BCG SFR	Instantaneous star-formation rate of the BCG, computed as the sum of the individual SFRs of all gas cells within the BCG.
Central BH Mass	Mass of the central black hole associated with the BCG; excludes any surrounding gas reservoir and evolves monotonically via the simulation accretion prescription.
Central BH Accretion Rate	Instantaneous mass accretion rate onto the BCG's central black hole ($\dot{M}_{BH}$).

Table 4: Halo-scale and brightest cluster galaxy (BCG) observables extracted from the TNG-Cluster simulation. These properties trace the global structure of the halo as well as the stellar, star-forming, and black hole content of the central galaxy.

Property	Definition
Central Number Density	Central electron number density within a 3D aperture of $0.012\,r_{500c}$, computed directly as the mass-weighted mean of n_e.
Central Cooling Time	Cooling cooling time of the ICM core, $$t_{cool} = \frac{3}{2}\frac{(n_e + n_i)\,k_B T}{n_e\,n_i\,\Lambda(T,Z)}$$ evaluated as the mass-weighted mean within a 3D aperture of $0.012\,r_{500c}$.
Central Entropy	Central entropy defined by $K = k_B T n_e^{-2/3}$ measured in the same $0.012\,r_{500c}$ aperture.
α Slope	Logarithmic slope of the radial electron-density profile, $$\alpha = -\left.\frac{d\ln n_e(r)}{d\ln r}\right|_{r=0.04\,r_{500c}}$$ evaluated from a 3D profile with 50 logarithmic bins between 10^{-3} and $1.5\,r_{500c}$.
C_{phys}	Physical concentration defined as the ratio of X-ray luminosities in projected apertures: $$C_{phys} = \frac{L_X(r_p < 40\,kpc)}{L_X(r_p < 400\,kpc)}$$ computed from 2D X-ray maps in the 0.5–5 keV band.
C_{scaled}	Scaled concentration defined as the ratio of X-ray luminosities in scaled apertures: $$C_{scaled} = \frac{L_X(r_p < 0.15\,r_{500c})}{L_X(r_p < r_{500c})}$$ computed from the same 0.5–5 keV 2D maps.
Cosmic Time	Cosmic time at which the cluster is considered. We use cosmic time rather than redshift since it provides a linear measure of time and corresponds more directly to the nearly uniform time spacing of the TNG snapshots.
COM Offset	Distance between the most-bound particle (i.e. potential minimum) and the galaxy cluster's center of mass, normalized by the R_{200c}. This dimensionless offset traces the dynamical relaxation state of the galaxy cluster [8].
M_{12}	Ratio of the M_{500c} of the cluster's most massive to the second most massive galaxy within the same halo.

Table 5: ICM core and global dynamical observables. The ICM quantities follow the definitions of Lehle et al. [93], while the dynamical diagnostics capture the evolutionary and relaxed-ness state of each halo.

9.2 UNOBSERVABLES

We use the merger-event catalog of Lee et al. [91], in which cluster mergers are defined as interactions between a primary halo with mass $M_{500c} > 10^{14}\,M_\odot$ and a secondary subcluster with mass $M_{sub} > 10^{13}\,M_\odot$. Unlike the observable properties described in section 9.1, these quantities are inherently unobservable in real data: they describe the intrinsic dynamical parameters of cluster mergers, which are accessible only in simulations through knowledge of the full three-dimensional phase-space trajectories of halos. In particular, the catalog reports the following merger parameters at the time of first pericenter passage.

Property	**Definition**
Collision Time	Cosmic time (in Gyr) of the first pericenter passage between the main cluster and subcluster.
Pericenter Distance	Three-dimensional distance (in kpc) between the main- and subcluster centers at first pericenter.
Collision Velocity	Maximum relative peculiar velocity (in km/s) between the two halos at pericenter, derived from the difference in their SubhaloVel vectors.
Main Cluster M_{500c}	M_{500c} of the primary cluster at the time of first pericenter, i.e. the FoF halo mass within R_{500c}.
Subcluster Mass	Maximum bound mass (total particles and cells) of the subcluster, taken at the snapshot when its pre-pericenter mass reaches its peak value.
Merger Mass Ratio	Ratio of subcluster to main cluster mass at the snapshot where the subcluster mass peaks, $\mu = M_{sub}/M_{500c,coll}$.

Table 6: Merger-event parameters as defined in the catalog of Lee et al. [91]. These "unobservable" quantities are accessible in simulations but cannot be directly inferred in observations.

9.3 FINAL SAMPLE

Our working data set comprises the 352 *TNG-Cluster* primary zoom-in halos and their *main–progenitor* branches sampled at eight outputs spanning $0 \leqslant z \leqslant 1$ (snapshots $\{99, 91, 84, 78, 72, 67, 59, 50\}$). For each halo and snapshot we assemble:

- **Observables:** quantities in table 4 and 5 measured at that snapshot.
- **Merger parameters (unobservables):** for each merger event recorded for each halo in Lee et al. [91] we have different t_{coll} for all merger events it goes through. For each of the target halo, two immediate events at t_{snap} are chosen: the *last* merger with $t_{coll} \leqslant t_{snap}$ (the most recent past event) and the *next* merger with $t_{coll} > t_{snap}$ (the nearest future event). Thus, per snapshot there is at most one "last" and one "next" event, chosen as the closest in time on each side (no averaging across multiple events). When present, we record that event's merging parameters (table 9.2).

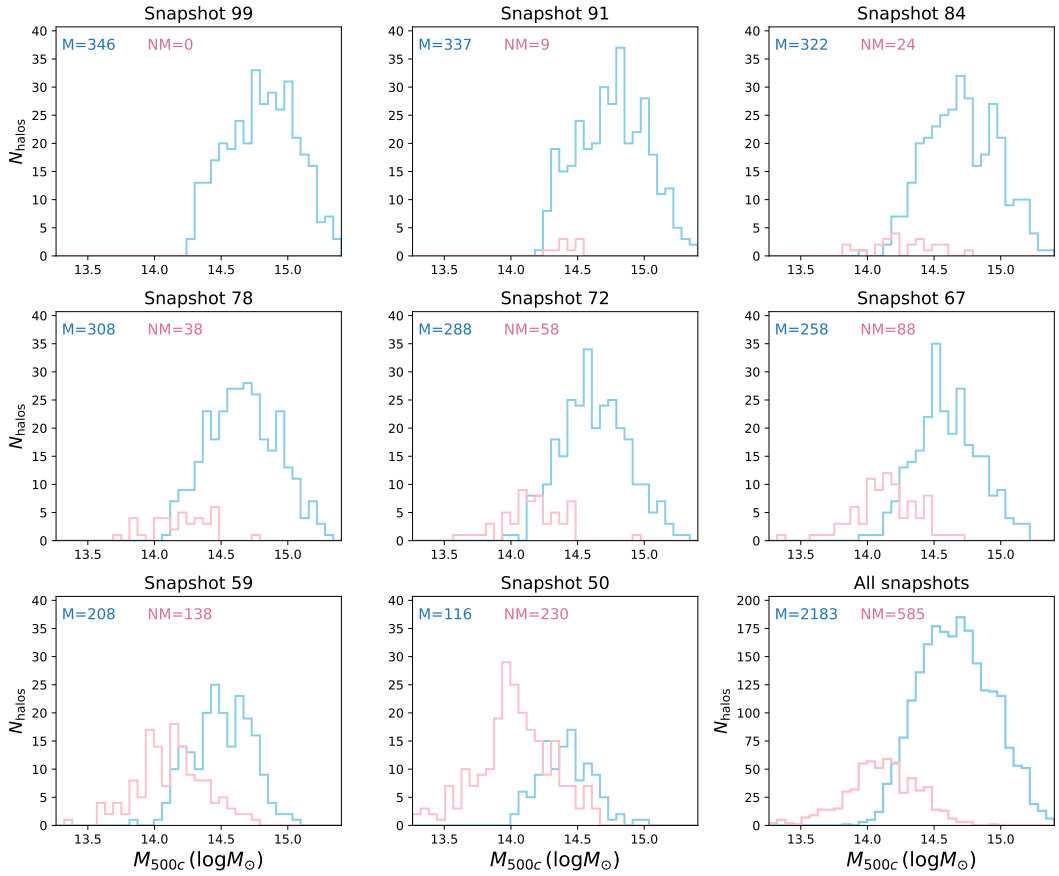

Figure 14: **Merging vs. non–merging halos across snapshots.** Step–histograms of halo mass at eight outputs (snapshots 99 to 50) split into systems whose last recorded merger occurred before that snapshot (MM; sky blue) and those without a prior merger in our window (NM; pink). Bins and x–limits are shared across snapshots; the bottom–right panel aggregates all outputs. Counts for MM and NM are annotated in each axis.

Figure 14 summarizes the mass demographics of merging (M; blue) and non–merging (NM; pink) halos at each snapshot. Each panel shows step–histograms of $\log_{10} M_{500c}$ with a common x–range and binning across snapshots to enable a fair, visual comparison; the counts for M and NM are printed in the upper left of each panel. The ninth panel stacks *all* snapshots. Two qualitative trends are evident;. First, the NM fraction increases toward earlier times (lower snapshot numbers / higher redshift), as expected because a smaller lookback window within $z \leqslant 1$ leaves fewer halos with a recorded *prior* merger. Second, at fixed snapshot the merger–flagged population increasingly dominates the high–mass tail, consistent with hierarchical growth in which the most massive systems have rich recent accretion histories.

10 SELECTION OF THE INPUTS

10.1 PREPROCESSING

Our learning pipeline (used identically for the baseline multilayer perceptron and for the conditional invertible neural network) requires a clean, aligned mapping from observable cluster properties to merger or unobservables properties (section 9.1-9.2). We therefore implement a deterministic preprocessing stage comprising: (i) sample alignment across catalogs, (ii) principled filtering of ill-defined targets and inputs, and (iii) variance–stabilizing transformations and standardization.

We combine the table of observable properties with the table of merger parameters. The merge is an inner join on a unique halo ID and snapshot, ensuring that each row corresponds to the same galaxy clusters observed at the same cosmic time in both spaces. Let $\mathbf{x} \in \mathbb{R}^p$ denote the vector of observables and $\mathbf{y} \in \mathbb{R}^q$ the vector of merger parameters; alignment produces paired samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$.

Not every cluster experiences a merger within the time window and mass ranges covered by the merger catalog; for such systems, unobservable properties as defined in section 9.2 do not exist. Including these rows would mix a well-posed conditional regression problem, $p(\mathbf{y} \mid \mathbf{x})$ with $\mathbf{y}$ defined at first pericenter, with no-event cases requiring a different statistical treatment. To preserve a single, physically coherent target definition, we discard rows lacking any merger parameter. This restriction yields a dataset of clusters with at least one resolved first pericenter passage, i.e., systems for which $\mathbf{y}$ is well-defined.

To stabilize scale and reduce skewness, we work in log space for all strictly positive target merger parameters: Main Cluster M_{500c}, Subcluster mass, Collision velocity, and Pericenter Distance. Concretely, we take the log of all the targets except the Collision Time and Merger Mass Ratio. An additional practical advantage is robustness to undefined values: in rare cases the merger catalog from Lee et al. [91], the Main Cluster M_{500c}, has a value of -1 because the group was not defined at the merger snapshot as explained in section 8.5. Such entries are undefined in log space and therefore become missing; our filtering will also drop these rows since they have missing values in log space. In effect, no unphysical negative values can leak into the training set, while valid positive values are consistently modeled in log space.

Several observables characterize the thermodynamic state of the ICM core (e.g., central electron density, entropy, cooling time, concentration measures) as discussed in section 9.1. These are defined on non–star-forming gas above a temperature threshold and within fixed or scaled apertures. If a system contains negligible hot gas within the relevant aperture (e.g., due to extreme feedback episodes or numerical sparsity), these quantities are undefined and missing. To preserve dimensional consistency, we therefor discard any rows missing any observable properties. This yields a feature space $\mathbf{x}$ that corresponds to well-measured physical quantities.

After filtering, we partition the dataset randomly into training, validation, and test subsets in an 80/10/10 split. For last mergers, this results in 1627/203/203 samples; for next mergers, 1548/194/193. These same indices will be saved and reused by both the MLP and the cINN to guarantee like-for-like comparisons. After partitioning is complete, we then apply per-dimension z-score standardization to both inputs and targets:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}_x}{\boldsymbol{\sigma}_x}, \qquad \tilde{\mathbf{y}} = \frac{\mathbf{y} - \boldsymbol{\mu}_y}{\boldsymbol{\sigma}_y}, \tag{10}$$

where $(\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x)$ and $(\boldsymbol{\mu}_y, \boldsymbol{\sigma}_y)$ are the empirical means and standard deviations on the training set. This rescaling ensures that all inputs and targets contribute on comparable numerical scales. Without such normalization, parameters with large absolute values or wide dynamic ranges could dominate the optimization, simply because of their units rather than their physical relevance. Standardization therefore improves numerical conditioning, balances the relative influence of different features, and supports more stable and efficient training for both the MLP and the cINN.

We archive (i) the standardized design matrix and target matrix, (ii) the identifiers linking each sample back to its halo and snapshot, and (iii) the scaling parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ required for inverse transformations and unit recovery. For model comparison and unbiased evaluation, scaling parameters are fitted on the training split only and then applied to validation and test data, preventing information leakage.

The resulting standardized pair $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ is used consistently by both stages of our pipeline: an initial MLP for feature ranking/selection and the subsequent cINN for posterior inference.

10.2 BASELINE MLP AND ENSEMBLE TRAINING

As a supervised baseline and for downstream feature screening, we train a multilayer perceptron (MLP) to predict the merger parameters from the standardized observable vector $\tilde{\mathbf{x}} \in \mathbb{R}^p$ (section 9.1). The model outputs $\hat{\mathbf{y}} \in \mathbb{R}^q$ in standardized target space; physical units are recovered via the inverse of the target standardization.

A Multilayer Perceptron (MLP) is a feed-forward neural network that defines a parametric mapping $f_\theta : \mathbb{R}^p \to \mathbb{R}^q$ from a p-dimensional input vector x (the observables) to a q-dimensional output vector $\hat{y}$ (the merger properties). The network consists of L layers of linear transformations and nonlinear activation functions, interleaved with batch-normalization blocks to stabilize training [56]:

$$\begin{aligned}
\mathbf{h}^{(0)} &= \mathbf{x}, \\
\mathbf{z}^{(\ell)} &= W^{(\ell)}\,\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}, \quad \tilde{\mathbf{z}}^{(\ell)} = \mathrm{BatchNorm}\big(\mathbf{z}^{(\ell)}\big), \\
\mathbf{h}^{(\ell)} &= \phi\big(\tilde{\mathbf{z}}^{(\ell)}\big), \quad \ell = 1, \ldots, L-1, \\
\hat{\mathbf{y}} = \mathbf{h}^{(L)} &= W^{(L)}\,\mathbf{h}^{(L-1)} + \mathbf{b}^{(L)},
\end{aligned}$$

where:

- $W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and $\mathbf{b}^{(\ell)} \in \mathbb{R}^{d_\ell}$ are the weight matrix and bias vector of layer ℓ, with $d_0 = p$, $d_L = q$, and $d_1 = \cdots = d_{L-1} = 256$ in our architecture.
- $\phi(\cdot) = \max(0, \cdot)$ is the Rectified Linear Unit (ReLU) activation.
- BatchNorm$(\cdot)$ denotes a learnable affine batch-normalization transform applied before each activation.
- $\theta = \{W^{(\ell)}, \mathbf{b}^{(\ell)}, \text{BatchNorm parameters}\}_{\ell=1}^{L}$ collects all trainable parameters.

In our implementation, the MLP consists of:

- an *input layer* of dimension $d_0 = 22$ (one node per observable feature),
- *three hidden layers* ($L - 1 = 3$) of equal width $d_1 = d_2 = d_3 = 256$, each block comprising:
  1. a bias-free fully connected transform $\mathbf{z}^{(\ell)} = W^{(\ell)}\mathbf{h}^{(\ell-1)}$,
  2. batch normalization on the 256 outputs,
  3. a Rectified Linear Unit activation (ReLU),

$$\mathbf{h}^{(\ell)} = \max\{0, \text{BatchNorm}(\mathbf{z}^{(\ell)})\}$$

- a *final output layer* of dimension $d_4 = 6$, implemented as a fully connected transform without activation, producing the six continuous merger predictions.

The training and optimization in this pipeline is implemented via PyTorch framework [121]. The MLP is trained via stochastic gradient descent with the Adam optimizer [79], weight decay 10^{-4}, batch size 256, employing a two-phase loss schedule and early stopping held out on a validation set. All training is conducted on the standardized feature and target matrices described in chapter 10.1. We now detail each component of the training pipeline:

1. **Phase I – MSE Warm-up:** We minimize the mean squared error (MSE) loss,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2,$$

   using an initial learning rate of $\eta = 10^{-3}$ for up to 100 epochs. The MSE objective heavily penalizes large residuals, guiding the weights into a parameter region that captures the dominant variance of each merger-history target.

2. **Phase II – MAE Refinement:** To mitigate the influence of outliers and ambiguous training samples, we switch to the mean absolute error (MAE) loss,

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_1,$$

   and reduce the learning rate to $\eta = 5 \times 10^{-4}$ for an additional 50 epochs. The MAE objective evenly weights all residuals, refining the model toward median-optimal predictions.

Within each phase, we monitor the loss on the held-out validation set at the end of every epoch. If the validation loss fails to improve for 20 consecutive epochs, training is stopped and the model reverts to the parameter state that achieved the lowest validation loss. Early stopping prevents overfitting, especially important during the MAE phase when the loss landscape is less steep—and caps unnecessary computation once meaningful gains cease.

To reduce sensitivity to the random initialization of weights and to improve predictive stability, we train an ensemble of seven MLP replicas, each initialized with a distinct random seed. Each replica undergoes the full two-phase training with independent early stopping. At inference time, predictions from the ensemble are aggregated via the element-wise median, which suppresses outlier model responses while preserving the central tendency of the learned mapping.

After ensemble aggregation, standardized predictions are inverse-transformed to physical units using the saved target scaler. We then compute the mean absolute error (MAE) of each of the merger-history outputs on the held-out test set. These MAE values serve as our baseline metrics, directly comparable to the cINN's MAP-based point estimates and informing the relative advantage of the probabilistic model.

10.3 SENSITIVITY ANALYSIS

Before training our conditional invertible neural network (cINN), we aim to identify and retain only the most informative subset of observables. Reducing the input dimensionality yields several scientific and practical benefits: it reduces the risk of overfitting, enhances interpretability of the learned mappings, and decreases computational cost.

To identify which observables are most informative for predicting the merger parameters, we quantify feature importance by permutation sensitivity on the held-out test set. The idea is to measure the degradation in predictive accuracy when the association between a single observable and the targets is destroyed, while preserving the marginal distribution of that observable and all correlations among the remaining features. For this reason we perform:

1. Train the MLP on the full set of standardized observables (as explained in section 9.1) to predict the six standardized merger-history targets.
2. Evaluate its baseline performance by computing the mean absolute error (MAE) on the held-out test set.
3. For each input feature, randomly permute its values across all test samples, thereby breaking any learned association, while leaving other features unchanged.
4. Recompute the MAE for each target on the permuted test set and record the increase in error relative to the baseline.
5. Rank the observables by the magnitude of their induced MAE increase, selecting the top k features that contribute most critically to accurate inference.

Let $\mathcal{T}$ denote the fixed test index set and $\hat{\mathbf{y}}(\mathbf{x})$ the ensemble-median MLP prediction in standardized target space (Section 10.2). We first compute a baseline error in *physical units*,

$$\mathbf{b} \in \mathbb{R}^q, \qquad b_k = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \left| y_{ik}^{\text{phys}} - \hat{y}_{ik}^{\text{phys}} \right|,$$

where the inverse of the target standardization is applied to both predictions and ground truth. Using physical units makes each component b_k directly interpretable for its target (e.g., Gyr for Collision Time, kpc for Pericenter Distance, etc.).

For each observable (feature) $j \in \{1, \ldots, p\}$ we form a perturbed test design $\tilde{\mathbf{X}}^{(j)}$ by applying an independent random permutation to the j-th column across test rows,

$$\tilde{X}_{ik}^{(j)} = \begin{cases} X_{\pi_j(i)\,j}, & k = j, \\ X_{ik}, & k \neq j, \end{cases} \qquad i \in \mathcal{T},$$

This operation preserves the distribution of feature j in the test set but breaks its joint dependence on the targets given the other features; all other columns remain unchanged. We then evaluate the ensemble-median prediction on $\tilde{\mathbf{X}}^{(j)}$, invert to physical units, and recompute the MAE vector,

$$m_k^{(j)} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \left| y_{ik}^{\text{phys}} - \hat{y}_{ik}^{\text{phys}}\big(\tilde{\mathbf{X}}_i^{(j)}\big) \right|.$$

The *sensitivity matrix* $S \in \mathbb{R}^{p \times q}$ is defined component-wise as the MAE increase,

$$S_{jk} = m_k^{(j)} - b_k = \Delta\text{MAE}_{j,k} \tag{11}$$

Large positive S_{jk} indicates that scrambling observable j substantially harms accuracy for target k, hence j is important for predicting k. Values near zero suggest little marginal contribution. Small negative entries can occur due to finite-sample noise or variance reduction from the permutation and are not truncated in our analysis.

Because each target is evaluated in its own physical units, raw sensitivity increments S_{jk} are not comparable across columns. We therefore convert each target column k into ranks by ordering observables in descending ΔMAE:

$$r_{jk} = 1 + \left| \{ j' : S_{j'k} > S_{jk} \} \right|, \qquad j = 1, \ldots, p, \quad k = 1, \ldots, q.$$

This yields a rank matrix $R \in \mathbb{R}^{p \times q}$, where smaller values indicate greater importance for the corresponding target.

To summarize feature utility across all targets, we compute the *mean rank* for each observable,

$$\bar{r}_j = \frac{1}{q} \sum_{k=1}^{q} r_{jk}, \qquad j = 1, \ldots, p,$$

The result can be seen in Figure 15. Using this we can select the conditioning set for the cINN model by taking the M observables with the smallest $\bar{r}_j$.

The top 8 selected inputs are: Gas Mass, Stellar Metallicity, M_{500c}, Central Entropy, BCG Total Mass, Cosmic Time, and COM Offset.

To visualize how much each selected observable affect each inferred *target*, we extract the corresponding rows from the permutation–based sensitivity matrix (equation 11) for the eight selected observables; e.g., M_{500c}, gas mass, stellar metallicity, BCG mass, central BH mass, central entropy, cosmic time, and COM offset). In Figure 16, for each target column d we plot a heatmap of $\log_{10}(|\Delta MAE_{j,d}|)$, where $\Delta MAE_{j,d}$ is the increase in test MAE when observable j is independently shuffled (higher $\log(|\Delta MAE|)$ shows stronger importance). Since ΔMAE is measured per target separately, it cannot be compared across different targets, and hence, each target has its own color bar.

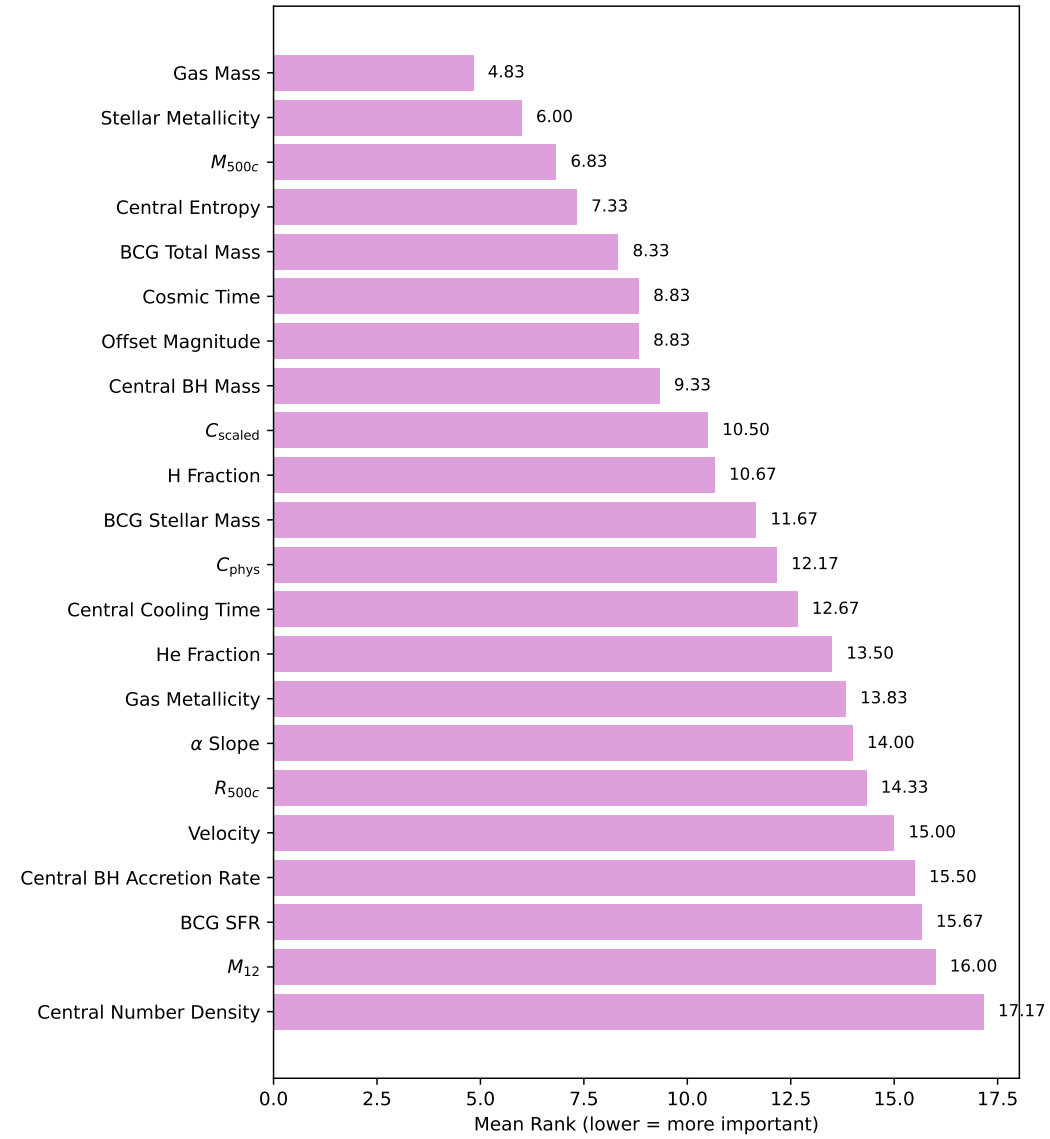

Figure 15: Mean rank of each observable, computed from the permutation sensitivity matrix by ranking features within each target and averaging across targets. The top eight features (smallest mean ranks) are selected as the conditioning set for the cINN.

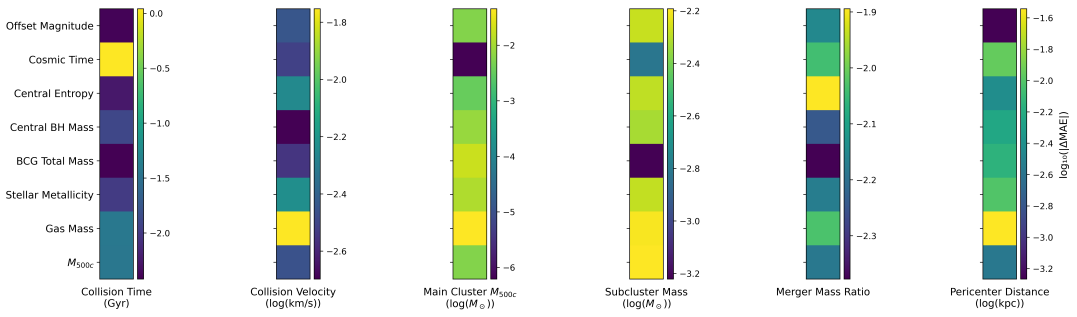

Figure 16: Permutation sensitivity for a selected set of observables (rows) across all targets (columns). Color encodes $\log_{10}(|\Delta\text{MAE}|)$ on the test split; larger values indicate a larger degradation in accuracy when that observable is destroyed, hence higher importance for that target.

11

CONDITIONAL INVERTIBLE NEURAL NETWORK

11.1 CINN MODEL ARCHITECTURE

The Conditional Invertible Neural Network (cINN) used in this thesis, is implemented using FrEIA (Framework for Easily Invertible Architectures) library [4, 5]. The library provides modular components for construcitng invertible architecture that enables a bijective mapping between target variable (x) and a latent space (embedding) z, conditioned on c. The main idea is to first learn a forward transformation that $f(x, c) = z$ where z lies in a latent space where standard probabilistic modeling (e.g., Gaussian likelihoods) can be applied.

This transformation is implemented via a sequence of normalizing flow blocks, which are invertible by design and have a tractable jacobian determinants. Learning such a mapping will allow us to get the conditional probability $p(x \mid c)$. The inverse transformation, $f^{-1}(z, c) = x$, generates target samples x that are conditioned on c. By doing this multiple times, the complete conditional posterior $p(x \mid c)$ is produced. rather than producing only a single point estimate. As a result, the model captures the uncertainty and possible multi-modality of the solution space.

To formalize this, let $X \in \mathbb{R}^{D_x}$ be the target variable and C the conditioning variable. The cINN defines an invertible, differentiable transformation:

$$Z = f(X \mid C), \quad X = f^{-1}(Z \mid C), \tag{12}$$

where Z is the latent variable in a space where a simple base density $p_Z(z)$ (e.g., a standard Gaussian) is assumed. For each fixed c, the mapping $x \mapsto z$ is bijective, and its local volume change is described by the Jacobian matrix

$$J_f(x \mid c) = \frac{\partial f(x \mid c)}{\partial x} \in \mathbb{R}^{D \times D}. \tag{13}$$

The relationship between the densities of X and Z follows from the multivariate change-of-variables theorem. For any measurable set $\mathcal{A} \subset \mathbb{R}^D$,

$$\int_{\mathcal{A}} p_{X|C}(x \mid c)\, dx = \int_{f(\mathcal{A} \mid c)} p_Z(z)\, dz. \tag{14}$$

An infinitesimal volume element transforms as

$$dz = |\det J_f(x \mid c)|\, dx, \tag{15}$$

which yields the pointwise density transformation:

$$p_{X|C}(x \mid c) = p_Z\big(f(x \mid c)\big)\ |\det J_f(x \mid c)|\,. \tag{16}$$

Equation (16) expresses the fact that probability mass is preserved under the mapping, up to a scaling given by the local volume change.

If we assume the base density to be standard Gaussian,

$$p_Z(z) = (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\|z\|^2\right), \tag{17}$$

then the conditional log-likelihood becomes

$$\log p_{X|C}(x \mid c) = \log p_Z\big(f(x \mid c)\big) + \log \left|\det J_f(x \mid c)\right| \tag{18}$$

$$= -\frac{1}{2}\|z\|^2 - \frac{D}{2}\log(2\pi) + \log \left|\det J_f(x \mid c)\right|. \tag{19}$$

Maximizing the likelihood is therefore equivalent to minimizing the negative log-likelihood (up to an additive constant):

$$\mathcal{L}(x, c) = \frac{1}{2}\|z\|^2 - \log \left|\det J_f(x \mid c)\right|, \quad z = f(x \mid c). \tag{20}$$

The first term encourages the latent variables to follow the chosen prior distribution, while the second term accounts for the volume change induced by the transformation f.

- **Input and Condition Nodes:** The cINN takes a target vector x as input and models its transformation through a series of invertible coupling blocks, while the condition vector c is injected at each block to condition the transformation.
- **Subnet Construction:** Each coupling block is parameterized by a small fully connected feedforward neural network, known as *subnet*, that computes the parameters of the transformation (e.g., shift and scale in affine couplings, or spline parameters in spline-based couplings).
- **Coupling Blocks:** CINNs are composed of a sequence of coupling blocks, each of which applies an invertible transformation to part of the input while leaving the rest unchanged. By alternating which part of the input is transformed across blocks, the network achieves full transformation of the input space. The number of blocks controls the expressivity and depth of the transformation.
- **Coupling Block Types:** Two main types of coupling transformations are commonly used: affine and rational quadratic spline (RQS). Affine couplings perform linear transformations and are easy to compute, while RQS couplings allow for more flexible and non-linear transformations, improving the model's expressiveness while preserving exact invertibility and tractable Jacobians.
- **Permutation Layers:** Permutation layers are inserted between the coupling blocks to rearrange the input dimensions. This prevents certain features from always being transformed or left unchanged, by ensuring that all coordinates are repeatedly updated.

The cINN model used in this thesis is instantiated with the following architectural and training choices, chosen after empirical experimentation:

- **Input and Condition Nodes:** In this work, the input x represents the scaled unobservable physical properties of galaxy clusters (as discussed in section 9.2), while the condition c can corresponds to either; 1- The observable properties of galaxy clusters (discussed in section 9.2) which will be further explored in this part (as in part iv) or 2- the image representations extracted from a SimCLR-trained encoder (as will be discussed in part v, vi, and vii).

Each block in the cINN receives this conditional input, allowing the model to learn how the distribution of physical parameters varies with changes in the representation space.

- **Subnet:** Each transformation function within the coupling blocks is parameterized by a fully connected feedforward MLP with 3 hidden layers, each containing 256 hidden units and ReLU activations. This architecture was selected to provide enough power to model complex transformations, while maintaining computational efficiency. Their depth and width were chosen through empirical tuning to balance model capacity and overfitting risk.

- **Coupling Blocks:** The model architecture comprises 8 sequential coupling blocks. Each block splits the input into two vectors and applies an invertible transformation to one half, while leaving the other half unchanged. This pattern alternates across layers to ensure that all dimensions are eventually transformed. The choice of eight blocks was selected based on practical experimentation to achieve a high capacity model without excessive depth.

- **Coupling Block Types:** We use **rational quadratic spline (RQS)** coupling transformations as introduced by Durkan et al. [44]. RQS transformations are highly expressive and support smooth, non-linear mappings, making them ideal for learning complex distributions. Unlike affine couplings, which are limited to linear transformations, RQS blocks capture variations in the conditional data distribution while retaining tractable likelihood computation and exact invertibility.

- **Permutation Layers:** Permutation operations are used between coupling blocks to improve mixing of variables across the network. In our configuration, fixed random permutations (i.e., hard permutations) are used, meaning a single randomly drawn permutation matrix is generated at initialization and applied consistently throughout training and inference. This ensures that all input dimensions participate in the transformation across the network's depth.

Within the FrEIA frame work, there are two primary input structures: an `InputNode`, which receives the target vector $x \in \mathbb{R}^{D_x}$, and a `ConditionNode`, which encodes the associated conditioning variable $c \in \mathbb{R}^{D_c}$ (observable properties or learned representation) [4]. These nodes are not just passive data containers and actually define the structure of the computation graph. The input variable is linked sequentially to a series of invertible transformations (e.g., the coupling blocks), while the conditioning variable is passed into every coupling block, ensuring that the learned mapping is explicitly conditioned on c.

The model stores the input dimensions (D_c and D_x), which will later be required for coupling blocks later in the architecture. Invertible architectures based on coupling transformations require input variables to lie within bounded and numerically stable domains. Since our target properties are already scaled (in section 10.1), this step is not necessary here and is skipped.

The model is made of a chain of multiple coupling blocks, where each of them is responsible for implementing an invertible transformation on the input vector. The model's objective is to learn an invertible function f_θ, parameterized by neural

network weights θ, that maps the input variable to a latent representation (embedding) $\mathbf{z} = f_\theta(\mathbf{x}|\mathbf{c})$. Once f_θ is learned, the inverse mapping $x = f_\theta^{-1}(c, z)$ can be used to generate plausible samples of x conditioned on c.

Each coupling block transforms split the target vector into two subsets: x_a, which remains unchanged in that block and x_b which is transformed. The transformation of x_b is parametrized by a small feedforward neural network, or subnet, which takes the unchanged portion x_a concatenated with the condition vector c. The subnet then outputs the transformation parameters, such as scaling factors and shifts in affine transitions, or bin width, heights and derivatives in RQS coupling that are applied to x_b. It is important to note that the transformations are constructed such that their inverse and Jacobian determinant can be computed analytically. This property allows the model to computes the loss as expressed in the equation 20.

Each subnet is constructed as a fully connected multi-layer perceptron (MLP) composed of L layers, where $L \geqslant 1$, with activation functions in between. For this thesis, the subnets are composed of three linear layers followed by non-linear ReLU activations. Its architecture can be summarized as follows:

$$\text{Linear}(D_{\text{in}}, D_{\text{hidden}}) \to \text{ReLU} \to \text{Linear}(D_{\text{hidden}}, D_{\text{hidden}}) \to \text{ReLU} \to \text{Linear}(D_{\text{hidden}}, D_{\text{out}})$$

where $D_{\text{in}} = \dim(\mathbf{x}_a) + \dim(\mathbf{c})$ is the total number of input features, D_{hidden} is the width of the hidden layers, and D_{out} is the number of transformation parameters required for $\mathbf{x}_b$. To improve stability in the early stages of training, the weights of the final layer are initialized to zero to ensure that the initial transformation is close to the identity mapping.

In this part, the conditional input $\mathbf{c}$ consists of a 8-dimensional representation space because fo the selected 8 observable inputs in section 9.1, while the input feature vector $\mathbf{x} \in \mathbb{R}^6$ is split evenly by the coupling block such that $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^3$. As a result, the subnet receives $D_{\text{in}} = 3 + 8 = 11$ input features.

The output dimensionality D_{out} depends on the type of transformation used in the coupling block. In the case of rational quadratic spline (RQS) transformations, each transformed input dimension requires K bin widths, K bin heights, and $K - 1$ derivatives to define the piecewise monotonic spline, and as a result $D_{\text{out}} = \dim(\mathbf{x}_b) \cdot (3K - 1)$. For our configuration with $K = 10$ spline bins and $\dim(\mathbf{x}_b) = 3$, we obtain $D_{\text{out}} = 3 \cdot (3 \cdot 10 - 1) = 87$. The hidden layer width of the subnet is fixed to $D_{\text{hidden}} = 256$ across all subnetworks.

Thus, in our implementation, each subnet used in a coupling block consists of the following sequence of layers:

$$\text{Linear}(11, 256) \to \text{ReLU} \to \text{Linear}(256, 256) \to \text{ReLU} \to \text{Linear}(256, 87)$$

As mentioned, the two coupling blocks that have been implemented and experimented with are:

AFFINE COUPLING. In affine coupling layers, the subnet predicts element wise shift and scale parameters for the subset $\mathbf{x}_b$, conditioned on $\mathbf{x}_a$ and the external condition $\mathbf{c}$. The transformation is then given by:

$$\mathbf{x}'_b = \mathbf{x}_b \odot \exp\left(s(\mathbf{x}_a, \mathbf{c})\right) + t(\mathbf{x}_a, \mathbf{c}), \quad \mathbf{x}'_a = \mathbf{x}_a, \tag{21}$$

where $s(\cdot)$ and $t(\cdot)$ are the scale and shift functions, respectively, and $\odot$ denotes elementwise multiplication. This coupling is computationally efficient and has a closed-form inverse:

$$\mathbf{x}_b = \left(\mathbf{x}'_b - t(\mathbf{x}_a, \mathbf{c})\right) \odot \exp\left(-s(\mathbf{x}_a, \mathbf{c})\right).$$

Moreover, the Jacobian of the transformation is triangular, and its log-determinant simplifies to the sum of the predicted scale outputs:

$$\log \left| \det \frac{\partial x'}{\partial x} \right| = \sum_j s_j(x_a, c).$$

SPLINE COUPLING. While affine coupling suffices for many applications, it is limited to transformations that are globally linear per dimension. To increase the expressiveness of the model, we employ *rational quadratic spline coupling* [44], a more flexible alternative that replaces the affine transformation with a monotonic, piecewise-defined spline function [44].

Same as the affine coupling block, the RQS coupling blocks also operate on the inputs by partitioning the input into x_a, remains unchanged and is used to condition the transformation of the second subset x_b. The conditional input, will be concatenated with x_a, and preparing it to be passed to the subnet that will generate the spline parameters [44].

The core transformation is based on a learnable piecewise rational quadratic spline, constructed from a set of bins that will partition the input. Each transformed dimension is mapped through its own spline, where the spline's shape is parametrized by a set of width, height and derivatives. For K bins, each spline segment requires $3K - 1$ parameters per transformed channel. These parameters are predicted subnet as discussed, which takes as input the conditioning features (i.e., the x_a and the external conditions c) and outputs the transformation parameters for each element of x_b. The unnormalized width, heights and derivatives are saved as a flattened tensor named θ [44].

For preserving the numerical stability and guarantee invertibility, each bin width and height is required to be greater or equal to 10^{-3}, therefor preventing vanishing segments and maintaining a monotonic mapping. Similarly, the derivatives at bin edges are constrained to remain above a minimum value of 10^{-3}, ensuring that the spline remains smooth and invertible throughout. To ensure each bin width and height is strictly positive and nonzero, the unnormalized values are passed through a softmax and scaled [44]:

$$w_i = \epsilon_w + (1 - \epsilon_w \cdot K) \cdot \text{softmax}(\theta_w)_i, \quad h_i = \epsilon_h + (1 - \epsilon_h \cdot K) \cdot \text{softmax}(\theta_h)_i$$

where K is the number of bins, and ϵ_w, ϵ_h are the minimum bin width and height thresholds, respectively. This normalization ensures that the sum of all widths (or heights) spans the entire bounded interval and avoids degenerate cases.

The cumulative widths and heights are then computed to obtain the bin edges. The cumulative values are padded and scaled to fit within the bounds [44]:

$$\text{cumwidths} \in [\text{left}, \text{right}], \quad \text{cumheights} \in [\text{bottom}, \text{top}]$$

And in the end, same as the height and width, the derivatives at each bin edge are obtained by applying the *softplus* function to the unnormalized derivatives, with a minimum derivative added to maintain monotonicity [44].

For each input value, the algorithm determines which bin it falls into. Based on the selected bin index, local parameters for each input—such as the bin width(Δx), height(h), slope (δ), and edge derivatives—are extracted for the subsequent transformation. Depending on the direction of the flow forward or inverse, two sets of equations are used [44]:

Forward transformation, we have the input x and we want the $y = f(x)$, and we computes the output as a smooth nonlinear mapping from input values :

$$\theta = \frac{x - x_{\text{left}}}{\Delta x}, \quad \theta(1-\theta) = \theta \cdot (1-\theta) \tag{22}$$

$$f(x) = y_{\text{left}} + \frac{h \cdot (\delta\theta^2 + d_{\text{left}} \cdot \theta(1-\theta))}{\delta + (d_{\text{left}} + d_{\text{right}} - 2\delta) \cdot \theta(1-\theta)} \tag{23}$$

For the inverse transformation, we have the $y = f(x)$, and we want to get the $x = f^{-1}(y)$. For that we use the equation 23 and solves a quadratic equation to recover the input given the output. This is achieved by finding the root of a spline inversion equation of the form [44]:

$$a\theta^2 + b\theta + c = 0 \tag{24}$$

$$\begin{aligned} a &= (y - y_{\text{left}})(d_{\text{left}} + d_{\text{right}} - 2\delta) + h(\delta - d_{\text{left}}) \\ b &= h \cdot d_{\text{left}} - (y - y_{\text{left}})(d_{\text{left}} + d_{\text{right}} - 2\delta) \\ c &= -\delta(y - y_{\text{left}}) \end{aligned} \tag{25}$$

The root θ is then computed, and from here, the solution x, can be calculated simply as [44]:

$$x = x_{\text{left}} + \theta\Delta x, \quad \theta = \frac{2c}{-b - \sqrt{b^2 - 4ac}} \tag{26}$$

For both forward and inverse transformations, the log-determinant of the Jacobian is computed analytically. The Jacobian is important in calculating the loss as in equation 20. The derivative of the spline is computed using the chain rule and and for the forward direction we have [44]:

$$\frac{\partial f(x)}{\partial x} = \frac{d_{\text{right}}\theta^2 + 2\delta \cdot \theta(1-\theta) + d_{\text{left}}(1-\theta)^2}{\delta + (d_{\text{left}} + d_{\text{right}} - 2\delta) \cdot \theta(1-\theta)} \tag{27}$$

For mixing the information and avoiding the risk of the model learning axis-aligned or degenerate transformations, a permutation operation is applied between the succsessive coupling layers. We use a fixed random permutation matrix, by drawing a uniform permutation over the input dimensions. This reorders

the input features before they are passed into the next coupling block, ensuring that different subsets of the input vector are designated as the conditioning and transformed parts in each layer. By doing so, it prevents any particular dimension from consistently remaining in the same role (e.g., always being conditioned on but never transformed). By alternating the transformed channels in each layer, we ensures that all dimensions participate in the learned mapping over the course of multiple coupling blocks.

To construct the invertible neural network architecture, the model defines a computational graph using the FrEIA [4]. The architecture begins with the declaration of an input node representing the observable variables—in this case, a six-dimensional vector corresponding to the unobservable astrophysical parameters that the model is designed to infer. Alongside this input node, a condition node is initialized to carry the conditional information, which remains fixed across all coupling layers and informs the transformation at each stage.

The core of the model consists of a sequence of eight invertible coupling blocks, each responsible for applying a bijective transformation to a subset of the input dimensions. These blocks are constructed using RQS coupling layers, which model complex, flexible nonlinear transformations while maintaining exact invertibility and efficient Jacobian computation. For each block, a subnet is used to predict the spline parameters—bin widths, heights, and derivatives—based on the concatenation of one half of the input features and the conditioning embedding. Each RQS block operates on a split of the input, leaving one half unchanged while transforming the other half based on the learned spline map. Between successive blocks, the input channels are permuted using a fixed random permutation (a hard permutation), ensuring that each dimension is eventually transformed and contributing to global expressivity. After the final block, an output node marks the termination of the transformation pipeline. The entire architecture is assembled into a `GraphINN` object, which supports forward and inverse evaluation and provides access to all trainable parameters for optimization during training.

Summary

In summary, the cINN implemented in this work is designed to learn an invertible mapping between a set of unobservable physical properties $\mathbf{x} \in \mathbb{R}^6$ and a latent space variable $\mathbf{z}$ that is encouraged to follow a standard normal distribution, conditioned on an 8 dimensional observable properties of galaxy clusters ($\mathbf{c} \in \mathbb{R}^8$) in this part, or a 512-dimensional embedding $\mathbf{c} \in \mathbb{R}^{512}$ derived from X-ray and Radio in the upcoming parts of v, vi, and vii. The model achieves this through a composition of 8 invertible transformations, each realized via RQS coupling blocks.

At each coupling block, the input vector $\mathbf{x}$ is split into two subsets of variables, denoted by (x_a, x_b), where $x_a \in \mathbb{R}^3$ and $x_b \in \mathbb{R}^3$. The subset x_a is passed unaltered through the coupling layer but is used, in concatenated with the conditional vectorc, to compute a set of transformation parameters θ that are applied to x_b. Specifically, the concatenated vector (x_a, c) is passed into a small neural network referred to as the subnet.

The subnet is a multilayer perceptron composed of three fully connected layers. The first layer takes the input of size $d_{x_a} + d_c$, maps it to a hidden embedding, applies a nonlinearity (ReLU), and passes it through two more linear layers. The

output of the subnet is a vector of dimension $3K - 1$ (where $K = 10$ is the number of spline bins); 3 parameters are learned per bin (widths, heights, and derivatives), minus one due to the last bin for derivatives. The resulting parameter vector is reshaped to match the shape of x_b so that a separate spline is applied per transformed dimension.

Each x_b component is then transformed using RQS, a flexible and invertible transformation defined over a bounded interval. The spline transformation maps each input through a piecewise function characterized by learned bin widths, heights, and derivatives. These parameters are regularized to enforce invertibility: bin widths and heights are normalized using a softmax and constrained to be no smaller than a minimal threshold (1e-3), and the derivatives are bounded from below using a softplus transformation.

For the inputs, the transformation computes a spline coordinate $\theta = \frac{x - x\_left}{\Delta x}$, identifies the appropriate bin, and applies the forward or inverse transformation depending on the training mode. The Jacobian determinant of this transformation is computed (as in equation 27) analytically and accumulated for use in the training loss.

After each coupling transformation, a fixed permutation of the channel dimensions is applied to the transformed vector. This shuffling ensures that all variables are eventually transformed across different coupling layers, even though each block modifies only one half of the input. By altering which subset of variables plays the role of x_a and x_b at each layer, and by reordering them, the model avoids learning degenerate axis-aligned transformations and encourages global information mixing.

The eight coupling blocks are composed sequentially, forming the complete bijective mapping $f : \mathbf{x} \mapsto \mathbf{z}$ conditioned on $\mathbf{c}$. Across the sequence of blocks, all six components of $\mathbf{x}$ undergo nonlinear transformations, resulting in a final output z in $\mathbb{R}^6$.

After the final coupling block, an output node is appended to signify the end of the flow graph. This node collects the transformed data, which is interpreted as latent variables $\mathbf{z}$. Finally, the condition node is appended to the graph to register its use, ensuring the graph remains complete and properly structured. The graph is then compiled into a full invertible model using FrEIA's `GraphINN` class, which manages data flow between nodes and automatically tracks invertible mappings. The result is a fully invertible embedding $z = f(x; c)$ [4].

And finally, the model is trained using maximum likelihood estimation under the assumption that the transformed variables follow a standard Gaussian distribution. The loss is computed as the negative log-likelihood from equation 20, where $z = f(x; c)$ is the output of the flow and $\det J_f$ is the Jacobian determinant accumulated across all coupling blocks. The conditional structure allows the model to learn complex, multi-modal posterior distributions over $\mathbf{x}$, given the observed embedding $\mathbf{c}$.

11.2 TRAINING

We train the cINN by maximum likelihood on paired targets and conditions $(\mathbf{x}, \mathbf{c})$, using the negative conditional log–likelihood (NLL) as the objective. Given a mini–batch $\{(\mathbf{x}_i, \mathbf{c}_i)\}_{i=1}^{B}$ from the training split, the loss minimized at each step is:

$$\widehat{\mathcal{L}}_{\mathrm{NLL}}(\theta) = -\frac{1}{B}\sum_{i=1}^{B} \log p_\theta(\mathbf{x}_i \mid \mathbf{c}_i) \tag{28}$$

$$= \frac{1}{2B}\sum_{i=1}^{B} \|\mathbf{z}_i\|^2 - \frac{1}{B}\sum_{i=1}^{B} \log|\det J_{f_\theta}(\mathbf{x}_i \mid \mathbf{c}_i)|\,. \tag{29}$$

where $\mathbf{z}_i = f_\theta(\mathbf{x}_i \mid \mathbf{c}_i)$ and J_{f_θ} is the Jacobian of the forward flow (section 11.1). Equation 28 is the stochastic mini–batch estimate of the negative log–likelihood derived in Equation 20.

As described in section 10.1, observables and targets are standardized and split 80/10/10 into train/validation/test. PyTorch `DataLoader` yields batches in the order $(\mathbf{x}, \mathbf{c})$ (targets first, conditions second), and we use a batch size of 256. At construction, the trainer instantiates the cINN with the training tensors (targets and conditions) and the architectural choices (number of blocks, subnet widths, spline settings) are those described in the section 11.1.

During training, we add small isotropic Gaussian noise to both targets and conditions in standardized space,

$$\mathbf{x} \to \mathbf{x} + \sigma\,\varepsilon_x, \qquad \mathbf{c} \to \mathbf{c} + \sigma\,\varepsilon_c, \qquad \varepsilon_x, \varepsilon_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \sigma = 0.01.$$

This acts as a mild data augmentation and regularizer for flows, discouraging brittle solutions and improving numerical stability of the Jacobian terms. This is specially important for time parameters (Cosmic Time in observables) because of their discrete nature, since they are measured in simulations as snapshots provided by the TNG-Cluster simulation as explained in section 9.1. Gradients are clipped to norm 5.0 each step to prevent rare exploding updates.

We use AdamW to minimize equation 28 with [79]:

$$(\text{learning rate},\ \beta_1,\ \beta_2,\ \epsilon,\ \text{weight decay}) = (5{\times}10^{-4},\ 0.9,\ 0.999,\ 10^{-6},\ 10^{-5})$$

A `reduce_on_plateau` learning–rate scheduler is also applied [121]. This scheduler, reduces the learning rate by a factor of 0.8 if the validation loss does not improve for 20 epochs (threshold 10^{-4}).

Training runs for 2000 epochs. Each epoch iterates over mini-batches (with Gaussian noise), computes equation 28, backpropagates, takes an AdamW step, and then evaluates Eq. 28 on the validation split in evaluation mode (no Gaussian noise). We save checkpoints at initialization, every 20 epochs, and at the final epoch. Learning–rate updates for `ReduceLROnPlateau` are applied after validation. From here on, for postprocessing and results (chapter 12) use the *last* checkpoint.

11.3 INFERENCE AND POSTPROCESSING

After training, the cINN provides an exact, normalized model for the conditional density $p(\mathbf{x} \mid \mathbf{c})$ via the invertible map

$$\mathbf{z} = f(\mathbf{x} \mid \mathbf{c}), \qquad \mathbf{x} = f^{-1}(\mathbf{z} \mid \mathbf{c}), \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Postprocessing proceeds in the following order;

(1) Fix the conditioning inputs. Given a batch of conditions $\mathbf{C} \in \mathbb{R}^{N \times D_c}$ (observables or learned embeddings), we will characterize $p(\mathbf{x} \mid \mathbf{c}_i)$ for each row $\mathbf{c}_i$.

(2) Draw posterior samples by inversion. For each $\mathbf{c}_i$, draw n_{sam} i.i.d. latent samples $\mathbf{z}^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and push them through the inverse flow:

$$\mathbf{x}_i^{(s)} = f^{-1}\big(\mathbf{z}^{(s)} \mid \mathbf{c}_i\big), \qquad s = 1, \dots, n_{\text{sam}}.$$

(3) Summarize each posterior with a MAP point. For a fixed condition $\mathbf{c}$ let $\{x_d^{(s)}\}_{s=1}^{n_{\text{sam}}}$ be the Monte Carlo samples of the d-th target coordinate drawn from $p(\mathbf{x} \mid \mathbf{c})$ and converted to physical units. We estimate the posterior marginal density of that parameter with a one–dimensional Gaussian kernel density estimator (KDE):

$$\widehat{p}(x_d \mid \mathbf{c}) = \frac{1}{n_{\text{sam}}\, h} \sum_{s=1}^{n_{\text{sam}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_d - x_d^{(s)})^2}{2h^2}\right),$$

where $h > 0$ is the bandwidth (smoothing scale). The *marginal* maximum–a–posteriori (MAP) estimate for coordinate d is then:

$$x_d^{\text{MAP}}(\mathbf{c}) \;\approx\; \arg\max_{x_d}\; \widehat{p}(x_d \mid \mathbf{c}),$$

computed by evaluating $\widehat{p}$ on a uniform grid over the sampled range and taking the maximizer.

(4) Build empirical priors. For visualization and sanity checks, we also estimate *empirical priors* per merger parameter using the *same* 1D Gaussian KDE applied to the test targets (unconditional on $\mathbf{c}$). Let $D_d^{\text{test}} = \{x_{i,d}\}_{i=1}^{N_{\text{te}}}$ be the test values of coordinate d in physical units. Then

$$\widehat{p}_{\text{emp}}(x_d) = \frac{1}{N\, h} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_d - x_{i,d})^2}{2h^2}\right).$$

For plotting, we normalize each $\widehat{p}_{\text{emp}}$ to unit peak. These empirical priors are *not* used in training, they simply summarize what stays in the test dataset and serve as a baseline against which to compare the learned posteriors $p(\mathbf{x} \mid \mathbf{c})$.

Implementation notes. We use Gaussian kernels with bandwidths chosen empirically (e.g., $h = 0.3$ for MAP KDEs and $h = 0.5$ for empirical priors) and evaluate on a dense grid (512 points) over the sampled range. All sampling and KDEs are carried out after converting to physical units, so MAPs and any credible summaries are directly interpretable (e.g., Gyr, kpc). Since we evaluate the MAP on a uniform grid of $G = 512$, this can impose a finite resolution $\Delta \approx (x_{\max} - x_{\min})/(G - 1)$ in each coordinate, i.e., very sharp modes are snapped to the nearest grid node. However, With $G = 512$, the resulting quantization error is far smaller than the gaussian error added during training, so its impact on reported MAPs is negligible.

This workflow yields, for each $\mathbf{c}$, a set of posterior samples $\mathbf{x}^{(s)}$ capturing uncertainty and possible multi-modality, along with a single, comparable point estimate $\widehat{\mathbf{x}}_{\text{MAP}}$ for downstream metrics and plots.

(5) Empirical prior modes. In addition to visualizing the empirical priors, we can also extract their most probable values by taking the argmax of the KDE estimate

$\widehat{p}_{emp}(x_d)$ along a uniform evaluation grid. These modes provide simple summary statistics of the unconditioned dataset and serve as interpretable benchmarks. Table 7 lists the empirical prior modes obtained for each merger parameter, together with their units. This can further be used for evaluation on the performance of the cINN.

Table 7: Empirical prior modes estimated via Gaussian KDE from the full dataset, for last and next mergers.

Parameter (units)	**Last merger**	**Next merger**
Collision Time (Gyr)	8.7019	12.4484
Collision Velocity (log(km/s))	3.2688	3.2647
Main Cluster M_{500c} ($\log(M_\odot)$)	14.4541	14.4956
Subcluster Mass ($\log(M_\odot)$)	13.3764	13.4232
Merger Mass Ratio	0.2149	0.2034
Pericenter Distance (log(kpc))	2.3968	2.4977

12 RESULTS AND DISCUSSIONS: SCALAR CONDITIONING

12.1 POSTERIOR DISTRIBUTION

We visualize conditional posteriors $p(\mathbf{x} \mid \mathbf{c})$ for a subset of test clusters using the trained cINN. From the saved test indices we randomly select $n_{\text{rows}} = 15$ galaxy clusters. For each selected cluster, or in other words, condition $\mathbf{c}_i$, we draw $n_{\text{sam}} = 1000$ posterior samples $\{\mathbf{x}_i^{(s)}\}_{s=1}^{n_{\text{sam}}}$ via the inverse flow (section 11.3), convert all samples to physical units, and display one target (merger parameter) per column. Each row corresponds to one randomly chosen cluster out of the test sample, annotated on the left with its `HaloID` and redshift z. Each column shows a different merger parameter (the unobservable of section 9.2). Within each panel four elements are overlaid (constructed as in section 11.3):

- a gray *prior* distribution curve for the merger parameter (derived as discussed in section 11.3). This curve shows the distribution of values across the test set of the galaxy clusters.
- a blue curve showing the posterior distribution of the merger parameter for the randomly chosen test galaxy cluster. The posterior distribution, is the immediate product of our cINN training.
- a gold/yellow vertical line, representing the MAP (maximum–a–posterior) estimate from the learned posterior distribution.
- a red vertical line at the ground-truth value, which is the direct outcome of the TNG-Cluster simulation as discussed in section 9.2.

It is important to note that both prior and posterior KDEs are peak-normalized, so absolute curve heights are not comparable across panels; inference relies on the *relative* location and sharpness of the blue posterior versus the gray prior and the red truth. Moreover, n_{sam} controls Monte Carlo smoothness of the posterior KDE; we use $n_{\text{sam}} = 1000$ in all panels.

For reading Figure 17, when the blue posteriors are (i) visibly narrower than the gray priors and (ii) vary across rows, indicate that the conditioning variables $\mathbf{c}$ (scalar observables) provide cluster-specific information about the merger parameters. Narrow posteriors point to precise inference, while having the MAP laying close to the ground-truth shows its accuracy. The performance can be seen to be the strongest for Collision Time and also pronounced for the Main Cluster M_{500c}; it is weaker—but still present—for Collision Velocity. Collision Time have the most precise posteriors; meaning that they are strongly contracted with respect to the prior. In contrast, Merger Mass Ratio, Subcluster Mass, and Pericenter Distance, do not have accurate MAP estimates and often retain posteriors that are broader and are similar across different clusters reflecting intrinsic degeneracies.

It also can be seen that the gold MAP lines lie close to the red ground truths (accurate inference) when contraction is strong as in Collision Time, Main Cluster

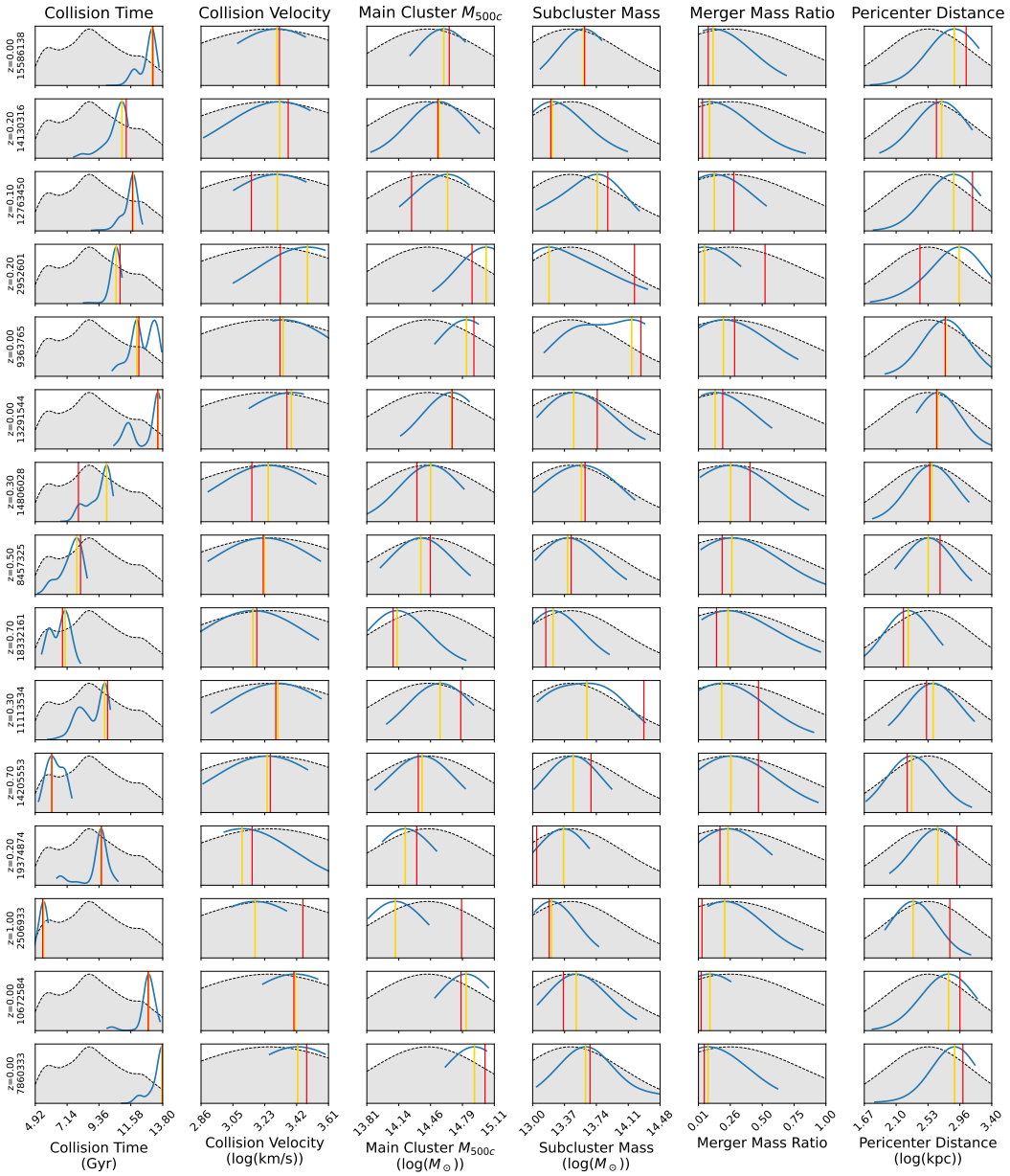

Figure 17: Conditional posterior distribution for 15 randomly selected test galaxy clusters (rows) out of 203, across all target merger parameters (columns). Gray: prior distribution over the test split (KDE). Blue: predicted posterior KDE for each galaxy cluster predicted by the cINN. Gold: MAP estimate. Red: ground truth from TNG-Cluster.

M_{500c}, and Collision Velocity. The largest MAP–truth discrepancies occur where posteriors are broad (Merger Mass Ratio, Subcluster Mass, Pericenter Distance), which is the desirable behavior: the model expresses uncertainty rather than overconfidently committing to an incorrect value. Figure 17 illustrates 15 cases; the next sections quantify these patterns over the full test set.

12.2 PREDICTION PERFORMANCE OF THE CINN

Figure 18 summarizes, for each merger parameter x_d, how the *full* posterior $p(x_d \mid \mathbf{c})$ compares to the simulation ground truth across the test set. We first partition the ground-truth axis into $B = 15$ equal-width bins. For every test galaxy cluster in a given truth bin we draw $n_{sam} = 500$ posterior samples with the inverse flow (Section 11.3) and histogram those samples into the *same* binning along the vertical axis. The result is a 15×15 count matrix shown as a heatmap in value space, with the white diagonal marking $y = x$ showing the ideal case. Overlaid black curves report the posterior median (solid) and central 10–90% quantiles (dashed) as functions of the truth. In the ideal case, probability mass concentrates in a narrow ridge along the diagonal with tight quantile bands.

Figure 19 complements the distributional view with *point estimates*. The top row scatters the MAP against the ground truth for each target (pink line: $y = x$), together with running medians (solid black line) and 10–90% envelopes in truth bins. The bottom row replaces the vertical axis with the relative error $\Delta = (\text{MAP} - \text{truth})/\text{truth}$ to make scale differences explicit, again showing the medians in solid black line and two dashed lines containing 80% of the data. Concentration near the identity in the top row and near $\Delta = 0$ in the bottom row indicates accurate predictions; the envelope width visualizes dispersion.

Systematic vertical offsets of the median reflect *bias*, formally $b(x) = \mathbb{E}[\hat{x} \mid x] - x$, so a median curve above $y = x$ indicates positive bias (overestimation) and below indicates negative bias (underestimation). However, departures from $y = x$ encode distinct effects. A bending of the distribution's center (or MAPs in our case) toward the sample's mode is *shrinkage*/regression-to-the-mean: when $\mathbf{c}$ is only partially informative, extreme truths are pulled toward the global mode. This can be seen in when the median line of posterior distributions or the MAP estimate, flips near the global mode of the target distribution (table 7). Widening 10–90% bands with |truth| indicate *heteroscedasticity*—uncertainty growing in certain regimes (e.g., very small or large values). In the relative-error panels, percentage errors can inflate where the denominator approaches zero (as in the case of Merger Mass Ratio); those regions should be interpreted with care.

Finally, both figures depend mildly on the analysis hyperparameters (B, n_{sam}): increasing the number of truth bins B gives finer resolution along the axes but spreads a fixed sample budget over more bins—reducing counts per bin and thus increasing variance (noise $\propto 1/\sqrt{\text{count}}$)—whereas increasing the posterior draws per object n_{sam} boosts those counts and correspondingly suppresses Monte Carlo fluctuations in the histograms/KDEs, stabilizing the MAP/quantile curves.

Taken together, the stacked posterior heatmaps (Figure 18) and the MAP–vs–truth / relative–error views (Figure 19) provide a consistent population-level assessment of calibration (median vs. $y = x$) and dispersion (quantile widths), as well as point-estimate accuracy. We summarize the per-target behavior below;

COLLISION TIME. Among all targets, Collision Time exhibits the best calibration: the posterior median closely tracks the identity line with comparatively tight 10–90% bands. The MAP–vs–truth relation mirrors this, but the relative errors cluster at $\sim$ 20–30% with a mild positive bias (overestimation). Part of this inflation reflects the discrete snapshot sampling of time, for which percentage errors are less forgiving than absolute errors; shrinkage toward the dominant snapshot spacings also contributes.

COLLISION VELOCITY. Calibration is good in the high-velocity regime (e.g., $\log(v) \geqslant 3.2\,\mathrm{km/s}$), with narrow posterior bands along $y = x$. At lower velocities the posterior widens (heteroscedasticity) but remains centered. Notably, the MAP bias changes sign around the global mode $v \approx 10^{3.2}\,\mathrm{km/s}$: below the mode we see mild overestimation, above it mild underestimation—canonical shrinkage toward the modal scale. Errors are modest overall, typically $< 10\%$ (often $< 5\%$ at high v).

MAIN CLUSTER M_{500c}. For $\log M_{500c} \geqslant 14.4$ the posterior median and quantiles lie close to $y = x$. Below this scale the median bends upward and the bands widen, indicating weaker conditioning information and shrinkage toward the global mode at around $\log M_{500c} \approx 14.5\,M_\odot$. Despite this, MAP errors are small: 5% at low masses and approaching 0 at the high-mass end.

SUBCLUSTER MASS. Posteriors deviate from $y = x$ with increasing spread toward high masses, consistent with limited information and sample imbalance. The MAP median departs from the identity with a sign flip near the mode $M_{sub} \approx 10^{13.5}\,M_\odot$: mild underestimation below the mode and mild overestimation above it, with amplitudes typically within $\pm 5\%$—again, shrinkage toward the modal scale.

MERGER MASS RATIO. This is the most challenging parameter. Neither posteriors nor MAPs align with $y = x$; errors can be large (up to several hundred percent) at small ratios. This reflects both weak conditioning signal and the sensitivity of percentage errors when the denominator is small. The bounded nature of the ratio ($0 < \mu \leqslant 1$) and skewed target distribution further complicate learning.

PERICENTER DISTANCE. Calibration is moderate: the posterior median shows a gentle S-shaped departure from $y = x$, and MAP errors switch sign near the modal scale $d_{peri} \approx 10^{2.53}$ kpc (positive below, negative above), with typical magnitudes $\sim \pm 20\%$.

Overall, the cINN contracts uncertainty and tracks the identity where $\mathbf{c}$ is informative (time, velocity, high-mass M_{500c}, and to lower extents Pericenter Distance), and it transparently expresses ambiguity (wider posteriors and larger MAP dispersion) where the inverse problem is intrinsically weakly constrained (mass ratio, subcluster mass). The observed sign flips around the target-wise modal scales are consistent with regression-to-the-mean induced by partial information and data imbalance.

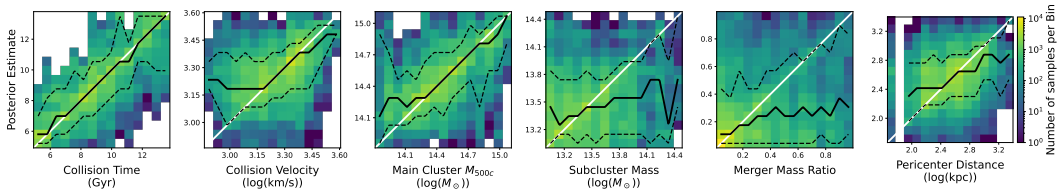

Figure 18: Posterior versus ground truth per target (merger parameter) on the 203 test galaxy clusters. Each panel shows a 2D histogram of posterior samples (vertical) binned on the bins of ground-truth (horizontal), with shared logarithmic color scale. White line: $y = x$. Black solid line: posterior median per ground-truth bin; black dashed lines: 10-90% posterior quantiles. A well-calibrated, accurate model concentrates mass near the diagonal with narrow quantile bands. Here we use $B = 15$ truth bins and draw $n_{\text{sam}} = 500$ posterior samples per test object.

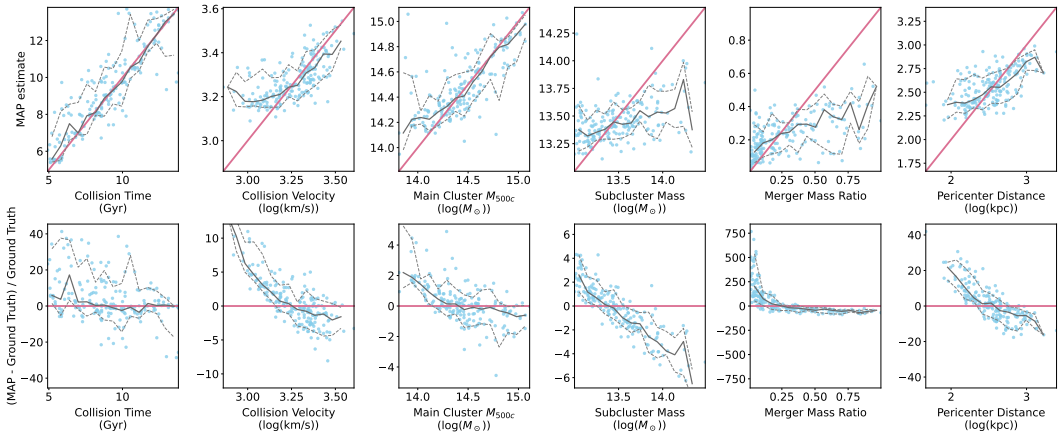

Figure 19: Per-target (merger parameter) MAP accuracy (top) and relative error (bottom) over the 203 test clusters. Top: scatter of MAP vs. truth with $y = x$ (ideal), plus median (solid) and 10–90% (dashed) MAP within truth bins. Bottom: relative error $\Delta = 100(\text{MAP} - \text{Truth})/\text{Truth}$ vs. truth with the same bin-wise summaries. Tight bands near the diagonal/zero indicate accurate, well-calibrated predictions; curvature or wide bands reveal bias.

12.3 CROSS CORRELATIONS

In the previous sections, we have tested the power of cINN to return the merger parameter of galaxy clusters given a set of observables. In this section, we want to see whether our cINN model is also able to learn the cross correlation between the merger parameters. To inspect cross–target dependencies learned by the cINN, we visualize all pairwise relations among the target merger parameters in a corner plot. We draw $n_{sam} = 200$ posterior samples per object, and create a scatter plot of the posteriors (blue), MAP (gold), and the ground truth (red). In addition to this, on the diagonal, we plot the 1D KDEs distribution of the posterior, MAPs, and ground truths in their respective colors.

Alignment of gold and red clouds indicates accurate MAPs; a blue cloud elongated along the red truth locus signals that the posterior captures the correct correlation. Systematic offsets between gold and red trends indicate bias; dispersed or multi–clumped blue structure indicates residual ambiguity or multi–modality in $p(\mathbf{x} \mid \mathbf{c})$.

Under ΛCDM, cluster halos are approximately self-similar, so a few first-order scalings organize the pairwise relations among merger parameters:

- **Host mass vs. cosmic time.** Clusters grow hierarchically; hence later cosmic times correspond, on average, to larger M_{500c}. A *positive* trend between Collision Time and M_{500c} is expected.
- **Mass sets size and speed scales.** With

$$R_{500c} \propto \left[\frac{M_{500c}}{\rho_c(z)}\right]^{1/3}, \quad v \sim \left(\frac{GM_{500c}}{R_{500c}}\right)^{1/2},$$

  we expect global *positive* correlations:

$$M_{500c} \uparrow \Rightarrow\ r_p\,(\mathrm{kpc}) \uparrow, \qquad M_{500c} \uparrow \Rightarrow\ v_{coll} \uparrow . \tag{30}$$

  Thus the (M_{500c}, r_p) and (M_{500c}, v_{coll}) panels should show rising loci.
- **Subcluster mass and mass ratio.** By definition $M_{sub} = \mu\, M_{500c}$. In log-space, at fixed M_{500c} the (μ, M_{sub}) panel is a near-linear band with positive slope; across the population it broadens due to the scatter in the M_{500c}.
- **Pericenter vs. time.** Because typical orbits sample similar *fractional* radii ($r_p \sim \xi R_{500c}$), and R_{500c} increases as clusters grow, later cosmic times tend to be associated with larger physical r_p at the population level (again, with wide scatter from orbital diversity).

The correlation structure visible in Fig. 20 demonstrates that the posterior samples, the MAP estimates, and the ground-truth values collectively reproduce the qualitative trends anticipated from the scaling arguments outlined above. In particular, the posterior clouds trace the expected pairwise dependencies among merger parameters, while the MAP points and ground truths align along the same loci, confirming that the inferred distributions capture not only the correct one-dimensional marginals but also the underlying cross-target correlations.

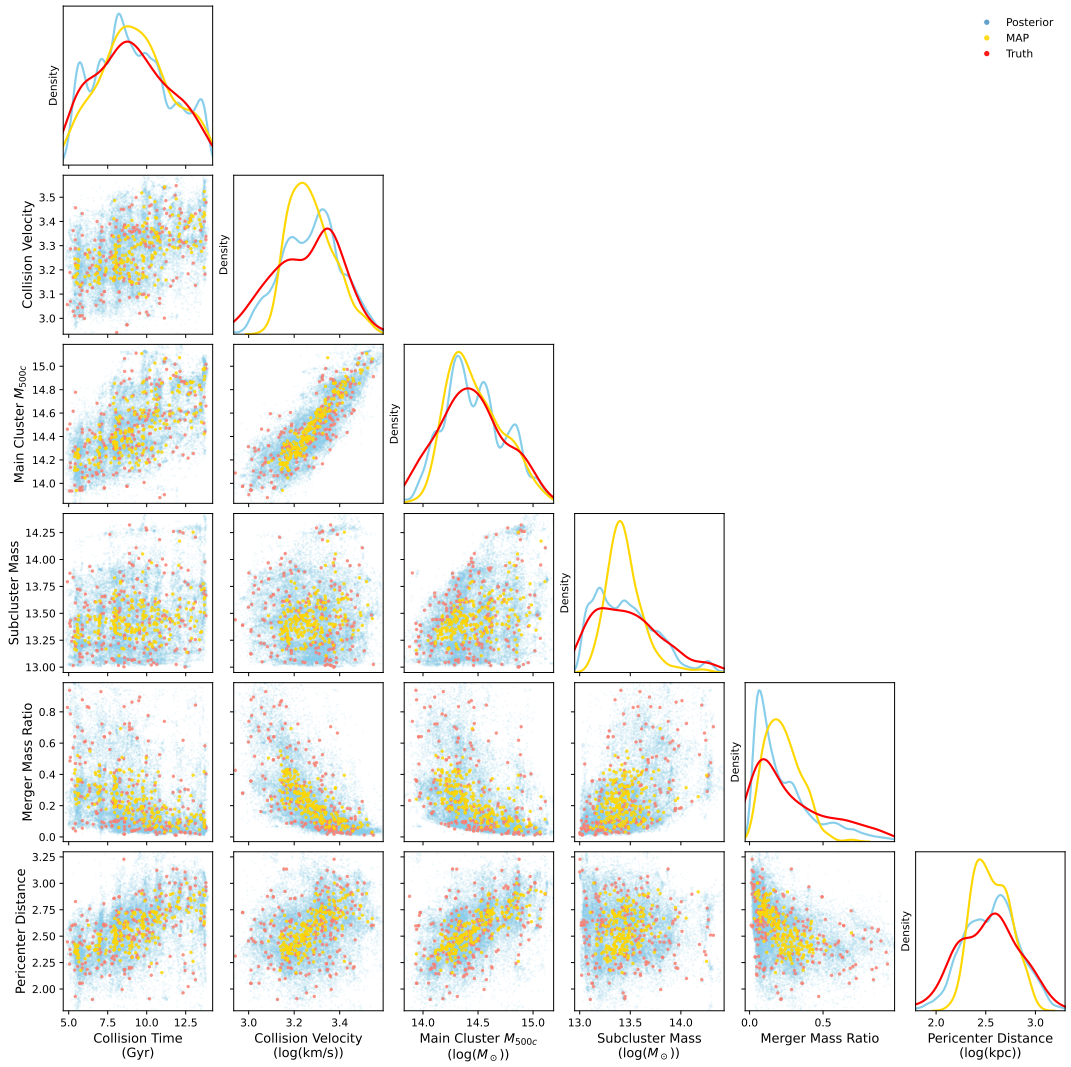

Figure 20: Corner plot across all merger parameters of the 203 test galaxy clusters. Diagonal: marginal KDEs of posterior (blue), MAP distribution (gold), and ground truth (red). Lower triangle: pooled posterior samples (200 posteriors for each test sample)(blue), MAPs (gold), and truths (red) for each test object. The plot exposes learned cross–target structure, MAP accuracy, and any residual multi–modality.

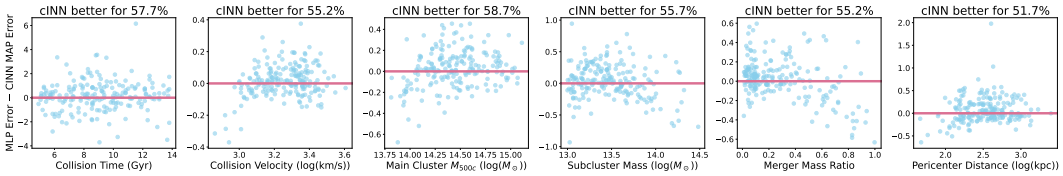

Figure 21: Δ-error scatter plot comparing the prediction error of the deterministic MLP with the cINN maximum a posteriori (MAP) estimates. Each subplot corresponds to one merger parameter. The x-axis shows the ground truth value, while the y-axis shows $\Delta\varepsilon = |\hat{y}_{\text{MLP}} - y| - |\hat{y}_{\text{MAP}} - y|$. Values above the pink dashed zero line indicate improved accuracy of cINN MAP estimates compared to the MLP. The percentage of test points with $\Delta\varepsilon > 0$ is annotated on top of each subplot.

12.4 MLP VS. CINN PERFORMANCE

To evaluate the benefit of the conditional invertible neural network (cINN) compared with a deterministic multilayer perceptron (MLP), we compute and visualize the Δ-*error scatter plot* for each physical merger parameter. For every test cluster and each target dimension d, we define the error difference

$$\Delta\varepsilon_d = \left|\hat{y}_{\text{MLP},d} - y_d\right| - \left|\hat{y}_{\text{MAP},d} - y_d\right|,$$

where $\hat{y}_{\text{MAP},d}$ is the MAP estimate obtained from cINN posterior samples after inverse transformation to physical units. Positive values ($\Delta\varepsilon_d > 0$) indicate that the cINN MAP estimate is more accurate than the MLP prediction for that data point; negative values indicate the opposite.

Each subplot corresponds to one target parameter: the x-axis shows the ground-truth physical value, and the y-axis shows $\Delta\varepsilon_d$. A dashed horizontal zero line separates regions where the cINN improves over the MLP (above) from regions where the MLP performs better (below). In addition, the percentage of test points with $\Delta\varepsilon_d > 0$ is reported at the top of each subplot as a compact, easy-to-interpret summary statistic. As can be seen from Figure 21, across all merger properties, the cINN MAP estimates tend to reduce error compared to the baseline MLP.

12.5 NEXT–MERGER INFERENCE

We repeat the scalar–conditioned analysis of section 12.1 and 12.2, now targeting the *next* merger (future event) rather than the last. The plotting protocol is unchanged: we visualize per–object posteriors $p(\mathbf{x} \mid \mathbf{c})$ as a grid (constructed as in section 11.3) and we summarize population–level calibration/accuracy with 2D posterior–vs–truth histograms and MAP–vs–truth/relative–error panels.

POSTERIOR GRIDS. Figure 22 shows $n_{\text{rows}} = 15$ randomly selected test clusters (rows) and one target per column, with prior (gray), posterior (blue), MAP (gold), and truth (red). Relative to the *last*–merger case (Fig. 17), the shapes are qualitatively similar but generally broader—especially for Collision Time—reflecting the added uncertainty inherent in forecasting forward in time. MAP markers remain close to the truths where posteriors contract (Collision Time, velocity, main–cluster mass), and deviate more where posteriors are wide (mass ratio, extreme pericenters), as desired.

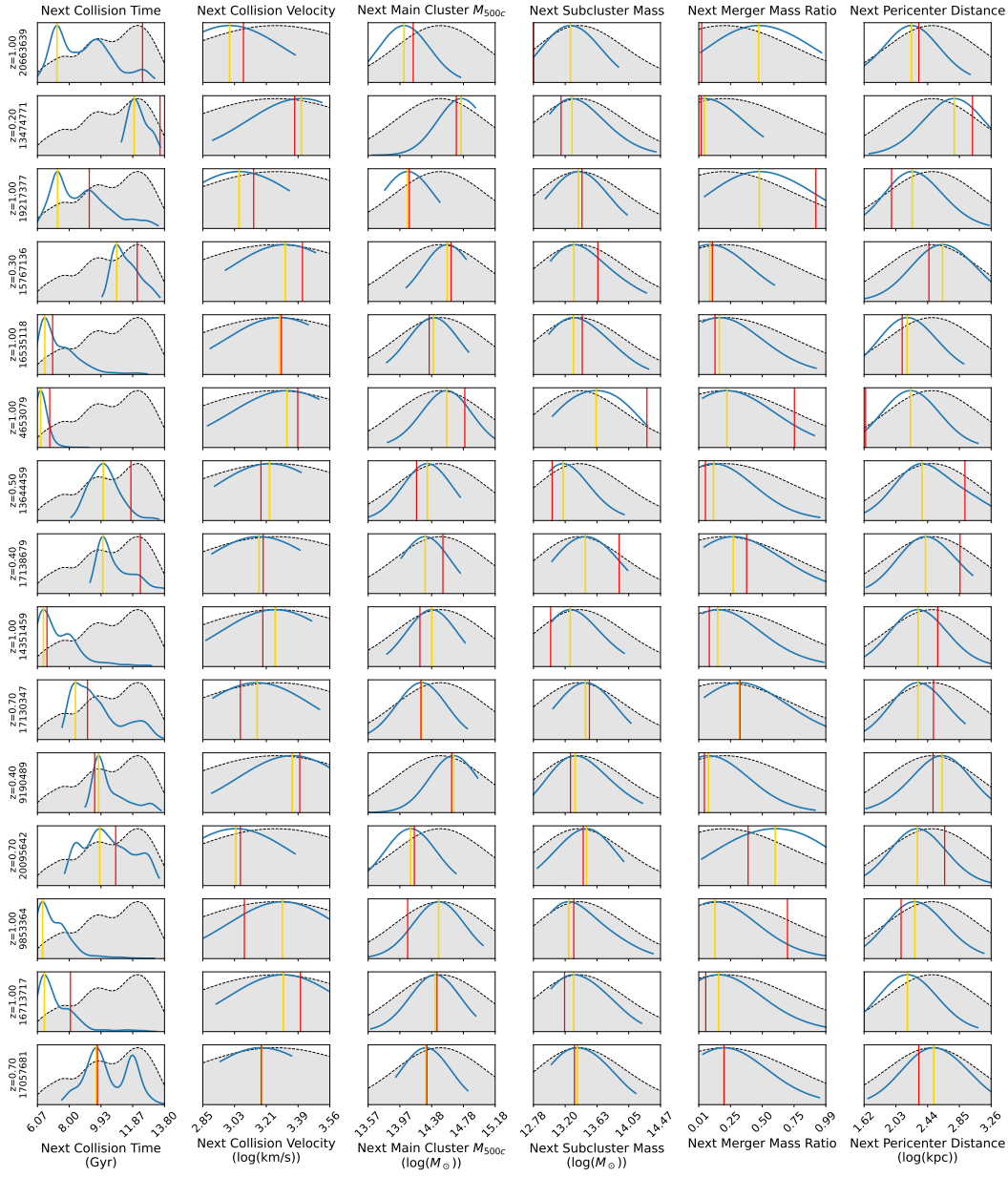

Figure 22: Scalar–conditioned posteriors for the *next* merger (15/193 test clusters; construction identical to Fig. 17). Compared to the last–merger case, posteriors are broader—most visibly for Collision Time—yet MAPs remain close to the truths where contraction is strong (for Collision Time, Collision Velocity and Main Cluster M_{500c}).

POPULATION–LEVEL PERFORMANCE. Figures 23 and 24 repeat the calibration/accuracy summaries with $B = 15$ truth bins and $n_{sam} = 500$ posterior draws per object. Heatmaps remain aligned with the diagonal $y = x$, and median curves (solid black) and 10–90% bands (dashed) retain the same qualitative trends as for the last–merger, albeit with slightly wider bands (particularly for Collision Time and Pericenter Distance). The MAP–vs–truth medians largely follow $y = x$ for Collision Time, velocity, and main–cluster mass; relative–error panels show modest, interpretable dispersion.

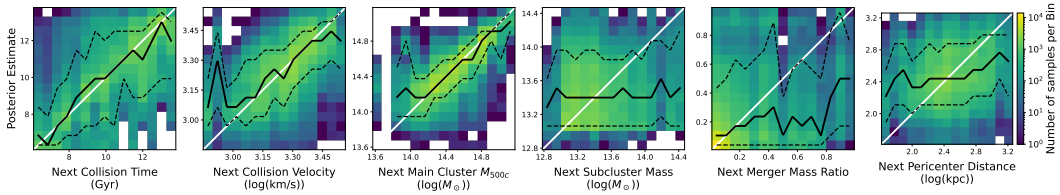

Figure 23: Next–merger: posterior vs. truth per target (scalar conditioning) across 193 test clusters. Construction as in Fig. 18 with $B = 15$ and $n_{sam} = 500$. Medians (solid) and 10–90% bands (dashed) remain close to $y = x$, with broader bands than the last–merger case—most visibly for Collision Time.

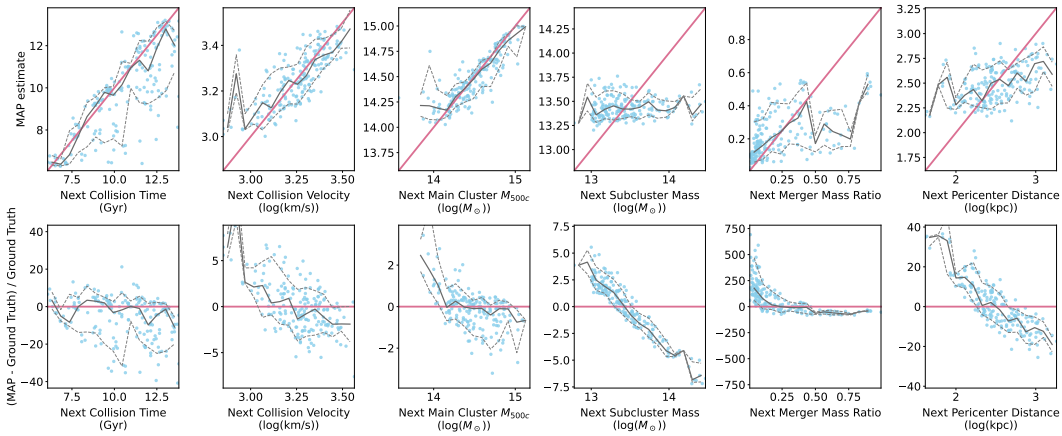

Figure 24: Next–merger: per–target MAP accuracy (top) and relative error (bottom) across 193 test clusters. Medians (solid) lie near $y = x$ (top) and near $\Delta = 0$ (bottom) for Collision Time, velocity, and main–cluster mass; envelopes are broader than for the last–merger, consistent with increased forecasting uncertainty.

We summarize the results observed in Figures 24 and 23; signs refer to $(\text{MAP} - \text{truth})/\text{truth}$:

- *Collision Time:* broader posteriors compared to last merger; errors mostly in $[-40\%, 20\%]$ with a predominantly negative bias (underestimation). The larger spread vs. the last–merger is consistent with snapshot discretization and the intrinsic difficulty of predicting future timing.
- *Collision Velocity:* tighter than the last–merger; errors $\approx \pm 5\%$ (previously $\sim [-5\%, 10\%]$). Median tracks $y = x$ well in velocities above the 10^3 km/s.
- *Main Cluster* M_{500c}*:* small dispersion; errors $\approx \pm 2\%$ with medians on the identity unless below $10^{14} M_\odot$.
- *Subcluster Mass:* similar to the last–merger; errors $\approx \pm 4\%$ and mild, mode–centered shrinkage, but the median do not follow the identity line.
- *Merger Mass Ratio:* same qualitative limitation as last merger—relative errors inflate near zero (bounded fraction); but the MAP median follows the identity line in ratios below 0.5.
- *Pericenter Distance:* slightly broader than the last–merger; errors mostly in $[-20\%, 20\%]$, with gentle S–shaped median and sign change around the modal scale.

Overall, the next–merger results mirror the last–merger findings but with wider posteriors and slightly larger errors where the forward prediction is intrinsically harder (time and pericenter). Calibration (medians near $y = x$) and the MAP accuracy remains good for Collision Time, Collision Velocity, and Main Cluster M_{500c}, and the model transparently expresses uncertainty where the conditioning is less informative or the target is intrinsically ill–behaved (mass ratio, extreme pericenters).

Part V

INFERRING THE MERGER HISTORIES OF GALAXY CLUSTERS VIA CONDITIONAL INVERTIBLE NEURAL NETWORKS AND CONTRASTIVE LEARNING: X-RAY MAPS

13 SIMULATION DATA AND X-RAY MAPS CONSTRUCTION

The main source of X-ray emission from galaxy clusters is thermal bremsstrahlung (free-free emission), that is the result of the deceleration of a free electron in the Coulomb field of an ion, producing a continuum spectrum. The X-ray spectrum, also shows a contribution from line emissions which arises from highly ionized electron transition from higher to lower energy orbits that can seen in elements such as oxygen (O), magnesium (Mg), silicon (Si), and iron (Fe) [99]. Although these metal emission lines contribute less to the X-ray luminosity than the bremsstrahlung, but prominent in softer X-ray energies (below ~2 Kev) and gives insights into the chemical compositions state of the ICM [16].

For this thesis, we use the intrinsic X-ray maps from the main 352 primary-zoom halos of the TNG-Cluster simulations used in Nelson et al. [113]. The X-ray emissivity of each gas cell is computed using its density, temperature, and metallicity following the APEC collisional ionization equilibrium model, including both continuum and line emission [154]. The maps are constructed by projecting the emissivity through the adaptively-sized Voroni gas cells using a cubic-spline kernel integration scheme as detailed in Nelson et al. [113].

The maps chosen for this thesis represent a field of view of $4\,R_{200c}$ (i.e., $\pm 2\,R_{200c}$ from the cluster center) and a line-of-sight depth of $2\,R_{200c}$, ensuring that both the core and the surrounding outskirts of the cluster are captured. The final maps have a pixel resolution of 2000×2000, providing high spatial detail. The projections are also taken along three orthogonal axes of the simulation box ($\hat{x}$, $\hat{y}$, and $\hat{z}$). Since the clusters are oriented randomly, theses projections gives us independent and statistically random viewing angles for each cluster. Given the triaxial nature of clusters, the three projections can be treated as different sample. As can be seen in Figure 25, the projected surface brightness looks very different across the three axes.

Figure 25 shows the X-ray surface brightness maps of four halos across three projections, representing the variety across two key classification: dynamical state and cool-core structure. The top two rows, represent dynamically relaxed clusters, while the bottom two rows are non-relaxed clusters. The first and third row also represent strong cool-core (SCC) clusters, and the second and fourth row are non cool-core clusters (NCC). The classification into SCC and NCC is based on the criterias defined in Lehle et al. [93]. The criteria used here is based on the cooling time, where a system is classified as SCC if its central cooling time $t_{cool} < 1$ Gyr, and NCC if $t_{cool} \geqslant 7.71$ Gyr.

The dynamic state of each cluster is determined according to Ayromlou et al. [8] within two main criteria: a cluster is relaxed if (i) the distance between its center of mass and its most bound particle is less than $0.1\,R_{200c}$, and (ii) the mass ratio between the central subhalo and the host cluster exceeds 0.85. The clusters in Figure 25 satisfies both of the conditions to be considered relaxed or non-relaxed.

In Figure 26, the evolution of the same four clusters shown in Figure 25 at $z = 0$, 0.2, 0.5, and 1 along the $\hat{x}$ projection axis. It can be seen that the relaxed SCC

systems tend to preserve their peaked surface brightness distribution, while non-realxed and NCC clusters, often show more change across the cosmic time.

In the end, our final sample is consisted of the intrinsic X-ray luminosity maps for the 352 primary-zoom halos and their main progenitors spanning eight full snapshots in the redshift range $0 \leqslant z \leqslant 1$. By taking three different projection for each of them, the dataset is includes 8448 maps.

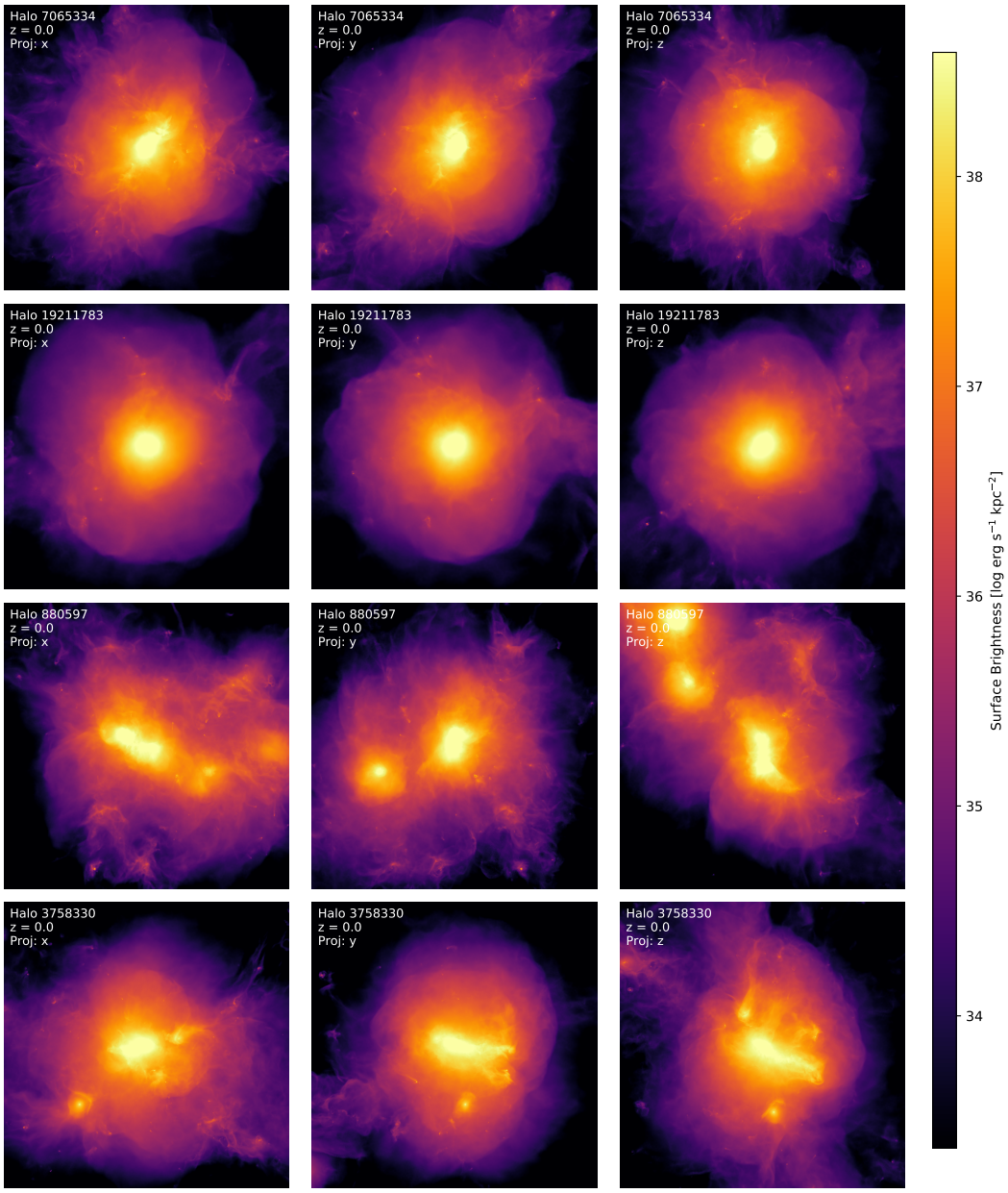

Figure 25: X-ray surface brightness maps (in $\log_{10}(\mathrm{erg\,s^{-1}\,kpc^{-2}})$) for representative clusters across three projections (columns) and four classification groups (rows): relaxed SCC, relaxed NCC, non-relaxed SCC, and non-relaxed NCC. Each projection axis reveals a different morphology due to the triaxial nature of clusters. These differences highlight the statistical independence of the projections, which we treat as separate data points in later analysis.

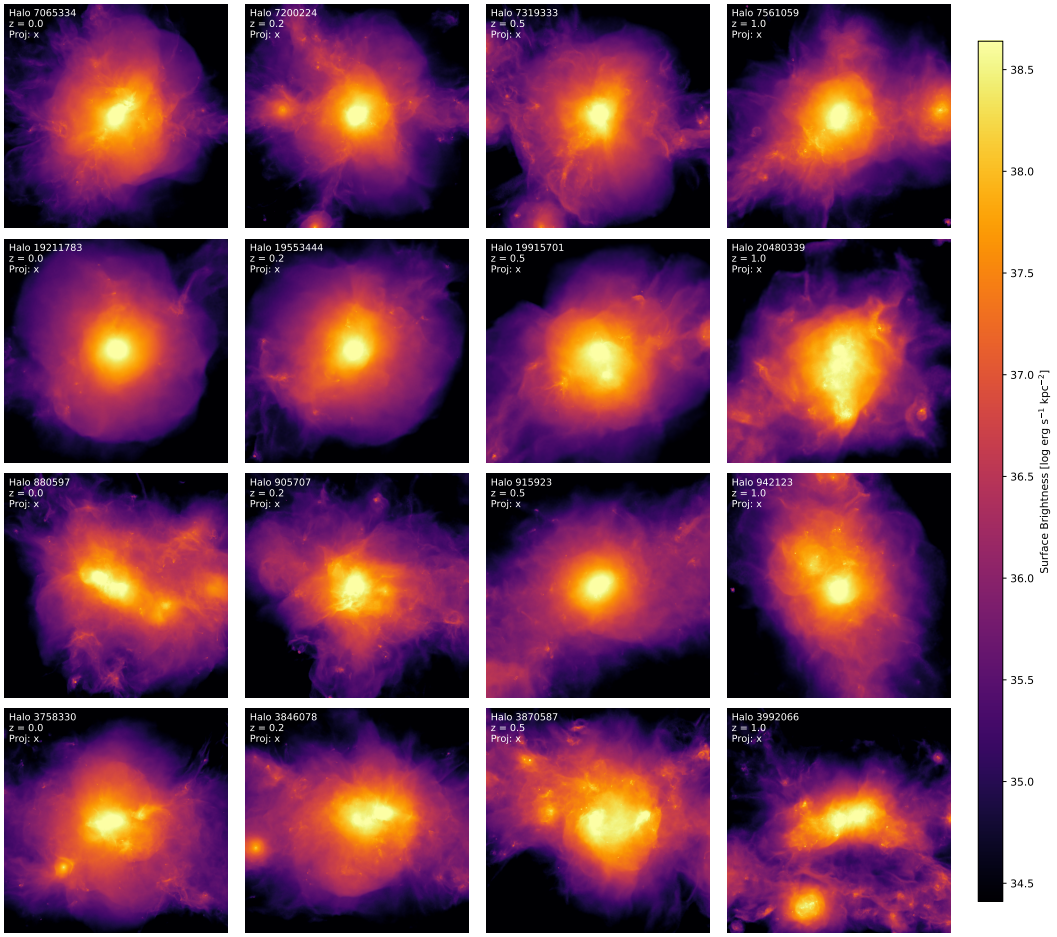

Figure 26: Evolution of the X-ray surface brightness maps (along the $\hat{x}$-axis) for the same four halos shown in Figure 25. Each row corresponds to one halo and the columns show snapshots at redshifts $z = 0$, 0.2, 0.5, and 1. Relaxed SCC clusters maintain regular and centrally concentrated morphologies, while NCC and non-relaxed clusters display disturbed, asymmetric, and evolving structures.

14 CONTRASTIVE LEARNING ON X-RAY MAPS

14.1 DATA PREPROCESSING (X-RAY)

The first step in our contrastive learning pipeline is converting the intrinsic X-ray maps explained in Section 13 into normalized FITS images. This is done to prevent our model to exploit the brightness differences between maps, and instead, emphasize on the structure over absolute brightness. Each halo at a given snapshot, the projected X-ray luminosity is processed into a two dimensional image along a projection axis, and these maps are normalized using the following method which is similar to what is used in Chadayammuri et al. [28].

Each image is normalized such that the 99th percentile pixel value in the central 10% area of the image is set to 1. For getting this central 10% region, a square with length of 31.6% (i.e., $\sqrt{0.1}$) of the full image size is chosen within which the 99th percentile pixel value is chosen. Following this, all pixel values 4 orders of magnitude or more fainter than the central 99th percentile value are set to zero. Our final image, will have all of its pixel values scaled between the minimum value (0) and maximum value (1).

This preprocessing step offers several advantages. From an astrophysical perspective, it effectively suppresses irrelevant background noise and mitigates the influence of extreme dynamic range across images, thereby enhancing the visibility of key morphological features such as cluster cores and merging substructures. From a machine learning perspective, normalizing the input in this way reduces the variance across the dataset, which is known to stabilize contrastive learning objectives and improve convergence. Moreover, by removing global intensity information and enforcing scale invariance, we ensure that the network learns representations based purely on spatial structure rather than being biased by absolute brightness, which in our context is a nuisance parameter rather than a meaningful signal.

The processed maps are then saved as normalized FITS files, which are the input for the self-supervised training pipeline described in the following sections. Each FITS file is a single channel normalized X-ray map of a cluster with its own snapshot, halo ID and projection axis.

14.2 INPUT HANDLING AND DATA AUGMENTATION

To prepare for data augmentation, each normalized array should be first converted into a standard image format to support standard computer vision libraries, such as torchvision [121]. For this purpose each normalized array is first rescaled to $[0, 255]$ range, and then converted to 8-bit unsigned integers. This will result in a single-channel grayscale image that will go through a sequence augmentation that can simulate observational diversity (the set of augmentations that are used here, are the same set of augmentations used in Chadayammuri et al. [28]).

Instead of using a predefined augmentation pipelines (e.g, the ones created for SimCLR in lightly.ai [178]), we design our own set of augmentations customized for this thesis. The augmentation pipeline has three main categories:

SHAPE INVARIANCES:

- **Random horizontal and vertical flips:** Each image is independently flipped along the vertical and horizontal axes with a probability of 0.5 to ensure that the model is insensitive to mirror symmetries.
- **Random rotations:** Random rotation in the range $[-90°, +90°]$ applied to images enhancing model's rotational invariance.

GEOMETRIC TRANSFORMATIONS:

- **Zoom:** Random affine scaling with zoom factors sampled from the range $[0.1, 0.15]$.
- **Affine translations:** Random translations up to 25% of the image size along both axes, accounting for off-centered targets or misalignment.

TEXTURE INVARIANCES:

- **Gaussian blur:** Kernel blur applied with standard deviation randomly sampled in the interval $[0.001, 1.0]$, introducing a range of blurring effects corresponding to variable image sharpness.
- **Gaussian noise:** Additive noise sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, where the standard deviation σ is defined as $\sigma = \mu_{\text{image}}/\text{SNR}$. Here, μ_{image} denotes the mean intensity of the input image, and the signal-to-noise ratio (SNR) is randomly drawn from a uniform range between 4 (relatively strong noise) and 8 (small perturbations and milder noise) for each image. The noise level is proportional to image brightness, maintaining realistic perturbation magnitudes across the dataset.

To generate training pairs for constrastive learning, each image has to go through two random augmentations from the list above. This process is implemented using the `MultiViewTransform` class from the lightly library, where two randomly selected augmentation of each image, make a pair of transformed views (v_1, v_2) [178]. These pairs, remain semantically matched and serve as positive samples for contrastive loss optimization which will be explained in detail in 14.3.

After the augmentations, the images are turned in to PyTorch tensors, which are rescaled back to the range of $[0, 1]$, and normalized by subtracting a mean of $\mu = 0.50$ and divided by the standard deviation of $\sigma = 0.25$. This normalization helps stabilizing the training procedure and a more efficient convergence [56].

14.3 SIMCLR: A CONTRASTIVE LEARNING FRAMEWORK

To learn robust representations of galaxy cluster X-ray maps without using any labels, we use the SimCLR algorithm (Simple Framework for Contrastive Learning of Visual Representations) [29], a self-supervised contrastive learning method. The

core idea of the SimCLR is to train a convolutional neural network to maximalize the similarity between different augmented views of the same images and bring them closer in the representation space, and push the views of different images apart from each other.

In SimCLR, as mentioned in section 14.2, each input goes through two random augmentation and creates a pair of correlated views. These views are passed through am encoder network (mainly a convolutional neural network) to get the representation space. These representations further undergo a small projection head which brings them into a space where the contrastive loss is applied. The contrastive loss function, as the name suggests, encourages the model to maximalize the similarity between the positive pairs while minimizing the similarity with the rest (negative pairs) within one batch [29].

Mathematically, given a batch of N images, SimCLR creates 2N augmented views (2 random augmentations per each image) and treats each positive pair (i,j) as similar, and the remaining $2N-2$ are considered as negative pairs. The contrastive loss function can be achieved mathematically as the normalized temperature-scaled cross-entropy loss (NT-Xent):

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{31}$$

where $\text{sim}(z_i, z_j)$ is the cosine similarity between the projected features z_i and z_j, and τ is a temperature parameter controlling the sharpness of the distribution.

The SimCLR architecture consists of three main components:

1. **Encoder**: A convolutional neural network (here is ResNet-18) that maps input images to high-dimensional representation space.
2. **Projection head**: A multi-layer perceptron (MLP) that projects these representations into a lower-dimensional latent space where the contrastive loss is applied.
3. **Contrastive loss**: The NT-Xent loss that optimizes the latent space by contrasting positive and negative pairs within the batch.

SimCLR does not need a specific architectures or memory banks, and relies mainly on large batch size to provide diverse negative samples within each training step. This framework, enables us to learn a representation space that captures relevant structure information. In this work, SimCLR is used on X-ray maps (this part), radio maps (part vi), and both (part vii). This result in a representation space, which will be the input to the conditional invertible neural network used for posterior inference, which will be described in the next sections.

In this work, we use a modified ResNet-18 architecture, which is pretrained on ImageNet, with the classification head removed. Since our inputs are single-channel FITS images, the first convolutional layer is adapted to accept a single channel rather than three (RGB). This is done by replacing the first layer with a convolutional filter that initialize it by average the original RGB weights across the channel dimensions. The resulting encoder $f(\cdot)$ maps input images into high-dimensional feature vectors.

Following the encoder, we add a non-linear projection head $g(\cdot)$, which maps the features to a 128-dimensional latent space where the contrastive loss is applied.

We use the projection head implementation from the `lightly` library, consisting of two fully connected layers with a ReLU non-linearity in between [178].

14.4 TRAINING PROCEDURE

Our contrastive learning model uses SimCLR framework (section 14.3 which is implemented by `PyTorch Lightning` enabling training across multiple components; data augmentation, model definition and optimization [178]. Training begins by initializing the SimCLR model class. As mentioned in section 14.3, ResNet-18 is the backbone that is used as the encoder with its classification head removed. The first convolutional layer is redefined to accept single channel FITS images as input.

Each FITS image goes through a PyTorch-Compatible dataset class that loads the single-channel images and applies two randomly selected augmentations (section 14.2). This results in a pair correlated views of (v_1, v_2), which makes the positive sample. The augmented samples are organized into batches, which is configures with a custom collate function to return the view pairs and their corresponding filenames. For each batch of size N, the model receives two tensors representing the positive pairs of the same set of N images. We use a batch size of 64, four worker threads for parallel data loading. To make sure that each training epoch sees a varied set of samples, the dataset is shuffled at the start of each epoch.

During training, each augmented view goes through the shared encoder network based on the ResNet-18 network. As mentioned in section 14.3, ResNet-18 is the backbone that is used as the encoder with its first convolutional layer modified for single-channel fits images. The classification head is removed, and the final output feature vector from the network, which will be used in the rest of the thesis, as it captures the morphological features learned by the model.

The representation space will further pass through the projection head, which is consisted of two fully connected layers, to produce a 128-dimensional representation space. This projection head, maps the high-dimensional encoder features into a representation space where the contrastive loss (equation 31) will be applied. After the training, the projection head is discarded, and the encoder is used as a frozen feature extractor to generate representation space for the next steps, including conditional inference of cluster properties.

Given a batch of N samples, the NT-Xent loss is calculated over 2N views (2 augmentations per each). The augmented views are passed through the encoder and the projection head, resulting in an embedding (representation) pair of z_0 and z_1. These are then passed to the NT-Xent loss (equation 31), where for each positive pair, the remaining $2(N-1)$ embeddings (representations) in the batch serves as the negative. This encourages the positive pairs to have similar embeddings (representations) while pushing apart the remaining negative pairs sampled from the current batch. This will shape the encoder's representation space to reflect meaningful structural similarity without using any labels [29].

Because SimCLR pretraining is entirely self-supervised; using only within-batch positives (two augmentations of the same sample) and negatives (the remaining $2(N-1)$ views). Therefor, it does not require a supervised train/val/test split for optimizing the NT-Xent objective. No labels are consumed, and hard negatives are sampled from the current batch, so the representation learning stage can leverage the full unlabeled corpus without risking label leakage. This practice is standard in

contrastive/self-supervised pipelines and is also used directly in Chen et al. [29]. We therefore pretrain the encoder on all intrinsic maps and only introduce train/-val/test partitions in next steps, when training and evaluating the downstream cINN to ensure unbiased assessment.

The optimization process is handled by stochastic gradient descent (SGD) with a momentum coefficient of 0.9 and weight decay of 5×10^{-4} for L2 regularization. A cosine annealing learning rate scheduler is used to reduce the learning rate smoothly from an initial value of 0.06 over the course of 100 epochs. The training loop is executed on a single GPU using bfloat16 mixed-precision computation, which improves memory efficiency while preserving training stability.

The training pipeline is orchestrated using the `Trainer` module from PyTorch Lightning. During training, model checkpoints are saved periodically based on the contrastive training loss with the best three and the final model saved to the disk. After 100 epochs, the final model weights, preserving the trained representation space, are saved for further steps.

14.5 REPRESENTATION EXTRACTION AND POSTPROCESSING (X-RAY)

When the training is complete, we extract and analyze the learned representation space. This process involves three primary stages: generating a representation space from the trained model, projecting it into a two dimensional space using UMAP, and visualizing the representation space via grid and nearest-neighbor plots.

For generating the representation space, the trained SimCLR encoder will be used to extract the vector representation for each input FITS image. We first load the trained SimCLR model, and each test image is passed through the same preprocessing pipeline (section 13. During the inference, the batches are forwarded through the encoder, while the projection head (used exclusively during training) is discarded. The result of this, is a 512 dimension feature vector (representation) for each input, which is flattened into a 1D vector. Representations for each batch are also concatenated across all the batches and ℓ_2-normalized (i.e., each embedding vector $\mathbf{v}_i$ is transformed to $\tilde{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}$) to ensure consistent scale and fair distance-based evaluations.

Since the representation space has a dimension of 512, for visualization we apply the UMAP (Uniform Manifold Approximation and Projection), to projects the high-dimensional representation space into a two dimensional space. UMAP is a non-linear technique applied for reducing the dimensions that preserves both local and global structure modeling the high-dimensional data manifold and optimizing a low-dimensional graph layout [102].

For visualizing we can use grid visualization to help reveal spatial clustering patterns. For this purpose, we normalize the UMAP coordinates to a $G \times G$ grid and assign each image to each corresponding grid. For normalization we use the min-max normalization:

$$x_i^{\text{norm}} = \left\lfloor \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} \cdot (G-1) \right\rfloor, \quad y_i^{\text{norm}} = \left\lfloor \frac{(y_i - y_{\min})}{(y_{\max} - y_{\min})} \cdot (G-1) \right\rfloor$$

This makes sure that each 2D UMAP is assigned to a unique cell on the grid within $[0, G-1] \times [0, G-1]$. A subset of images are chosen randomly to be shown

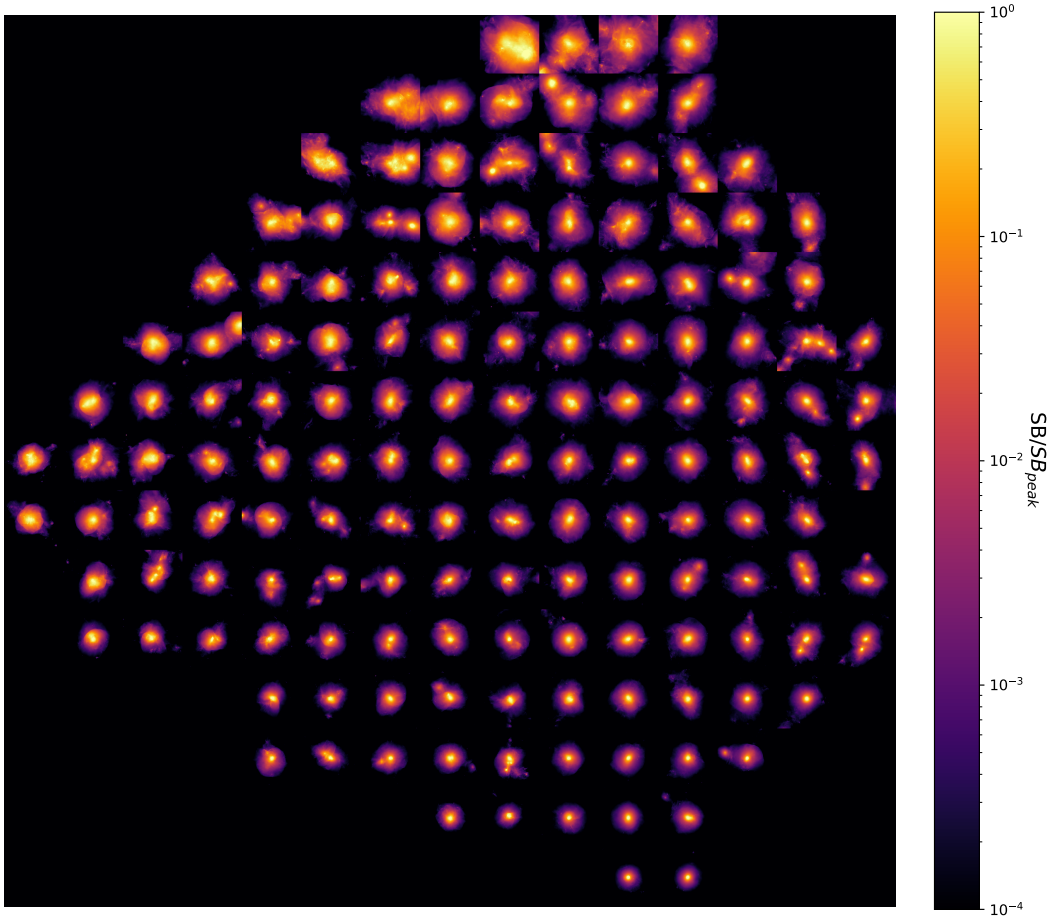

Figure 27: Grid visualization of the learned representation space of X-ray map representation space. Each image corresponds to a UMAP-projected point in the representation space. Clusters with similar morphological features tend to occupy adjacent cells, revealing locally smooth organization in the representation space.

on the grid. These images are then placed on the grid, with their positions corresponding to their grid coordinates.

Figure 27 shows the resulting grid visualization for the learned representation space of the X-ray maps, with grid size $G = 15$. Clear groupings of visually similar X-ray morphologies emerge, suggesting that the encoder has learned to capture meaningful structural information. In particular, adjacent cells often contain clusters with comparable core brightness or elongation, indicating that the representation space is smooth and preserves astrophysically relevant features.

This procedure yields a mainfold in which morphologically similar clusters tend to be placed in neighboring cells, reflecting local continuity in the learned representation space. By visualizing the representation space in this way, we can directly inspect whether clusters with related structural features are organized coherently.

In Figure 27, we see few trends: most of the merging galaxy clusters appear on top while relaxed clusters appear mainly the bottom. It also can be seen that from bottom right to bottom left we go from cuspy to flatter radial profiles, and going from top left to bottom right, the transition from merging system with similar sized clusters to different component sizes on the top right.

To further evaluate the structure of the learned representation space, we perform nearest-neighbor analysis. For a set of randomly selected images, we retrieve their k-nearest neighbors using Euclidean distance in the 512 dimensional representation space. This allows us to evaluate by eyes whether the model has learned meaningful, morphologically coherent representations. Because embeddings (representation space) are ℓ_2-normalized, Euclidean distance is monotonically related to cosine dissimilarity ($\|\mathbf{u} - \mathbf{v}\|^2 = 2(1 - \cos\theta)$), making it an appropriate choice for neighborhood queries. And since the nearest neighbors are chosen based on their distance on the 512 dimensions, it evaluates the representation space without the distortions introduced by nonlinear dimensionality reduction such as UMAP.

In figure 28, we display 5 nearest neighbors for 6 randomly chosen X-ray maps. We can see that the nearest neighbors share similar X-ray morphologies. The nearest neighbor also displays clusters with similar morphologies as nearest neighbors of each other. These visualizations provide intuitive evidence that the SimCLR-based model captures meaningful similarities between galaxy clusters. The learned representation space show local continuity, reflect high-level structure, and form a promising basis for downstream tasks such as clustering, anomaly detection, or supervised fine-tuning.

So far, all of our evaluations have been done without using access to any labels. For the last evaluation plot, we will use the physical properties of the galaxy clusters used in this training. These properties are the observable and unobservable (merger) properties from table 4 and 5 in Figures, 29, and 30, and from table 6 in Figures 31 and 32.

For each selected property, we visualize the 2D UMAP projection of the representation space using a hexagonal binning plot, with the color of each bin representing the value of the physical quantity. For each bin, the average value of the label is computed and shown using a continuous color map. Taken that during the training, our model had no access to any labels (properties), this analysis can show us whether our representation space correlated with key properties of galaxy clusters.

The emergence of smooth, label-correlated gradients over the representation space manifold provides strong evidence that the self-supervised model captures latent physical relationships. These results motivate further downstream tasks, such as cINN only using these representation spaces instead of any observables which we will further discover in the next chapter.

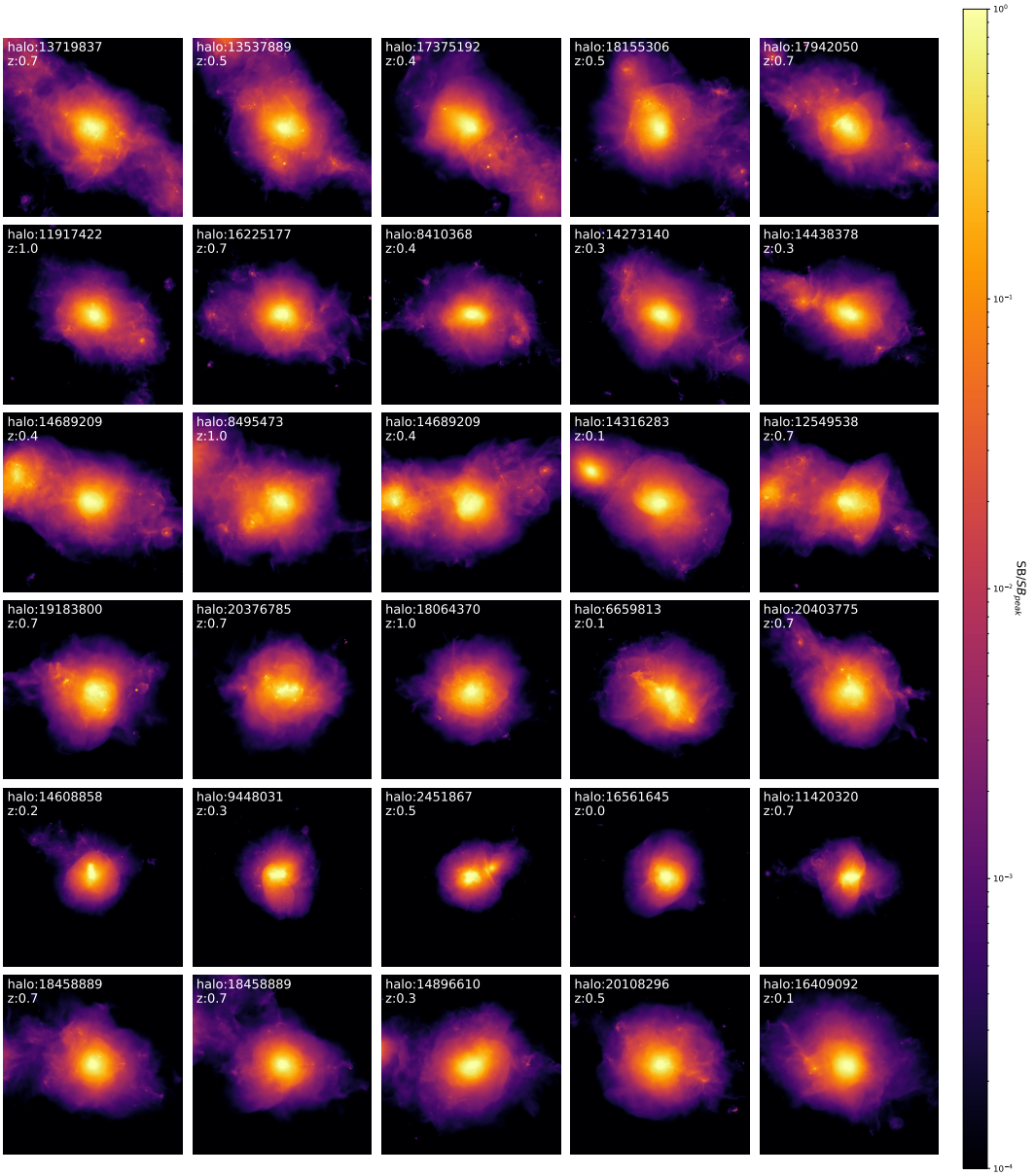

Figure 28: Nearest-neighbor retrieval in representation space. Each row shows one anchor FITS image (far left) and its $k = 4$ nearest neighbors. The learned representations capture structural similarities, with visually similar X-ray morphologies grouped together.

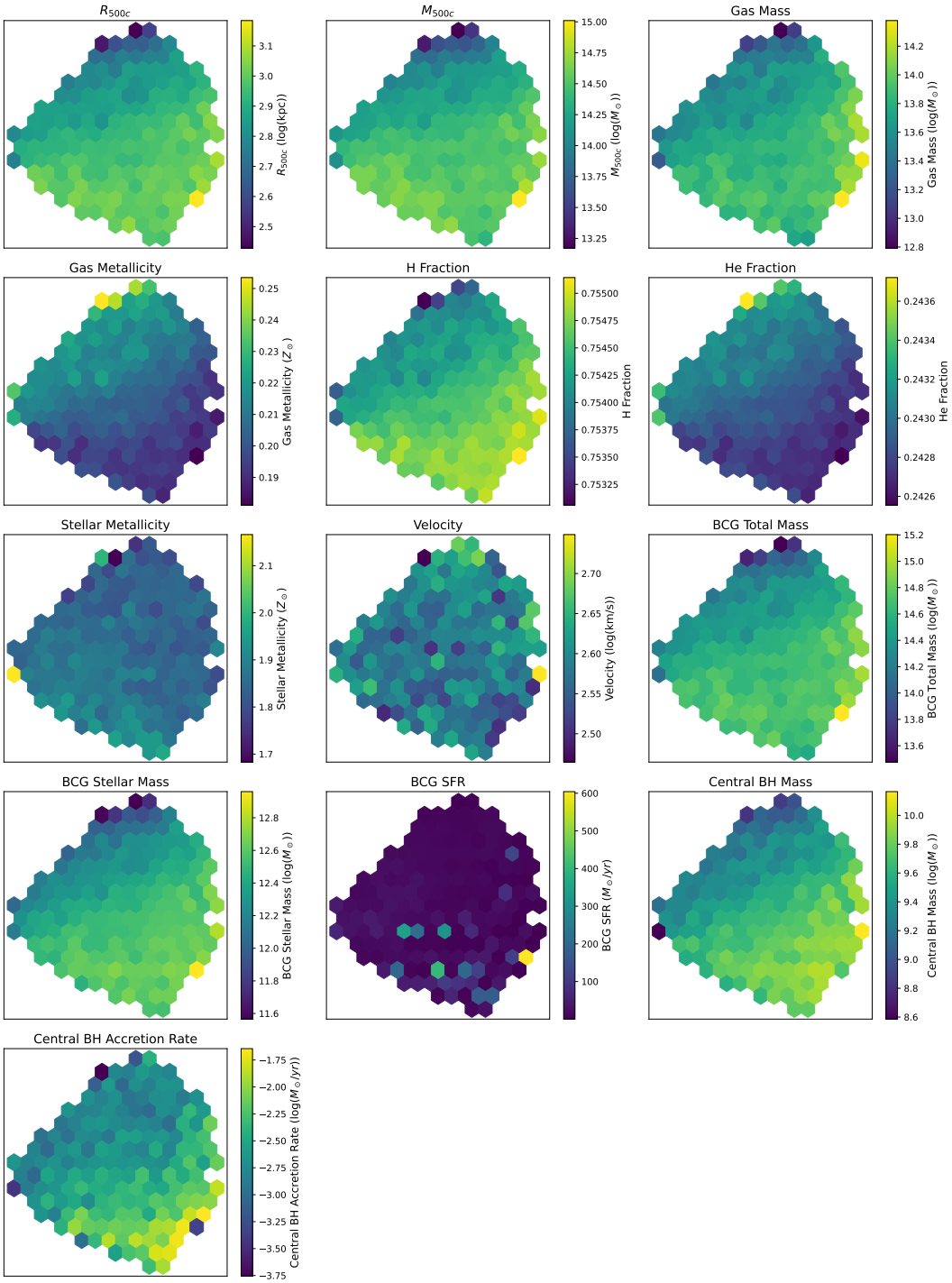

Figure 29: UMAP projection of representation space learned from SimCLR training on X-ray maps, colored by mean binned values of halo and BCG observables (Table 4). Smooth gradients across the manifold indicate that the SimCLR representation encodes global scaling relations, despite the model being trained without access to labels.

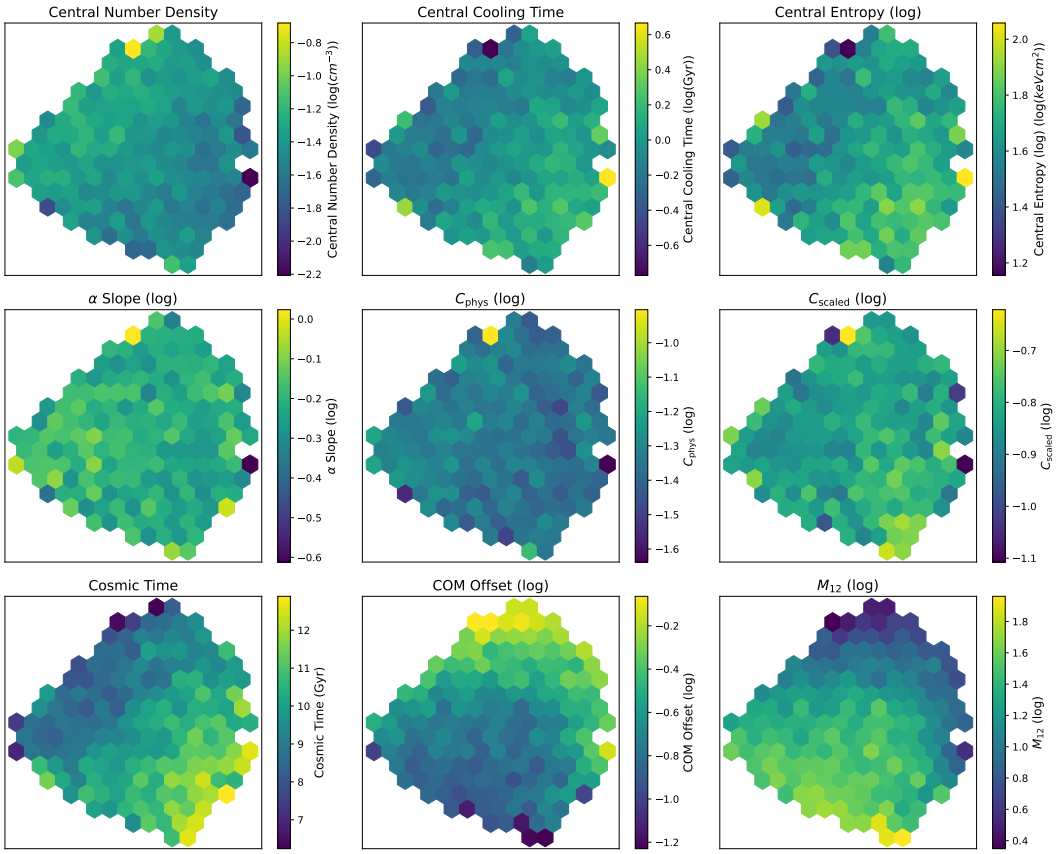

Figure 30: UMAP projection of X-ray representation space, colored by binned mean values of ICM core and dynamical properties (Table 5). Clear trends, show that the representation space captures thermodynamical and dynamical state information.

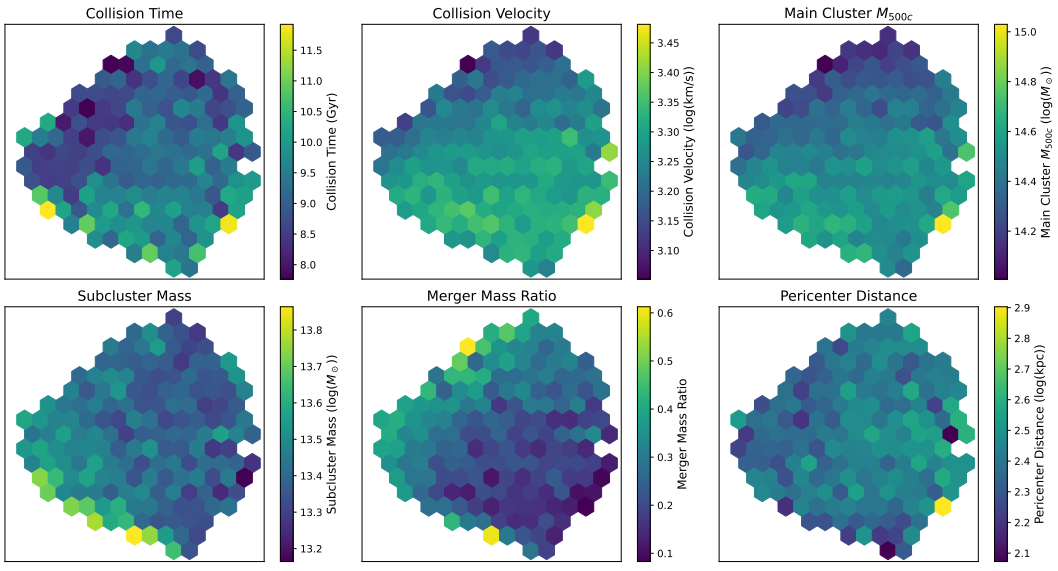

Figure 31: UMAP projection of X-ray representation space, colored by the binned mean values of last–merger parameters (Table 6). Strong coherent gradients suggest that the representation space retains signatures of recent merger activity in the cluster morphologies.

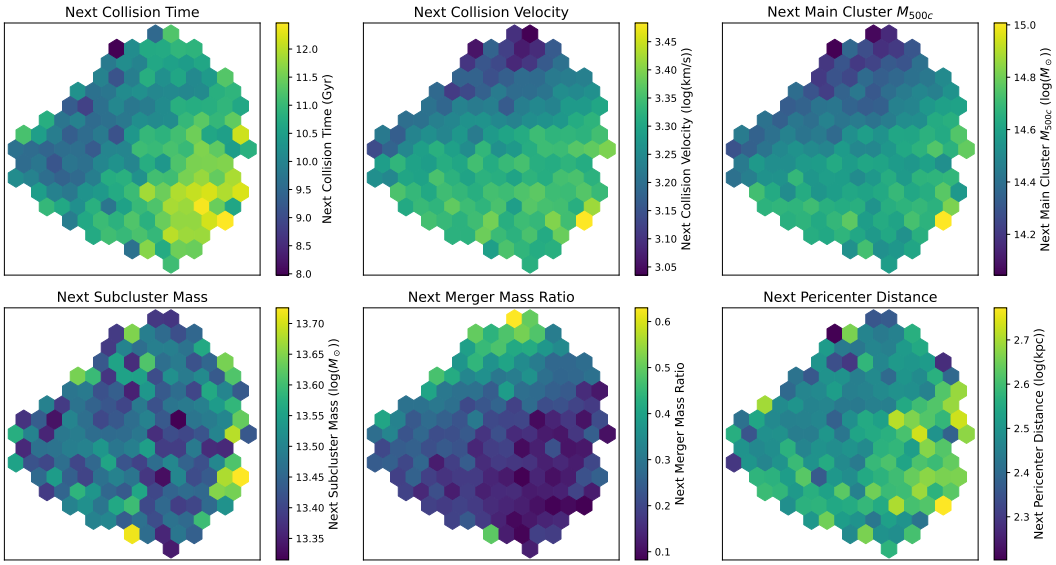

Figure 32: UMAP projection of X-ray representation space, colored by the binned mean values of next–merger parameters (Table 6). The presence of smooth structures indicates that the representation space also encodes information predictive of upcoming merger events.

15

CONDITIONAL INVERTIBLE NEURAL NETWORK WITH X-RAY REPRESENTATION

15.1 DATA PREPROCESSING FOR CINN

In this chapter we try to infer the unobservable properties of merging galaxy clusters by using conditional invertible neural networks (cINNs). The required inputs for this pipeline is the learned representation space from our contrastive learning method (chapter 14), and the derived physical properties of mergers (described in section 9.2). In this section we will take a look at the complete preprocessing pipeline used for preparing the data.

We start by filtering the merging parameters (table 6) similar to what was explained in section 10.1, our sample is made of the galaxy clusters who have gone through at least one merger event based on its definition in 9.2. After filtering, same as section 10.1, the values are scaled to zero mean and unit variance, by subtracting the mean and dividing them by their standard deviation for each target dimension :

$$Y_j^{\text{scaled}} = \frac{Y_j - \mu_j}{\sigma_j} \tag{32}$$

where μ_j and σ_j are the mean and standard deviation of the j-th target variable across all clusters and $\mathbf{Y}_j$ and $\mathbf{Y}_{\text{scaled}}$ are the raw and scaled target respectively.

This transformation ensures that all of the target variables are on a similar scale, which helps the model to train more efficiently and reduces the chance of certain variables dominating the learnign process only for having a larger value. Unlike techniques such as min-max normalization (which forces the values in a fixed range), standardization keeps the original distribution function, which is very important when modeling physical processes.

And in the end, similar to section 10.1, to make future interpretation and evaluations possible, both the standardized target values and the fitted StandardScaler object (containing the original means and standard deviation) are saved. This allows us to convert the model's output back to their original values (before scaling).

The representation space that will be used as the inputs are derived from the SimCLR-trained encoder (14), and saved as an array of shape $(N_{\text{proj}}, D_{\text{emb}})$ where each row corresponds to one projected view of a galaxy clusters. Each of these embeddings (representations) also have a matching filename that encodes the halo ID, snapshot number and the projection axis of the original FITS image. To link these image embeddings (representation) with their target values, a unique key for each data based on its haloID, snapshot number and projection is created.

On the other hand, target merger parameters and metadata (halo ID and snapshot) are initially defined per galaxy cluster without accounting the different projections. To create a one-to-one correspondence between each image embedding (representation) and its physical parameters, the meta data and the target arrays are replicated three times - once for each projection. This results in a fully aligned

dataset with shape (N, D_{emb}) for the embeddings and (N, D_{target}) for the target values, where N is the total number of projected images after replication.

The learned representations generated by passing each image through the encoder network are loaded and matched with their associated halo identifiers, snapshots, and projection axes using a filename normalization routine. This ensures that embeddings are correctly aligned with their respective metadata entries. The embeddings are then stacked into a matrix $\mathbf{E}_{rep} \in \mathbb{R}^{N \times D}$, where D denotes the embedding dimensionality ($D = 512$).

To normalize the representation space and ensure consistent scaling across dimensions, each feature is standardized producing a zero mean, unit-variance version $\tilde{\mathbf{E}}_{rep}$ which ensures that clusters is not biased by the scale differences across features. After that a 80/10/10 split is applied by randomly partitioning the galaxy cluster into training, validation, and test groups. Of the 8448 galaxy clusters, 6192 have experienced a past merger and 5916 will undergo a future merger. These totals correspond to splits of 4953/619/620 for last-merger samples and 4732/592/592 for next-merger samples.

To construct a Mixture-of-Experts (MoE) framework that reflects the structure of the learned representation space, we apply unsupervised clustering [148]. Since our representation space has been scaled, any scale-based biases across different feature dimensions is prevented, and all features are treated equally when measuring the distance. The clustering is done by using the **K-Means algorithm** with $k = 10$ experts, which helps dividing the high-dimensional representation space into smaller and more manageable regions. It is important to note that the clustering step is performed only on the training portion of the dataset. This ensures that the expert definitions are based solely on the data seen during training, preventing any data leakage and the keeping the evaluation reliable [94].

Once the expert centroids are computed using the training data, each data point, whether in the training, validation, or test set, will be assigned to the nearest expert label based on its proximity to the learned centroids. These labels define the mechanism used to direct each sample to the corresponding expert during the MoE training phase. While the full dataset receives expert labels, each expert model is trained only on those training samples whose representations lie within its assigned expert. This strategy allows each expert to specialize in modeling a specific region of the representation while preserving the integrity of the data split and supporting generalization to unseen data.

15.2 MODEL ARCHITECTURE AND TRAINING

The model used in this section, is the exact model used in Part iv and section 11.1. The only main differences will be the dimensionality of the input which would change the subnet input structure.

In this part, the conditional input $\mathbf{c}$ consists of a 512-dimensional representation space from contrastive learning (see Section 14), while the input feature vector stays the same ($\mathbf{x} \in \mathbb{R}^6$) is split evenly such that $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^3$. As a result, the subnet receives $D_{in} = 3 + 512 = 515$ input features.

The output dimensionality D_{out}, is also similar to Part iv, which is $D_{out} = 3 \cdot (3 \cdot 10 - 1) = 87$, where RQS blocks, with $K = 10$ are used. The hidden layer also stays similar with $D_{hidden} = 256$ across all subnetworks.

Thus, in our implementation, each subnet used in a coupling block consists of the following sequence of layers:

$$\text{Linear}(515, 256) \to \text{ReLU} \to \text{Linear}(256, 256) \to \text{ReLU} \to \text{Linear}(256, 87)$$

The training objective is similar to what was described in the scalar training section 11.2; we maximalize the conditional likelihood of the targets igven the conditioning input, by minimizing the NLL loss (equation 28). The training is similar to the scalar pipeline but with the difference that here the condition is the learned representation space. Also in contrast to the scalar pipeline, where the cINN is trained on all clusters, we use a Mixture of Experts (MoE) strategy [148].

Using the MoE training paradigm, a separate conditional invertible neural network (cINN) is trained for each local region of the representation space. Specifically, the conditional space $\mathbb{R}^{D_c}$ is partitioned into $M = 10$ clusters via k-means clustering, using only the training embeddings (as explained in section 15.1) [94, 148]. Each expert $m \in 0, \ldots, M-1$ is assigned the subset of training samples whose conditional embeddings are closest to the m-th expert center. This strategy encourages each expert to specialize in modeling a localized conditional density $p_m(x \mid c)$, improving accuracy in heterogeneous regions of the input space.

Each expert $m \in 0, \ldots, M-1$ is trained independently, so that training yields an ensemble of specialized cINNs rather than a single global model. For each expert, the training and validation loaders are restricted to the training and validation indices belonging to that cluster, while test samples remains defined globally. A minimum threshold of 50 training samples per cluster was adopted to avoid unstable optimization in poorly populated experts; however, in practice all experts in our dataset exceeded this threshold, so no experts were discarded.

For expert m, the training dataset consists of input target pairs (x, c). During each epoch, the training data is divided into batches of size 32 and are passed through the model. For each batch, the forward transformation $z = f(x, c)$ and the the log-determinant $\log|\det(\partial f/\partial x)|$ (section 11.2) are computed to evaluate the NLL loss in accordance with Equation 28. The gradients are backpropogated and used to update the parameters of the expert network using the AdamW optimizer with a learning rate of 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an ϵ value of 10^{-6} [79]. To ensure training stability, gradients are clipped to a maximum ℓ_2 norm of 5.0.

The learning rate is adjusted dynamically using a `ReduceLROnPlateau` scheduler [121], which reduces the learning rate by a factor of 0.8 if the validation loss does not improve for 5 consecutive epochs, with a minimum improvement threshold of 10^{-4}. Each expert is trained for 200 epochs, and at each epoch, the model is evaluated on the validation set using the NLL loss (but with gradient computations disabled). In the end, the model with the lowest validation loss is retained. The training and validation losses are logged over time to monitor the convergence behavior and detect signs of overfitting.

Once training is complete, the model can be used for probabilistic inference. Given a new galaxy clusters, a feature space in the representation space c, the KMeans that was used during preprocessing is reused, and each galaxy cluster is assigned to the expert of the nearest stored center in the representation space. Then the corresponding expert will be used to get multiple latent samples $z \sim \mathcal{N}(0, I)$ and invert them via $x = f^{-1}(z, c)$ to obtain conditional samples from $p(x \mid c)$.

15.3 INFERENCE AND POSTPROCESSING

Once the conditional invertible neural network (cINN) or the ensemble of expert models in the Mixture of Experts (MoE) framework has been trained, the learned mapping $f(x, c)$ can be inverted to perform probabilistic inference on the merger parameters x conditioned on new input features c. Postprocessing begins by generating samples from the learned conditional distribution, $f^{-1}(z, c) = x$, and by doing this multiple times, we get the $p(x \mid c)$.

Similar to what we had in Section 11.3, to approximate the posterior distribution $p(x \mid c)$ using a trained cINN model, the sampling procedure is straightforward: for each representation c_i, we draw $z_j \sim \mathcal{N}(0, I)$ for $j = 1, \ldots, N$, and compute $x_j = f^{-1}(z_j, c_i)$. In the MoE setting, however, the conditional space is partitioned into M experts. Each representation c_i, is assigned to the nearest expert center using Euclidean distance, and the corresponding expert is used to generate posterior samples.

The rest of the postprocessing, including calculating the MAP, the prior distribution, etc. is similar to what was explained in the postprocessing part for scalar conditions, as in section 11.3.

16

RESULTS AND DISCUSSIONS (X-RAY CONDITIONING)

16.1 POSTERIOR DISTRIBUTIONS WITH X-RAY CONDITIONING

We repeat the visualization of conditional posteriors $p(\mathbf{x} \mid \mathbf{c})$ for a subset of test galaxy clusters, now conditioning on the learned X-ray representation (embedding) $\mathbf{c}$ rather than scalar observables (as in section 12.1). From the saved test indices we randomly select $n_{\text{rows}} = 15$ clusters from the test set; for each condition $\mathbf{c}_i$ we draw $n_{\text{sam}} = 1000$ posterior samples $\{\mathbf{x}_i^{(s)}\}_{s=1}^{n_{\text{sam}}}$ via the inverse flow (Sec. 11.3), map samples to physical units, and arrange one target per column. Rows correspond to distinct clusters (annotated with `HaloID` and redshift z on the left), columns to merger parameters (Section 9.2). Within each panel we overlay the same four elements defined in Section 11.3: a gray prior KDE (test-set marginal for context), a blue posterior KDE for the chosen cluster, a gold MAP vertical line, and a red ground-truth line. As before, prior and posterior KDEs are peak-normalized (only shapes and locations are comparable), and n_{sam} controls Monte Carlo smoothness (we use $n_{\text{sam}} = 1000$ here as well).

Compared to the scalar-conditioned case (Figure 17), the X-ray representation space provides a richer conditioning signal: the blue posteriors contract around the red truths across *all* targets, including those previously challenging (Merger Mass Ratio, Subcluster Mass, Pericenter Distance), meaning that the inference is more precise for most merger properties. The contraction is stronger than scalar conditioning in most cases, except Collision Time. MAP markers (gold) coincide with the truths in most cases, indicating improved identifiability under representation conditioning. In other words, although the posteriors remain comparatively wide, they are well-centered on the ground truth, yielding accurate (low-bias) MAP estimates despite residual uncertainty.

16.2 PREDICTION PERFORMANCE OF THE CINN CONDITIONED ON THE LEARNED REPRESENTATION SPACE FOR X-RAY MAPS

We repeat the evaluation of section 12.2, now conditioning on the learned representation space of X-ray maps. Figures 34 and 35 are constructed identically: for each target x_d we bin the ground–truth axis into $B = 20$ equal–width bins and, for every test object in a bin, draw $n_{\text{sam}} = 500$ posterior samples with the inverse flow (Sec. 11.3). Stacking those samples into the *same* binning yields a 20×20 heatmap in value space. On the figures we have overlaid the white diagonal: $y = x$, with posterior medians (solid black) and 10–90% quantiles (dashed). Figure 35 shows MAP vs. truth with bin–wise medians and 10–90% envelopes, and the corresponding relative errors $\Delta = 100(\text{MAP} - \text{truth})/\text{truth}$.

Relative to scalar conditioning in section 12.2, the learned representation space of X-ray maps provides a stronger, more discriminative context: posterior ridges are slightly thinner in most targets except Collision Time and their median follows the $y = x$. Although it does not align exactly on the identity line, but the

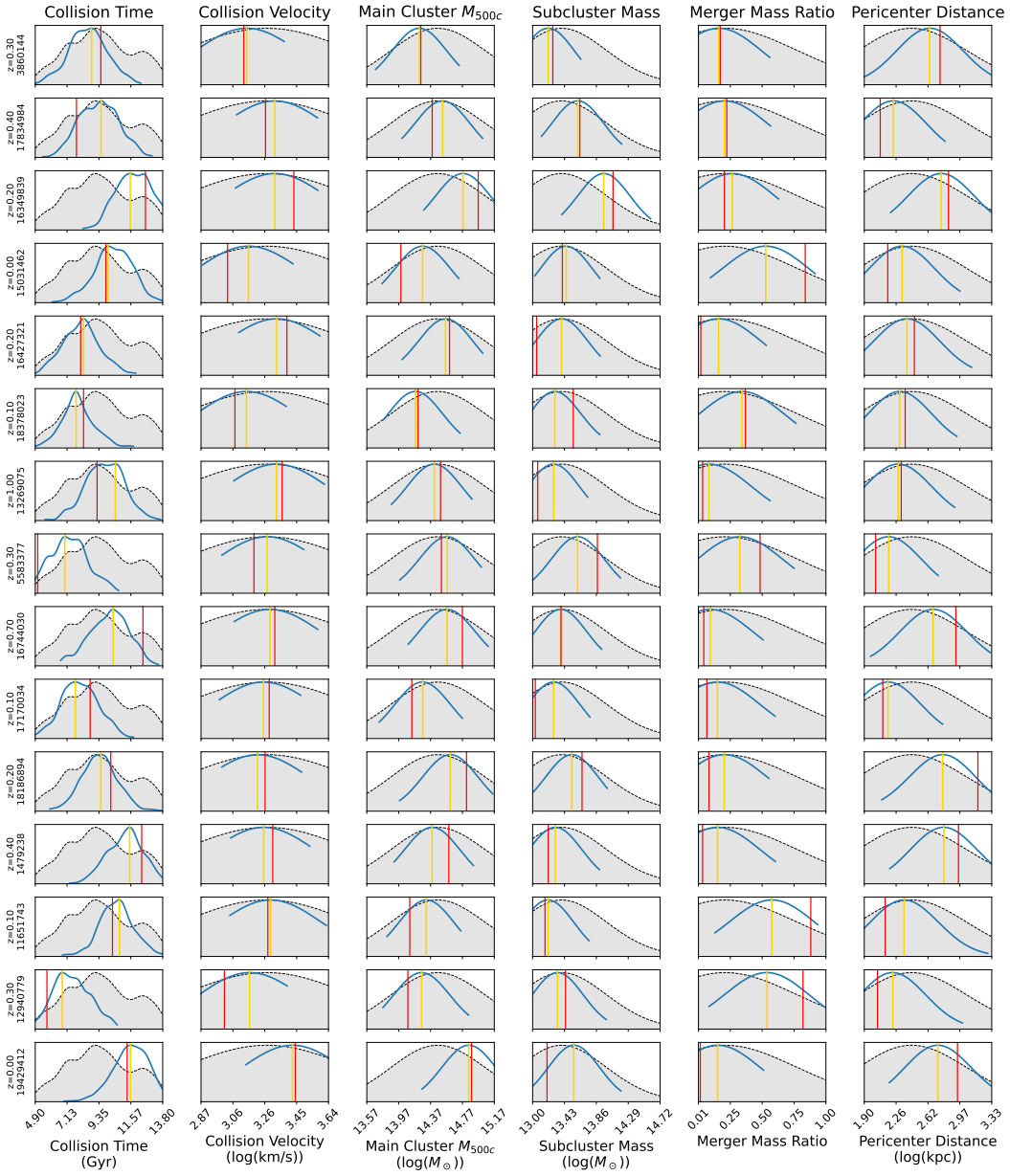

Figure 33: X-ray Representation conditioned posterior grids for 15 randomly selected test clusters (rows) from 620 test samples, across all target merger parameters (columns). Gray: prior KDE over the test split; blue: posterior KDE conditioned on the embedding; gold: MAP estimate (vertical line); red: ground truth (vertical line). Construction mirrors Figure 17, now with the learned SimCLR representation space of X-ray maps as the conditioner.

linear increasing of the median line and the percentile lines shows better performance than the scalar conditioning in previously challenging parameters. MAP medians lie close to the identity for *all* targets, and the errors are mainly small for most cases but Collision Time and Merger Mass Ratio which is high due to the discrete nature of time, and approaching zero values. A mild, uniform *shrinkage* (regression–to–the–mean) remains visible as a slight bending of medians toward the global modal scale, but its amplitude is small.

Here we summarize the per-merger parameter cINN inference results:

- *Collision Time:* Both the posterior and the MAP median, and their percentile lines follows the identity line. While the scatter of the posteriors are slightly larger than the scalar version, the MAP scatter is lower with typical relative errors within $[-20, 40]$ % with only a faint positive bias at the extremes. The relative higher error compared to the rest of the merger parameters is expected because time is sampled at discrete snapshots; percentage errors are less forgiving under discretizations, even when absolute errors are small.

- *Collision Velocity:* The calibration in Collision Velocity is better than the scalar conditioning performing better in lower velocities, with slightly tighter posteriors. The MAP estimation shows lower scatter than posteriors, but both have their median and percentile lines increasing linearly almost along the identity lines without showing any signs of heteroscedasticity for lower velocities (as of the case of scalar conditioning). The MAP also has errors mainly $\leqslant \pm 5\%$ which is slightly smaller than the scalar conditioning. The median shows a small percentile regression toward the global modal scale at $\log_{10} v \approx 3.2\,\mathrm{km/s}$ (as in the scalar case), with slight overestimation below the mode and underestimation above it.

- *Main Cluster* M_{500c}*:* The posterior calibration is performing almost similar to scalar conditioning, with slightly better performance in lower masses and also tighter posteriors. The MAP estimaton also performs similar to scalar condition with narrower quantiles around lower masses, but similar relative MAP error $\sim \pm 2\%$. A very small bend toward the global mode at $\log_{10}(M_{500c}/M_\odot) \approx 14.5$ remains visible, but its amplitude is negligible relative to the total dynamic range.

- *Subcluster Mass:* Significantly stronger performance compared to scalar conditioning for higher masses, with better calibration, tighter posterior and quantiles for both posteriors and MAP estimations. Relative MAP errors typically within $\sim \pm 2\%$. The median crosses $y = x$ near the modal scale $\log_{10}(M_{sub}/M_\odot) \approx 13.5$, showing the same mode-centered shrinkage pattern seen with scalar conditioning, but at much reduced magnitude.

- *Merger Mass Ratio:* Calibration performs significantly stronger toward higher ratios with tighter posterior distributions and without heteroscedasticity. However, it is still the hardest due to its fractional/bounded nature; errors are larger in relative terms at small ratios (denominator effect), but the MAP median nonetheless increase linearly almost close to the identity lines.

- *Pericenter Distance:* Stronger posterior calibration and tighter posteriors with tighter envelope compared to the scalar conditioning. The MAP estimation

is increasing linearly almost along the identity line with MAP errors around $\pm 10\%$. A gentle S-shaped median and a sign flip in the MAP bias around the modal scale $\log_{10}(d_{peri}/kpc) \approx 2.4$ indicate mild shrinkage toward the mode.

Overall, learned representation space of X-ray map conditioning yields posterior calibrations that are close to ideal (narrow, diagonal–aligned 2D histograms) and MAP accuracies that are uniformly strong, with only small, interpretable regression–to–the–mean effects across all merging parameters. The great performance can be explained by the relatively smooth gradient in the representation space manifold, as in figure 31.

16.3 CROSS CORRELATIONS: X-RAY CONDITIONED INFERENCE

In addition to the scalar conditioning explored above, we now assess whether the cINN trained on X-ray maps is able to learn the cross correlations among merger parameters. As in Section 12.3, we visualize all pairwise relations between the merger parameters in a corner plot. For each test object, we draw $n_{sam} = 200$ posterior samples, and plot the pooled posterior realizations (blue), MAP estimates (gold), and ground truths (red). On the diagonal, we include the one-dimensional KDEs of the corresponding marginals for posterior, MAP, and ground truth.

The interpretation of this visualization follows the same principles as in the scalar case. Alignment between gold and red clouds indicates accurate MAP recovery, while elongated blue posterior structure aligned with the red locus indicates that the model has captured the correct correlation between parameters. Systematic displacements of gold relative to red point to bias, and dispersed or multi-clumped blue structures reflect residual ambiguity or multi-modality in $p(\mathbf{x} \mid \mathbf{c})$.

The correlation structure evident in Fig. 36 is consistent with the expectations summarized in Section 12.3. Posterior samples, MAP estimates, and ground-truth values collectively reproduce the qualitative trends anticipated from ΛCDM scaling arguments. In particular, the X-ray conditioned inference successfully recovers not only the correct one-dimensional marginals but also the underlying cross-target correlations among merger parameters.

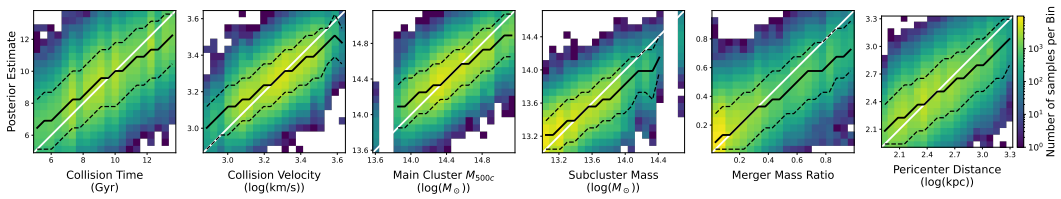

Figure 34: Posterior versus ground truth for merger parameters under X-ray representation conditioning. Construction as in Figure 18 with $B = 20$ and $n_{sam} = 500$ over the 620 test clusters. The white diagonal represents $y = x$. Black solid lines: posterior medians; black dashed lines: 10–90% quantiles. The histograms are rather wide, but the median mainly follows the diagonal without signs of heteroscedasticity, indicating relatively good calibration and dispersion control. The conditioning input is the representation space learned via SimCLR on intrinsic X-ray maps, as explained in Chapter 14.

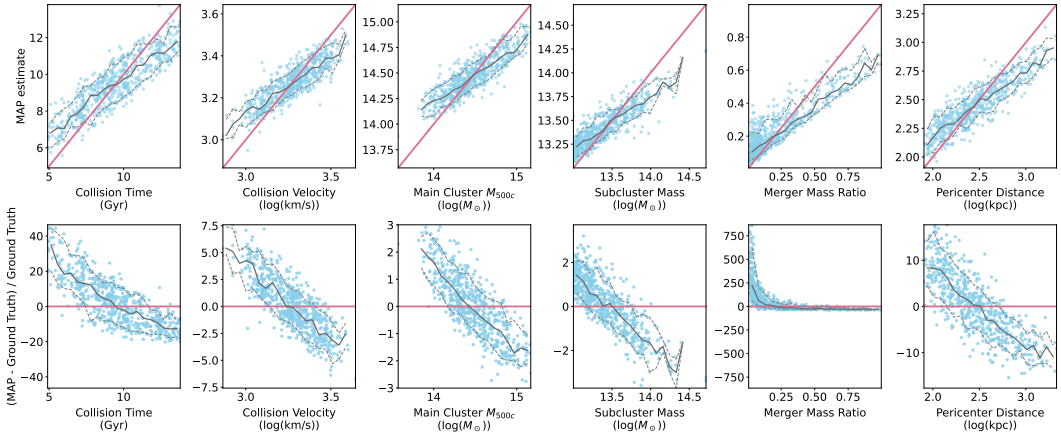

Figure 35: MAP performance of the 620 test clusters under X-ray embedding conditioning. *Top*: MAP estimate versus ground truth, with bin wise medians (black solid) and 10-90% quantiles (black dashed). The pink diagonal represents the $y = x$. *Bottom:* relative MAP error $\Delta = 100(\text{MAP} - \text{truth})/\text{truth}$, with the same line style. The pink horizontal line represents the ideal case of zero relative error. Median lie near $y = x$ (top) and near $\Delta = 0$ (bottom), with tight 10–90% envelopes. The conditioning input, is the representation space learned via SimCLR on intrinsic X-ray maps as explained in Chapter 14

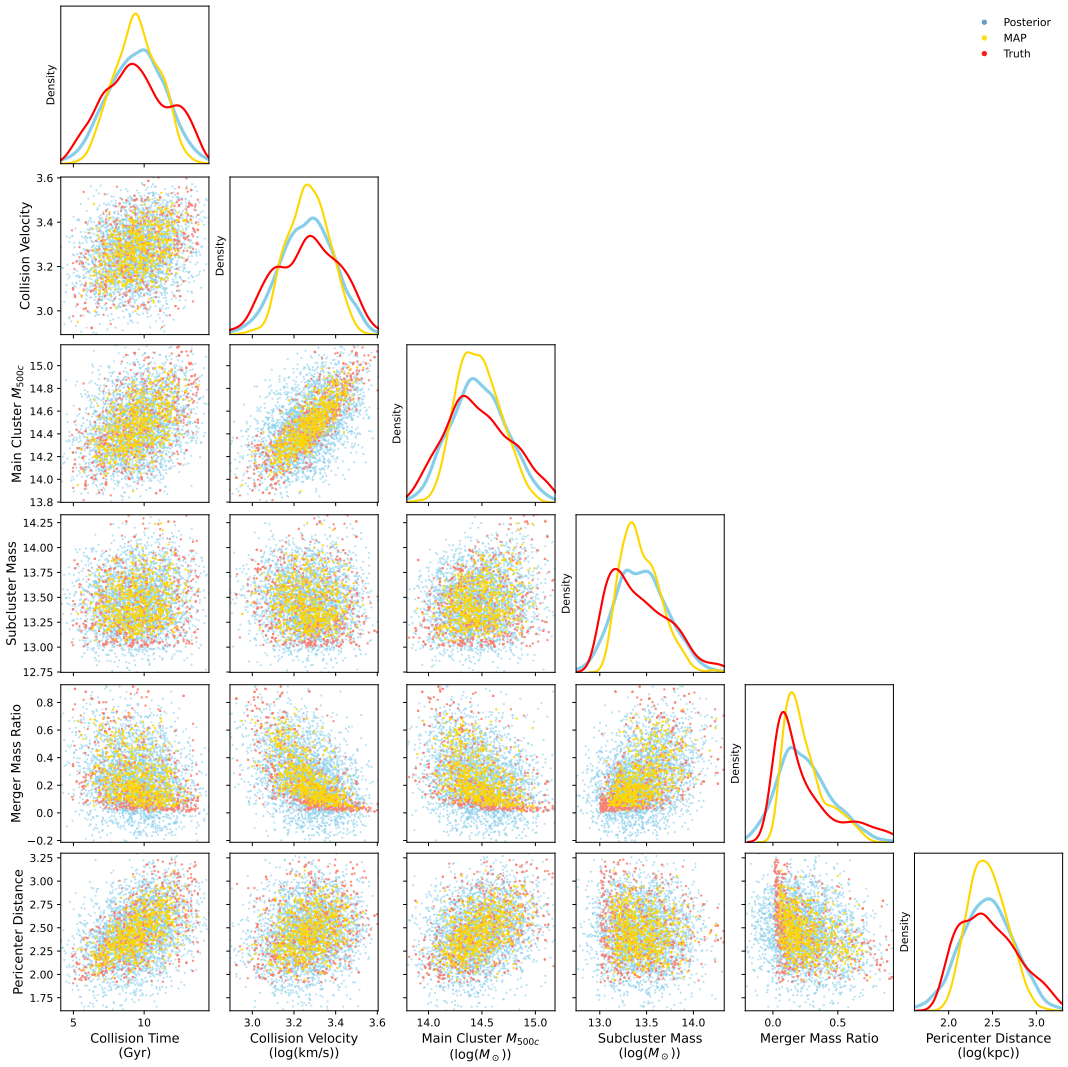

Figure 36: Corner plot across all merger parameters for X-ray conditioned inference. Diagonal: marginal KDEs of posterior (blue), MAP distribution (gold), and ground truth (red). Lower triangle: pooled posterior samples (blue), MAPs (gold), and truths (red) for each test object. The figure demonstrates that the cINN trained on X-ray maps' representation sapce captures both the marginal distributions and the cross-target correlations among merger parameters.

16.4 NEXT–MERGER INFERENCE WITH X–RAY REPRESENTATION SPACE CONDITIONING

We repeat the X–ray–conditioned analysis for the *next* merger (future event), using the same plotting protocol as in Sections 16.1 and 16.2. For each selected test cluster we sample $n_{sam} = 1000$ draws from $p(\mathbf{x} \mid \mathbf{c})$ via the inverse flow (section 11.3) and visualize per–object posteriors as in the previous section.

Figure 37 shows 15 randomly chosen test clusters (rows) and one target per column, with prior (gray), posterior (blue), MAP (gold), and ground truth (red). Relative to the *last*–merger case (Figure 33), the overall picture looks the same but with slightly broader posteriors, reflecting the increased uncertainty when forecasting forward in time. Aside from this broadening, the qualitative behavior mirrors the last–merger results.

Figures 38 and 39 summarize calibration and point–estimate accuracy across the test set using $B = 20$ truth bins and $n_{sam} = 500$ posterior draws per object (as before). The 2D histograms remain aligned with $y = x$; the posterior medians (solid black) increase linearly close to the identity line but with regression to the mean stronger than the last merger. The 10–90% bands (dashed) increase linearly but are slightly wider than in the last–merger case, again most visibly for Collision Time.

The MAP estimate has lower scatter than the posteriors. The MAP median increase linearly but again with regression to the mean across all merger parameters. The MAP errors are almost the same as the last merger case, with the difference that the MAP error in Collision Time is slightly lower than the last merger case; in next merger the relative map error is $\pm 20\%$ while in the last merger, the error was in range $[-20, 40]\%$. It seems like that the MAP estimates for early mergers, are more accurate for future merger events than previous merger history. The reason for this can hide in the mode of next mergers lying in later times ($\sim$ 12.4 Gyr) than the last merger ($\sim$ 8.7 Gyr) as can be seen in table 7, and since Collision Time like the rest of merger parameters suffers from regression to the mean, this can be the cause for larger error.

Per–target error ranges (next–merger, X–ray). Quantitatively, errors are *very similar* to the last–merger representation space case (Sec. 16.2), with only mild broadening where expected:

- *Collision Time:* The calibration is slightly worse than the last mergers, but with the slightly wider quantile bands. The MAP estimate's performance is similar to the last merger, with the difference that we have relative errors $\approx \pm 20\%$ which is smaller of earlier mergers. This is expected, as for the last merger, the recorded mergers are for earlier universe, which can be the reason of this uncertainty.

- *Collision Velocity:* Similar calibration and percentile bands compared to the last merger, with MAP errors mainly $\leqslant \pm 5\%$ which is just very slightly lower than last merger.

- *Main Cluster* M_{500c}*:* Both Posterior calibration and the MAP estimation performance are similar to the last merger.

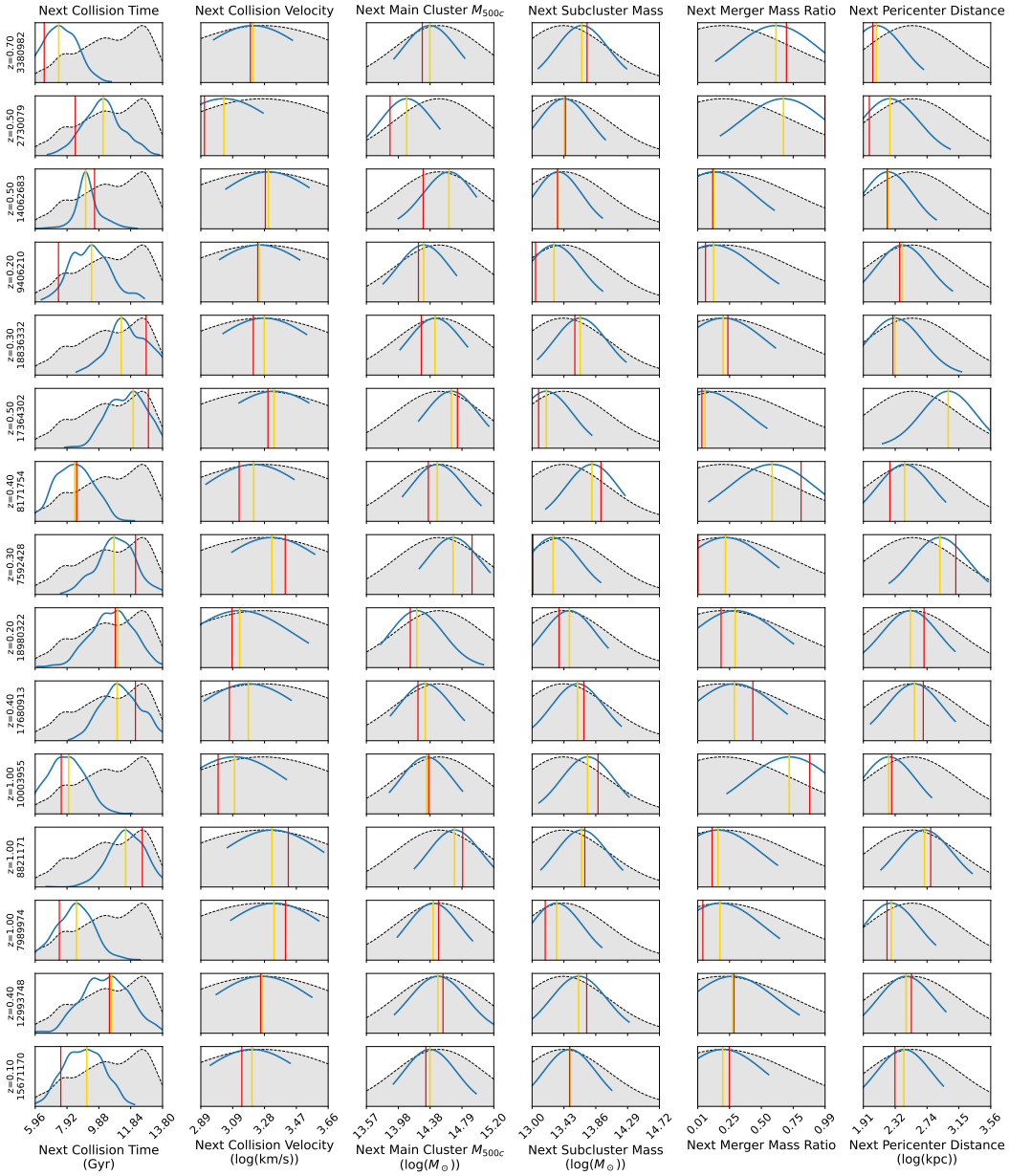

Figure 37: X–ray representation space–conditioned posteriors for the *next* merger (15/592 test clusters). Construction mirrors Figure 33. Gray: prior KDE; blue: posterior KDE; gold: MAP estimate; red: ground truth. Compared to last–merger inference, posteriors are slightly broader, particularly for Collision Time, consistent with increased forward-prediction uncertainty.

- *Subcluster Mass:* Both Posterior calibration and the MAP estimation performance are similar to the last merger.
- *Merger Mass Ratio:* remains the most delicate (fractional, bounded); relative errors inflate at small ratios, but the MAP estimation performance and the posterior calibration remains identical to the last merger.
- *Pericenter Distance:* Both Posterior calibration and the MAP estimation performance are similar to the last merger.

In short, conditioning on the X–ray representation space continues to yield relatively well–calibrated next–merger inferences. The only systematic change with respect to the last–merger is a modest, physically expected widening of the posterior bands, mainly for Collision Time.

16.5 DISCUSSION

X–ray representation space yield well–calibrated posteriors whose medians track $y = x$ across targets (Figures 34, 35). Typical MAP scatter is: with characteristic MAP error ranges: Collision Time ($\sim [-20, 40]\%$), Collision Velocity ($\leqslant \pm 5\%$), Main Cluster M_{500c} and Subcluster Mass ($\sim \pm 2\%$), Pericenter Distance ($\sim \pm 10\%$). Mass Ratio remains the most delicate (bounded/fractional) with larger relative errors at small ratios (Figs. 34, 35). As seen in Figure 33, posteriors are contracted relative to the empirical prior and vary across different clusters, indicating that X-ray representation carries genuine conditioning signal. Posterior medians show gentle curvature toward modal scales (classical regression-to-the-mean), but the calibration diagnostics remain sound: median posterior increase linearly along the identity line with small offsets. This indicates the representation space of X-ray maps carry real information rather than merely reproducing prior structure.

While we see wider posteriors (compared to what we achieve in part vi) across most merger parameters, the envelopes containing 80% of the posteriors were consistently small, and the MAP estimation relative error is reliably low. This performance however can be expected from relative smooth transition in Figure 31 across most merger parameters.

WHAT "SMOOTH TRANSITIONS" MEASURE. Coloring the representation space by a target and observing gradual color gradients ("smooth transitions") indicates that nearby representations tend to share similar target values (Figure 31). This is reassuring, but there is a caveat that applies: UMAP is a *nonlinear 2D projection* of a 512-D space and may distort distances and gradients, especially globally; even if a gradient looks ragged in 2D, the *full* 512-D geometry can still be well organized. Hence, imperfect smoothness in UMAP does not necessarily imply poor conditioning signal for the inference model. And in the end, the cINN consumes the full representation space, not the UMAP, so imperfect smoothness in projection is not diagnostic of poor conditioning.

WHAT THE CINN ACTUALLY LEARNS. The cINN, as explained completely in chapter 11.1, is a single, *joint* density model over the full target vector $\mathbf{x} \in \mathbb{R}^6$ (collision time, velocity, masses, mass ratio, pericenter) conditioned on the repre-

sentation $\mathbf{c} \in \mathbb{R}^D$. It learns an invertible map $f : (\mathbf{x}, \mathbf{c}) \mapsto \mathbf{z}$ with tractable Jacobian so that

$$\log p(\mathbf{x} \mid \mathbf{c}) = \log p_Z\big(f(\mathbf{x}, \mathbf{c})\big) + \log\left|\det \frac{\partial f}{\partial \mathbf{x}}\right|,$$

with $p_Z = \mathcal{N}(0, I)$. Each coupling block splits $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_b]$ and transforms $\mathbf{x}_b$ via a *conditional rational–quadratic spline* whose parameters are predicted from $(\mathbf{x}_a, \mathbf{c})$; masks and permutations alternate across blocks, yielding a flexible factorization $p(\mathbf{x} \mid \mathbf{c}) = p(x_{\pi_1} \mid \mathbf{c})\, p(x_{\pi_2} \mid x_{\pi_1}, \mathbf{c}) \cdots$. Thus, *all targets are modeled jointly* and their correlations are built into the likelihood.

WHY ACCURACY CAN REMAIN HIGH DESPITE RAGGED UMAP GRADIENTS. Because the cINN learns $p(\mathbf{x} \mid \mathbf{c})$ jointly, it exploits *cross-target structure* that was shown to hold in Figure 36. If, for a given region of $\mathbf{c}$-space, some parameters (e.g., masses, velocity) are tightly organized and strongly correlated with others (e.g., pericenter, time), the flow can leverage those correlations to sharpen posteriors even when the UMAP color map for a specific parameter looks less smooth. In other words, the relevant information may be distributed across several parameters and along directions that UMAP compresses.

The observed well-calibrated, narrow posteriors and small MAP scatter pattern for many targets under X-ray conditioning despite less crisp 2D gradients for Collision Time and Pericenter Distance. This is consistent with (i) informative structure residing in the *full* 512-D representation space, and (ii) the cINN's ability to encode the joint covariance of the targets (merger parameters).

REGRESSION TO THE MEAN VS. CALIBRATION. Gentle curvature of MAP bin medians toward the population mode reflects finite-sample regularization and overlapping morphologies: many distinct merger settings produce similar images, so the posterior mean/median bends toward high-density regions. The cINN rightly spreads probability mass across these alternatives (wide intervals), yet still places the mode near the truth when certain features anchor a high-likelihood explanation. Our diagnostics show that this shrinkage coexists with good *calibration*: median posterior curves follow $y = x$, and the envelopes including 80% of the posteriors are noticeably tighter around the identity line. Hence the broadening or shrinkage is a faithful expression of uncertainty, not a modeling flaw.

WIDE POSTERIORS AND SMALL MAP ERRORS. We have seen in Figures 34 and 35, wide posteriors and small MAP errors. The width reflect genuine non-identifiability in the image-to-physics mapping, which can include projection ambiguities, snapshot discretizations of times, and overlapping morphologies can all admit multiple plausible solutions. Our cINN captures this by allocating probability mass across those alternatives. At the same time, the MAP of that distribution can sit very close to the truth because some features in the representation space still anchor a high likelihood explanation. a low MAP error does *not* imply the parameter is tightly constrained, decision-making should use the full posterior, not just the point estimate. Conversely, preferring a sharper but miscalibrated posterior could hide real degeneracies. Our results therefore indicate that the model is both *accurate* (mode near the truth) and *honest* about uncertainty (wide intervals where the data are intrinsically ambiguous).

ROLE OF THE MIXTURE-OF-EXPERTS (MOE). The MoE partitions **c**-space and trains *local* joint flows. This reduces global function complexity, and could separate flat core and cuspy profiles, merging clusters with similar or different mass components, and relaxed or disturbed clusters. This could improves calibration and sharpness where a single flow would blur multi-modal structure. Expert domains align with coherent bands in the representation space, supporting this interpretation.

INTRINSIC MAPS, TRANSFER, AND AUGMENTATION. Because our inputs are *intrinsic* (no PSF/beam, background, Poisson statistics, or uv-coverage), we rely on a physics-aware SimCLR augmentation policy (Section 14.2) to inject observing-like nuisance variation while preserving merger morphology. flips/rotations build orientation invariance; affine zoom and translations mimic centering/scale errors; Gaussian blur approximates PSF/beam smearing; and additive Gaussian noise with SNR sampled in $[4, 8]$ introduces variable depth. That said, augmentation is necessary but not sufficient: realistic forward modeling (e.g., X–ray: Poisson counting, instrumental/background components, exposure/vignetting, PSF convolution, and bandpass/K–corrections) and/or unlabeled fine-tuning on real data remain essential for full simulation to observation transfer.

NEXT MERGER. The next-merger results closely mirror the last-merger case as discussed above, but with expected modest broadening, most visibly for time of collision. Similar to last merger, the median remains close to the identity lines, however with a stronger regression to the mean. Overall, X-ray conditioning, as expected form Figure 47, remains reliable for forecasting, with honest uncertainty inflation of posteriors when forward prediction in intrinsically harder.

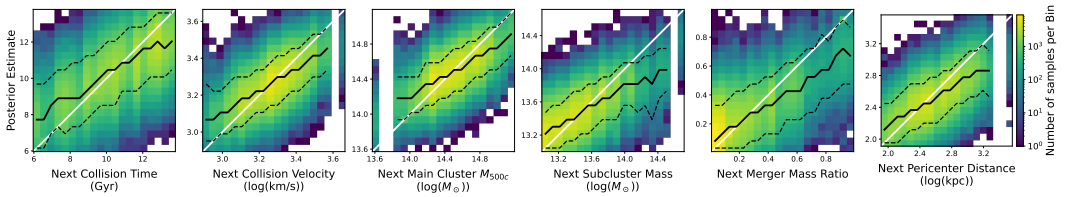

Figure 38: Next–merger posterior versus truth for each merger parameter under X-ray representation conditioning for the 592 test clusters. Same construction as Fig. 34 with $B = 20$, $n_{sam} = 500$. White diagonal: $y = x$. Black solid lines: posterior medians; black dashed lines: 10–90% quantiles. Medians remain close to $y = x$; bands are modestly broader than for the last–merger, chiefly for Collision Time.

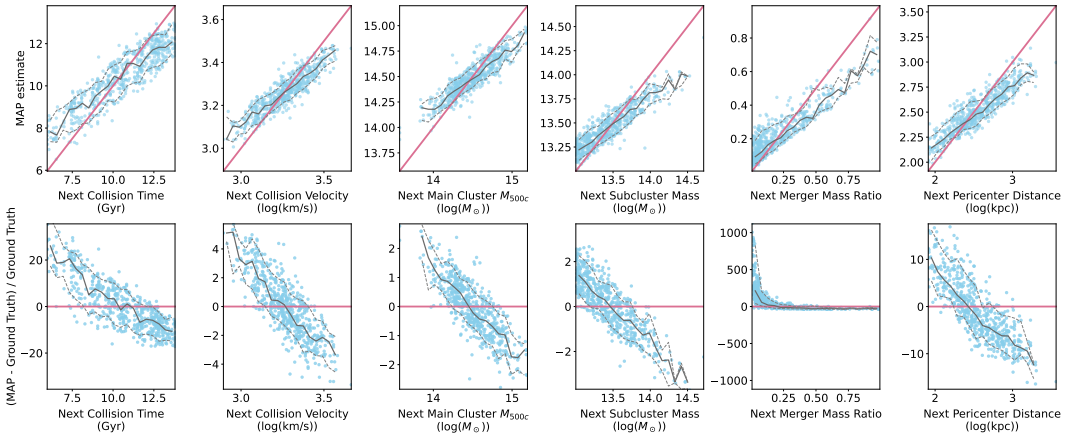

Figure 39: Next–merger MAP performance under X-ray representation conditioning for the 592 test clusters. *Top:* MAP verus truth with bin-wise medians (black solid) and 10-90 % quantiles (black dashed). Pink diagonal: $y = x$. *Bottom:* relative MAP error with the same line styles, and pink horizontal $\Delta = 0$ line. Medians lie near $y = x$ (top) and $\Delta = 0$ (bottom); envelopes are comparable to the last–merger case, with slight widening for Collision Time.

Part VI

INFERRING THE MERGER HISTORIES OF GALAXY CLUSTERS VIA CONDITIONAL INVERTIBLE NEURAL NETWORKS AND CONTRASTIVE LEARNING: RADIO MAPS

17

SIMULATION DATA AND RADIO MAPS CONSTRUCTION

The diffuse radio emission from galaxy clusters arises from synchrotron radiation produced by relativistic electrons spiraling in μG-level magnetic fields of the intracluster medium (ICM). These cosmic-ray electrons (CRe) are (re-)accelerated at collisionless shocks generated during cluster mergers, via diffusive shock acceleration (DSA), yielding steep power-law spectra and highly polarized, elongated features in projection (radio relics) that are typically found in the cluster outskirts. In projection along the merger axis, such shock-related features can mimic centrally located, halo-like morphologies, underscoring the importance of projection effects when classifying radio structures [52, 179].

Relativistic electrons with Lorentz factor $\gamma \gg 1$ in a magnetic field B radiate synchrotron power

$$P_{\rm syn} = \frac{4}{3}\sigma_T\, c\gamma^2\beta^2 U_B, \qquad U_B = \frac{B^2}{8\pi},$$

peaked near the critical frequency

$$\nu_c = \frac{3}{4\pi}\gamma^2\frac{eB_\perp}{m_e c},$$

where $B_\perp$ is the field component perpendicular to the particle velocity. If the electron distribution is a power law $N(E)dE \propto E^{-s}dE$, the monochromatic emissivity scales as

$$j_\nu \propto n_{\rm CRe} B_\perp^{1+\alpha}\nu^{-\alpha}, \qquad \alpha = \frac{s-1}{2}.$$

Diffuse, elongated radio relics are widely interpreted as synchrotron emission from relativistic electrons (CRe) accelerated at collisionless merger shocks and radiating in μG intracluster magnetic fields. Lee et al. [91] model this by assuming diffusive shock acceleration (DSA) injects a power-law CRe spectrum at each shock with slope set by the shock compression (hence Mach number). They then integrate the time-dependent CRe spectrum, including synchrotron and inverse-Compton cooling, to obtain the monochromatic synchrotron emissivity j_ν; the local magnetic field $\mathbf{B}$ is taken directly from the MHD simulation. The implementation follows the analytic formalism of Hoeft and Brüggen [64], with a fixed electron acceleration efficiency used to normalize the CRe energy density. The adopted normalization corresponds to an upper–limit radio power for given shock and field strengths and implies order-percent conversion of the shock kinetic energy flux into CRe for $\mathcal{M} \sim 2$–5 shocks.

In the TNG-Cluster simulations, shocks are detected on–the–fly during the Arepo simulation with the conservative shock finder of Schaal and Springel [143]. Shock zones are first flagged by (i) converging flow ($\nabla\cdot\mathbf{v} < 0$), (ii) alignment of ∇T and $\nabla\rho$ to reject contacts, and (iii) a minimum jump corresponding to $\mathcal{M} > 1.3$. The shock surface cell is then the location of maximum compression; the shock normal is defined from ∇T (second–order accuracy), and upstream/downstream states are

sampled as the first cells outside the zone along ± the normal [118]. From these, the Mach number and the shock kinetic energy dissipation rate is computed as:

$$E_{diss} = \frac{1}{2}\rho_1 (\mathcal{M} c_{s,1})^3 A \delta(\mathcal{M}) \tag{33}$$

where ρ_1, $c_{s,1}$ are upstream density and sound speed, A is an effective shock area, and $\delta(\mathcal{M})$ is the Rankine–Hugoniot thermalization efficiency.

In diffusive shock acceleration (DSA), electrons crossing a shock front gain energy through repeated scattering, and as a result they are injected with a power-law distribution in energy, as in equation 34 where σ is the compression ratio and $\gamma = 5/3$ for the ICM. The slope of this distribution depends on the shock strength, which is usually expressed in terms of the Mach number. For very strong shocks($\mathcal{M} \gg 1$), the spectrum is relatively flat ($\sigma \to 4, s = 2$), while for weaker shocks the spectrum is steeper ($s > 2$), meaning there are fewer high-energy electrons [42].

$$n_e(E) \propto E^{-s}, \qquad s = \frac{\sigma + 2}{\sigma - 1}, \qquad \sigma = \frac{(\gamma+1)\mathcal{M}^2}{(\gamma-1)\mathcal{M}^2 + 2} \tag{34}$$

There is also a natural upper limit to the electron energies. Acceleration competes with energy losses from synchrotron radiation (emission in the presence of magnetic fields) and from inverse Compton scattering (up-scattering of photons, mainly from the cosmic microwave background). Where these losses balance the acceleration, the spectrum cuts off at a maximum energy. Once electrons leave the shock, they are carried downstream with the post-shock flow. As they move away, they gradually lose energy through radiation. Importantly, these losses are stronger for high-energy electrons, because the cooling rate increases with the square of the electron energy [75].

The cooling strength depends on both the magnetic field in the intracluster medium and on the background radiation field of the Universe as in equation 35. The cosmic microwave background can be thought of as an effective magnetic field, which grows stronger at higher redshift. Together, these determine how quickly the electron spectrum steepens downstream of the shock [76].

$$C_{cool} = \frac{\sigma_T}{6\pi m_e c}(B^2 + B_{CMB}^2), \qquad B_{CMB} \simeq 3.24\ \mu G\,(1+z)^2, \tag{35}$$

The population of shock-accelerated electrons is commonly normalized by assuming that a fixed fraction ξ of the shock's dissipated thermal energy is transferred into non-thermal cosmic-ray electrons [64]. A value of $\xi = 0.05$ is often adopted [91]; when combined with the dissipation efficiency, this corresponds to an order-percent conversion of the shock kinetic energy flux into relativistic electrons for Mach numbers $\mathcal{M} \sim 2$–5. This normalization should be regarded as an upper limit, since weak shocks are likely over-luminous under this choice [128].

The resulting synchrotron emission at an observing frequency ν_{obs} can be expressed in the form derived by Hoeft and Brüggen [64],

$$\begin{aligned} \frac{dP}{d\nu} = {} & 5.2 \times 10^{23}\,\mathrm{W\,Hz^{-1}} \left(\frac{\xi}{0.05}\right)\left(\frac{E_{diss}}{10^{44}\,\mathrm{erg\,s^{-1}}}\right)\left(\frac{B}{\mu G}\right)^{\frac{s}{2}+1} \\ & \times \left[\left(\frac{B}{\mu G}\right)^2 + \left(\frac{B_{CMB}}{\mu G}\right)^2\right]^{-1} \left(\frac{\nu_{obs}}{1.4\ \mathrm{GHz}}\right)^{-s/2} \Phi(\mathcal{M})\,. \end{aligned} \tag{36}$$

This relation shows how the radio power depends on the fraction of dissipated energy injected into electrons, the shock energy flux, the magnetic field, the observing frequency, and the Mach number. The function $\Phi(\mathcal{M})$ describes the efficiency of electron acceleration, approaching unity for strong shocks and declining steeply for $\mathcal{M} \leqslant 3$ [64].

Equation 36 is evaluated on the shock-surface cells of TNG-Cluster, using the local Mach number ($\mathcal{M}$), dissipated energy flux (E_{diss}), and magnetic field strength ($\mathbf{B}$) [91]. This procedure by construction produces very thin emitting layers, as the radio emissivity is assigned directly to the shock surface; additional spatial broadening due to downstream advection or radiative ageing is not explicitly included. The underlying assumption is that the shock properties remain approximately constant over a typical cooling time of order 10^8 yr, which is characteristic of GHz-emitting electrons in microgauss-level magnetic fields. Intrinsic surface-brightness maps are then obtained by line-of-sight projection, with three orthogonal views shown for each system to illustrate the impact of projection effects [91]. Unless stated otherwise, from here on the emissivity is evaluated at the $\nu = 1.4$ GHz, a standard reference frequency for radio-relic studies (also similar to VLA bands); the model is frequency-scalable and can be re-evaluated at low frequencies (e.g., LOFAR bands) without changing the underlying methodology.

Figure 40 showcases the morphological diversity of intrinsic radio emission across three orthogonal projections for four halos of TNG-Cluster spanning classes commonly encountered in observations: a double relic, a single relic, an inverted (center–convex) relic, and a system with no detectable diffuse emission at our intrinsic sensitivity. All panels use a fixed square field of view of 5000 kpc (i.e. 5 Mpc across, ± 2500 kpc about the center) and are binned onto a 200×200 grid. This corresponds to a uniform pixel scale of 25 kpc per pixel in the image plane. Operationally, the emissivity attached to shock cells within the window is summed into these pixels and divided by the pixel area to yield intrinsic surface brightness, enabling uniform, like-for-like visual comparison across systems and projections.

Single relics appear when asymmetries in the merger (mass ratio, impact parameter), the shock strengths, and/or the 3D sheet geometry place only one bright shock within the field of view, or when projection suppresses the surface brightness of the counterpart along the chosen line of sight [91, 179] (as seen in top row of figure 40). *Double relics* arise naturally after core passage, when two counter–propagating merger shocks bracket the potential minimum and extend roughly perpendicular to the mass/X–ray elongation, producing the textbook, tangential pair seen in second row of figure 40. It also clearly can be seen that relics are truly dependent on their projection direction, with the relics no longer being observed on the z projection. *Inverted relics* (convex toward the cluster center) are a geometric/projection effect of curved shock sheets in complex mergers: multi–body encounters or subsequent infall can bend and compress pre–existing shocks so that, for some orientations, the brightest edge faces inward (third row of figure 40). Finally, *non–detections* occur when no sufficiently strong shock lies within the FOV at that epoch, when shocks are viewed too face–on (short line–of–sight path through the emitting sheet), or when intrinsic emissivity is low due to weak Mach numbers and/or modest magnetic fields; outcomes expected in a cosmological population even without instrumental effects [52, 91] (last row of figure 40).

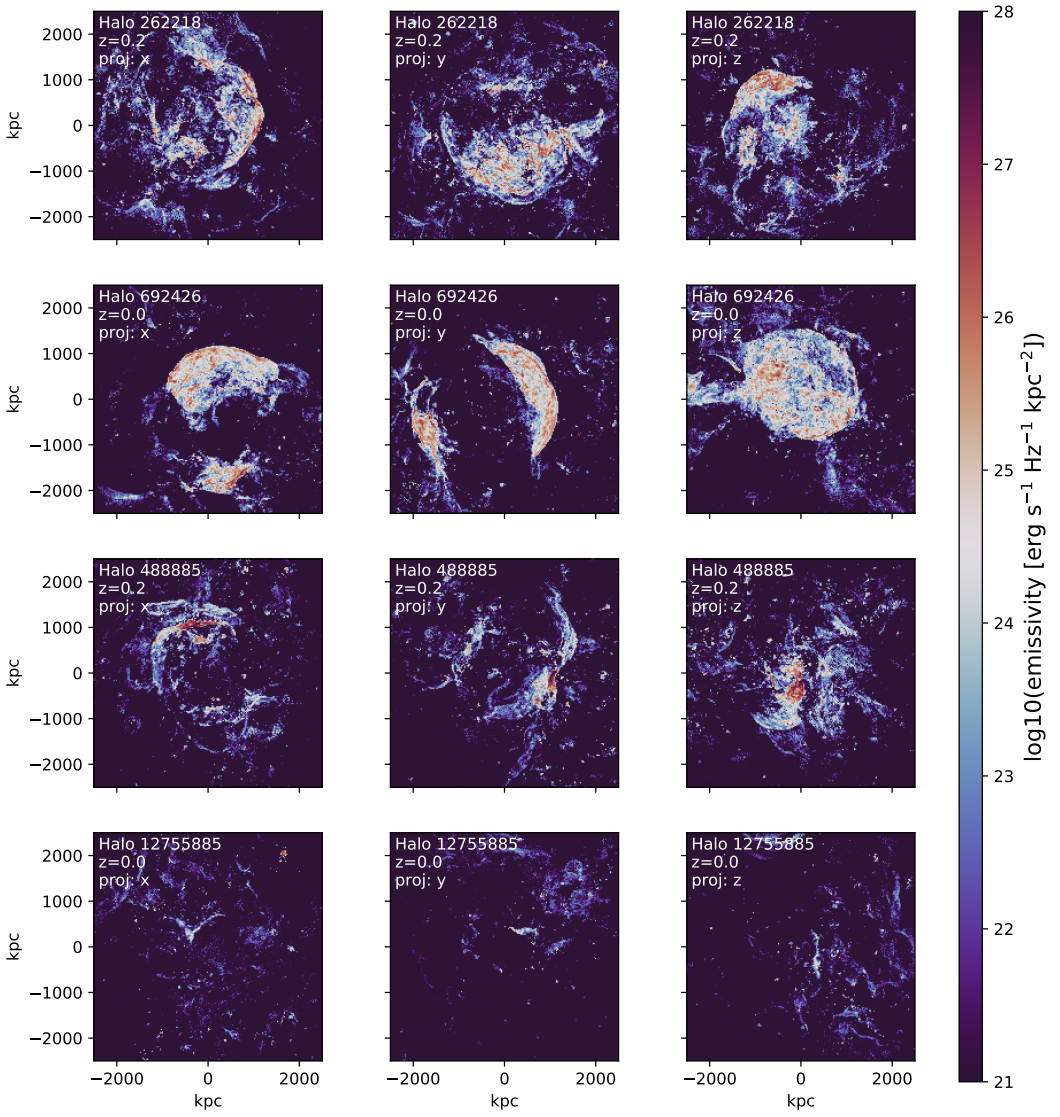

Figure 40: **Diversity of intrinsic radio morphologies in TNG-Cluster.** Rows (top to bottom): *single* relic, *double* relic, *inverted*/center-convex relic, and a *non-detection* case. Columns show three orthogonal projections ($\hat{x}$, $\hat{y}$, $\hat{z}$), highlighting strong orientation effects. All panels use a 5000 kpc field of view binned on a 200×200 grid. All radio maps are intrinsic, constructed based on the data of Lee et al. [91].

The fixed three–projection layout highlights the strong role of orientation and triaxiality in shaping the appearance of shock–related structures [91]. Across all rows, the three orthogonal projections make clear that the same 3D shock geometry can project to markedly different 2D morphologies: a double system may appear single, a thin ridges can fragment into multiple layers and so on.

At a fixed observing frequency (for example, $\nu = 1.4$ GHz), radio relics become fainter with time after the shock passes. This happens because both the population of relativistic electrons and the magnetic field, which together produce the synchrotron emission, evolve in a way that lowers the emissivity j_ν. Fresh electrons are only injected at the moving shock front; once the shock has passed, the electrons left behind simply drift downstream and lose energy [64, 179].

Their energy losses are mainly due to synchrotron radiation and inverse Compton scattering off the CMB, which act more strongly on high-energy electrons ($\dot{\gamma} = -C_{cool}\gamma^2$ with C_{cool} given by Equation 35). As a result, the electron spectrum develops a break that shifts to lower energies with time. The frequency corresponding to this spectral break also decreases with time [75, 76].

$$\nu_b(t) \propto \frac{B}{\left(B^2 + B_{CMB}^2\right)^2} t^{-2},$$

For frequencies $\nu > \nu_b$, the radio spectrum steepens and the emissivity drops rapidly, a behavior known as aging [19].

In parallel, adiabatic expansion of the post–shock plasma lowers particle energies and n_{CRe}, while the initially compressed/amplified magnetic field relaxes, weakening the $B^{1+\alpha}$ leverage in $j_\nu \propto n_{CRe}\, B^{1+\alpha}\, \nu^{-\alpha}$ [19, 64].

As merger shocks travel outward into the lower-density cluster outskirts, their kinetic energy flux decreases and their ability to inject new relativistic electrons becomes less efficient. This is because the Mach number ($\mathcal{M}$) tends to be smaller and the energy dissipated per unit area is reduced, leading to a weaker supply of freshly accelerated particles. At the same time, geometric effects reduce the observed brightness: the emitting layer is very thin, and as the shock front curves and moves beyond the field of view, the line-of-sight depth through this layer becomes smaller [75].

Combining these effects, e.g., radiative aging (losses scaling as $\propto \gamma^2$), adiabatic and magnetic field evolution, reduced injection, and geometric dilution, radio relics fade steadily at a fixed observing frequency ν_{obs}. The dimming is even faster at higher redshift, since the energy density of the CMB increases as $(1+z)^4$ as expressed in equation 35 [76, 179].

In Figure 41 we follow the same four systems (in $\hat{x}$ projection) across cosmic time ($z = 0$, 0.2, 0.5, and 1), emphasizing the transient nature of merger shocks and their radio signatures. It can be seen that, bright relic-like arcs appear after pericenter, when the merger-driven shocks are launched, but they gradually fade and become less distinct as the system evolves. Both the peak surface brightness and the sharpness of the arcs decrease with time, reflecting the radiative ageing of shock-accelerated electrons: once the shock front has passed, synchrotron and inverse Compton losses (Eq. 35) shift the spectral break to lower frequencies, reducing the emissivity at 1.4 GHz. The fading is therefore a genuine evolutionary effect, rather than a change in viewing geometry.

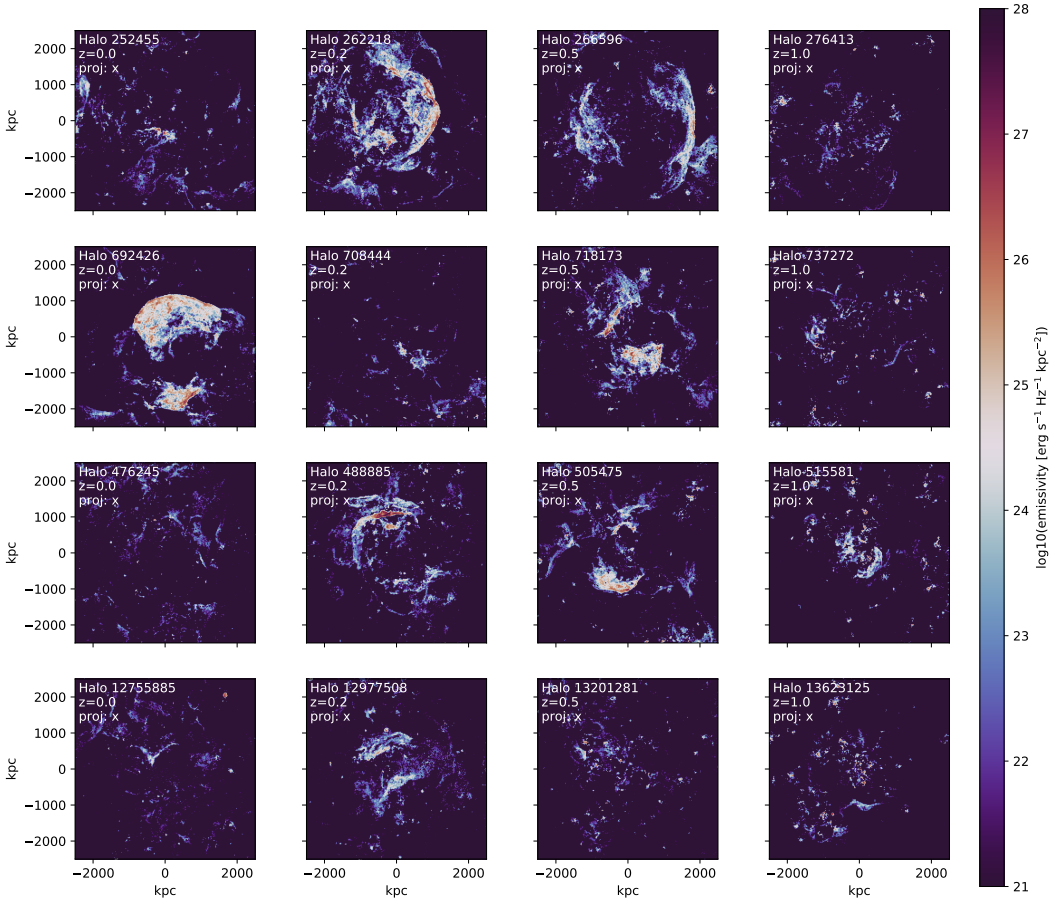

Figure 41: Evolution of the radio surface brightness (along the $\hat{x}$-axis) for the same four halos shown in Figure 40, at redshifts $z = 0$, 0.2, 0.5, and 1. Merger-driven radio features are transient: they tend to brighten following first pericenter, evolve in shape as shocks propagate through the ICM, and may fade on Gyr timescales as accelerated electrons cool and shocks weaken. Projection effects can transform elongated relics into apparently halo-like morphologies when viewed along the collision axis.

18 CONTRASTIVE LEARNING ON RADIO MAPS

18.1 DATA PREPROCESSING (RADIO)

For this thesis, similar to X-ray maps we adopt the *intrinsic* radio surface-brightness maps published by Lee et al. [91] for the same 352 primary zoom halos of TNG-Cluster as used in part v. We consider 3 orthogonal projections across $(\hat{x}, \hat{y}, \hat{z})$ at eight snapshots (99, 91, 84, 78, 72, 67, 59, 50) spanning $0 \leqslant z \leqslant 1$. Since clusters are oriented randomly in the simulation box, this means that viewing angles of the maps are random. As seen in chapter 17, due to the triaxial nature of radio maps, the three projections provide independent, and statistically distinct viewing angles. This yields a total of $352 \times 8 \times 3 = 8448$ maps, which will be the input to the contrastive learning pipeline.

To enable a direct comparison with the X-ray maps in part v and for using in Part vii, we adopt the same projection setup for the radio: a square field of view of $4R_{200c}$ (i.e., $\pm 2R_{200c}$ from the cluster center). We render maps at a pixel resolution of 200×200.

For creating the normalized fits files, empty bins are floored prior to taking logarithms to avoid infinity values. Surface brightness is obtained by dividing by the pixel area so that the final units correspond to $\mathrm{erg\,s^{-1}\,Hz^{-1}\,kpc^{-2}}$. We adopt a fixed global intensity stretch for all maps: we compute $\log_{10}$ of the surface brightness and clip to $[21, 28]$ (in dex). For normalizing the fits file for the next steps, all the values are linearly mapped to $[0, 1]$ and saved as FITS images.

The rest of the preprocessing and augmentations is similar to what was explained in Part v.

18.2 REPRESENTATION EXTRACTION AND POSTPROCESSING (RADIO)

We repeat the representation–space analysis for the *radio* maps in Figure 42, using the same pipeline as in Section 14.5. In brief, after training we use the SimCLR *radio* encoder (projection head discarded) to extract 512–dimensional representations for each test FITS image, applying the identical preprocessing used for radio maps (see Section 18.1). For visualization, the resulting 2D coordinates are placed on a $G \times G$ grid (here $G = 20$) using the same normalization–and–assignment procedure described in Section 14.5.

The learned representation space exhibits the same desirable structure observed for X–ray: clear neighborhoods, smooth transitions, and coherent large–scale trends. Tiles with extended, diffuse synchrotron emission (e.g., halo/relic–like morphologies) tend to cluster together, mainly on the outer (right) border (prior to the tail (upper 2/3)) of the UMAP. Whereas more radio quiet or weak–emission systems (relaxed–like) occupy the inner (left) border of the UMAP. Across the grid, as we go from right to left, we see gradual variations; from radio relics, to weak radio halo emission to no radio emission, and the same holds for going from upper left to lower right, which gradually decreases the strength of radio emissivity.

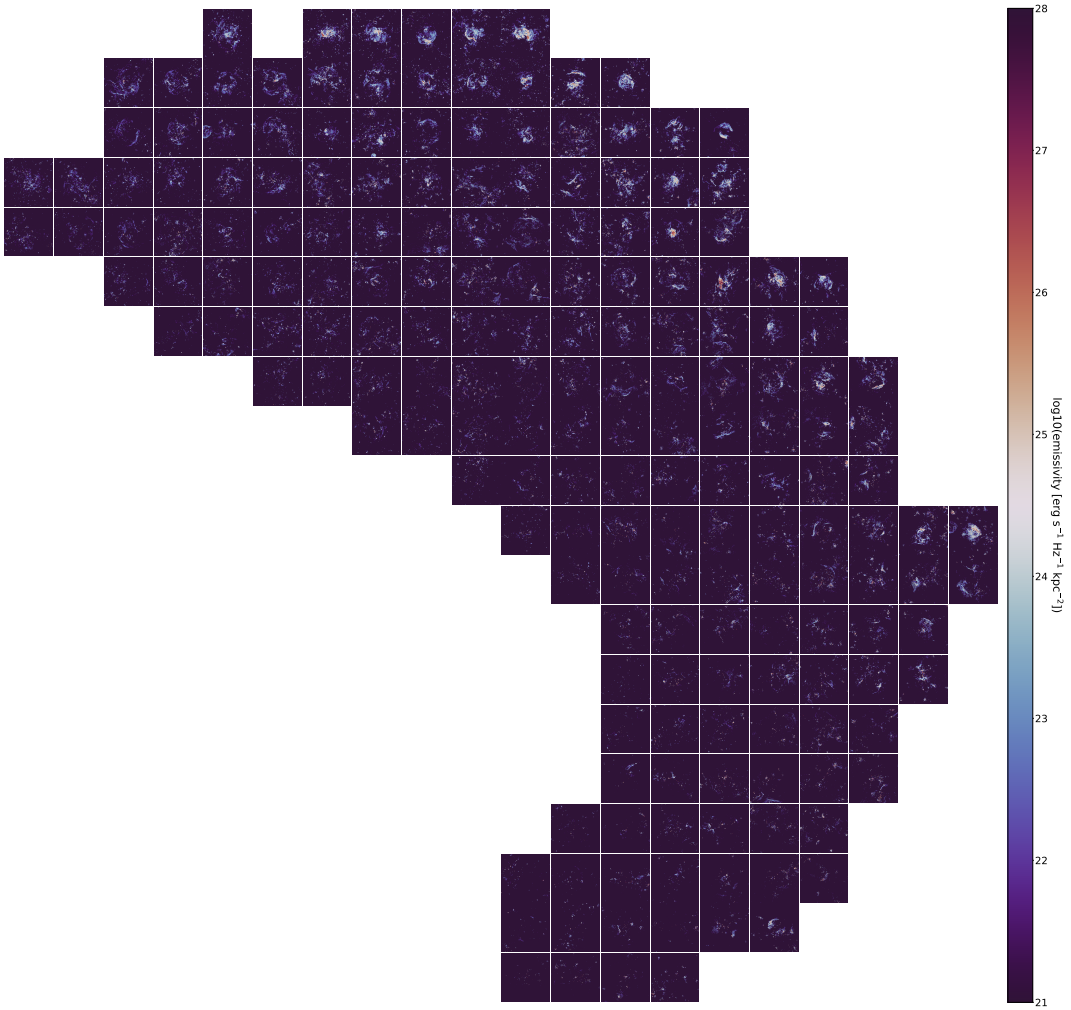

Figure 42: Grid visualization of the UMAP of the SimCLR learned *radio* representation. As in Figure 27, morphologically similar systems populate neighboring cells, indicating a smooth, astrophysically meaningful organization of the representation.

To further probe neighborhood consistency, we perform a nearest–neighbor (NN) analysis directly in the 512–D representation (as in Section 14.5). For a set of randomly chosen anchors we retrieve their k nearest neighbors (Euclidean distance on ℓ_2–normalized vectors; monotonically related to cosine similarity) and display them side–by–side. Because this operates in the original feature space (not the UMAP plane), it tests the semantic organization learned by the encoder without the distortions of nonlinear projection. As can be seen in Figure 43, the neighbors almost have the same radio emissivity pattern; radio halos, and radio relics, and clusters without any radio emissivity are grouped together. This indicates that the representation space captures astrophysically meaningful structure and that the radio representation space is locally smooth and well organized.

Following the same pipeline as in Section 14.5, we probe the *radio* representation space by coloring the 2D UMAP with astrophysical labels using hexagonal bin–averages. This label–free training / label–aware probing assesses whether the radio encoder organizes clusters coherently with respect to independent diagnostics.

These properties are the observable and unobservable (merger) properties from Table 4 and 5 in Figures, 44, and 45, and from table 6 in Figures 46 and 47. Overall, the radio UMAP overlays display smooth, label–correlated gradients for most of the observables and (last/next) merger parameters, mirroring (and in places sharpening),the structure seen with radio. This corroborates that the radio encoder learns astrophysically meaningful features that the cINN can exploit downstream.

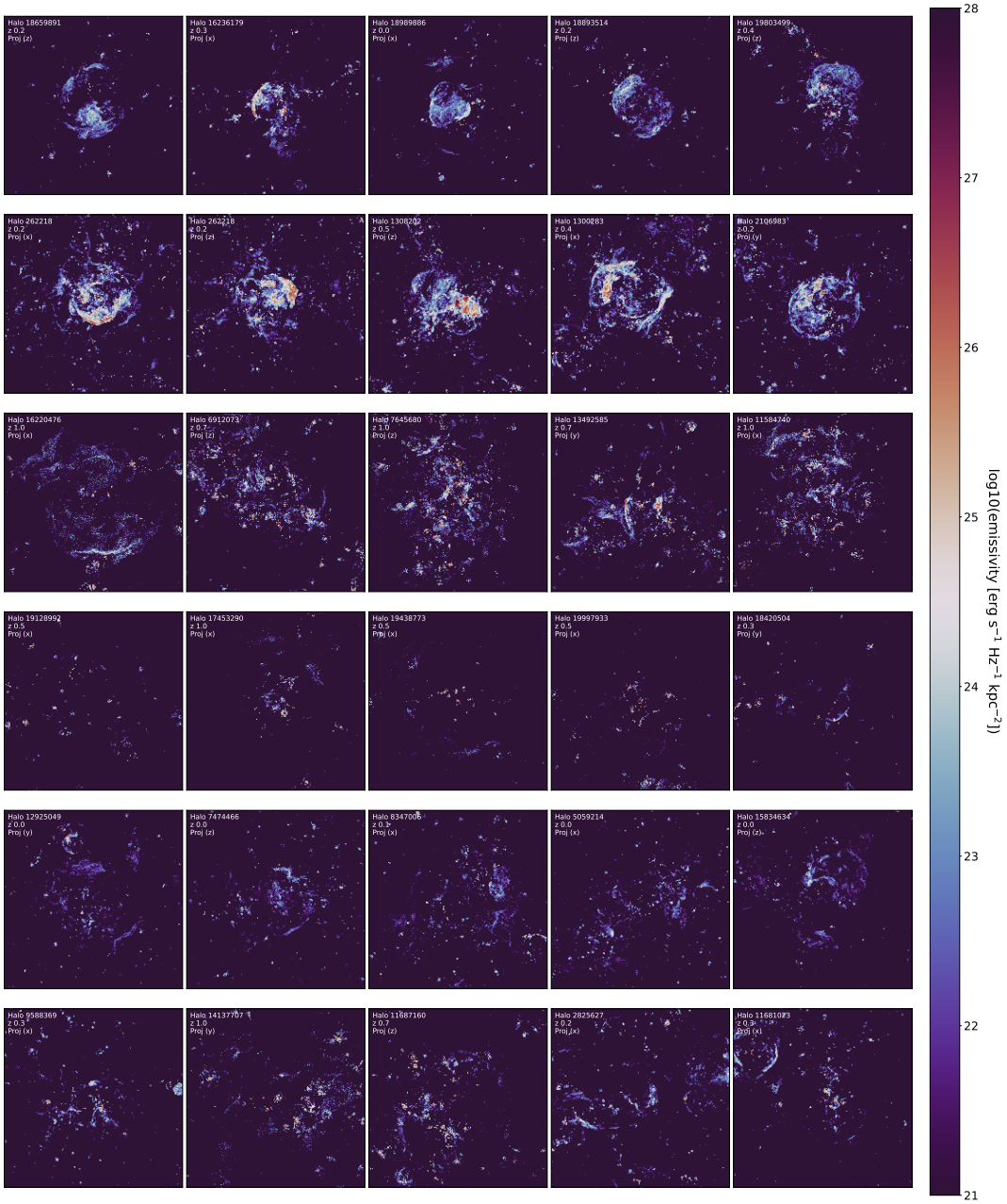

Figure 43: Nearest–neighbor retrieval in the *radio* embedding. Each row shows one anchor map (far left) and its $k = 4$ nearest neighbors in representation space. Neighbors share salient radio morphology (extent, elongation, texture), corroborating the semantic coherence of the learned representation.

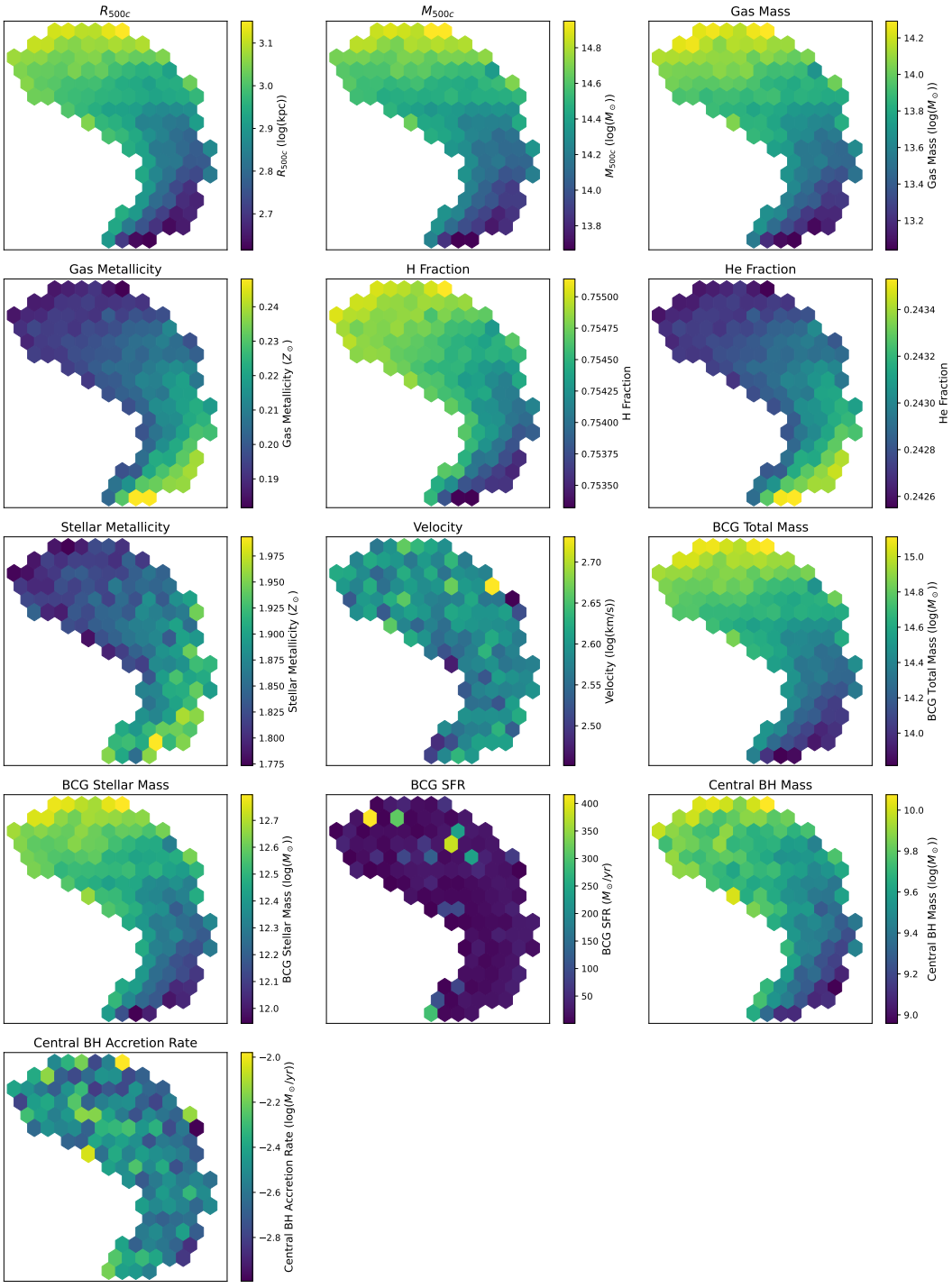

Figure 44: Radio 2D representation (UMAP) colored by the binned mean of **halo/BCG observables** (Table 4). Smooth, coherent gradients indicate that the self-supervised representation encodes global halo–BCG scaling relations despite label–free training.

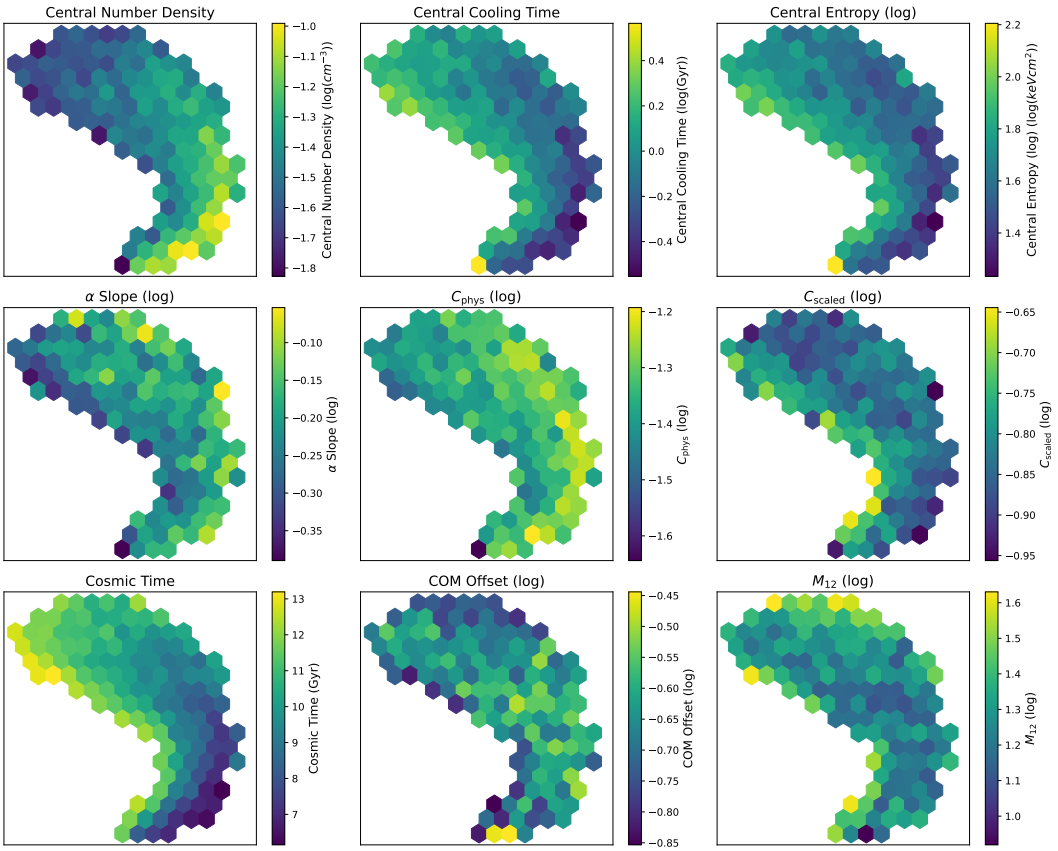

Figure 45: Radio representation (UMAP) colored by the binned mean of **ICM core and dynamical diagnostics** (Table 5). Clear trends show that the representation space captures thermodynamical and dynamical state information.

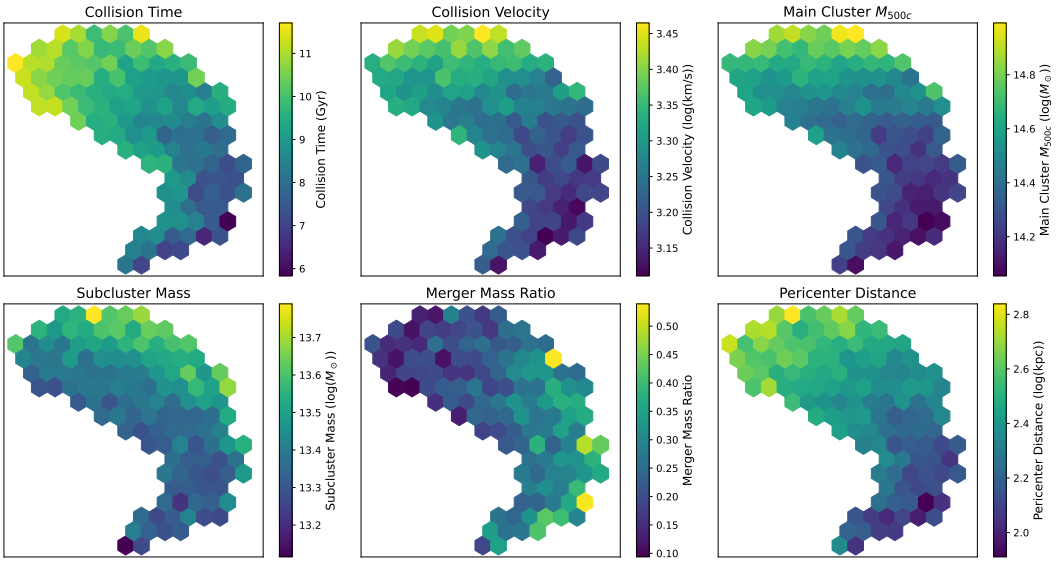

Figure 46: Radio representation (UMAP) colored by the binned mean of **last–merger parameters** (Table 6). Pronounced, ordered gradients suggest that radio morphology retains a clear imprint of recent merger activity relevant for downstream inference.

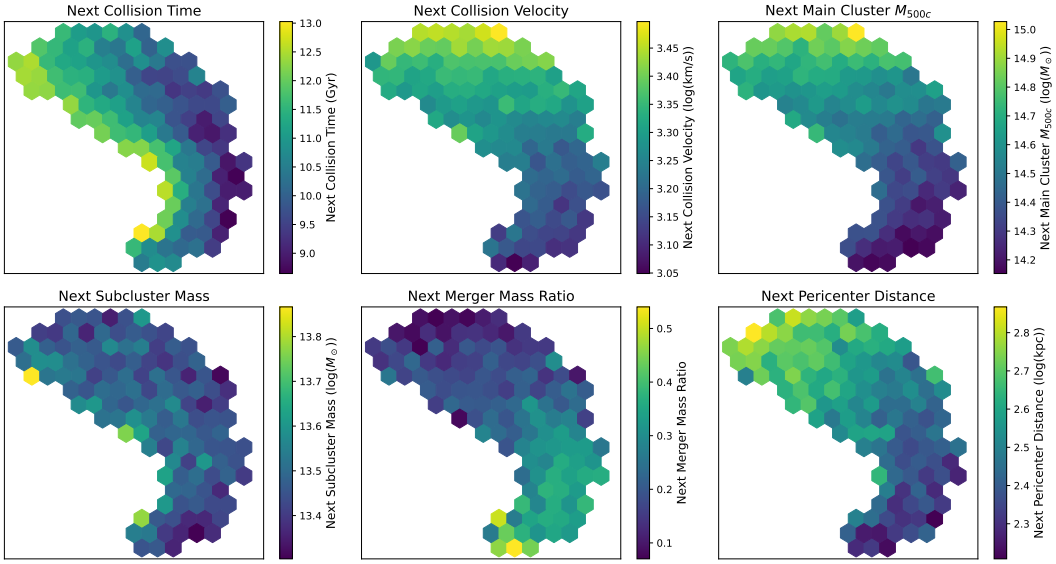

Figure 47: Radio representation (UMAP) colored by the binned mean of **next–merger parameters** (Table 6). The presence of smooth structures indicates that the representation also carries information predictive of upcoming merger events.

19

RESULTS AND DISCUSSIONS (RADIO CONDITIONING)

19.1 POSTERIOR DISTRIBUTIONS WITH RADIO CONDITIONING

We repeat the visualization of conditional posteriors $p(\mathbf{x} \mid \mathbf{c})$ for a subset of test clusters, now conditioning on the learned *radio* representation space $\mathbf{c}$ (see section 18.2) for radio maps. From the saved test indices we randomly select $n_{\text{rows}} = 15$ clusters; for each condition $\mathbf{c}_i$ we draw $n_{\text{sam}} = 1000$ posterior samples $\{\mathbf{x}_i^{(s)}\}_{s=1}^{n_{\text{sam}}}$ via the inverse flow (Section 15.3), map samples to physical units, and arrange one target per column. Rows correspond to distinct galaxy clusters (annotated on the left with `HaloID` and redshift z); columns correspond to merger parameters (section 9.2). Within each panel we overlay the same four elements defined in Section 11.3: a gray prior KDE (test-set marginal for context), a blue posterior KDE for the selected cluster, a gold MAP vertical line, and a red ground-truth line. As before, prior and posterior KDEs are peak-normalized (only shapes and locations are comparable), and n_{sam} controls Monte Carlo smoothness (we use $n_{\text{sam}} = 1000$).

Qualitatively, the radio representation space yield posterior behavior matching the X-ray case (Figure 33) but performing better; the blue posteriors contract strongly around the red ground truths across *all* targets. The contraction is visibly tighter and stronger in radio representation conditioning compared to X-ray, which means that the radio representation space, provided stronger conditioning, or in other words, more organized unobservable patterns. In addition to this, MAP markers (gold) typically coincide with the truths (accurate), and posterior widths are uniformly narrow (precise), indicating that the learned representation space of radio maps (section 18.2) provides a rich, discriminative conditioning signal for the cINN.

19.2 PREDICTION PERFORMANCE OF THE CINN CONDITIONED ON THE RADIO MAPS' REPRESENTATION SPACE

We repeat the test–set evaluation of sections 12.2 and 16.2, now conditioning on the learned *radio* representation space (18). Figures 49 and 50 are constructed identically to their X-ray counterparts (section 16.2): for each target x_d we bin the ground–truth axis into $B = 20$ equal–width bins and, for every test object in a bin, draw $n_{\text{sam}} = 500$ posterior samples with the inverse flow (Section 15.3). Stacking these into the *same* binning yields a 20×20 heatmap in value space with the white diagonal $y = x$; black curves show the posterior median (solid) and 10–90% quantiles (dashed). Figure 50 shows MAP vs. truth with bin–wise medians (solid black line) and 10–90% envelopes(dashed black lines), and the corresponding relative errors $\Delta = 100(\text{MAP} - \text{truth})/\text{truth}$.

Relative to X-ray conditioning (Section 16.2), the radio representation provides a stronger conditioning signal: posterior distributions (Figure 49) concentrate more tightly along the diagonal, with the median lines following the identity lines strongly with very small regression to the mean. MAP medians very closely fol-

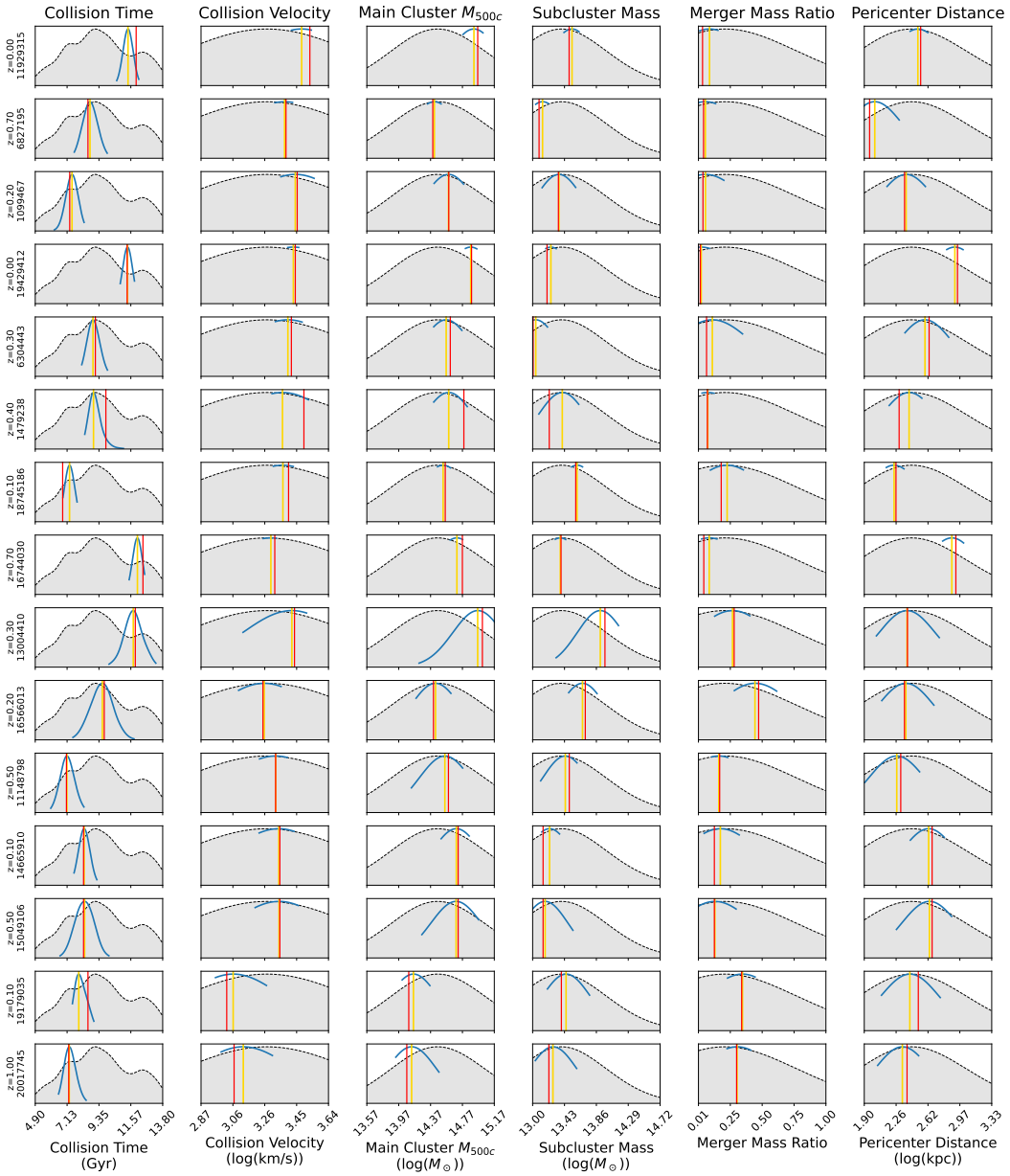

Figure 48: Posteriors conditioned on Radio maps' learned representation space for 15 randomly selected test galaxy clusters (rows) out of 620 across all target merger properties (columns. Gray: prior KDE over the test split; blue: posterior KDE; gold: MAP (vertical line); red: ground truth (vertical line). Construction mirrors the X-ray case in figure 33.

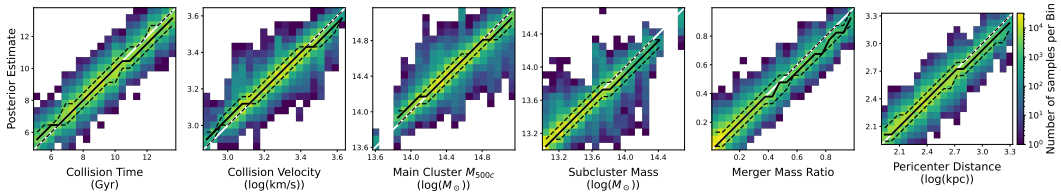

Figure 49: Posterior versus ground truth per target for radio representation conditioning across all 620 test clusters. Construction as in figures 34 and 18 with $B = 20$ and $n_{sam} = 500$. The white diagonal shows $y = x$. Black solid: posterior median; black dashed: 10–90% quantiles. Compared to X-ray (Figure 34), posteriors are *tighter*, across *all* merger parameters. The calibration is also performing very strong (and better than X-ray representation conditioned) tracking $y = x$ even more closely with a very negligible regression to the mean.

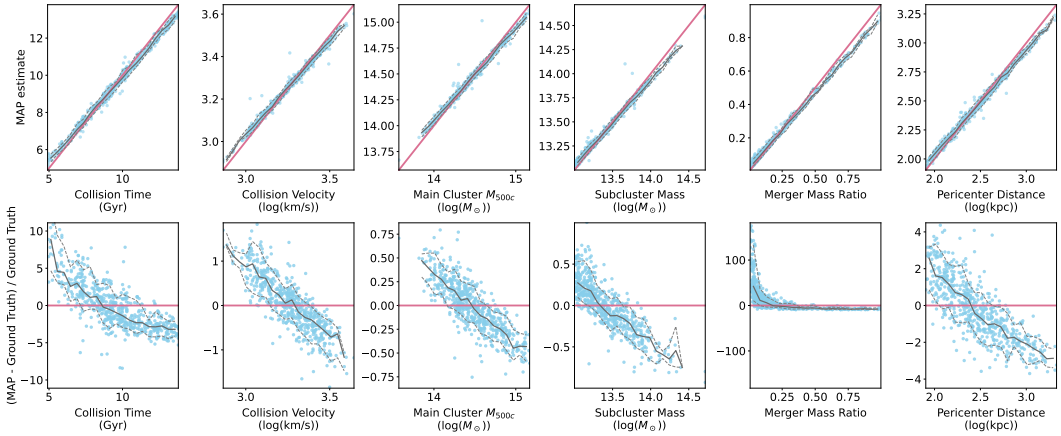

Figure 50: Per merger parameter MAP accuracy (top) and relative error (bottom) under radio representation conditioning across all 620 test clusters. *Top:* MAP vs. truth with bin–wise black solid medians and black dashed 10–90% envelopes; the pink diagonal indicates $y = x$ (perfect agreement). Bottom: relative MAP error $\Delta = 100(\text{MAP} - \text{Truth})/\text{truth}$ with the same line style, with pink horizontal line marking $\Delta = 0$. Medians lie near the identity line with mainly small error ranges (except merger mass ratio) around $\Delta = 0$ (bottom) with very tight 10–90% envelopes. All error ranges are smaller than in the X-ray case (Figure 35).

lows $y = x$ for *all* targets, with very small errors across all merger parameters except lower mass ratios. A very mild, uniform shrinkage (regression–to–the–mean) is still visible as a slight bending toward the global modal scales, but with reduced amplitude compared to X-ray representation conditioning across all targets.

- *Collision Time:* Posteriors accumulate strongly with very tight percentile lines around the $y = x$ line with its median line following it very nicely with very small bending due to the regression to the mean. MAP estimations are also performing very well; MAPS are all around the identity line with very small scatter, and MAP errors mostly in the range $[-5\%, 10\%]$. The relatively larger percentage spread (vs. other targets) remains consistent with discrete snapshot timing, where percentage errors are less forgiving despite small absolute offsets. Compared to X-ray representation conditioning, the posteriors are significantly tighter, with very little scatter of MAP estimation, and lower relative error ($\sim 75\%$ improvement).

- *Collision Velocity:* Similar to Collision Time, calibration is very strong with small envelopes. Similarly, MAP estimates are also performing very strong with medians essentially on the identity, with MAP errors $\leqslant \pm 1\%$. We still have a very slight mode–centered shrinkage as in other targets/conditioning. Compared to X-ray representation conditioning the posteriors are more tighter especially in both lower and higher velocities, with significantly smaller scatter of MAP estimation and relative errors ($\sim$ 80% improvement in extremes).

- *Main Cluster* M_{500c}*:* Posterior distribution is again compact with strong calibration, and tight quantile bands. MAP estimations also are along the $y = x$ line with small scatter and very small relative error of $\leqslant \pm 0.5\%$. Regression to the mean also exist similar to rest of the merger parameters. Compared to X-ray conditioning performance, the posteriors are distributed tighter, however, still less strong than Collision Time, Pericenter Distance and Merger Mass Ratio. The MAP estimation performance has also significantly less scatter with lower relative error ($\sim$ 75% improvement in extremes).

- *Subcluster Mass:* Posteriors accumulate over the identity line with median line and percentile lined following the identity line closely. MAP estimations are around the $y = x$ with very small scatter, tight bands an MAP errors $\leqslant \pm 0.5\%$. A minimal curvature due to regression to the mean can be seen. Compared to cINN conditioned on X-ray representation, posterior distribution are tighter (again not as strong as the Collision Time, Merger Mass Ratio and Pericenter Distance), but its quantile lines are much more concentrated. Also the MAP estimations, have less scatter, and lower relative error ($\sim$ 75% improvement in extremes).

- *Merger Mass Ratio:* Very tight posterior distribution with well-behaving median and small bands. still the most challenging due to its bounded/fractional nature; with MAP estimation's relative errors inflating at small ratios (denominator effect), but median trends follow $y = x$ closely over the bulk of support. Radio representation conditioning provides a significantly less scattered Posterior and MAP estimations distributions, with still high but lower relative error at small ratios ($\sim$ 87% improvement in extremes).

- *Pericenter Distance:* Strong calibration with tight posterior distribution and quantiles. MAP estimation performance is also very strong with MAPS concentrated around the identity line and relative errors $\sim \pm 3\%$. Compared to X-ray, the calibration is stronger, posteriors and its quantiles are more tightly concentrated around $y = x$. There is also less scatter is MAP estimations with lower relative error ($\sim$ 75% improvement in extremes).

In summary, radio representation conditioning yields tight, diagonal–aligned posteriors and uniformly small quantile bands, significantly stronger performance than X-ray conditioning across all targets, with particularly evident tighter posteriors for collision time, Merger Mass Ratio, and Pericenter Distance. The MAP estimations also consistently lie close to the identity lines with very small scatter, and small relative errors (except Merger Mass Ratio) with improvements over the X-ray representation conditioning. case particularly evident for collision time

and Pericenter Distance in posterior distribution, and lower MAP error across all merger parameters.

19.3 CROSS CORRELATIONS: RADIO CONDITIONED INFERENCE

In addition to the scalar and X-ray conditioning explored in previous parts, we now assess whether the cINN trained on radio maps is able to learn the cross correlations among merger parameters. As in Section 12.3, we visualize all pairwise relations between the target merger parameter in a corner plot. For each test object, we draw $n_{\text{sam}} = 200$ posterior samples, and plot the pooled posterior realizations (blue), MAP estimates (gold), and ground truths (red). On the diagonal, we include the one-dimensional KDEs of the corresponding marginals for posterior, MAP, and truth.

The interpretation of this visualization follows the same principles as in the scalar and X-ray cases. Alignment between gold and red clouds indicates accurate MAP recovery, while elongated blue posterior structures aligned with the red loci indicate that the model has captured the correct correlations between parameters. Systematic displacements of gold relative to red point to bias, and dispersed or multi-clumped blue structures reflect residual ambiguity or multi-modality in $p(\mathbf{x} \mid \mathbf{c})$.

The correlation structure evident in Figure 51 is consistent with the expectations summarized in Section 12.3. Posterior samples, MAP estimates, and ground-truth values collectively reproduce the qualitative trends anticipated from ΛCDM scaling arguments. Compared to the scalar and X-ray cases, the posterior distributions are notably less scattered, indicating that the radio-conditioned inference provides tighter constraints on the merger parameters. Moreover, the diagonal KDEs of posterior, MAP, and ground truth are nearly identical, highlighting the high fidelity of the recovered marginal distributions. In particular, the model recovers not only the correct one-dimensional marginals but also the underlying cross-target correlations, demonstrating that it has learned the physical connections between synchrotron radio observables and the dynamical state of cluster mergers.

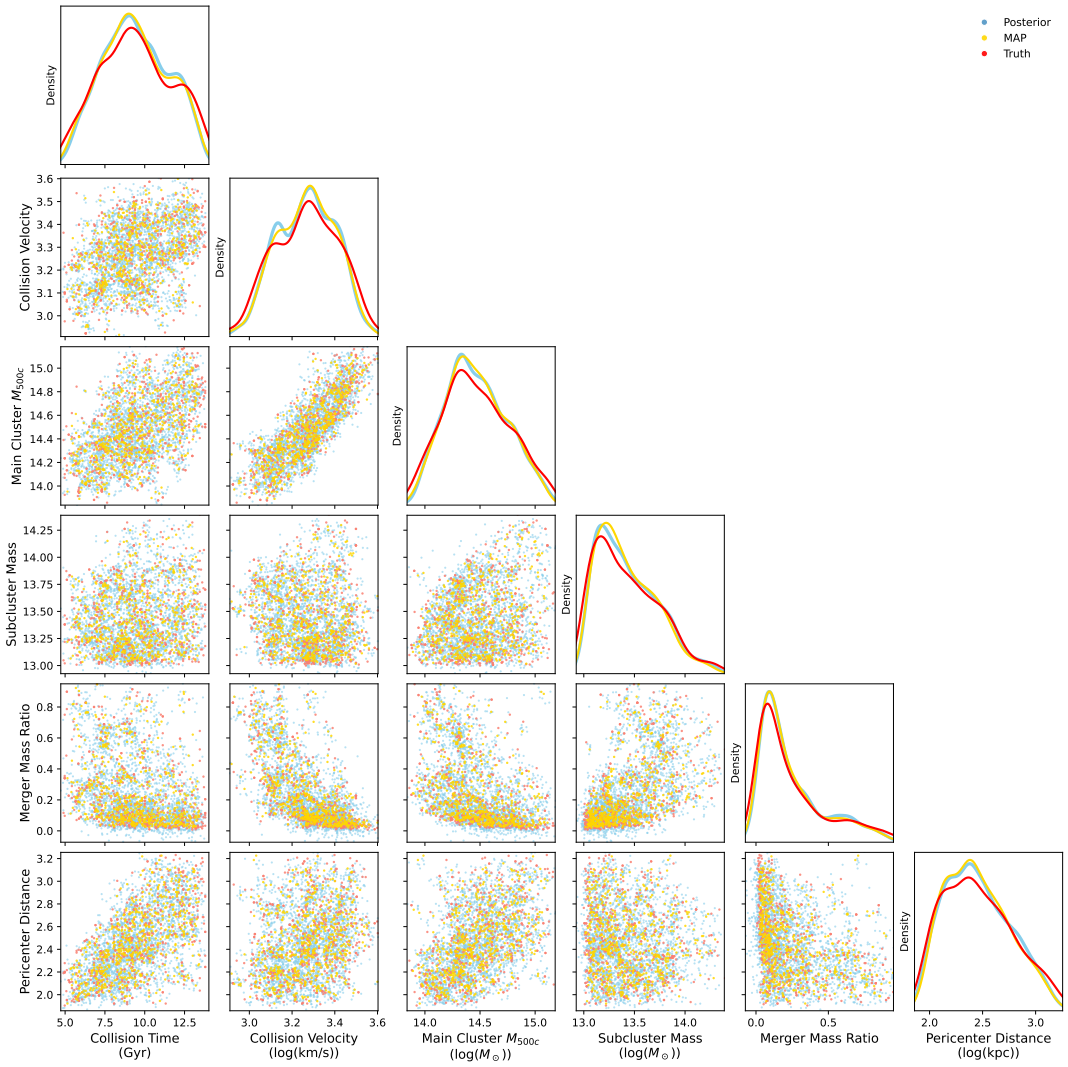

Figure 51: Corner plot across all target merger properties for radio conditioned inference across the 620 test clusters. Diagonal: marginal KDEs of posterior (blue), MAP distribution (gold), and ground truth (red). Lower triangle: pooled posterior samples (blue), MAPs (gold), and truths (red) for each test object. The figure demonstrates that the cINN trained on radio maps captures both the marginal distributions and the cross-target correlations among merger parameters.

19.4 NEXT–MERGER INFERENCE WITH RADIO REPRESENTATION CONDITIONING

We repeat the radio–conditioned analysis for the *next* merger (future event), combining the per–object posterior grids and population–level assessments in a single section. For each selected test cluster we draw $n_{sam} = 1000$ samples from $p(\mathbf{x} \mid \mathbf{c})$ via the inverse flow (Section 11.3) and visualize one target per column as in the last–merger chapter.

Figure 52 shows 15 randomly chosen test clusters (rows) with prior (gray), posterior (blue), MAP (gold), and truth (red) overlays. As in the last–merger radio case (Figure 48), posteriors remain sharply concentrated around the truths across *all* targets; the only systematic change is a mild broadening of the blue ridges for timing–related quantities (most notably collision time, and slightly for Pericenter Distance and Merger Mass Ratio), consistent with the increased uncertainty of forward prediction in time. MAP estimate markers continue to coincide with the red truths in the majority of panels.

Figures 53 and 54 are constructed identically to their last–merger counterparts: we use $B = 20$ truth bins and $n_{sam} = 500$ posterior draws per object, stack samples into the same binning to form a 20×20 heatmap in value space, and overlay the white identity $y = x$, the posterior median (solid black), and 10–90% bands (dashed). The histograms remain narrow and closely aligned with $y = x$; median curves hug the diagonal, and MAP medians and relative–error envelopes stay tight. The radio-conditioned last merger inference (19.1), yields tighter posteriors across most target, especially evident in Collision Time, Merger Mass Ratio and Pericenter Distance. The MAP estimation errors remain mainly similar to the last merger, except of having a lower error in earlier mergers which is consistent with what we would expect for predicting features for events in earlier universe.

Numerically, the next–merger errors are *comparable* to the last–merger radio case, with at most slight broadening for posterior distributions:

- *Collision Time:* Perfomance in posterior distribution and MAP estimation is similar to the last merger, except that the next merger posteriors and its quantile bands are modestly wider than in the last–merger, but still tighter than X–ray.
- *Collision Velocity:* Nearly identical calibration and MAP estimation performance and errors as the last merger csae, with slightly wider posteriors.
- *Main Cluster* M_{500c}*:* similar calibration and MAP estimation performance to last merger.
- *Subcluster Mass:* Calibration remains strong with MAP estimation performance comparable to the last merger.
- *Merger Mass Ratio:* remains the most sensitive (bounded/fractional); relative errors inflate at small ratios, but similar performance ot last merger; the MAP and posterior median follows $y = x$ closely with slightly wider posterior and quantile bands.
- *Pericenter Distance:* Calibration and MAP estimation performance is similar to the last merger with slightly wider posterior compared to the last merger.

In summary, radio representation conditioning sustains tight, diagonal–aligned posteriors and uniformly small MAP errors for the *next* merger, mirroring the last–merger performance with only a small, physically expected slight widening for all parameters, more evident in Collision Time, Merger Mass Ratio, and Pericenter Distance. The radio advantage over X–ray persists across all targets, in the next merger, as well as the last merger predictions.

19.5 DISCUSSION

Conditioning on radio representation space yields the tightest posteriors and the smallest MAP scatter across *all* targets (Figures 49, 50). Typical MAP error ranges are narrower: Collision Time ($\sim [-5, 10]\%$), Collision Velocity ($\leqslant \pm 1\%$), Main Cluster M_{500c} and Subcluster Mass ($\leqslant \pm 0.5\%$), Pericenter Distance ($\sim \pm 3\%$); Mass Ratio still broadens at small values but improves overall. The regression to the mean, still exist, however, it is very negligible and less noticeable than conditioning on X-ray representation. Posterior median follow along the identity line with negligible offset, and tight envelopes which is an improvement from X-ray conditioning. Also as can be seen in Figure 48, posteriors are contracted significantly stronger relative to the prior. This improved performance, can be explained by very smooth transitions of merger parameters in the UMAP in Figure 46, which could help the cINN by giving stronger conditioning signals.

WHAT THE RADIO REPRESENTATION SPACE CAPTURES. The radio encoder organizes maps primarily by shock–driven morphology: presence/absence of relics or halos, bilateral symmetry of double relics, relic curvature and thickness, projected separation from the cluster center, and large–scale anisotropy. These features respond to recent pericenter passage, shock Mach number/speed, and viewing angle, so color–gradients on UMAP for collision time, velocity, and pericenter distance tend to be crisp. As with X-ray, UMAP is a nonlinear 2D projection of a 512-D geometry; smooth bands in 2D are supportive and can be a more informative 512-D conditioner.

WHY RADIO EXCELS ON KINEMATICS. Synchrotron brightness is concentrated in thin, edge–brightened shock sheets. After pericenter, outward-moving shocks produce elongated, high-contrast relics whose projected radius, curvature, and bilateral symmetry can encode time since passage and instantaneous shock speed Lee et al. [92]. This tight visual coupling geometry and merger timing/velocity could translates into narrower posteriors and smaller MAP scatter for those targets when conditioning on the radio representation space.

HIGHLY ORGANIZED REPRESENTATION PACE → INFORMATIVE NEIGHBORHOODS. The SimCLR representation space of radio maps, Figures 46 and 42, shows smooth, monotonic gradients for merger labels across all merger parameters (for Merger Mass Ratio it is not as strong). This indicates that shock-driven morphology (e.g., relic curvature and thickness, bilateral symmetry, etc.) varies coherently with time since pericenter, collision velocity, pericenter distance, and indirectly mass components. Although UMAP is a 2D projection of a 512-D space, the consistent banding and locality suggest that the full geometry is also strongly predictive.

JOINT LEARNING IN THE CINN. The cINN still models the *joint* density $p(\mathbf{x} \mid \mathbf{c}_{\text{radio}})$. Inside each coupling block, spline parameters for a subset of targets are predicted from the remaining targets and the $\mathbf{c}_{\text{radio}}$. Thus, covariances such as time and velocity of collision, pericenter distance, and mass–kinematics links are encoded explicitly as was shown in Figure 51. Even when a single property (e.g, Merger Mass Ratio) shows a less uniform UMAP gradient in some regions, strong organization of the rest of the parameters (e.g, velocity and masses) in the representation space allows the flow to sharpen $p(d_{\text{peri}} \mid \mathbf{c})$ via cross–target structure.

REGRESSION TO THE MEAN VS. CALIBRATION. Shrinkage of MAP bin medians toward modal scales is weaker than in the X-ray case, consistent with stronger cues. While it is mild, it still can indicate that the tighter radio posteriors are not over-confident: improved sharpness coincides with maintained calibration.

REPRESENTATION SPACE VS. RADIO SCALARS. Hand-crafted relic descriptors (length, curvature, separation, polarization fraction) capture only a slice of the morphology and are fragile to apertures, thresholds, and beam. Conditioning the cINN on such scalars could produces broader, less well-calibrated posteriors than conditioning on the learned radio representation. The representation space retains topology (single vs. double relics), bilateral symmetry, thickness gradients, and halo–relic coexistence that scalars miss, explaining its downstream advantage.

ROLE OF THE MIXTURE-OF-EXPERTS (MOE). Routing in representation space naturally could separates regimes such as (i) double-relic systems, (ii) single-relic plus halo, (iii) halo-dominated/weak-shock, and (iv) radio-quiet morphologies. Expert-local flows reduce averaging across these heterogeneous regimes, yielding sharper, better-calibrated posteriors and preserving potential multi-modality that a single global flow would blur.

INTRINSIC MAPS, AUGMENTATION, AND RADIO-SPECIFIC CAVEATS. Our inputs are *intrinsic* (no beam convolution, uv-coverage, correlated noise, or Faraday effects). To promote robustness of the learned representation, we apply a physics-aware SimCLR augmentation policy (Section 14.2): random flips and rotations (to remove arbitrary sky orientation), affine zoom and translations (to desensitize centering/scale), Gaussian blur with $\sigma \in [10^{-3}, 1.0]$ (to mimic beam smearing), and additive Gaussian noise with SNR $\sim \mathcal{U}(4, 8)$ (to emulate depth variations). These augmentations help the encoder ignore nuisances and focus on merger morphology, which we see as stronger prior–posterior contraction and cleaner posterior–vs–truth ridges. That said, they *do not* substitute for radio-specific instrument effects or plasma microphysics: our emissivity prescription (shock-based injection, thin-sheet emission, no explicit downstream re-acceleration/diffusion) and simulated B-fields can emphasize thin, high-contrast rims; in observations, uv-sampling, beam convolution, spectral ageing, depolarization, and RFI/masking will broaden and reshape features, and inverse-Compton losses scale as $(1+z)^2$. Hence, while augmentation improves robustness, instrument-aware forward modeling and sim-to-real validation/fine-tuning remain essential before deploying these radio-conditioned posteriors on survey data.

NEXT MERGER. Radio conditioning inference sustains its best overall performance for forecasting; posteriors remain narrow and diagonal across targets, MAP medians track the identity line, and error envelopes are nearly identical to the last merger case, with only a mild broadening mainly for time of collision. Calibration patterns and small MAP scatters are preserved, indicating that the radio morphology, as can be seen in Figure 47, continues to to anchor merger parameters when predicting the next merger event.

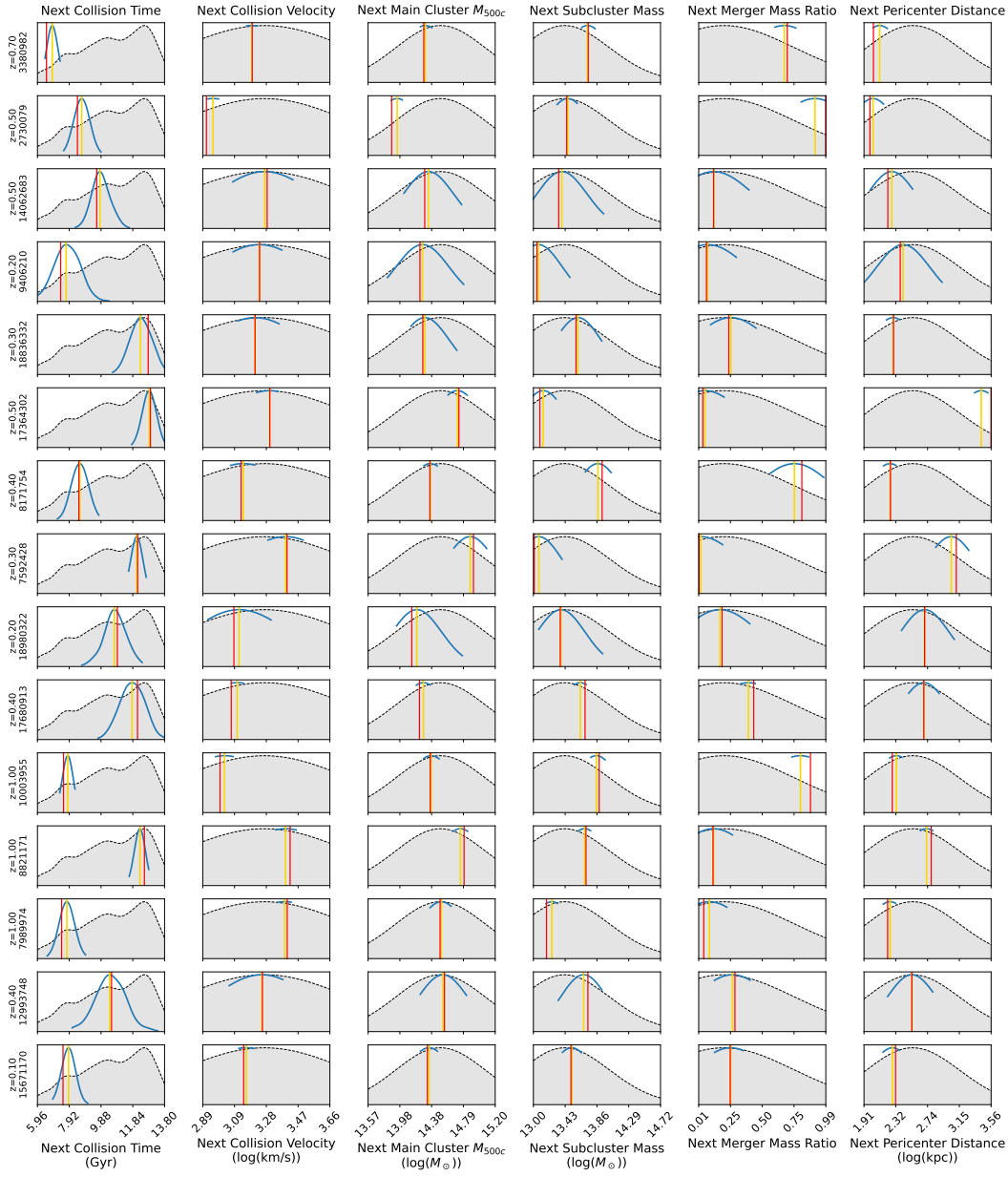

Figure 52: Next merger posteriors conditioned on the radio representation (15 randomly test clusters from 592). Construction mirrors Figure 48. Gray: prior; blue: posterior; gold: MAP; red: ground truth. Posteriors remain well–concentrated around the truths; relative to the last–merger, bands are only slightly broader, mainly for collision time.

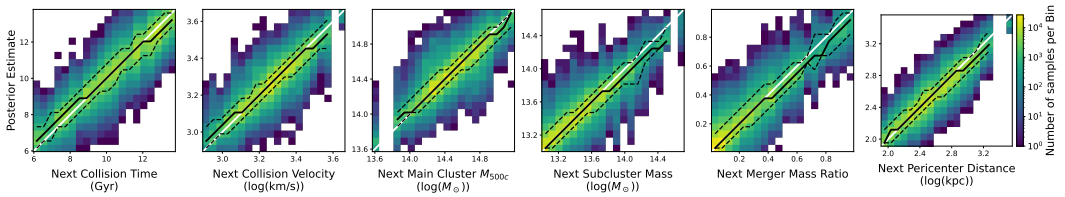

Figure 53: Next–merger: posterior vs. truth per merger parameter conditioned on radio representation across 592 test clusters. Same construction as Figure 49 with $B = 20$, $n_{sam} = 500$. The white diagonal shows $y = x$. Black solid: posterior median, black dashed: 10-90 % quantiles. Medians track the diagonal with most broadening relative to the last merger case (chiefly for collision time).

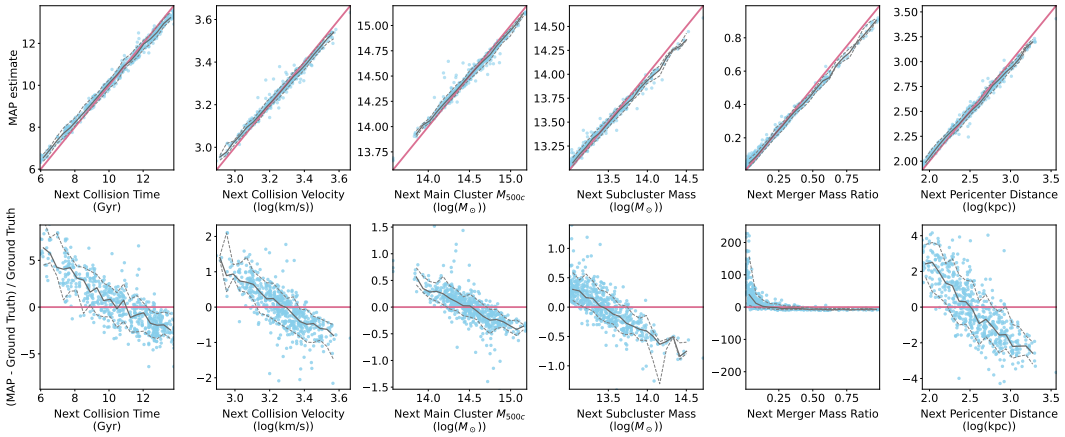

Figure 54: Next–merger MAP estimation performance per merger parameter across 592 test clusters; MAP accuracy (top) and relative error (bottom) under radio representation conditioning. *Top:* MAP vs. truth with black solid medians and black dashed $10-90\%$ envelopes; the pink diagonal marks $y = x$. *Bottom:* relative MAP error Δ with the same line styles; the pink horizontal line marks $\Delta = 0$. Medians lie near the identity line with comparable envelopes and Error ranges.

Part VII

INFERRING THE MERGER HISTORIES OF GALAXY CLUSTERS VIA CONDITIONAL INVERTIBLE NEURAL NETWORKS AND CONTRASTIVE LEARNING: X-RAY + RADIO MAPS

20

CONTRASTIVE LEARNING ON PAIRED X-RAY–RADIO MAPS

This chapter extends the single-modality contrastive pipeline of Chapter 14 to *paired* galaxy-cluster maps, where each training example comprises a co-registered X-ray map and a radio map of the same halo, snapshot, and line-of-sight. To avoid repetition, we retain the notation, optimization details, and loss formulation introduced earlier (Sections 14.3 and 14.2), and focus here on the changes specific to multi-channel inputs.

We (i) construct filename-matched X-ray–radio pairs, (ii) harmonize their spatial resolution without hallucinating detail, (iii) stack them into a two-channel tensor compatible with the SimCLR framework, and (iv) apply channel-consistent augmentations that preserve cross-modal correspondence. The resulting representations are evaluated with the same postprocessing tools (nearest-neighbor retrieval, UMAP visualization) described in Chapter 14.

20.1 DATA PREPROCESSING AND AUGMENTATION (X-RAY–RADIO)

Let $\mathcal{A}$ and $\mathcal{B}$ denote two FITS roots (X-ray and radio, respectively). We first form pairs using their names, making sure that the radio and X-ray maps for each galaxy cluster, at a certain snapshot and projection are paired together. Individually, X-ray maps (Section 14.1) and radio maps (Section 18.1) are pre-normalized to $[0, 1]$ by construction.

However, the problem is that X-ray and radio maps have different grids: X-ray maps are 2000×2000, radio maps are 200×200. To form a two-channel tensor on a *common* grid without inventing high-frequency content, we need to *resize* them. Since the radio maps have a lower resolution, we *downsample* the X-ray map from 2000×2000 to 200×200. Downsampling uses area interpolation (anti-aliased average pooling), which best preserves photometric scale.

Let $X \in [0, 1]^{H_X \times W_X}$ and $R \in [0, 1]^{H_R \times W_R}$ be the arrays. After resizing, we have $\widetilde{X}, \widetilde{R} \in [0, 1]^{H_R \times W_R}$ on a shared grid, and we stack them as a two-channel tensor

$$\mathbf{x} = \left[\widetilde{X}, \widetilde{R}\right] \in [0, 1]^{2 \times H_R \times W_R}.$$

We follow the same augmentations of Section 14.2, but adapt it to paired inputs. Let $\phi(\cdot; \omega)$ be an augmentation with random parameters ω. For SimCLR we draw two independent sets ω_1, ω_2 and form two *correlated* views:

$$\mathbf{v}^{(1)} = \phi\left(\mathbf{x}; \omega_1\right), \qquad \mathbf{v}^{(2)} = \phi\left(\mathbf{x}; \omega_2\right).$$

After applying the same list of augmentations from section 14.2, both views are provided to the encoder. As in the single-modality case, all augmentations are implemented via a `MultiViewTransform` wrapper from lightly.ai [178] that returns $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$.

20.2 SIMCLR WITH TWO-CHANNEL INPUTS

We use the same SimCLR formulation as Section 14.3. In brief, for a batch of N paired images, two random augmentations produce 2N views. Each view is encoded by a convolutional backbone, projected by an MLP head, and optimized with the NT-Xent contrastive objective (Eq. 31).

For the backbone, we retain a ResNet-18 encoder pre-trained on ImageNet but replace the first convolution with a two-channel. For this purpose, weights are initialized by averaging the pre-trained RGB kernel across the color dimension and repeating along the new channel axis; all subsequent layers are unchanged. The projection head is a 2-layer MLP mapping the 512-dimensional pooled features to a 128-dimensional contrastive space [29, 178].

For training, we use SGD with momentum 0.9, weight decay 5×10^{-4}, cosine annealing from an initial learning rate of 0.06 over 100 epochs. The procedure is similar to what was explained in Chapter 14.

20.3 EMBEDDING EXTRACTION AND POSTPROCESSING (JOINT X–RAY + RADIO)

At test time we discard the projection head and use the encoder as a frozen feature extractor, exactly as in Chapter 14. Each paired input is preprocessed by the *same* resize and standardization policy as in training, then passed through the backbone and globally pooled to obtain a 512-dimensional descriptor.

We construct a *joint* representation by fusing the modality–specific encoders introduced in Sections 14.5 (X–ray) and 18.2 (radio). Concretely, we extract 512–D embeddings from each trained SimCLR encoder, apply the same preprocessing and ℓ_2 normalization used in the single–modality pipelines. For visualization we project the joint codes to two dimensions with UMAP and place the coordinates on a G×G grid (we use G = 15), following the exact grid–assignment procedure already described in section 14.5.

For visualizing the joint FITS tiles, we render the same FITS images used in Parts v (X–ray) and vi (radio). Because the native image sizes differ (X–ray 2000×2000 vs. radio 200×200), we first downsample X–ray to the radio grid. Following that, we build an RGB composite with R=radio, B=X–ray, G=0, and convert to `uint8`. In this encoding, radio-only emission appears red, X–ray-only emission appears blue, and co-spatial structures appear magenta.

The joint UMAP exhibits the desirable traits seen in both X–ray (Figure 27) and radio (Figure 42); smooth local neighborhoods and coherent global trends, while adding complementary structure. If the UMAP is divided into three bands from top left to bottom right; the right band is mainly consisted of cuspy X-ray maps, with little to no radio emission, the middle band, is mainly consisted of merging clusters with clusters having stronger radio emission on top and weaker on the bottom, and the final band, are the clusters with flat core X-ray profile, again with the top part, exhibiting radio relics, and lower parts having mainly weaker radio emission.

Nearest–neighbor panels (queried directly in the joint representation space, not in UMAP) confirm semantic consistency beyond 2D projection: X-ray profiles relaxed or merging with strong radio relics/halo grouped together; and the same

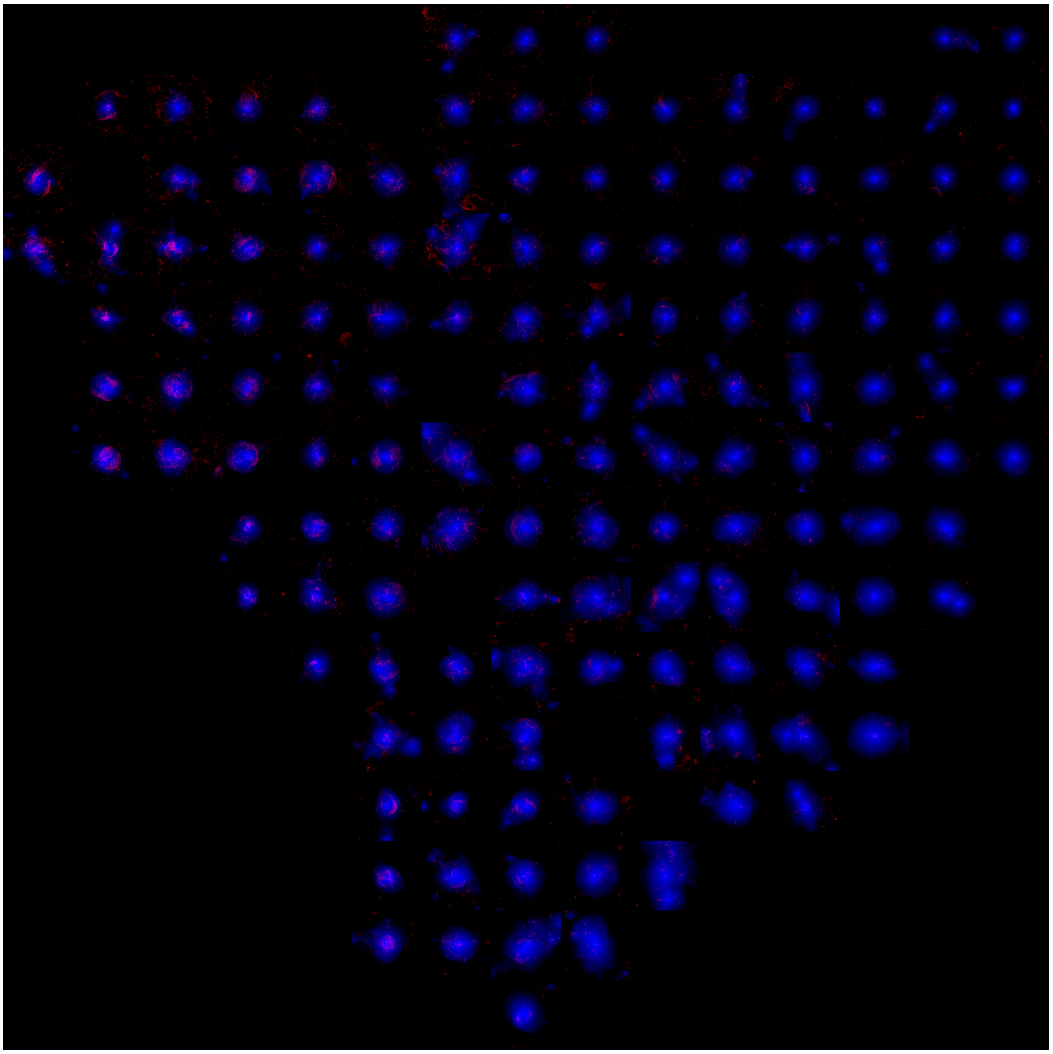

Figure 55: Grid visualization of the **joint** (X–ray + radio) representation (UMAP to 2D, $G = 15$). Each tile is the RGB composite (R=radio, B=X–ray, G=0) corresponding to a projected point from the 512-D joint representation space.

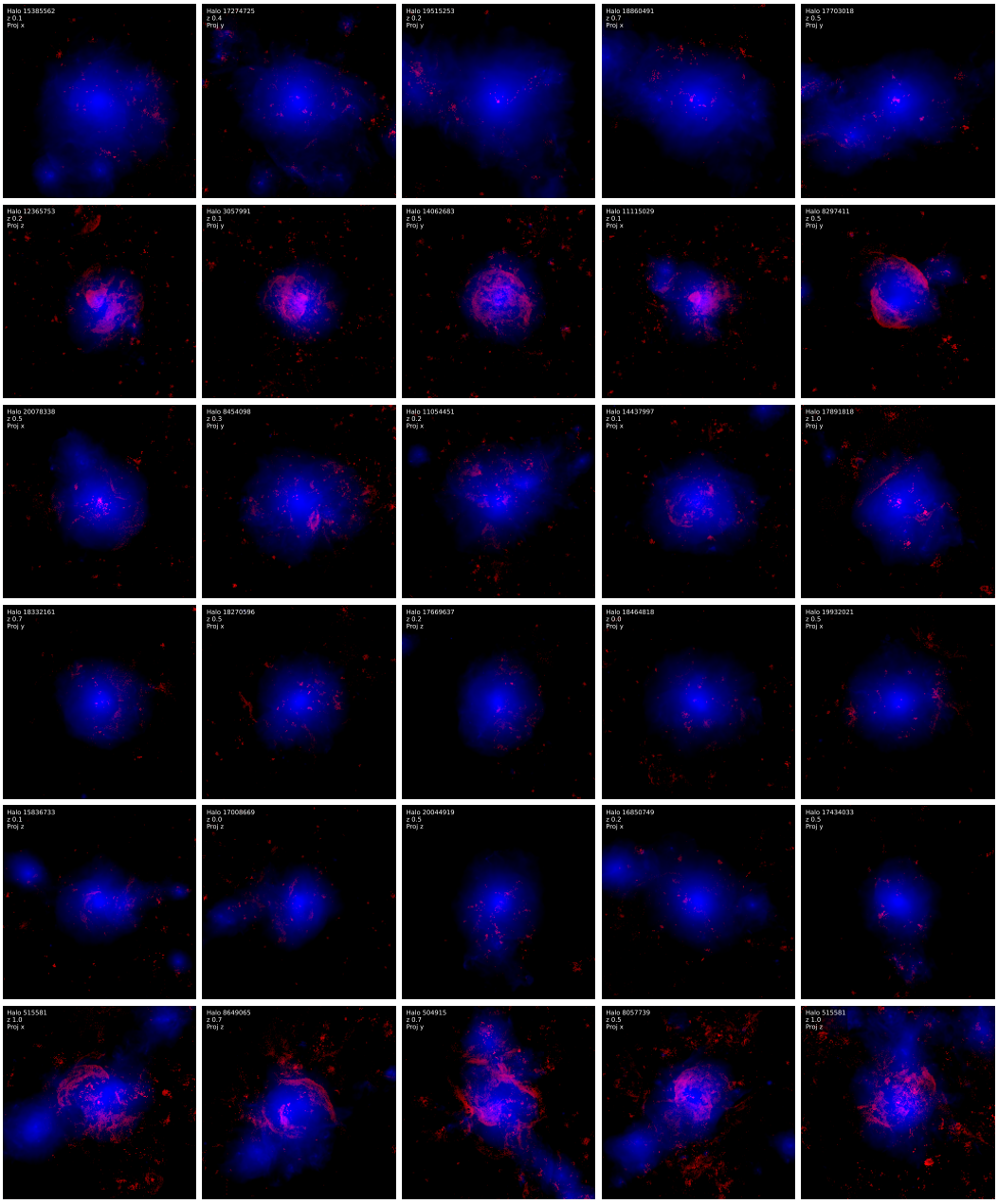

Figure 56: Nearest–neighbor retrieval in the joint representation. Each row shows one anchor (far left) and its $k = 4$ nearest neighbors queried in the 512 dimensional representation space. Each image is the RGB composite (R=radio, B=X–ray, G=0)

for similar x-ray profiles with weak or no radio emission. This mirrors the single–modality findings while tightening neighborhoods through complementary cues.

Across grid views, nearest–neighbor queries, and label-aware hexbin overlays, the joint (X–ray + radio) representation is locally smooth, globally structured, and label-correlated. The joint representation integrates complementary thermal and non-thermal cues, yielding coherent neighborhoods and crisp gradients that motivate its use as a strong conditioner for downstream cINN inference in the joint–conditioning experiments.

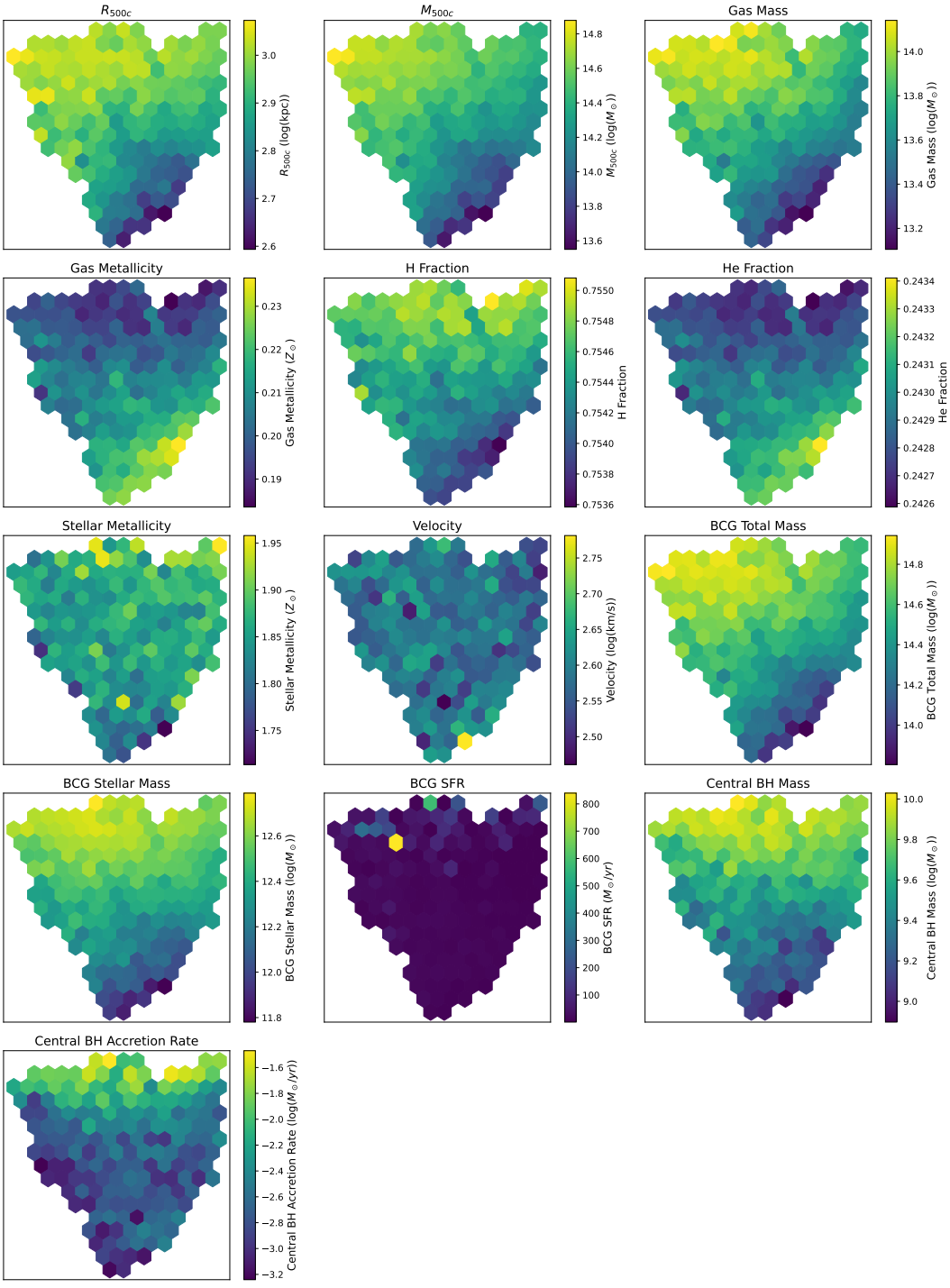

Figure 57: Joint representation (UMAP) colored by binned means of halo/BCG observables (Table 4). Smooth, monotonic gradients indicate that the representation encodes global halo relations despite label-free training.

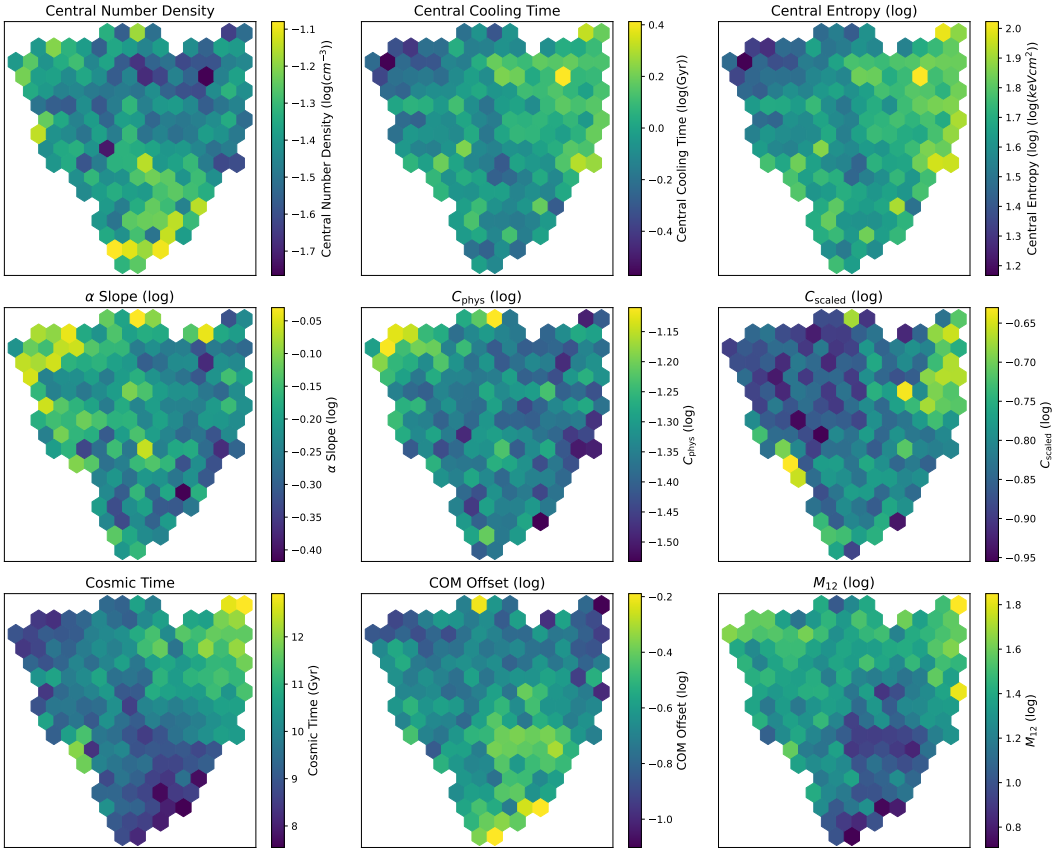

Figure 58: UMAP projection of joint (X–ray + radio) representation space, colored by binned mean values of ICM core and dynamical properties (Table 5). Clear trends show that the representation space captures thermodynamical and dynamical state information.

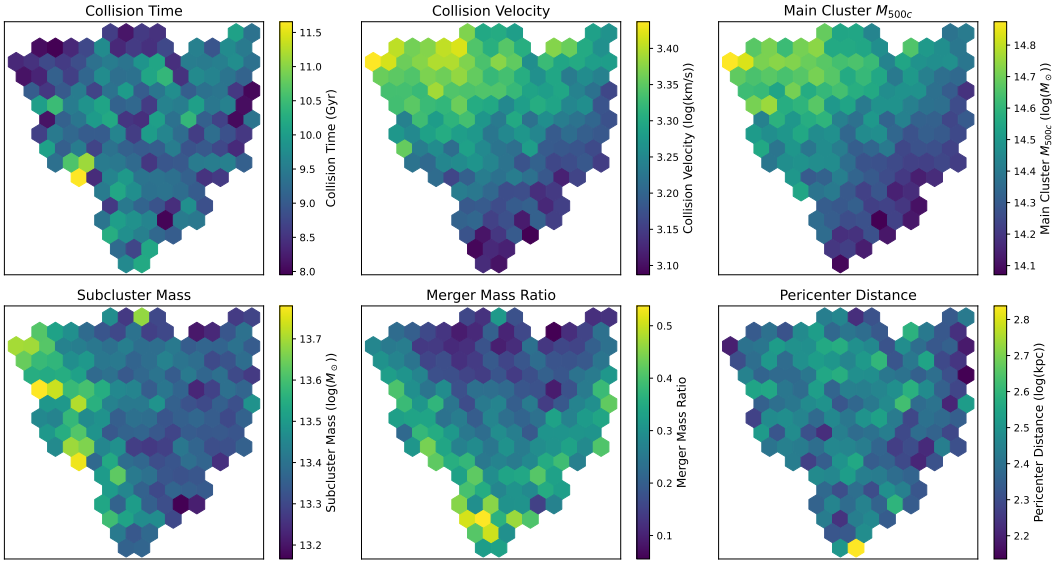

Figure 59: UMAP projection of joint (X–ray + radio) representation space, colored by the binned mean values of last–merger parameters (Table 6). Strong coherent gradients suggest that the representation retains signatures of recent merger activity in the cluster morphologies.

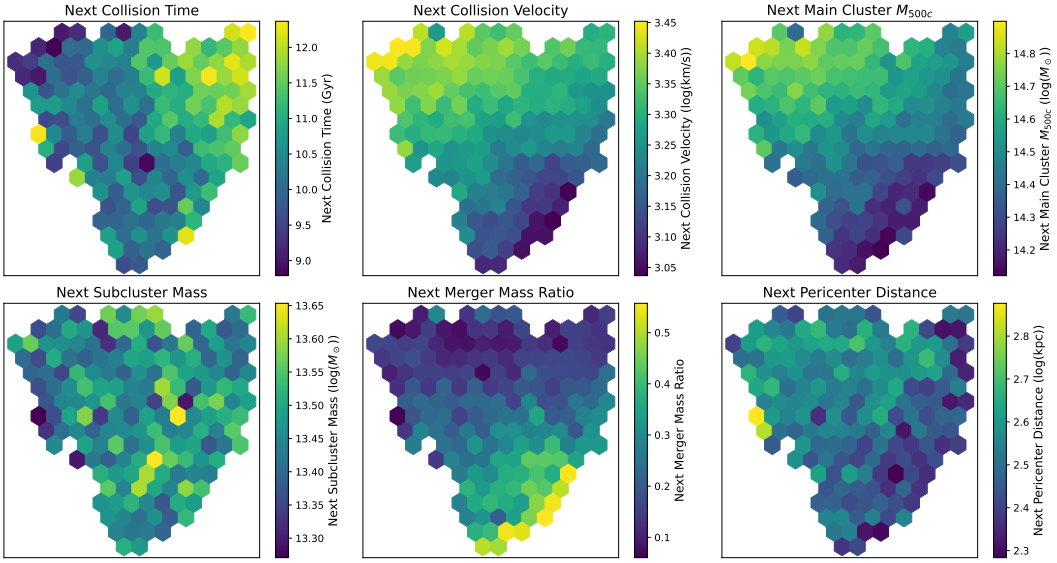

Figure 60: UMAP projection of joint (X–ray + radio) representation space, colored by the binned mean values of next–merger parameters (Table 6). The presence of smooth structures indicates that the representation space also encodes information predictive of upcoming merger events.

21

RESULTS AND DISCUSSIONS (JOINT X-RAY AND RADIO CONDITIONING)

21.1 POSTERIOR DISTRIBUTIONS WITH JOINT X-RAY AND RADIO REPRESENTATION CONDITIONING

We now condition the cINN on the *joint* representation space derived from both X-ray and radio maps, rather than on either modality alone. As before, we randomly select $n_{\text{rows}} = 15$ test clusters, and for each embedding (representation) $\mathbf{c}_i$ we draw $n_{\text{sam}} = 1000$ posterior samples $\{\mathbf{x}_i^{(s)}\}_{s=1}^{n_{\text{sam}}}$ via the inverse flow (Section 11.3). Figure 61 shows the resulting posterior grids: gray test-set prior KDEs, blue posterior KDEs, gold MAP estimate, and red ground-truth vertical lines.

Relative to the unimodal cases, the joint conditioning produces posterior KDEs that are narrower than in the X-ray-only case, but still noticeably broader than under radio conditioning. However, similar to both cases of X-ray and Radio, the gold MAP lines, are close to the red-ground truth derived from TNG-Cluster. In summary, joint conditioning was not less/more precise and accurate as the radio/X-ray only conditioning; or in other words, having an intermediate performance.

21.2 PREDICTION PERFORMANCE OF THE CINN CONDITIONED ON JOINT X-RAY AND RADIO REPRESENTATIONS

Figures 62 and 63 summarize posterior calibration and MAP performance across the test set using the same construction as before ($B = 20$, $n_{\text{sam}} = 500$). The 2D posterior-truth histograms show relatively narrow ridges aligned with the identity line; medians track $y = x$ with mild regression-to-the-mean that is weaker than in the X-ray case but stronger than in radio. MAP estimates lie close to the identity for most parameters, reducing scatter compared to X-ray-only conditioning but not achieving the compactness of radio conditioning. The regression to the mean problem is still present, similar to both X-ray and Radio conditioned cINN.

Calibration is slightly better than X-ray conditioning, with median posterior lines more closely following the ground truth, but does not reach the near-perfect alignment of radio representation space. The quantile lines are also tighter than X-ray, and wider compared to the radio conditioning. MAP estimates remain consistently closer to the truth than in the X-ray case, though with somewhat larger scatter than for radio-only conditioning. Following this, we discuss the performance per merger parameter:

- *Collision Time:* Posterior calibration is improved compared to X-ray conditioning, though still broader than radio. MAP errors are within $\pm 10\%$, showing moderate improvement over X-ray but weaker precision than radio. The posterior distributions are about as tight as those from X-ray, though broader than radio; their calibration (posterior mean) is only slightly improved over X-ray but remains worse than radio. For the MAP distribution, the scatter is smaller than X-ray yet wider than radio.

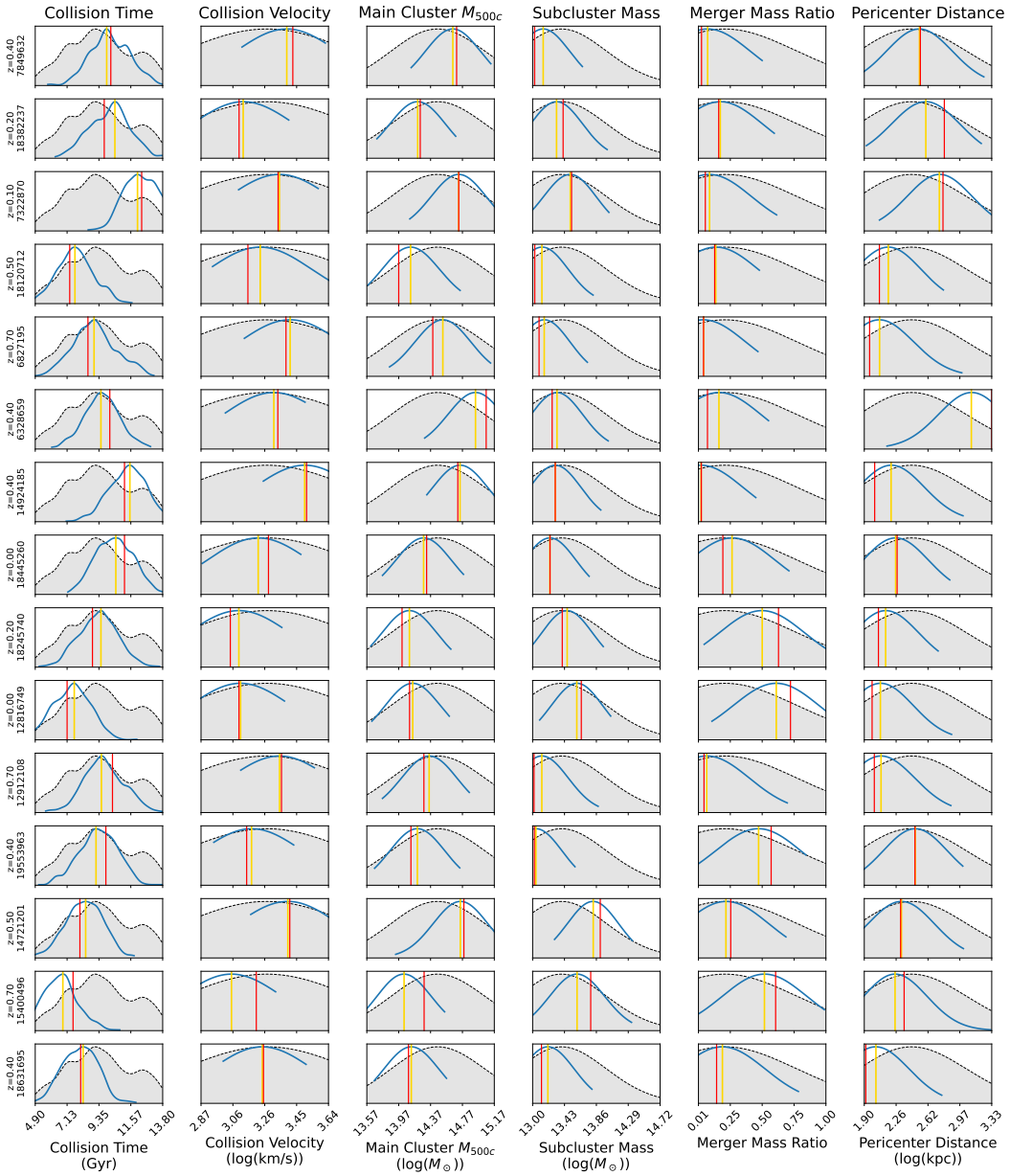

Figure 61: Joint X-ray + Radio representation conditioned posterior grids for 15 randomly selected clusters from 620 test clusters (rows) across all target merger properties (columns). Gray: prior KDE; blue: posterior KDE; gold: MAP; red: ground truth. Posterior contraction is stronger than X-ray conditioning alone but weaker than radio-only conditioning.

- *Collision Velocity:* Posterior distributions show a similar level of tightness compared to X-ray but are less compact than radio. Their calibration is marginally better than X-ray, though still weaker than radio. The MAP distributions reveal scatter that lies between X-ray (larger) and radio (smaller). MAP errors are around $\pm 2\%$, which makes the MAP estimation Performance is intermediate between the X-ray ($\pm 5\%$) and radio ($\pm 1\%$) cases. Similar to Collision Time, the variability in MAPS, along with its relative errors, is smaller than in X-ray but greater than in radio-based representations.

- *Main Cluster* M_{500c}*:* The tightness of the posterior distributions matches that of X-ray but is exceeded by the more concentrated radio case. Calibration is slightly superior to X-ray but inferior to radio. In terms of MAPs, the scatter is reduced relative to X-ray, yet remains broader than radio. Relative MAP errors $\leqslant 1\%$, showing slightly better performance than X-ray-only ($\sim \pm 2\%$) but looser than radio ($\sim \pm 0.5\%$). In MAPS, the scatter and relative error values fall below those of X-ray yet remain above those observed in radio-conditioned data.

- *Subcluster Mass:* We observe that the posterior distributions are as narrow as X-ray but not as narrow as radio. Their calibration is marginally improved compared with X-ray but does not reach the level of radio. For the MAP distributions, the scatter is below that of X-ray but greater than radio. The MAP errors around $\pm 1\%$. This is markedly better than X-ray conditioning ($\sim \pm 2\%$), though slightly less precise than radio-only ($\sim \pm 0.5\%$). Compared to X-ray, MAPS shows reduced scatter and relative errors, though these are still larger than what is seen in radio representations.

- *Merger Mass Ratio:* The posterior distributions maintain a comparable spread to X-ray, though they are wider than those from radio. Their calibration is just above X-ray in quality, while still below radio. The MAP scatter is positioned between X-ray (larger) and radio (smaller). As with all conditionings, this remains the most difficult parameter, particularly at low ratios. Joint conditioning reduces scatter compared to X-ray, but the intrinsic bounded nature of the ratio maintains relatively high fractional errors. MAPS exhibits less scatter and lower relative errors than X-ray, but higher levels of both than in radio-conditioned representation.

- *Pericenter Distance:* Posterior distributions are similarly tight as X-ray but looser than radio; it is also slightly better calibrated than X-ray but falls short of radio. Meanwhile, the MAP distributions show scatter that is smaller than X-ray but larger than radio. MAP errors are around $\pm 5\%$, again between the X-ray performance of $\pm 10\%$ for X-ray and $\pm 3\%$ for Radio. Calibration is stable but not as sharp as in radio-only conditioning. The scatter and relative errors in MAPS are diminished compared to X-ray, yet remain elevated relative to radio-conditioned representation.

In summary, joint representation (X-ray and Radio) conditioning yields diagonal–aligned posteriors, which are calibrated slightly better than X-ray and worse than Radio conditioned cINN. The posterior distributions, and its quintiles are slightly tighter than the X-ray conditioned, while radio-conditioned cINN, showed tighter posterior, and quantiles. The MAP estimations also consistently lie close to

the identity lines with smaller scatter, and relative errors (except Merger Mass Ratio) compared to X-ray and higher with respect to radio representation conditioned cINN.

21.3 CROSS CORRELATIONS: MIXED CONDITIONED INFERENCE

Finally, we assess whether the cINN trained on the combined set of X-ray and radio maps is able to learn the cross correlations among merger parameters. As in Section 12.3, we visualize all pairwise relations between the target merger parameters in a corner plot. For each test object, we draw $n_{\text{sam}} = 200$ posterior samples, and plot the pooled posterior realizations (blue), MAP estimates (gold), and ground truths (red). On the diagonal, we include the one-dimensional KDEs of the corresponding marginals for posterior, MAP, and truth.

The interpretation of this visualization follows the same principles as in the previous cases. Alignment between gold and red clouds indicates accurate MAP recovery, while elongated blue posterior structures aligned with the red loci indicate that the model has captured the correct correlations between parameters. Systematic displacements of gold relative to red point to bias, and dispersed or multi-clumped blue structures reflect residual ambiguity or multi-modality in $p(\mathbf{x} \mid \mathbf{c})$.

The correlation structure evident in Figure 64 is consistent with the expectations summarized in Section 12.3. Posterior samples, MAP estimates, and ground-truth values collectively reproduce the qualitative trends anticipated from ΛCDM scaling arguments. In particular, the mixed-conditioned inference successfully recovers not only the correct one-dimensional marginals but also the underlying cross-target correlations among merger parameters.

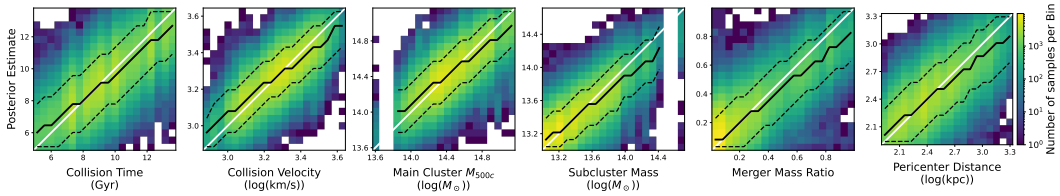

Figure 62: Posterior versus ground truth per target across all 620 test clusters (joint X–ray + radio conditioning). Each panel is a 20×20 2D histogram from $n_{sam} = 500$ draws per test object and $B = 20$ truth bins. The white diagonal shows $y = x$. Black solid: posterior median; black dashed: 10–90% quantiles. Ridges are narrower and medians track $y = x$ more closely than in X–ray-only conditioning, but remain broader than radio-only.

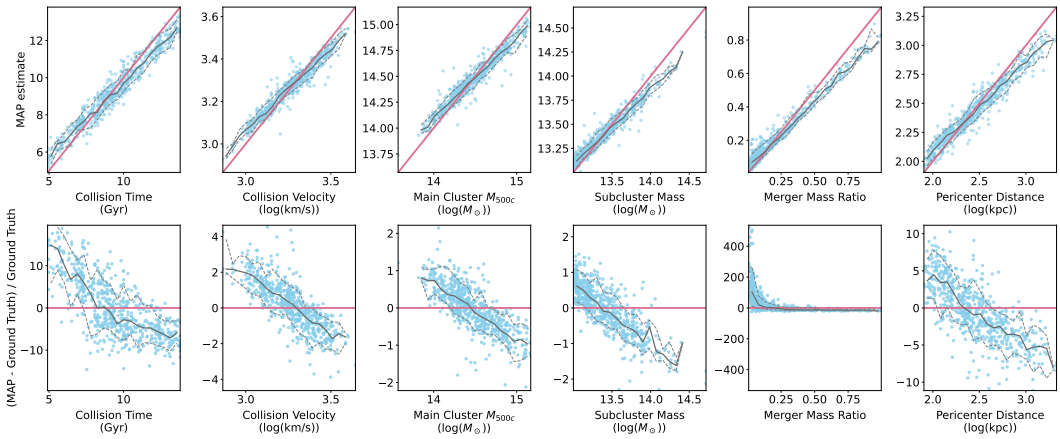

Figure 63: Per–target MAP accuracy (top) and relative error (bottom) under joint X–ray + radio conditioning across 620 test clusters. **Top:** MAP vs. truth with bin-wise medians (black solid) and 10–90% envelopes (black dashed); the pink diagonal marks $y = x$. **Bottom:** relative MAP error $\Delta = 100(\text{MAP} - \text{truth})/\text{truth}$ with the same line styles; the pink horizontal line marks $\Delta = 0$. Scatter and error envelopes are reduced relative to X–ray-only, but remain larger than radio-only.

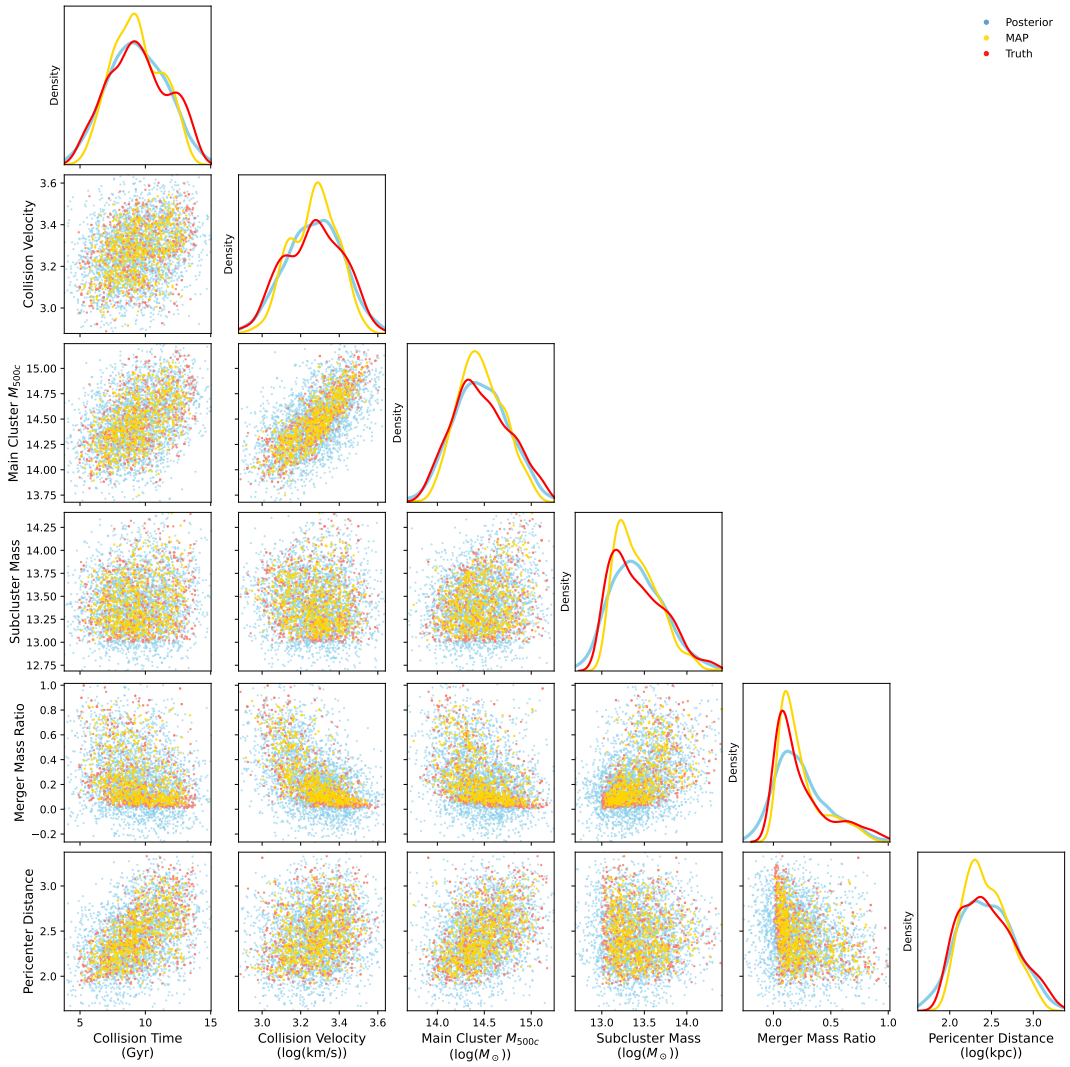

Figure 64: Corner plot across all merger properties for mixed conditioned inferenc across 620 test clusters. Diagonal: marginal KDEs of posterior (blue), MAP distribution (gold), and ground truth (red). Lower triangle: pooled posterior samples (blue), MAPs (gold), and truths (red) for each test object. The figure demonstrates that the cINN trained on both X-ray and radio maps captures the marginal distributions as well as the cross-target correlations among merger parameter.

21.4 NEXT-MERGER INFERENCE WITH JOINT X-RAY AND RADIO CONDITIONING

We repeat the joint conditioning analysis for the *next* merger. As expected, posterior distributions broaden compared to the last-merger case (Figure 65), particularly for Collision Time. Calibration remains slightly better than X-ray but worse than radio.

Figures 66 and 67 show that 2D posterior-truth histograms remain aligned with the identity, albeit with wider percentile bands than in the last-merger case, most notably for Collision Time. Nevertheless, MAP estimates maintain strong accuracy, with error ranges essentially identical to those in the last-merger case.

- *Collision Time:* The posterior and MAP estimates are consistent with the last–merger performance, though the next–merger posteriors and their quantile intervals are slightly broader. The MAP errors also remain $\pm 10\%$, almost similar to the last merger.
- *Collision Velocity:* Calibration, MAP accuracy (with MAP error $\pm 2\%$), and error levels are essentially the same as in the last–merger case, apart from modestly wider posteriors.
- *Main Cluster* M_{500c}*:* Calibration and MAP estimates align closely with last–merger results, with errors staying around $\pm 1\%$.
- *Subcluster Mass:* Calibration continues to be strong, with MAP estimates close with the last merger, and same relative MAP error of $\sim \pm 1\%$.
- *Merger Mass Ratio:* This parameter stays the most sensitive (bounded/fractional); relative errors increase at small ratios, but the overall performance matches the last–merger case. The MAP and posterior median track $y = x$ closely, with somewhat wider posteriors and quantile bands.
- *Pericenter Distance:* Calibration and MAP estimation are consistent with the last–merger case, again with slightly broader posteriors. Relative MAP estimation errors remain around $\pm 5\%$, similar to the last-merger case.

In summary, joint X-ray and radio conditioning improves over X-ray-only inference by reducing scatter and narrowing posteriors, but it does not reach the precision achieved by radio-only conditioning. Performance is consistently intermediate: better than X-ray, worse than radio, and largely stable between last- and next-merger predictions.

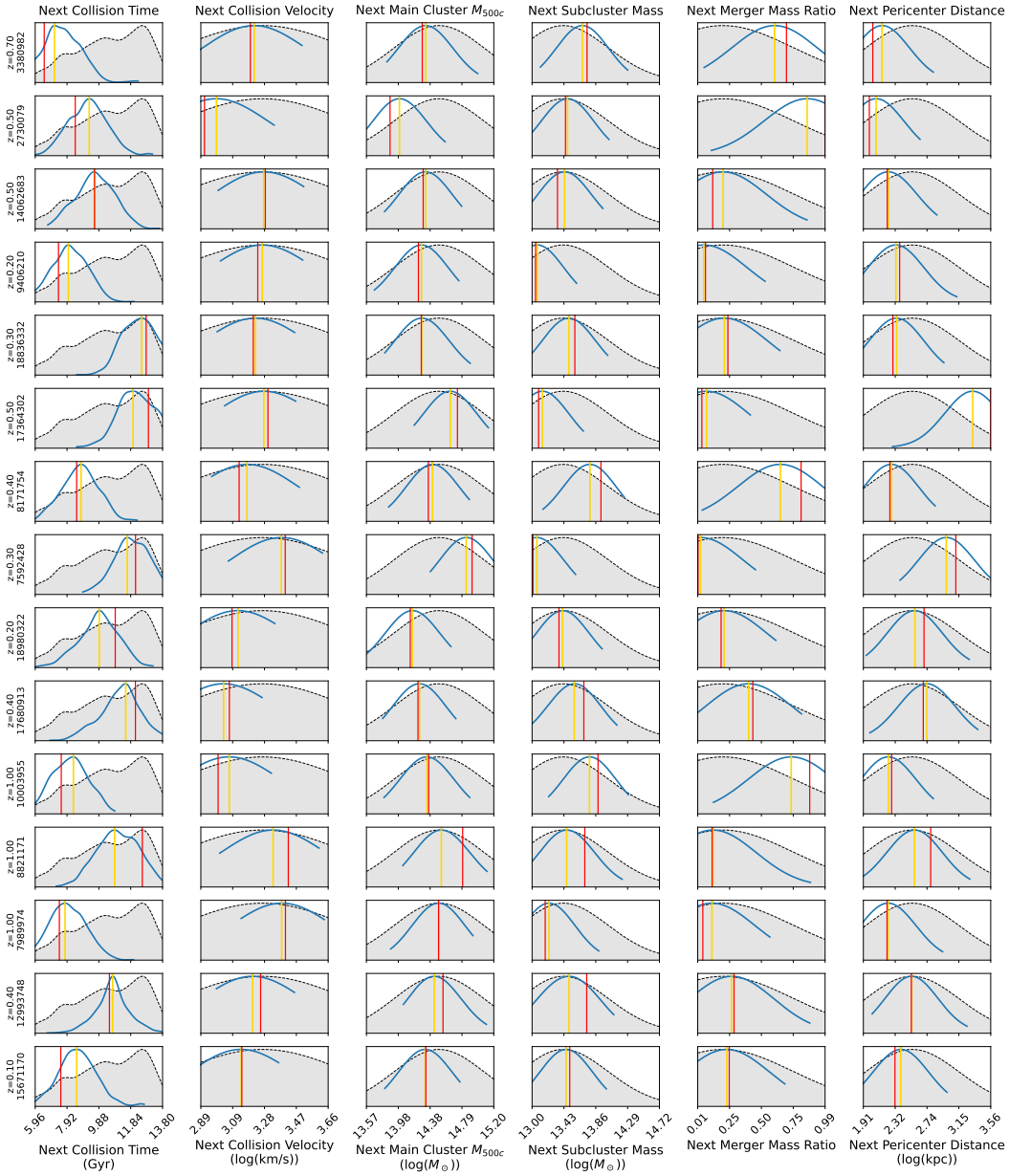

Figure 65: Next–merger: posterior distributions conditioned on the **joint X–ray + radio** representation (15/592 test clusters). Gray: prior; blue: posterior; gold: MAP; red: ground truth. Posteriors remain concentrated around the truths but are modestly broader than in the last–merger case, chiefly for Collision Time; performance remains intermediate between X–ray-only and radio-only.

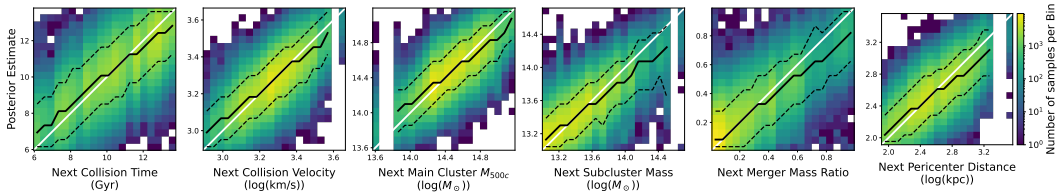

Figure 66: Next–merger: posterior vs. truth per target (joint X–ray + radio) across 592 test clusters. Same construction as the last–merger plot with $B = 20$ and $n_{sam} = 500$. The white diagonal shows $y = x$. Black solid: posterior median; black dashed: 10–90% quantiles. Bands broaden slightly relative to last–merger, most visibly for Collision Time, while calibration remains intermediate between X–ray and radio.

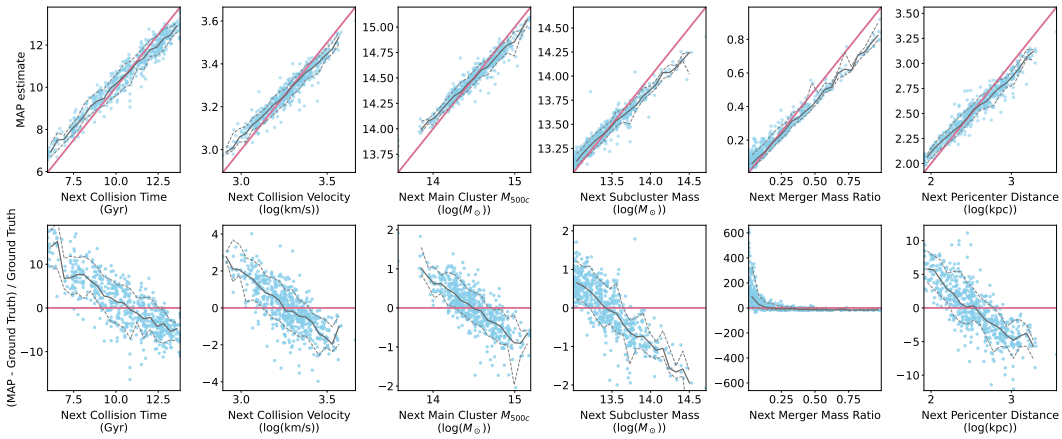

Figure 67: Next–merger: per–target MAP accuracy (top) and relative error (bottom) under joint X–ray + radio conditioning across 592 test clusters. **Top:** MAP vs. truth with medians (black solid) and 10–90% envelopes (black dashed); the **pink diagonal** marks $y = x$. **Bottom:** relative MAP error Δ with the same line styles; the **pink horizontal** line marks $\Delta = 0$. MAP scatter and error ranges remain close to the last–merger case and stay between X–ray-only and radio-only performance.

21.5 DISCUSSION

Conditioning on the *joint* X–ray+radio representation space yields performance that is consistently *intermediate* between radio-only and X–ray-only across merger parameters (Figures 62, 63). As can be seen in Figure 61, posteriors are contracted relative to the prior and vary across different clusters, with MAP estimation close to the ground truth. The contraction, and hence the precision, is not as strong as seen for radio, but similar and slightly stronger than X-ray, indicating that joint radio and X-ray representation carries genuine conditioning signal. Posterior–vs–truth heatmaps exhibit thin diagonal ridges with clear posterior contraction, and MAP–vs–truth bands show small scatter with typical relative MAP errors of: Collision Time ~ ±10%, Collision Velocity ~ ±2%, Main Cluster M_{500c}/Subcluster Mass ~ ±1%, Pericenter Distance ~ ±5%. Mass Ratio remains the most delicate at small values. Relative to X–ray alone, joint conditioning tightens posteriors and reduces outliers; relative to radio alone, posteriors are typically broader. This suggests that while thermal and non–thermal cues are complementary, in these simulations the intrinsic radio morphology is the most directly informative about merger parameters.

Although joint posteriors can be as wide as X–ray for some merger parameters, calibration is slightly stronger (straighter median curves), 10 and 90 % quantile lines are slightly smaller, and MAP scatter and relative error is noticeably reduced. In other words, we often see "wide but honest" posteriors with low MAP error: multiple solutions remain plausible, yet the posterior mode sits near the truth. This pattern is consistent with the relatively smooth gradients in Figure 59 and with the cINN exploiting *joint* structure in the full 512-D representation space, not the 2D UMAP. Also it can be noted that, in the cINN, the well-organized properties (e.g., masses, velocity) could help constrain weaker ones (e.g., pericenter, and collision time).

WHAT THE JOINT REPRESENTATION SPACE CAPTURES. The joint SimCLR encoder learns a shared representation space that fuses the *thermal* morphology; e,g., concentration/cuspiness, COM offsets, and multiple peaks, and *non-thermal* shock geometry, e.g., relic presence, curvature, bilateral symmetry and thickness. Our learned UMAP, is organized into three broad bands from top–left to bottom–right: (i) a *right* band dominated by cuspy X–ray morphologies with little or no radio emission; (ii) a *middle* band of disturbed, merging systems, where the *upper* locus is radio–bright (pronounced relics/halos) and the *lower* locus shows weaker non–thermal signal; and (iii) a *left* band populated by flatter X–ray profiles, again with the upper portions exhibiting relics and the lower portions fainter emission. This stratification is physically sensible: horizontal position tracks thermal core state and overall relaxation, while vertical position reflects shock–related radio geometry and brightness.

WHAT THE CINN LEARNS FROM THE JOINT CONDITIONING. The flow models the *joint* density $p(\mathbf{x} \mid \mathbf{c}_{\text{joint}})$ with conditional spline RQS couplings, effectively factorizing $p(\mathbf{x} \mid \mathbf{c}) = p(x_{\pi_1} \mid \mathbf{c}), p(x_{\pi_2} \mid x_{\pi_1}, \mathbf{c}) \cdots$. In effect, targets that are more tightly organized in the joint representation space (e.g. judging from the UMAP; Main Cluster's M_{500c}, Collision Velocity, to lower extent Subcluster Mass and

Merger Mass Ration) become informative parents for targets that are less directly organized (e.g, Pericenter Distance and Collision Time) through learned correlations as in Figure 21.3. This is visible as tight MAP–vs–truth bands and posterior contraction even when a single merger property shows less-smooth UMAP gradients.

MODELING CAVEAT: WHY CROSS–MODAL INCONSISTENCIES CAN ARISE. Even with *intrinsic* maps (no instrumental effects), thermal X–ray and non–thermal radio respond to different physics, timescales, and viewing geometries, so apparent "mismatches" are expected. X–ray surface brightness scales as $\int n_e^2 \Lambda(T, Z)\, d\ell$, emphasizing dense cores and pressure structure over Gyr-scale relaxation, whereas radio synchrotron highlights thin shock sheets where recently accelerated electrons radiate in μG fields, with emissivity set by local Mach number, B, and ageing losses. Consequently: (i) *projection*; relics brighten for edge-on shocks (long path length) but can be faint or halo-like face-on, while the X–ray, as a line-of-sight integral of n_e^2, can remain elongated or multi-peaked even when shocks project weakly; (ii) *timescale offset*; GHz synchrotron is transient ($\sim 10^8$–10^9 yr) whereas X–ray asymmetries and centroid shifts can persist $\geqslant$Gyr, so bright outer relics may coexist with comparatively symmetric cores (early post-pericenter), or disturbed X–ray structure may remain after radio emission fades; (iii) *microphysics and environment*; weak–Mach encounters, low magnetization, or higher redshift (stronger inverse–Compton losses with $B_{CMB} \propto (1+z)^2$) suppress radio even when X–ray disturbance is obvious, while strong outer shocks can light up relics despite relatively regular central X–ray isophotes. Two dataset choices can accentuate these effects: percentile/log intensity normalizations intentionally remove absolute flux (masking very faint features), and in joint experiments the X–ray is downsampled to the radio grid, reducing small–scale thermal detail. These ingredients explain cases like elongated X–ray isophotes without obvious relics (face-on/weak shocks, weak B, high z) or bright relics coexisting with seemingly symmetric X–ray cores (very recent passage, shocks at large radius), and they justify broader credible intervals where cross–modal cues are genuinely in tension.

REGRESSION TO THE MEAN AND CALIBRATION. Modest shrinkage toward the population mode (regression to the mean) persists in the joint conditioning, however, it is milder than in X–ray–only conditioning reflecting the added constraints contributed by radio features. Our calibration proxies support that these gains are reliability improvements rather than overconfidence: posterior–vs–truth heatmaps show better calibration with lower scatter in MAP estimation and relative error. Quantitatively, the joint 80% posterior envelopes are narrower than X–ray–only and typically wider than radio–only, matching the "intermediate" behavior seen in MAP scatter. Residual curvature is consistent with (i) genuine non-identifiability and (ii) the downsampling of X–ray to the radio grid, which slightly attenuates detail (which also can be seen in less smooth distribution in thermodynamical parameters in the UMAP (Figure 58)). Overall, the joint model remains *well calibrated*: it is slim where it should be and honestly wide where the data cannot uniquely resolve the merger physics.

ROLE OF THE MIXTURE-OF-EXPERTS (MOE). Routing in the joint space could separates coherent regimes, for example, flat cores with single and double radio relics or weaker radio emissions, disturbed cores with relics or radio halos, relaxed clusters with cuspy profile and low or quiet radio surface brightness. Expert-local flows reduce averaging across these regimes, improving sharpness and coverage where a single global model would blur multi-modality.

INTRINSIC MAPS, AUGMENTATION, AND FUSION CAVEATS. Joint inputs rely on *intrinsic* maps (no PSF/beam, uv-coverage, Poisson/background, or spectral ageing). Our physics-aware SimCLR augmentations (Section 14.2), e.g., flips/rotations for orientation invariance; affine zoom/translations for centering/scale; Gaussian blur for PSF/beam surrogates; additive Gaussian noise with for depth variation, help the joint representation space ignore nuisances and focus on morphology. Still, two caveats remain: (i) downsampling X–ray to the radio grid caps thermal spatial detail; a more faithful fusion would be multi-wavelength and instrument-aware; (ii) radio emissivity (shock-based, thin-sheet, simulated B) and the absence of observational systematics (beam, uv, ageing/depolo-rization) can amplify the radio advantage in simulation. Forward modeling and sim–to–real adaptation are therefore necessary steps before applying joint posteriors directly to survey data.

NEXT–MERGER (FORECASTING): JOINT X–RAY+RADIO CONDITIONING. The joint conditioner offers an intermediate next–merger forecasts: posterior bands are typically intermediate in between radio–only and X–ray–only, with MAP scatter larger than radio and smaller than X-ray only. As for the other modalities, timing is the most affected by forecasting, with broader credible intervals for Collision Time; nonetheless, posterior–vs–truth ridges remain diagonal and contracted relative to the prior, indicating that the joint representation space carries genuine predictive signal for upcoming merger events.

Part VIII

DISCUSSION AND RESULTS

22

RESULTS AND DISCUSSION

We construct a simulation dataset from the 352 zoom-in halos of TNG-Cluster, rendered in three orthogonal projection across 8 snapshots spanning $0 \leqslant z \leqslant 1$. For every galaxy cluster at snapshot-projection pair, we use the intrinsic thermal X-ray maps (from Nelson et al. [113]), and intrinsic non-thermal radio maps (from Lee et al. [91]). We further assemble merging event labels at both last and next merger events using first pericenter passage as the collision time definition (as derived by Lee et al. [91]). The merger parameters are: Collision Time, Main Cluster's M_{500c}, Subcluster Mass, Mass Ratio, Pericenter Distance and Collision Velocity. To compress high-resolution maps into informative descriptors, we train SimCLR encoders separately on X-ray, Radio and joint X-ray+radio maps. Due to the triaxial nature of galaxy clusters, the three projections are treated as independent samples, resulting in a total of 8448 maps. During training, a physics-aware augmentation suite, e.g., random flips/rotations, affine zoom and translations, Gaussian blur, and Gaussian noise, is applied to form positive pairs and minimize the NT-Xent loss (eq. 31). This yields to a 512-D representation space, that will serve as the conditioning vector $\mathbf{c}$ for inference in the next step. As a baseline without representation learning, we also build a scalar conditioner form observable properties, which will be directly passed as the conditioning vector $\mathbf{c}$.

For the next step, we use a conditional invertible neural network (cINN) that learns an invertible, bijective map $f : (\mathbf{x}, \mathbf{c}) \to \mathbf{z}$ with tractable Jacobian, where $\mathbf{z} \sim \mathcal{N}(0, I)$, and $\mathbf{x}$ is the target vector (merger parameters). To exploit the representation space's clustered structured, we partition the $\mathbf{c}$-space with k-means and train expert-local cINN. The cINN is a stack of eight rational-quadratic spline (RQS) coupling blocks, with permutations inserted in between. In each block, the target vector $\mathbf{x}$, is split into pass-through $\mathbf{x}_a$, and transformed $\mathbf{x}_b$. A small conditioner network (subnet) takes $[\mathbf{x}_a, \mathbf{c}]$ and outputs the RQS parameters of the transform on $\mathbf{x}_b$. Stacking these permuted blocks yields an effective autoregressive factorization of $p(\mathbf{x} \mid \mathbf{c})$ (i.e., $p(x_{\pi_1} \mid \mathbf{c})\, p(x_{\pi_2} \mid x_{\pi_1}, \mathbf{c}) \cdots$), allowing the model to capture cross-parameter dependencies. Training maximizes the conditional likelihood on simulated pairs $(\mathbf{x}, \mathbf{c})$ (equivalently, minimizing the NLL loss as in equation 20). At inference time, posterior samples are obtained by drawing $\mathbf{z} \sim \mathcal{N}(0, I)$ and inverting the flow, $f^{-1}(z, c) = x$, yielding samples from $p(\mathbf{x} \mid \mathbf{c})$.

This chapter synthesizes the empirical findings across the four conditioning setups; Scalar, X–ray, radio, and joint X–ray+radio maps, based on the common evaluation tools introduced earlier (posterior grids, posterior–vs–truth heatmaps, MAP–vs–truth with relative errors, and correlation recovery). We highlight where the self–supervised representations of maps carry the most information for merger physics, analyze systematic error patterns, and discuss implications for both astrophysics and practice.

22.1 EVALUATION SUMMARY AND PROTOCOL

We evaluate the conditional posteriors $p(\mathbf{x} \mid \mathbf{c})$ for *last–merger* and *next–merger* targets using:

1. per–object posterior distribution (blue posterior kernel density estimate (KDEs), gray test–set priors, gold MAP, red truth; e.g., Figs. 17, 33, 48, 61); the performance is accurate when the maximum a posteriori (MAP) estimate lies close to the ground truth, and precise when the posteriors are tightly contracted around the ground truth.

2. population heatmaps (histograms) of posterior draws in bins of ground truth versus truth with median and 10–90% bands (Figures 18, 34, 49, 62); calibrated posteriors is when the median line follows the identity line ($y = x$). Performance is precise when 10% and 90% percentile lines are narrowly around the identity line.

3. MAP–vs–truth with bin–wise envelopes and relative–error panels (Figures 19, 35, 50, 63); in the ideal case, the MAP etimate points are clustered along the identity line, with narrow quantile lines, and low relative errors centered near zero.

4. recovery of cross–target correlation recovery via corner plots (Figures 20, 36, 51, 64).

22.2 POSTERIOR CALIBRATION AND POINT–ESTIMATE ACCURACY

Scalar-conditioned

Conditioning the cINN on scalars yields informative posteriors but not across all of the targets, and in general, the inference performance is weaker compared to the representation-based conditioning. Posterior's calibration were only reliable for Collision Time with relative MAP error $\sim [-20, 40]\%$, and Main Cluster M_{500c} with MAP error of $\sim \pm 3\%$. Calibration is also reasonable at *high* collision velocities and *large* pericenter distances. However, elsewhere, especially for Subcluster Mass and Mass Ratio, the posterior medians depart from the identity line and the 10–90% bands widen (heteroscedasticity). This validates the central premise that reducing $10^5 - 10^6$ pixels to a few scalars creates an information bottleneck that drops multi–scale morphology and shock geometry, making inference across most targets fragile and less calibrated compared to using a learned representation space.

X–ray–conditioned (thermal morphology)

X–ray representation space yield well–calibrated posteriors whose medians track the identity line (the ideal case) across targets with modest shrinkage (regression to the mean) toward modal scales (Figures 34 and 35). Typical relative MAP error ranges are: Collision Time ($\sim [-20, 40]\%$), Collision Velocity ($\sim \pm 5\%$), Main Cluster M_{500c} and Subcluster Mass ($\sim \pm 2\%$), Pericenter ($\sim \pm 10\%$), while Mass Ratio remains the most delicate (bounded/fractional) with larger relative errors at small

ratios (Figures 34, 35). As seen in Figure 33, posteriors are contracted relative to the empirical prior and differ across different clusters, indicating that X-ray representation carries genuine conditioning signal rather than reproducing the prior.

This performance however can be expected from relative smooth transition in Figure 31. While the transition might not be as smooth in some merger parameter such as Collisioin Time, and Pericenter Distance, as explained in 11.1, the cINN receives the concatenation of condition (here is X-ray representation, and half of the merger parameters). Therefor, good prediction in some targets with their relative positions on the representation space, can compensate for this. That said, UMAP is only the projection, and there might be more organization in the 512 dimensional space, which is the condition used for the cINN.

Radio–conditioned (non–thermal morphology)

Conditioning on radio representation space yields the tightest posteriors across *all* merger parameters and the smallest MAP scatter relative to X–ray and joint conditioning (Figs. 49, 50). Typical relative MAP error ranges are narrower: Collision Time ($\sim [-5, 10]\%$), Collision Velocity ($\sim \pm 1\%$), Main Cluster M_{500c} and Subcluster Mass ($\sim \pm 0.5\%$), Pericenter Distance ($\sim \pm 3\%$); Mass Ratio still broadens at small values but improves overall. Regression to the mean is present but very weak compared to the rest of the conditioning: posterior median lie essentially on the identity line with negligible offsets, and the envelopes are correspondingly tight. As seen in Figure 48, posteriors contract much more strongly relative to the empirical priors, confirming that the radio embedding provides a high–information conditioner. This advantage is consistent with the very smooth, label–correlated gradients observed in the radio UMAP (Figure 46), which indicate well–organized neighborhoods in the full 512-D space and supply the cINN with stronger conditioning signals.

Joint X–ray+Radio conditioning

The joint setup performs *intermediately* between X–ray and radio map conditioning for most targets; with posteriors narrower than X–ray, and not as tight as radio. The relative MAP error ranges are: Collision Time ($\sim \pm 10\%$), Collision Velocity ($\sim \pm 2\%$), Main Cluster M_{500c} and Subcluster Mass ($\sim \pm 1\%$), Pericenter Distance ($\sim \pm 5\%$) (Figs. 62, 63). As in Figure 61, posteriors contract relative to the prior with a weaker contraction than radio, but slightly stronger than X-ray, again hinting that the joint representation carries genuine conditioning signal. While calibration modestly improves over X–ray alone conditioning, the envelopes are tighter and the MAP scatter is noticeably reduced. However, the radio conditioning, still has the strongest performance.

This pattern is consistent with the relatively smooth gradients in Fig. 59 and with the cINN exploiting joint structure in the full 512-D conditioner (not the 2-D UMAP): well-organized coordinates (e.g., masses, velocity) help constrain weaker ones (e.g., pericenter distance, collision time) via learned cross-target covariances. This suggests that while thermal and non–thermal cues are complementary, in these simulations the radio morphology is the most directly informative about merger parameters.

22.3 FORECASTING PERFORMANCE: NEXT–MERGER TARGETS

For *next–merger* inference, as discussed in Sections 12.5, 16.4, 19.4, and 21.4 we see that the results for each conditioning is similar to the last mergers, with radio conditioning having the best performance, followed by joint, X-ray and scalar conditioning. In most cases, the performance slightly degrades with the broadening being most visible for Collision Time. MAP error ranges remain close to their last–merger counterparts for radio and joint conditioning, indicating stable forward prediction; X–ray also remains competitive but with wider bands for timing.

22.4 CROSS–TARGET CORRELATIONS AND PHYSICAL CONSISTENCY

Correlation plots show that the cINN recovers the qualitative ΛCDM–consistent correlations among targets in all modalities (Figs. 36, 51, 64): e.g., links between total mass and collision velocity, pericenter and timing, and mass ratio are preserved. Posterior clouds align with truth loci; MAP clouds overlay without large bias.

Importantly, because the flow models $p(\mathbf{x} \mid \mathbf{c})$ jointly via conditional spline couplings that effectively factorize the density autoregressively, well–organized coordinates in the embedding could inform weaker ones. As a result, even when a single coordinate shows less smooth UMAP color gradients, the cINN can still tighten that coordinate's posterior by leveraging learned cross–target covariances. This is visible as narrower, better–aligned posterior contours than would be expected from the 2D embedding alone. Consistent with the modality ranking, radio–conditioned posteriors are generally the least dispersed, reflecting the sharper organization of information in the radio embedding; the joint conditioner sits between radio and X–ray, but similar to the rest of the cases, it still benefits from the same correlation–driven sharpening.

22.5 READING THE EMBEDDINGS

UMAP overlays of the learned representation spaces (X–ray, radio, joint) reveal smooth, label–correlated gradients for some parameters in both observables and merger parameters (Figs. 29–30-31; 44–45–46; 57–58–59). Radio displays especially smooth variations along most merger parameters. Joint embedding and X-ray, also show smooth transitions in some merger parameters (including collision velocity and masses), with Joint embeddings sharpening some of those trends further. That said, UMAP is a nonlinear 2D projection of a 512-D representation: it preserves some local neighborhoods but can distort global distances and smoothness. Consequently, ragged color maps in 2D do not imply poor conditioning. The cINN consumes the full high-dimensional embedding, and the observed posterior contraction and MAP accuracy reflect structure in that space, even when its 2D projection looks imperfect.

22.6 MIXTURE–OF–EXPERTS (MOE): SPECIALIZATION IN REPRESENTATION SPACE

Partitioning the conditional space with k–means (as explained in Section 15.1) and training per–cluster experts stabilizes optimization and encourages local spe-

cialization. By experimenting, MoE has proved to yield sharper posteriors due to the existence of heterogeneous regions of the embedding manifold and reduce failure cases where a single global cINN can overlook the extreme cases.

22.7 SYSTEMATIC ERROR PATTERNS AND FAILURE MODES

Across modalities we observe:

- **Regression to the mean.** In all cases the bin medians gently bend toward the population mode; the effect is strongest for X-ray and mildest for radio. This could arise from (i) finite data and overlapping morphologies, distinct merger states can look similar, so the conditional posterior spreads mass and its median drifts toward high-density regions; (ii) tail sparsity at extreme target values, which increases epistemic uncertainty and encourages conservative (central) estimates; and (iii) regularization effects in the flow, which favors the solution that are conservative across the population where training signal is weak.
- **Small–denominator effects.** Relative errors inflate for very small mass ratios even when absolute errors are modest.
- **Timing discretization.** Since the simulation data are recorded on snapshot base, it gives a discrete nature to time, so that the percentage errors are less forgiving despite tight calibration; radio maps conditioning reduces this sensitivity the most.

22.8 IMPLICATIONS AND OUTLOOK

The results establish a practical route from images to merger physics: label-free contrastive encoders compress high-resolution maps into informative representations, and cINN turns those into calibrated posteriors that respect degeneracies. For simulated intrinsic maps, radio morphology provides the most discriminative conditioning signal; while joint X–ray+radio improves reliability over thermal (x-ray) conditioning alone. Building on this foundation, several directions can make the method more realistic or more informative:

FROM SIMULATION TO SKY (INSTRUMENT-AWARE TRAINING). Replace intrinsic maps with instrumented mocks that, for X-ray data, include for example the point-spread function (PSF), spatially varying exposure and vignetting, energy-dependent effective area and bandpass, Poisson shot noise from source and background, and both instrumental or particle as well as astrophysical backgrounds; and that, for radio data, include for example the synthesized beam and primary-beam attenuation, uv coverage with weighting and deconvolution artifacts, correlated and thermal noise, bandpass and K-corrections to the observed frame, spectral ageing from synchrotron and inverse-Compton losses, and Faraday rotation and depolarization. Train on a mixture of intrinsic and instrumented mocks, then fine-tune SimCLR encoders on unlabeled survey cutouts using channel-consistent augmentations across frequency and polarization to reduce the simulation-to-real distribution shift.

MULTI-RESOLUTION, MULTI-WAVELENGTH FUSION. Preserve thermal and non thermal structure by fusing native-resolution X-ray and radio maps using multi-scale encoders such as feature pyramids or and cross-modal attention, instead of downsampling to the common low resolution. Incorporate additional tracers including SZ, weak-lensing mass maps, and optical galaxy density/kinematics with physics-aware co-registration that aligns astrometry, matches point spread functions and beams, and account for different pixel scale. Such multi-channel integration can exploit the complementary strengths of each observable to improve recovery of cluster mass profiles and merger geometry while reducing projection effects by linking information across wavelengths at their natural spatial scales

SCALABILITY AND DEPLOYMENT. Package the contrastive learning and MoE–cINN pipeline for survey processing (eROSITA, Chandra, XMM with LOFAR, MeerKAT, and VLA), exploiting amortized inference (milliseconds per object) and batched evaluation.

Part IX

APPENDIX

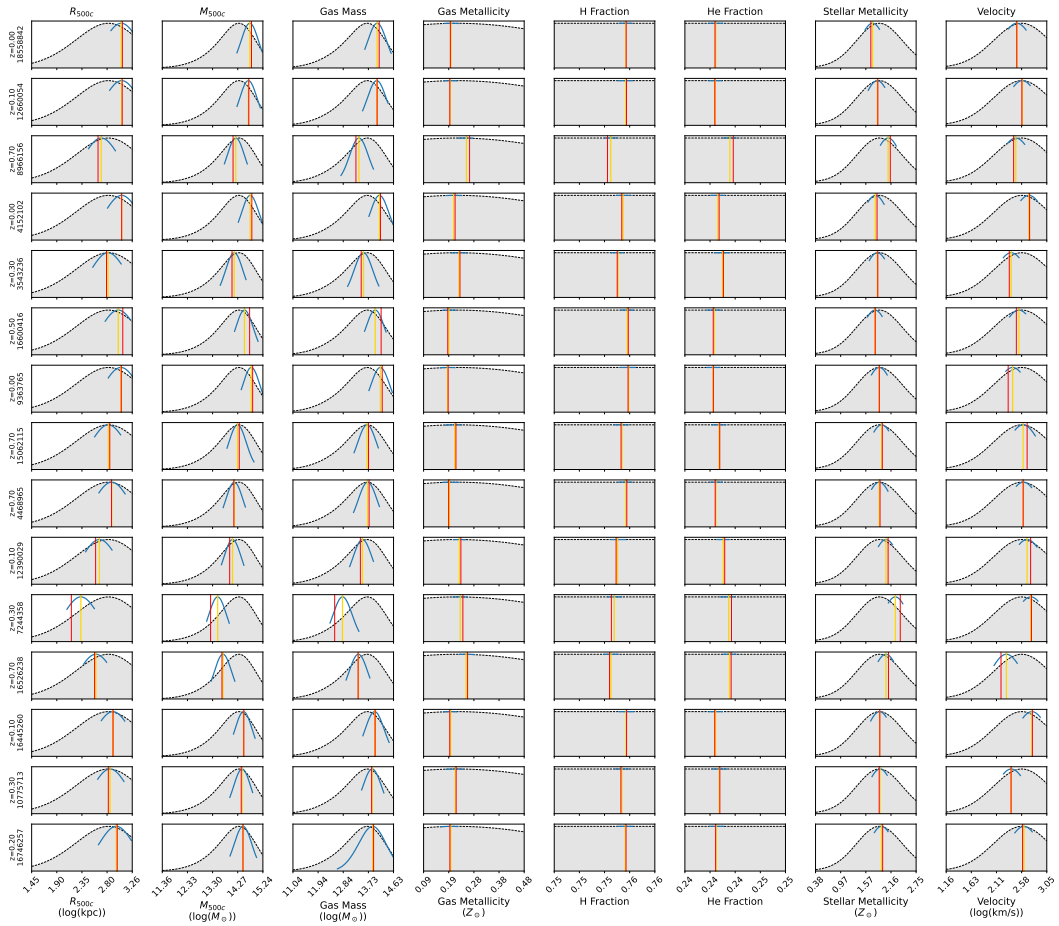

Figure 68: Posterior distributions of *halo-scale* observables inferred from X-ray embeddings (Table 4). Each panel corresponds to one observable; rows show 15 randomly selected test clusters. Gray: prior marginal distribution (KDE over the test set); blue: inferred posterior; gold: MAP estimate; red: ground truth. The strong overlap between posterior modes and true values indicates accurate calibration across R_{500c}, M_{500c}, gas mass, metallicity, and velocity.

.1 OBSERVABLES CONDITIONED ON X-RAY EMBEDDINGS

We assess how well the cINN, conditioned on the learned X-ray embedding $\mathbf{c}$, predicts the full set of observables $\mathbf{x}$ from Tables 4–5. We mirror the plotting schemes used in the main text: (i) per-object posterior grids (blue posterior KDE, gray prior KDE, gold MAP, red truth), (ii) posterior-vs-truth heatmaps (value-space $B = 15$ by and $n_{sam} = 500$), and (iii) MAP vs. truth with bin-wise medians and 10–90% envelopes, plus relative-error panels $\Delta = 100(\text{MAP} - \text{truth})/\text{truth}$. Across halo, BCG, ICM, and dynamical properties, posteriors are narrow and closely aligned with $y = x$, and MAP medians follow the identity with tight dispersion, consistent with the smooth relations seen in the hexbin maps for X-ray observables.

With interpretation of MAP error, we should be mindful of near zero values; Central Number Density, Central Cooling Time, Central Entropy, α Slope, Offset Magnitude, and M_{12}. The good performance is expected while we see smooth transition in the learned X-ray representation in figures 29 and 30.

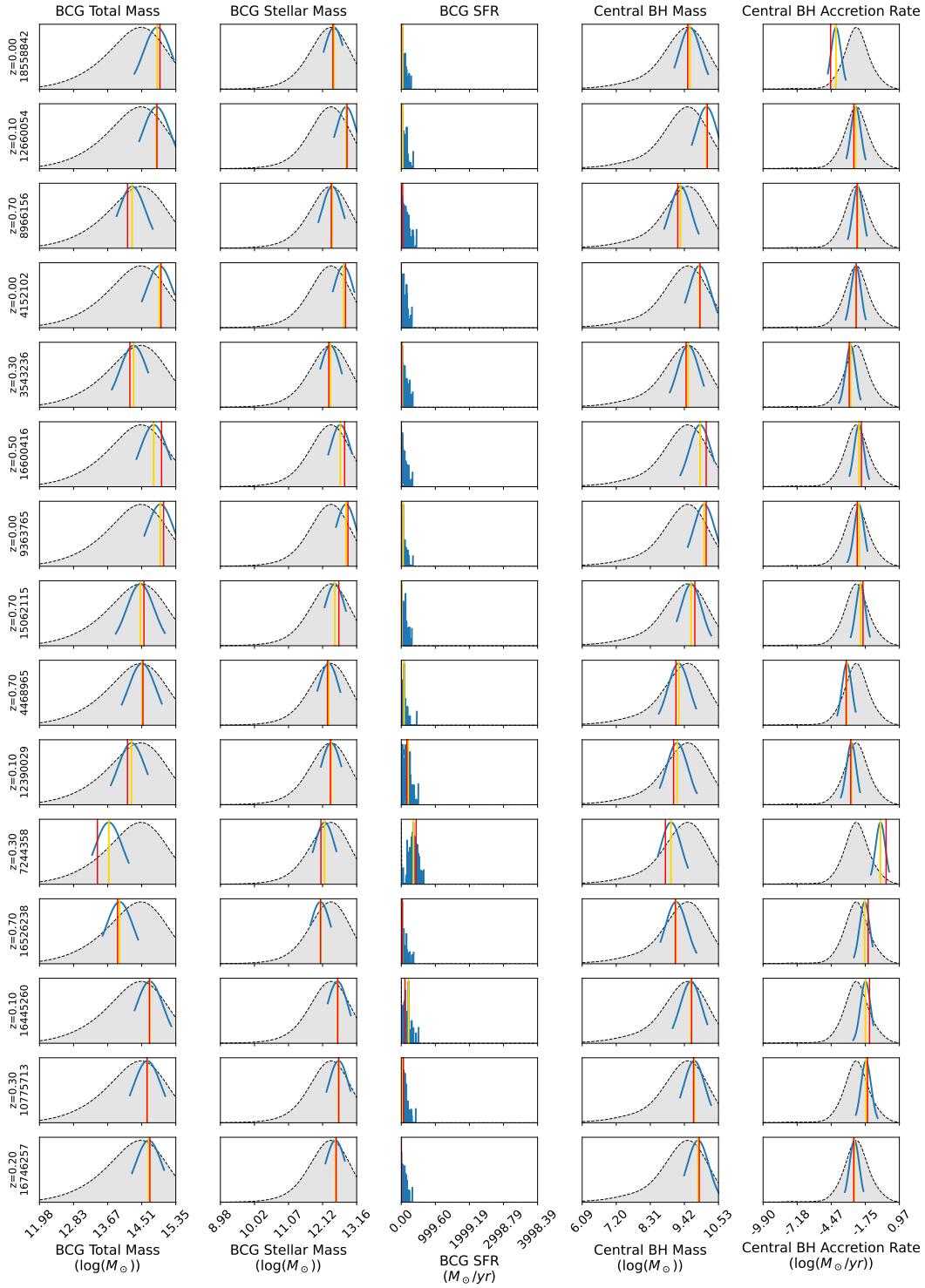

Figure 69: Posterior distributions of *BCG/BH* observables inferred from X-ray embeddings (Table 4). Same layout as Fig. 68. The posterior densities track the ground truth closely across BCG stellar mass, star formation rate, central black hole mass, and accretion rate. This demonstrate that the model is succesful in prediciting the BCG properties.

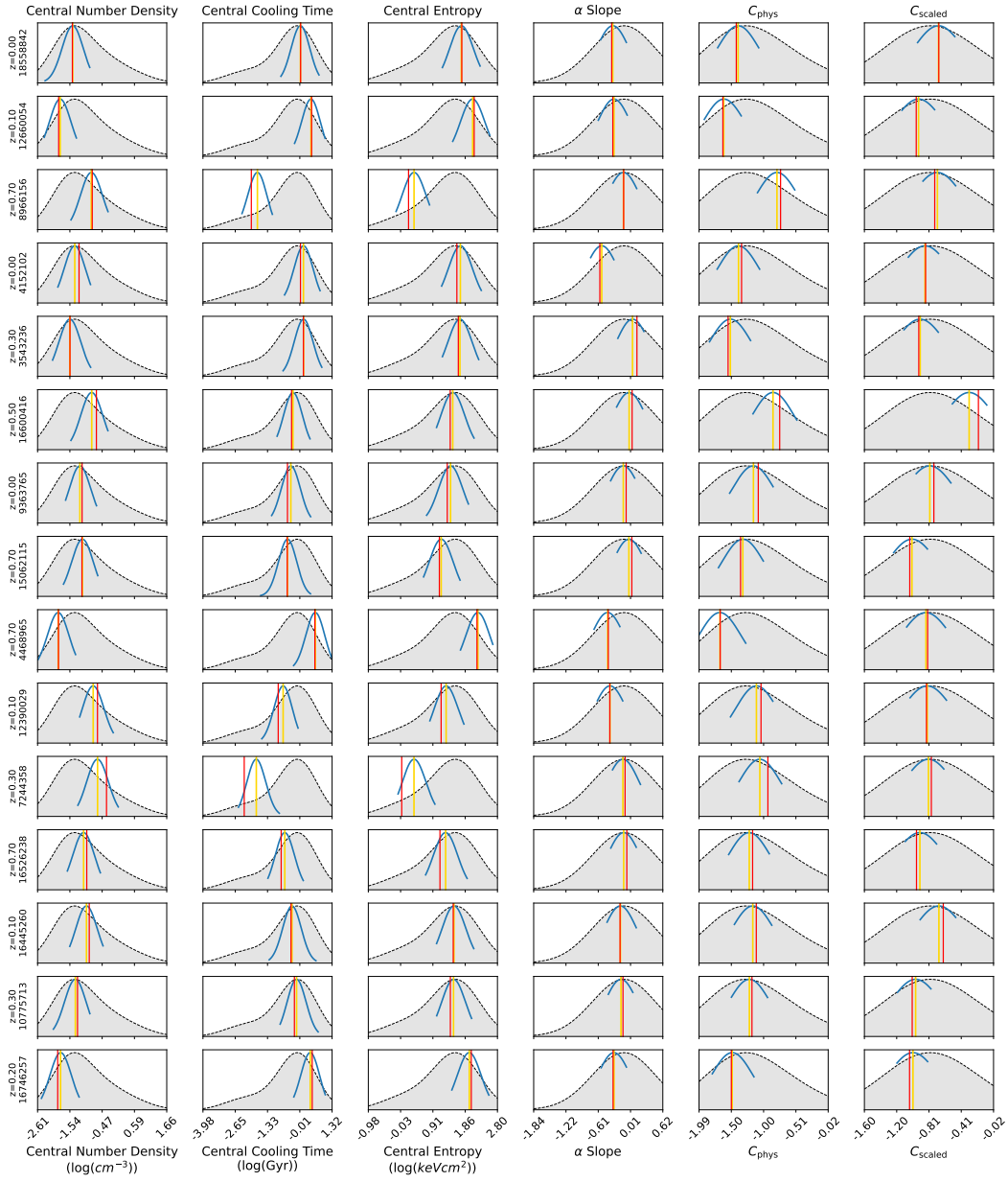

Figure 70: Posterior distributions of *ICM core* observables inferred from X-ray embeddings (Table 5). Same layout as Fig 68. The posteriors reproduce the true values for central electron density, cooling time, entropy, logarithmic slope α, and X-ray concentration indices (C_{phys}, C_{scaled}), demonstrating robust recovery of thermodynamical core structure.

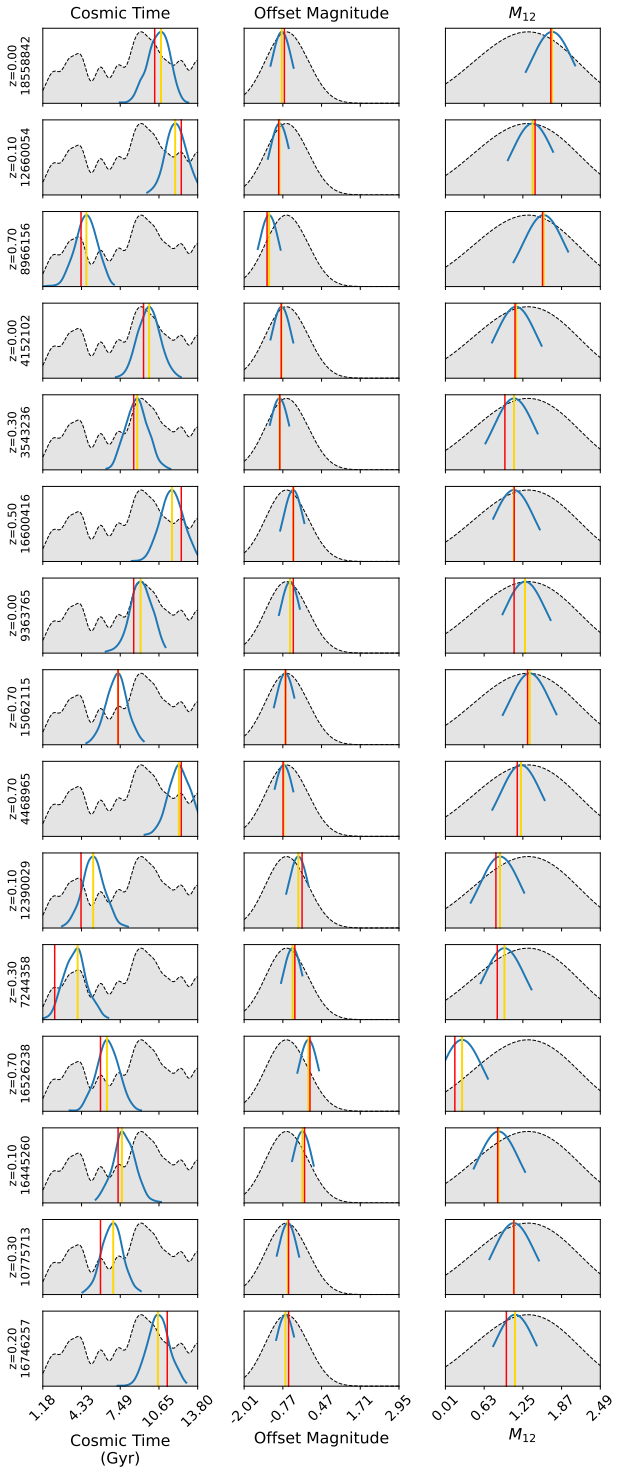

Figure 71: Posterior distributions of *dynamical state* observables inferred from X-ray embeddings (Table 5). Same layout as Fig 68. Inferred posteriors align well with the truth for cosmic time, center-of-mass offset, and the M_{12} merger statistic, supporting the method's ability to capture both structural and temporal aspects of cluster dynamics.

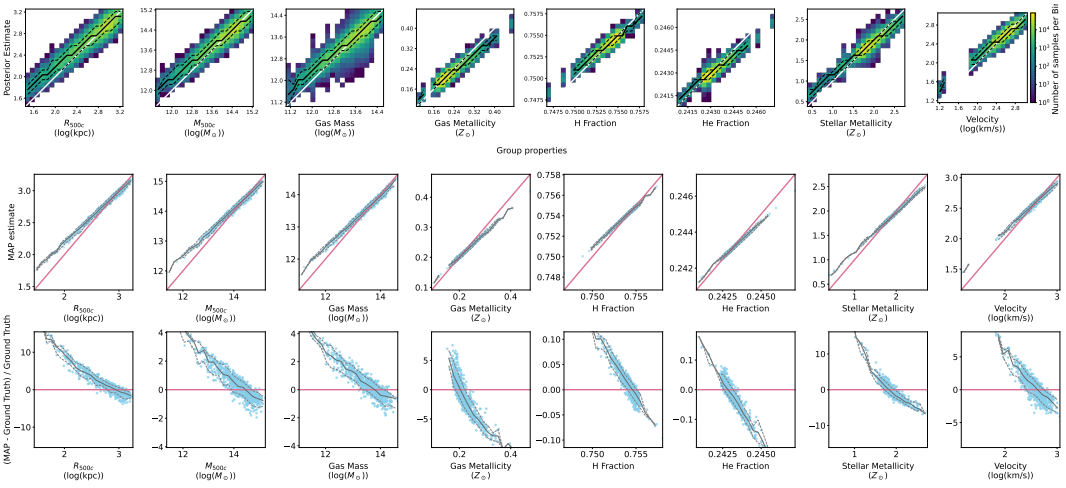

Figure 72: X-ray embedding → *halo-scale* observables (Table 4). (a) Posterior vs. truth heatmaps in value space (B = 15 bins, $n_{sam} = 500$ samples per object). White: $y = x$; black: posterior median (solid) and 10–90% quantiles (dashed). (b) MAP vs. truth (top) with $y = x$ in pink and bin-wise median (black solid) with 10–90% envelope (black dashed). Bottom: relative error $\Delta = 100(\text{MAP} - \text{truth})/\text{truth}$. Narrow, diagonal ridges and tight envelopes confirm small bias and dispersion across R_{500c}, M_{500c}, gas mass, metallicities, and velocity.

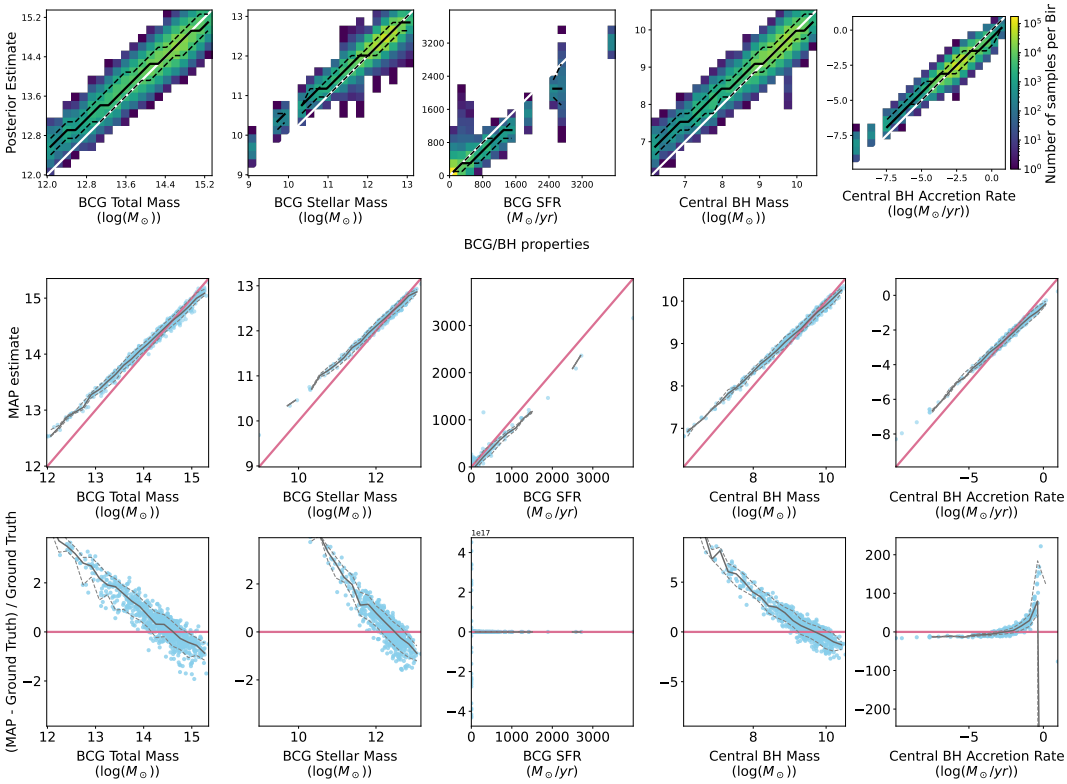

Figure 73: X-ray embedding → *BCG/BH* observables (Table 4). (a) Posterior vs. truth heatmaps (B = 15, $n_{sam} = 500$). White: $y = x$; black: median (solid) and 10–90% (dashed). (b) MAP vs. truth (top) and relative error Δ (bottom). Pink: $y = x$/zero; black: bin-wise median and 10–90%. Thin, diagonal ridges and tight error bands across BCG stellar mass, SFR, BH mass, and accretion rate indicate strong calibration.

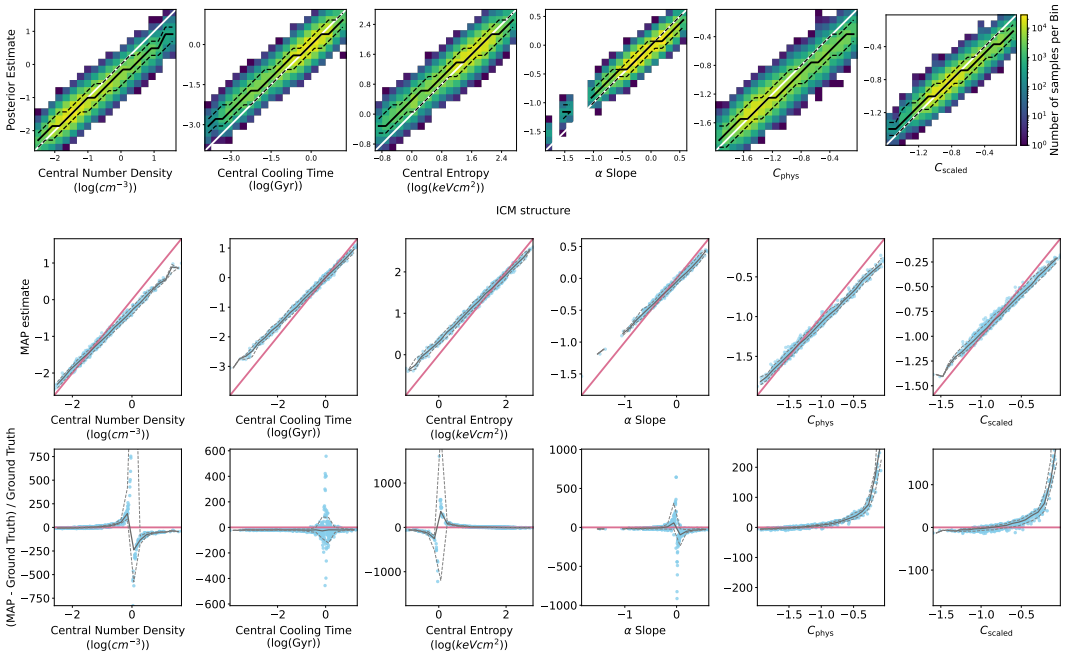

Figure 74: X-ray embedding → *ICM core* observables (Table 5). (a) Posterior vs. truth heatmaps (B = 15, n_{sam} = 500) for central number density, cooling time, entropy, α slope, and X-ray concentrations C_{phys}, C_{scaled}. White: y = x; black: median/quantiles. (b) MAP vs. truth (top) and relative error Δ (bottom). Bin-wise medians (solid) and 10–90% bands (dashed) remain tight with minimal curvature, confirming accurate point estimates across core diagnostics.

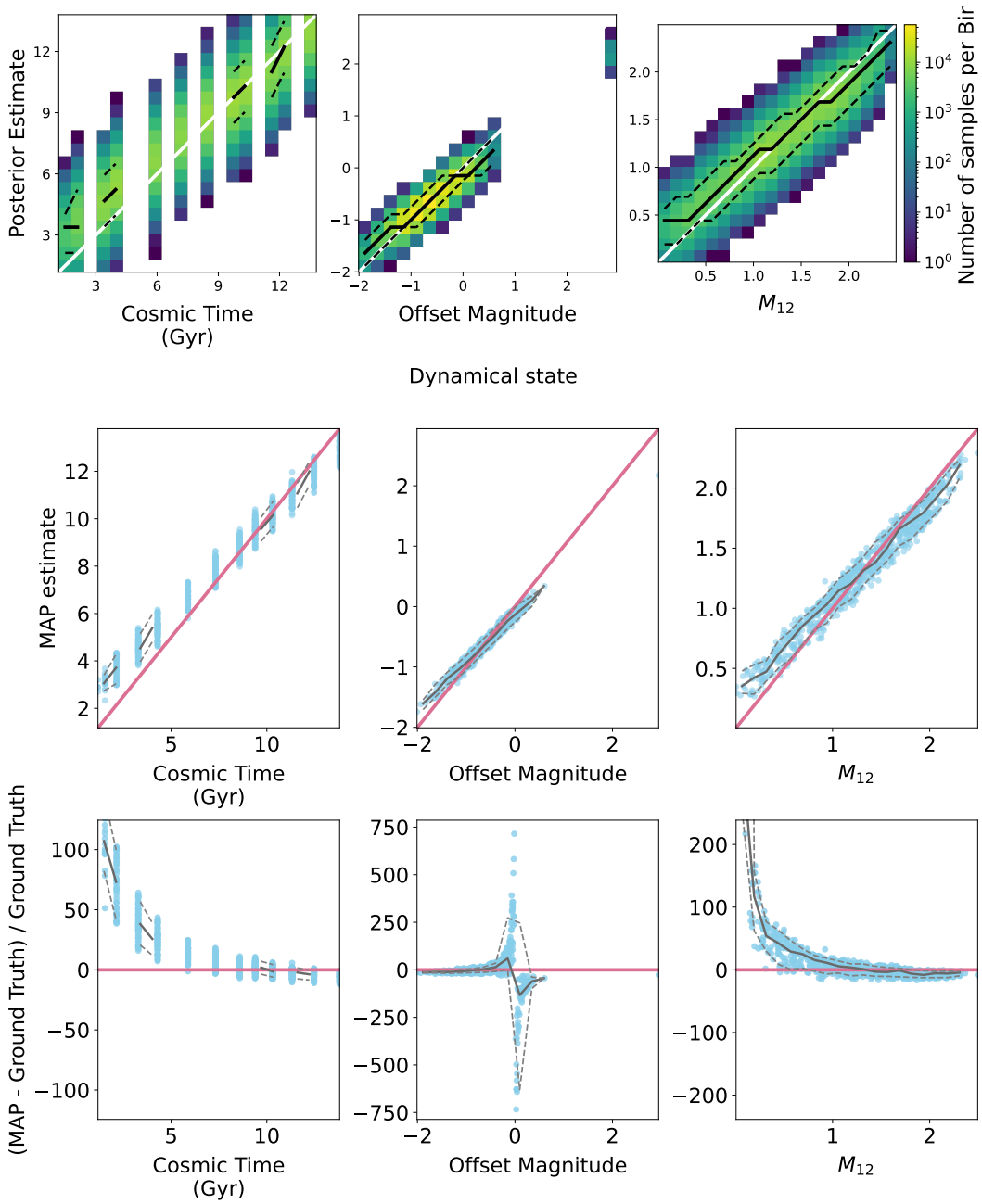

Figure 75: X-ray embedding → *dynamical state* observables (Table 5). (a) Posterior vs. truth heatmaps (B = 15, n_{sam} = 500) for cosmic time, COM offset, and M_{12}. White: $y = x$; black: median/quantiles. (b) MAP vs. truth (top) and relative error Δ (bottom). Medians lie near $y = x$ and $\Delta = 0$ with tight 10–90% envelopes. Small systematic bends are consistent with mild shrinkage near modal scales.

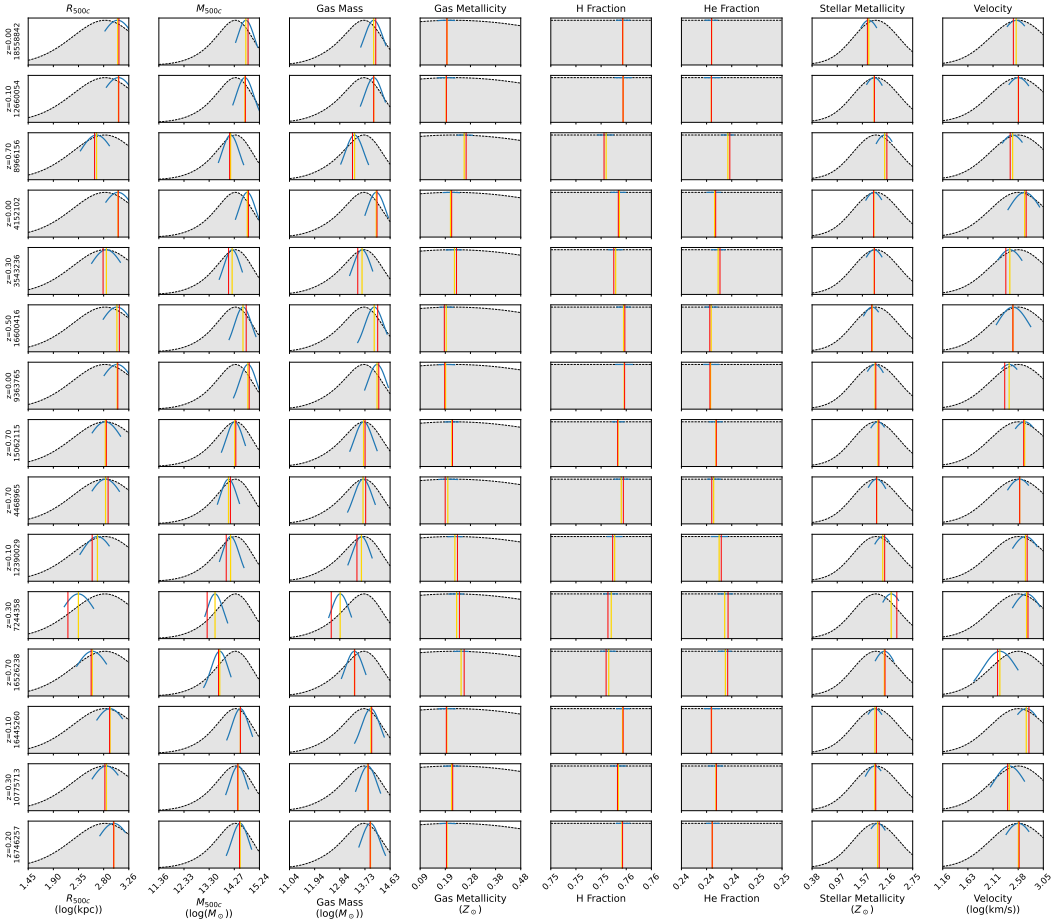

Figure 76: Posterior distributions of *halo-scale* observables inferred from radio embeddings (Table 4). Each panel corresponds to one observable; rows show 15 randomly selected test clusters. Gray: prior marginal distribution (KDE over the test set); blue: inferred posterior; gold: MAP estimate; red: ground truth. Strong overlap between posterior peaks and true values demonstrates accurate calibration for R_{500c}, M_{500c}, gas mass, metallicity, and velocity.

.2 OBSERVABLES CONDITIONED ON RADIO EMBEDDINGS

We now examine the performance of the cINN when conditioned on the representation learned from radio maps. The conditional input $\mathbf{c}_{\text{radio}}$ is obtained from the radio encoder, and the network is tasked with predicting the full set of cluster observables $\mathbf{x}$ listed in Tables 4–5.

For consistency, we employ the same visualization schemes used in the X-ray case: (i) posterior grids for randomly chosen clusters (posterior/prior KDEs, MAP, and ground truth), (ii) posterior–truth heatmaps constructed from $B = 15$ bins and $n_{\text{sam}} = 500$ posterior samples per object, and (iii) MAP vs. truth relations with bin-wise summaries and relative-error panels. These complementary diagnostics allow us to assess both global calibration and object-by-object predictive accuracy.

With interpretation of MAP error, we should be mindful of near zero values; Central Number Density, Central Cooling Time, Central Entropy, α Slope, Offset Magnitude, and M_{12}. The good performance is expected while we see smooth transition in the learned radio representation in figures 44 and 45.

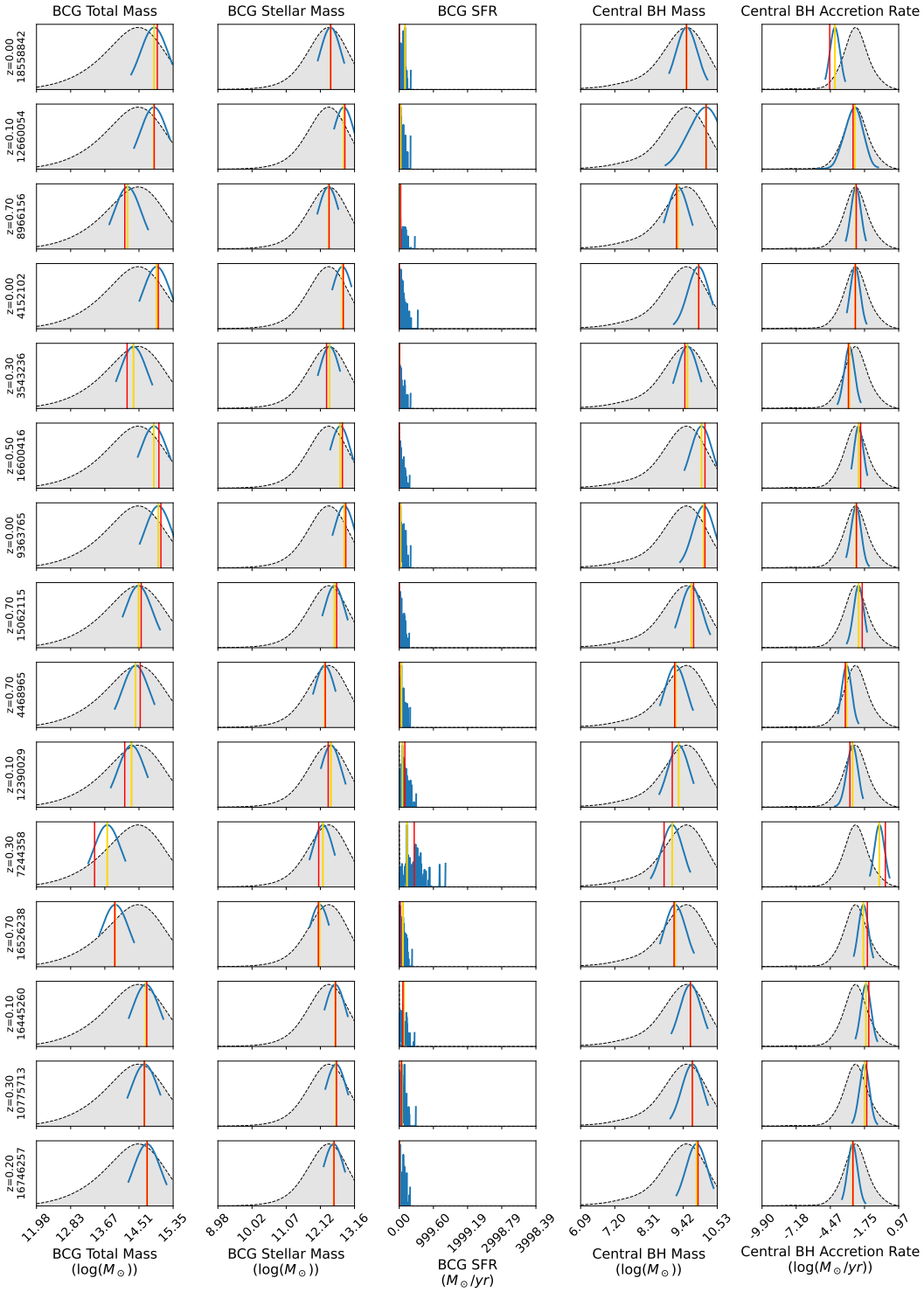

Figure 77: Posterior distributions of *BCG/BH* observables inferred from radio embeddings (Table 4). Same layout as Fig. 76. The posteriors reproduce the true values for BCG stellar mass, star formation rate, central black hole mass, and accretion rate, indicating that the radio features capture both stellar and AGN-related diagnostics.

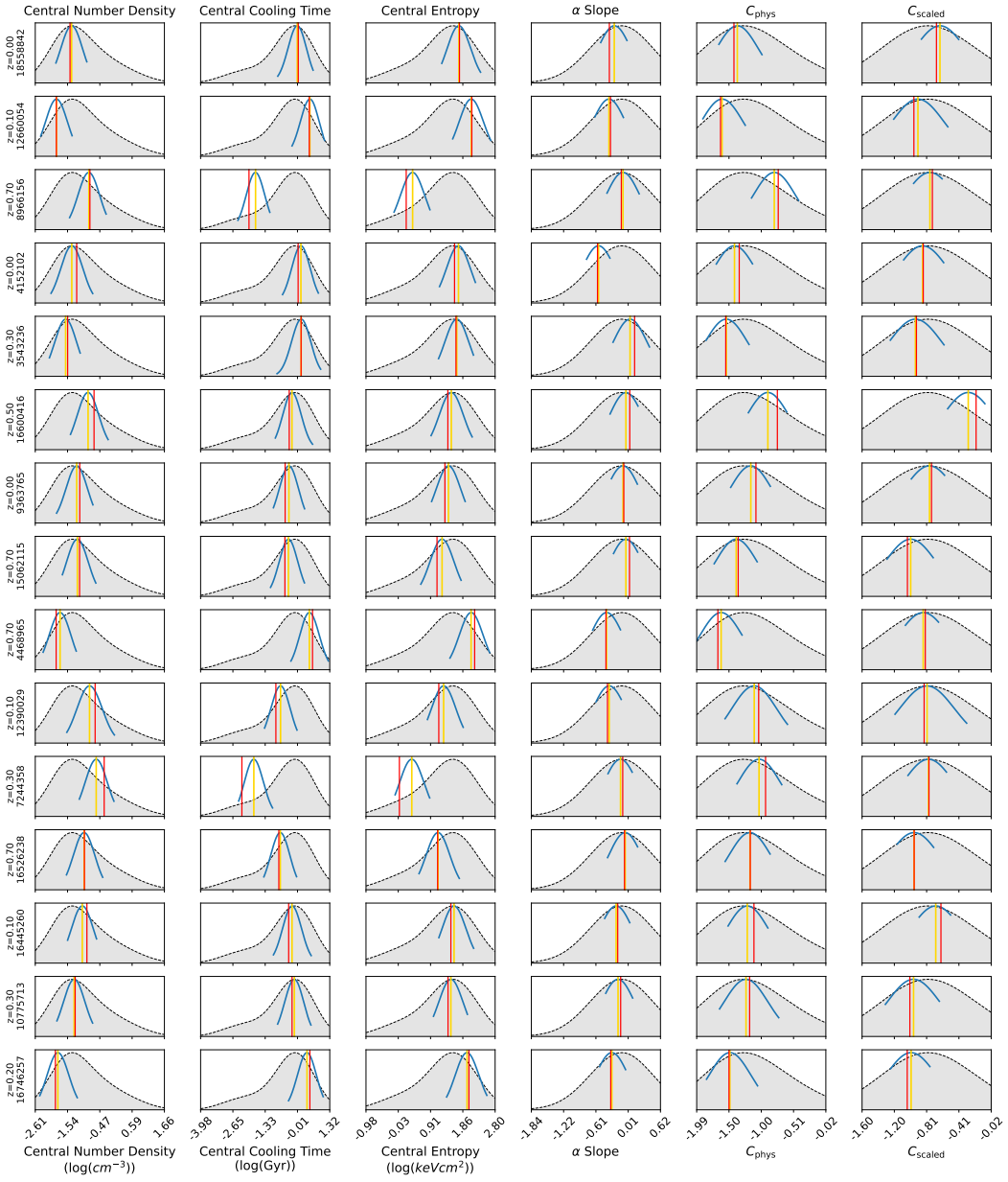

Figure 78: Posterior distributions of *ICM core* observables inferred from radio embeddings (Table 5). Same layout as Fig. 76. Inferred posteriors align well with the truth for central electron density, cooling time, entropy, logarithmic slope α, and concentration indices (C_{phys}, C_{scaled}), supporting the method's ability to constrain thermodynamical structure from radio morphology.

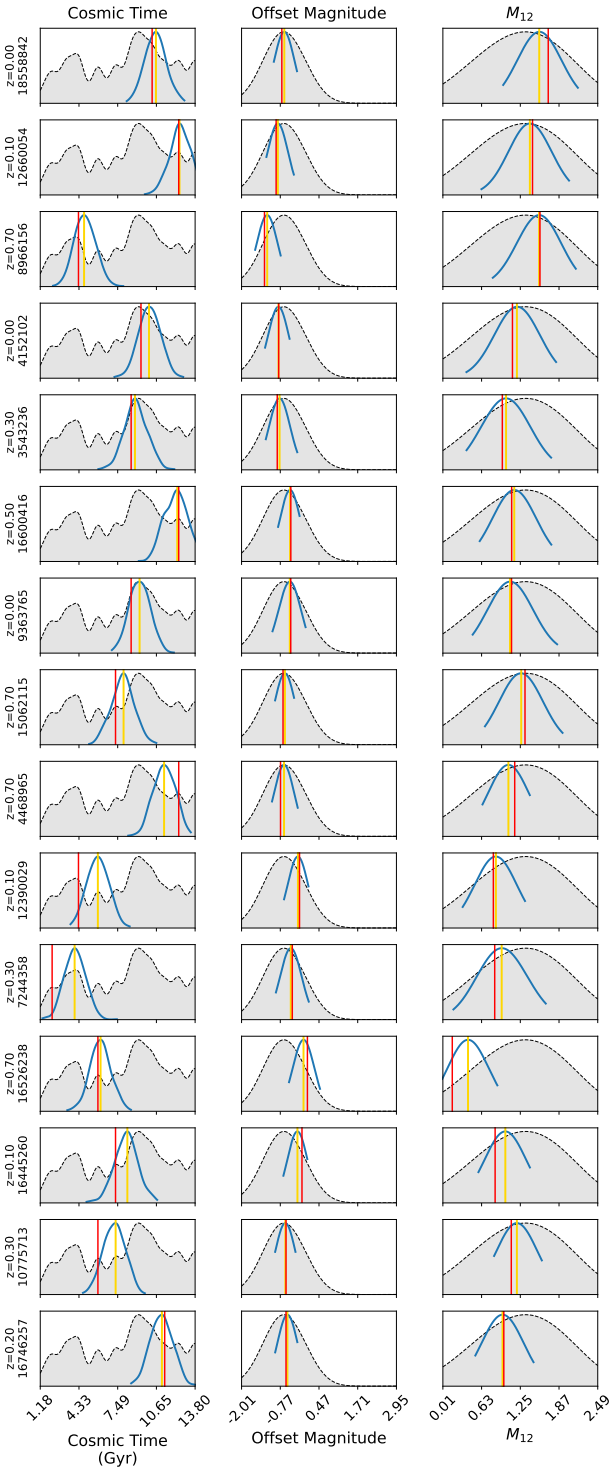

Figure 79: Posterior distributions of *dynamical state* observables inferred from radio embeddings (Table 5). Same layout as Fig. 76. Posteriors track the true values for cosmic time, center-of-mass offset, and the M_{12} merger statistic, demonstrating that the learned radio embedding encodes both temporal and structural aspects of cluster dynamics.

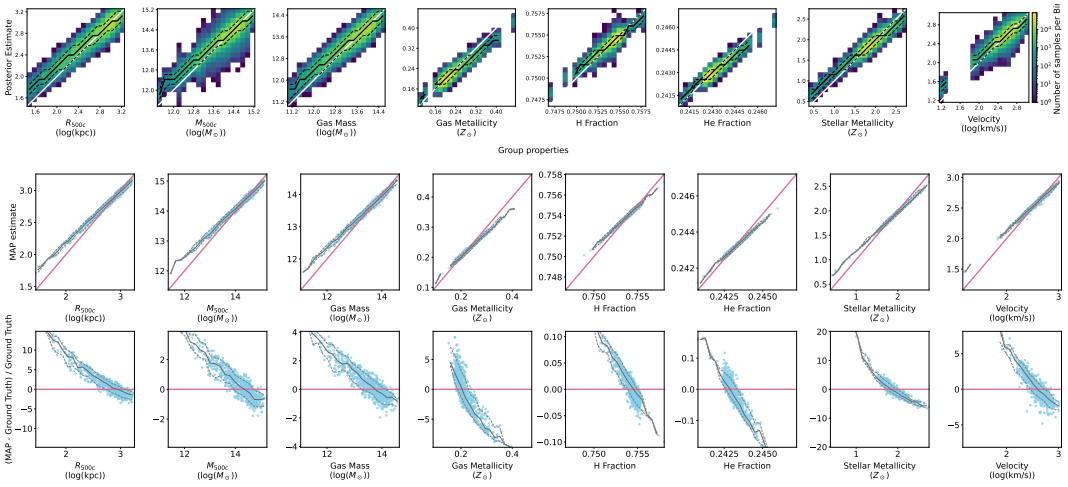

Figure 80: Radio embedding → *halo-scale* observables (Table 4). (a) Posterior vs. truth heatmaps (B = 15, n_{sam} = 500). White: y = x; black: posterior median (solid) and 10–90% quantiles (dashed). (b) MAP vs. truth (top) and relative error Δ (bottom). The tight alignment of MAP medians with y = x and narrow error envelopes confirms small bias and dispersion across R_{500c}, M_{500c}, gas mass, metallicities, and velocity.

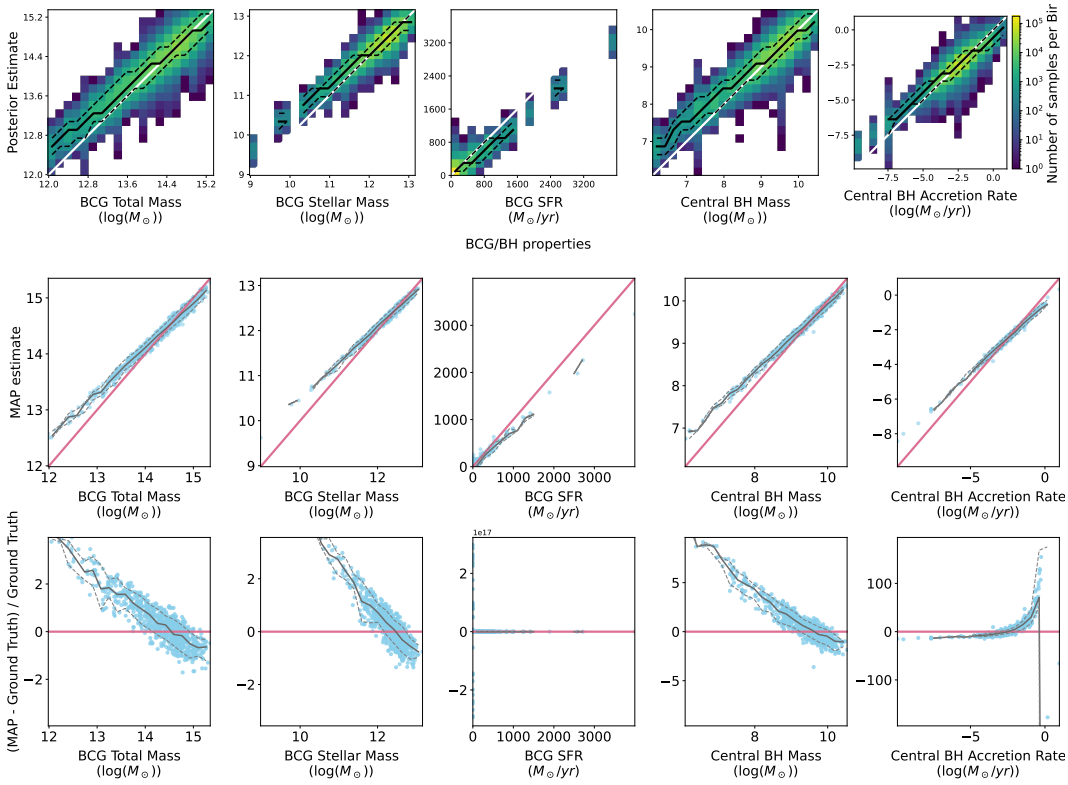

Figure 81: Radio embedding → *BCG/BH* observables (Table 4). (a) Posterior vs. truth heatmaps. White: y = x; black: median (solid) and 10–90% (dashed). (b) MAP vs. truth (top) and relative error Δ (bottom). Results show strong calibration with tight error distributions across BCG stellar mass, star formation rate, black hole mass, and accretion rate.

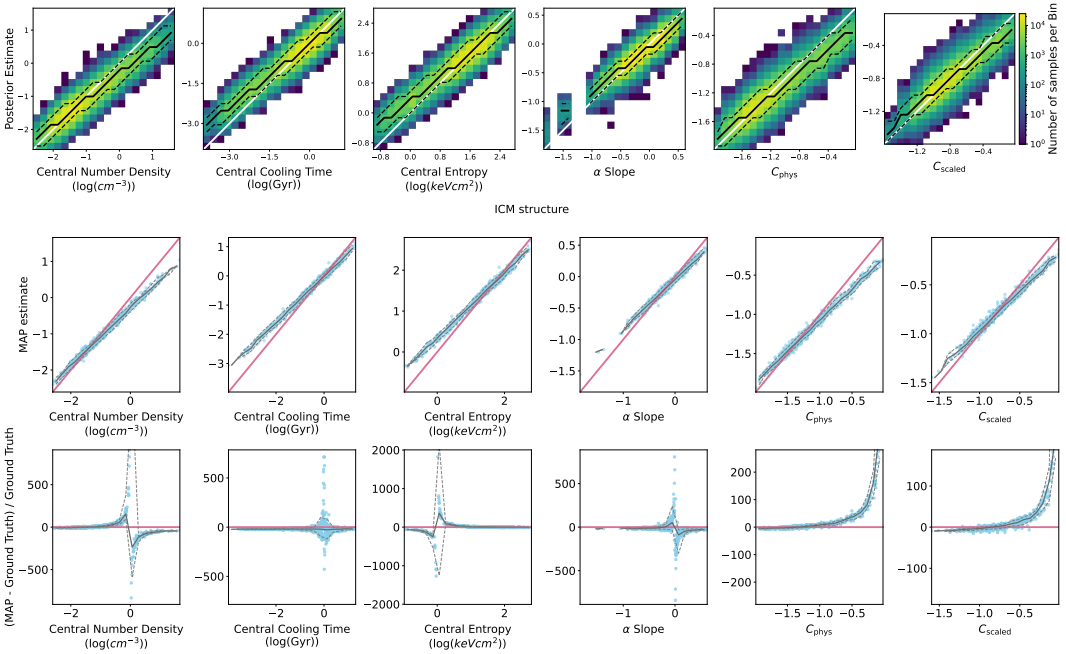

Figure 82: Radio embedding → *ICM core* observables (Table 5). (a) Posterior vs. truth heatmaps for central electron density, cooling time, entropy, slope α, and concentrations C_{phys}, C_{scaled}. White: $y = x$; black: median/quantiles. (b) MAP vs. truth (top) and relative error Δ (bottom). Small dispersion and nearly unbiased errors confirm robust predictions for ICM core properties.

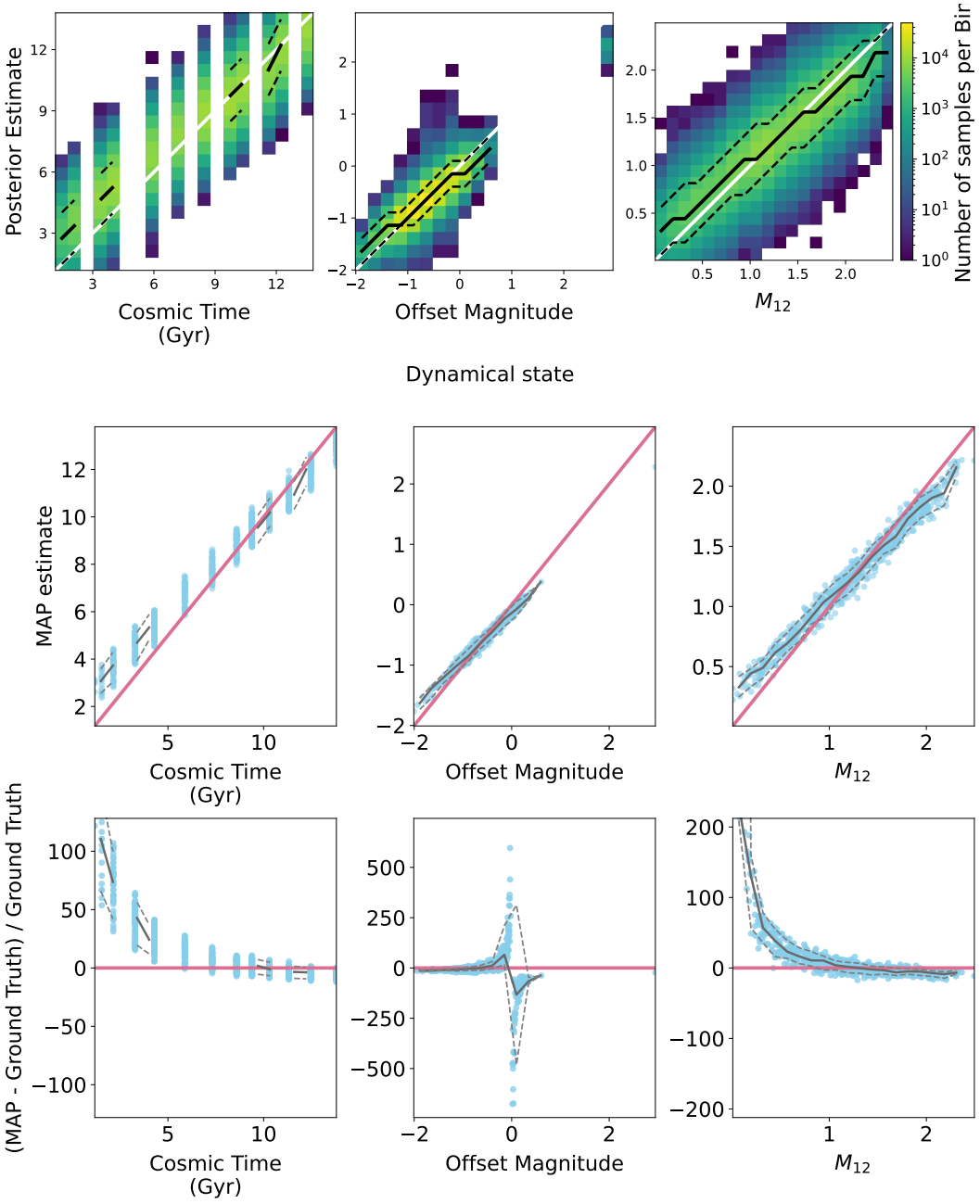

Figure 83: Radio embedding → *dynamical state* observables (Table 5). (a) Posterior vs. truth heatmaps for cosmic time, COM offset, and M_{12}. White: $y = x$; black: median/quantiles. (b) MAP vs. truth (top) and relative error Δ (bottom). MAP estimates track the identity with tight 10–90% envelopes, indicating reliable inference of cluster dynamical state from radio morphology.

BIBLIOGRAPHY

[1] Md Abul Hayat, Peter Harrington, George Stein, Zarija Lukić, and Mustafa Mustafa. "Estimating Galactic Distances From Images Using Self-supervised Representation Learning." In: *arXiv e-prints*, arXiv:2101.04293 (Jan. 2021), arXiv:2101.04293. DOI: 10.48550/arXiv.2101.04293. arXiv: 2101.04293 [astro-ph.IM].

[2] S. W. Allen, D. A. Rapetti, R. W. Schmidt, H. Ebeling, R. G. Morris, and A. C. Fabian. "Improved constraints on dark energy from Chandra X-ray observations of the largest relaxed galaxy clusters." In: 383.3 (Jan. 2008), pp. 879–896. DOI: 10.1111/j.1365-2966.2007.12610.x. arXiv: 0706.0033 [astro-ph].

[3] Steven W. Allen, August E. Evrard, and Adam B. Mantz. "Cosmological Parameters from Observations of Galaxy Clusters." In: 49.1 (Sept. 2011), pp. 409–470. DOI: 10.1146/annurev-astro-081710-102514. arXiv: 1103.4829 [astro-ph.CO].

[4] Lynton Ardizzone, Till Bungert, Felix Draxler, Ullrich Köthe, Jakob Kruse, Robert Schmier, and Peter Sorrenson. *Framework for Easily Invertible Architectures (FrEIA)*. 2018-2022. URL: https://github.com/vislearn/FrEIA.

[5] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W. Pellegrini, Ralf S. Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. "Analyzing Inverse Problems with Invertible Neural Networks." In: *arXiv e-prints*, arXiv:1808.04730 (Aug. 2018), arXiv:1808.04730. DOI: 10.48550/arXiv.1808.04730. arXiv: 1808.04730 [cs.LG].

[6] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. "Guided Image Generation with Conditional Invertible Neural Networks." In: *arXiv e-prints*, arXiv:1907.02392 (July 2019), arXiv:1907.02392. DOI: 10.48550/arXiv.1907.02392. arXiv: 1907.02392 [cs.CV].

[7] M. Arnaud, G. W. Pratt, R. Piffaretti, H. Böhringer, J. H. Croston, and E. Pointecouteau. "The universal galaxy cluster pressure profile from a representative sample of nearby systems (REXCESS) and the Y_{SZ} - M_{500} relation." In: 517, A92 (July 2010), A92. DOI: 10.1051/0004-6361/200913416. arXiv: 0910.1234 [astro-ph.CO].

[8] Mohammadreza Ayromlou, Dylan Nelson, Annalisa Pillepich, Eric Rohr, Nhut Truong, Yuan Li, Aurora Simionescu, Katrin Lehle, and Wonki Lee. "An atlas of gas motions in the TNG-Cluster simulation: From cluster cores to the outskirts." In: 690, A20 (Oct. 2024), A20. DOI: 10.1051/0004-6361/202348612. arXiv: 2311.06339 [astro-ph.GA].

[9] Nicholas M. Ball and Robert J. Brunner. "Data Mining and Machine Learning in Astronomy." In: *International Journal of Modern Physics D* 19.7 (Jan. 2010), pp. 1049–1106. DOI: 10.1142/S0218271810017160. arXiv: 0906.2173 [astro-ph.IM].

[10] J. M. Bardeen, J. R. Bond, N. Kaiser, and A. S. Szalay. "The Statistics of Peaks of Gaussian Random Fields." In: 304 (May 1986), p. 15. DOI: 10.1086/164143.

[11] James M. Bardeen, Paul J. Steinhardt, and Michael S. Turner. "Spontaneous creation of almost scale-free density perturbations in an inflationary universe." In: 28.4 (Aug. 1983), pp. 679–693. DOI: 10.1103/PhysRevD.28.679.

[12] David J. Barnes et al. "The Cluster-EAGLE project: global properties of simulated clusters with resolved galaxies." In: *Monthly Notices of the Royal Astronomical Society* 471.1 (June 2017), pp. 1088–1106. ISSN: 0035-8711. DOI: 10.1093/mnras/stx1647. eprint: https://academic.oup.com/mnras/article-pdf/471/1/1088/49203811/mnras_471_1_1088.pdf. URL: https://doi.org/10.1093/mnras/stx1647.

[13] Thomas Bayes and null Price. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S." In: *Philosophical Transactions of the Royal Society of London* 53 (1763), pp. 370–418. DOI: 10.1098/rstl.1763.0053. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1763.0053. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1763.0053.

[14] G. R. Blumenthal, S. M. Faber, J. R. Primack, and M. J. Rees. "Formation of galaxies and large-scale structure with cold dark matter." In: 311 (Oct. 1984), pp. 517–525. DOI: 10.1038/311517a0.

[15] Paul Bode, Jeremiah P. Ostriker, and Neil Turok. "Halo Formation in Warm Dark Matter Models." In: 556.1 (July 2001), pp. 93–107. DOI: 10.1086/321541. arXiv: astro-ph/0010389 [astro-ph].

[16] Hans Böhringer and Norbert Werner. "X-ray spectroscopy of galaxy clusters: studying astrophysical processes in the largest celestial laboratories." In: 18.1-2 (Feb. 2010), pp. 127–196. DOI: 10.1007/s00159-009-0023-3.

[17] J. Richard Bond, Lev Kofman, and Dmitry Pogosyan. "How filaments of galaxies are woven into the cosmic web." In: 380.6575 (Apr. 1996), pp. 603–606. DOI: 10.1038/380603a0. arXiv: astro-ph/9512141 [astro-ph].

[18] Alessandro Boselli and Giuseppe Gavazzi. "Environmental Effects on Late-Type Galaxies in Nearby Clusters." In: 118.842 (Apr. 2006), pp. 517–559. DOI: 10.1086/500691. arXiv: astro-ph/0601108 [astro-ph].

[19] Gianfranco Brunetti and Thomas W. Jones. "Cosmic Rays in Galaxy Clusters and Their Nonthermal Emission." In: *International Journal of Modern Physics D* 23.4, 1430007-98 (Mar. 2014), pp. 1430007–98. DOI: 10.1142/S0218271814300079. arXiv: 1401.7519 [astro-ph.CO].

[20] Greg L. Bryan et al. "ENZO: An Adaptive Mesh Refinement Code for Astrophysics." In: 211.2, 19 (Apr. 2014), p. 19. DOI: 10.1088/0067-0049/211/2/19. arXiv: 1307.2265 [astro-ph.IM].

[21] James S. Bullock and Michael Boylan-Kolchin. "Small-Scale Challenges to the ΛCDM Paradigm." In: 55.1 (Aug. 2017), pp. 343–387. DOI: 10.1146/annurev-astro-091916-055313. arXiv: 1707.04256 [astro-ph.CO].

[22] C. L. Carilli and G. B. Taylor. "Cluster Magnetic Fields." In: 40 (Jan. 2002), pp. 319–348. DOI: 10.1146/annurev.astro.40.060401.093852. arXiv: astro-ph/0110655 [astro-ph].

[23] John E. Carlstrom, Gilbert P. Holder, and Erik D. Reese. "Cosmology with the Sunyaev-Zel'dovich Effect." In: 40 (Jan. 2002), pp. 643–680. DOI: 10.1146/annurev.astro.40.060401.093803. arXiv: astro-ph/0208192 [astro-ph].

[24] Bradley W. Carroll and Dale A. Ostlie. *An introduction to modern astrophysics and cosmology*. 2006.

[25] Sean M. Carroll, William H. Press, and Edwin L. Turner. "The cosmological constant." In: 30 (Jan. 1992), pp. 499–542. DOI: 10.1146/annurev.aa.30.090192.002435.

[26] C. M. Casey, A. Cooray, P. Capak, H. Fu, K. Kovac, S. Lilly, D. B. Sanders, N. Z. Scoville, and E. Treister. "A Massive, Distant Proto-cluster at z = 2.47 Caught in a Phase of Rapid Formation?" In: 808.2, L33 (Aug. 2015), p. L33. DOI: 10.1088/2041-8205/808/2/L33. arXiv: 1506.01715 [astro-ph.GA].

[27] R. Cassano, S. Ettori, S. Giacintucci, G. Brunetti, M. Markevitch, T. Venturi, and M. Gitti. "On the Connection Between Giant Radio Halos and Cluster Mergers." In: 721.2 (Oct. 2010), pp. L82–L85. DOI: 10.1088/2041-8205/721/2/L82. arXiv: 1008.3624 [astro-ph.CO].

[28] Urmila Chadayammuri, Lukas Eisert, Annalisa Pillepich, Katrin Lehle, Mohammadreza Ayromlou, and Dylan Nelson. "ERGO-ML: A continuous organization of the X-ray galaxy cluster population in TNG-Cluster with contrastive learning." In: *arXiv e-prints*, arXiv:2410.22416 (Oct. 2024), arXiv:2410.22416. DOI: 10.48550/arXiv.2410.22416. arXiv: 2410.22416 [astro-ph.GA].

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A Simple Framework for Contrastive Learning of Visual Representations." In: *arXiv e-prints*, arXiv:2002.05709 (Feb. 2020), arXiv:2002.05709. DOI: 10.48550/arXiv.2002.05709. arXiv: 2002.05709 [cs.LG].

[30] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning." In: *arXiv e-prints*, arXiv:2011.10566 (Nov. 2020), arXiv:2011.10566. DOI: 10.48550/arXiv.2011.10566. arXiv: 2011.10566 [cs.CV].

[31] Yi-Kuan Chiang, Roderik A. Overzier, Karl Gebhardt, and Bruno Henriques. "Galaxy Protoclusters as Drivers of Cosmic Star Formation History in the First 2 Gyr." In: 844.2, L23 (Aug. 2017), p. L23. DOI: 10.3847/2041-8213/aa7e7b. arXiv: 1705.01634 [astro-ph.GA].

[32] I. Chiu et al. "Baryon content in a sample of 91 galaxy clusters selected by the South Pole Telescope at 0.2 <z < 1.25." In: 478.3 (Aug. 2018), pp. 3072–3099. DOI: 10.1093/mnras/sty1284. arXiv: 1711.00917 [astro-ph.CO].

[33] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." In: *arXiv preprint arXiv:1406.1078* (2014).

[34] A. O. Clarke et al. "LOFAR MSSS: Discovery of a 2.56 Mpc giant radio galaxy associated with a disturbed galaxy group." In: 601, A25 (May 2017), A25. DOI: 10.1051/0004-6361/201630152. arXiv: 1702.01571 [astro-ph.GA].

[35] Douglas Clowe, Maruša Bradač, Anthony H. Gonzalez, Maxim Markevitch, Scott W. Randall, Christine Jones, and Dennis Zaritsky. "A Direct Empirical Proof of the Existence of Dark Matter." In: 648.2 (Sept. 2006), pp. L109–L113. DOI: 10.1086/508162. arXiv: astro-ph/0608407 [astro-ph].

[36] Robert A. Crain et al. "The EAGLE simulations of galaxy formation: calibration of subgrid physics and model variations." In: 450.2 (June 2015), pp. 1937–1961. DOI: 10.1093/mnras/stv725. arXiv: 1501.01311 [astro-ph.GA].

[37] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. "Discovering Symbolic Models from Deep Learning with Inductive Biases." In: *arXiv e-prints*, arXiv:2006.11287 (June 2020), arXiv:2006.11287. DOI: 10.48550/arXiv.2006.11287. arXiv: 2006.11287 [cs.LG].

[38] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. "The evolution of large-scale structure in a universe dominated by cold dark matter." In: 292 (May 1985), pp. 371–394. DOI: 10.1086/163168.

[39] Sander Dieleman, Kyle W. Willett, and Joni Dambre. "Rotation-invariant convolutional neural networks for galaxy morphology prediction." In: 450.2 (June 2015), pp. 1441–1459. DOI: 10.1093/mnras/stv632. arXiv: 1503.07077 [astro-ph.IM].

[40] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. "Unsupervised Visual Representation Learning by Context Prediction." In: *arXiv e-prints*, arXiv:1505.05192 (May 2015), arXiv:1505.05192. DOI: 10.48550/arXiv.1505.05192. arXiv: 1505.05192 [cs.CV].

[41] A. Dressler. "Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies." In: 236 (Mar. 1980), pp. 351–365. DOI: 10.1086/157753.

[42] L. Oc. Drury. "REVIEW ARTICLE: An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas." In: *Reports on Progress in Physics* 46.8 (Aug. 1983), pp. 973–1027. DOI: 10.1088/0034-4885/46/8/002.

[43] Y. Dubois et al. "Dancing in the dark: galactic properties trace spin swings along the cosmic web." In: 444.2 (Oct. 2014), pp. 1453–1468. DOI: 10.1093/mnras/stu1227. arXiv: 1402.1165 [astro-ph.CO].

[44] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. "Neural Spline Flows." In: *arXiv e-prints*, arXiv:1906.04032 (June 2019), arXiv:1906.04032. DOI: 10.48550/arXiv.1906.04032. arXiv: 1906.04032 [stat.ML].

[45] Jaan Einasto. "Large scale structure of the Universe." In: *The Sun, the Stars, the Universe and General Relativity: International Conference in Honor of Ya.B. Zeldovich's 95th Anniversary*. Ed. by Remo Ruffini and Gregory Vereshchagin. Vol. 1205. American Institute of Physics Conference Series. AIP, Mar. 2010, pp. 72–81. DOI: 10.1063/1.3382336. arXiv: 0906.5272 [astro-ph.CO].

[46] Albert Einstein. "Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie." In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften* (Jan. 1917), pp. 142–152.

[47] Daniel J. Eisenstein and Wayne Hu. "Power Spectra for Cold Dark Matter and Its Variants." In: 511.1 (Jan. 1999), pp. 5–15. DOI: 10.1086/306640. arXiv: astro-ph/9710252 [astro-ph].

[48] Daniel J. Eisenstein et al. "Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies." In: 633.2 (Nov. 2005), pp. 560–574. DOI: 10.1086/466512. arXiv: astro-ph/0501171 [astro-ph].

[49] Lukas Eisert, Connor Bottrell, Annalisa Pillepich, Rhythm Shimakawa, Vicente Rodriguez-Gomez, Dylan Nelson, Eirini Angeloudi, and Marc Huertas-Company. "ERGO-ML: comparing IllustrisTNG and HSC galaxy images via contrastive learning." In: 528.4 (Mar. 2024), pp. 7411–7439. DOI: 10.1093/mnras/stae481. arXiv: 2310.19904 [astro-ph.GA].

[50] Lukas Eisert, Annalisa Pillepich, Dylan Nelson, Ralf S. Klessen, Marc Huertas-Company, and Vicente Rodriguez-Gomez. "ERGO-ML I: inferring the assembly histories of IllustrisTNG galaxies from integral observable properties via invertible neural networks." In: 519.2 (Feb. 2023), pp. 2199–2223. DOI: 10.1093/mnras/stac3295. arXiv: 2202.06967 [astro-ph.GA].

[51] A. C. Fabian. "Observational Evidence of Active Galactic Nuclei Feedback." In: 50 (Sept. 2012), pp. 455–489. DOI: 10.1146/annurev-astro-081811-125521. arXiv: 1204.4114 [astro-ph.CO].

[52] Luigina Feretti, Gabriele Giovannini, Federica Govoni, and Matteo Murgia. "Clusters of galaxies: observational properties of the diffuse radio emission." In: 20, 54 (May 2012), p. 54. DOI: 10.1007/s00159-012-0054-z. arXiv: 1205.1919 [astro-ph.CO].

[53] L. Gao, J. F. Navarro, C. S. Frenk, A. Jenkins, V. Springel, and S. D. M. White. "The Phoenix Project: the dark side of rich Galaxy clusters." In: 425.3 (Sept. 2012), pp. 2169–2186. DOI: 10.1111/j.1365-2966.2012.21564.x. arXiv: 1201.1940 [astro-ph.CO].

[54] Simona Giacintucci, Maxim Markevitch, Rossella Cassano, Tiziana Venturi, Tracy E. Clarke, and Gianfranco Brunetti. "Occurrence of Radio Minihalos in a Mass-limited Sample of Galaxy Clusters." In: 841.2, 71 (June 2017), p. 71. DOI: 10.3847/1538-4357/aa7069. arXiv: 1701.01364 [astro-ph.HE].

[55] R. A. Gingold and J. J. Monaghan. "Smoothed particle hydrodynamics: theory and application to non-spherical stars." In: 181 (Nov. 1977), pp. 375–389. DOI: 10.1093/mnras/181.3.375.

[56] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[57] Jean-Bastien Grill et al. "Bootstrap your own latent: A new approach to self-supervised Learning." In: *arXiv e-prints*, arXiv:2006.07733 (June 2020), arXiv:2006.07733. DOI: 10.48550/arXiv.2006.07733. arXiv: 2006.07733 [cs.LG].

[58] James E. Gunn and J. Richard Gott III. "On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution." In: 176 (Aug. 1972), p. 1. DOI: 10.1086/151605.

[59] Alan H. Guth. "Inflationary universe: A possible solution to the horizon and flatness problems." In: 23.2 (Jan. 1981), pp. 347–356. DOI: 10.1103/PhysRevD.23.347.

[60] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum Contrast for Unsupervised Visual Representation Learning." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9729–9738.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR*. June 2016, 1, p. 1. DOI: 10.1109/CVPR.2016.90. arXiv: 1512.03385 [cs.CV].

[62] Hitomi Collaboration et al. "The quiescent intracluster medium in the core of the Perseus cluster." In: 535.7610 (July 2016), pp. 117–121. DOI: 10.1038/nature18627. arXiv: 1607.04487 [astro-ph.GA].

[63] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (Nov. 1997), 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[64] Matthias Hoeft and Marcus Brüggen. "Radio signature of cosmological structure formation shocks." In: 375.1 (Feb. 2007), pp. 77–91. DOI: 10.1111/j.1365-2966.2006.11111.x. arXiv: astro-ph/0609831 [astro-ph].

[65] David W. Hogg. "Distance measures in cosmology." In: *arXiv e-prints*, astro-ph/9905116 (May 1999), astro–ph/9905116. DOI: 10.48550/arXiv.astro-ph/9905116. arXiv: astro-ph/9905116 [astro-ph].

[66] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks." In: *Neural Netw.* 4.2 (Mar. 1991), 251–257. ISSN: 0893-6080. DOI: 10.1016/0893-6080(91)90009-T. URL: https://doi.org/10.1016/0893-6080(91)90009-T.

[67] Svenja Jacob, Rüdiger Pakmor, Christine M. Simpson, Volker Springel, and Christoph Pfrommer. "The dependence of cosmic ray-driven galactic winds on halo mass." In: 475.1 (Mar. 2018), pp. 570–584. DOI: 10.1093/mnras/stx3221. arXiv: 1712.04947 [astro-ph.GA].

[68] N. Jeffrey et al. "Dark Energy Survey Year 3 results: Curved-sky weak lensing mass map reconstruction." In: 505.3 (Aug. 2021), pp. 4626–4645. DOI: 10.1093/mnras/stab1495. arXiv: 2105.13539 [astro-ph.CO].

[69] Danilo Jimenez Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows." In: *arXiv e-prints*, arXiv:1505.05770 (May 2015), arXiv:1505.05770. DOI: 10.48550/arXiv.1505.05770. arXiv: 1505.05770 [stat.ML].

[70] Wei Jin, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, and Jiliang Tang. "Self-supervised Learning on Graphs: Deep Insights and New Direction." In: *arXiv e-prints*, arXiv:2006.10141 (June 2020), arXiv:2006.10141. DOI: 10.48550/arXiv.2006.10141. arXiv: 2006.10141 [cs.LG].

[71] N. Kaiser. "On the spatial correlations of Abell clusters." In: 284 (Sept. 1984), pp. L9–L12. DOI: 10.1086/184341.

[72] N. Kaiser. "Evolution and clustering of rich clusters." In: 222 (Sept. 1986), pp. 323–345. DOI: 10.1093/mnras/222.2.323.

[73] Nick Kaiser and Gordon Squires. "Mapping the Dark Matter with Weak Gravitational Lensing." In: 404 (Feb. 1993), p. 441. DOI: 10.1086/172297.

[74] Hyesung Kang and Dongsu Ryu. "Diffusive Shock Acceleration at Cosmological Shock Waves." In: 764.1, 95 (Feb. 2013), p. 95. DOI: 10.1088/0004-637X/764/1/95. arXiv: 1212.3246 [astro-ph.HE].

[75] Hyesung Kang, Dongsu Ryu, and T. W. Jones. "Diffusive Shock Acceleration Simulations of Radio Relics." In: 756.1, 97 (Sept. 2012), p. 97. DOI: 10.1088/0004-637X/756/1/97. arXiv: 1205.1895 [astro-ph.HE].

[76] N. S. Kardashev. "Nonstationarity of Spectra of Young Sources of Nonthermal Radio Emission." In: 6 (Dec. 1962), p. 317.

[77] Neal Katz, David H. Weinberg, and Lars Hernquist. "Cosmological Simulations with TreeSPH." In: 105 (July 1996), p. 19. DOI: 10.1086/192305. arXiv: astro-ph/9509107 [astro-ph].

[78] Neal Katz and Simon D. M. White. "Hierarchical Galaxy Formation: Overmerging and the Formation of an X-Ray Cluster." In: 412 (Aug. 1993), p. 455. DOI: 10.1086/172935.

[79] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[80] Anatoly A. Klypin, Sebastian Trujillo-Gomez, and Joel Primack. "Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation." In: 740.2, 102 (Oct. 2011), p. 102. DOI: 10.1088/0004-637X/740/2/102. arXiv: 1002.3660 [astro-ph.CO].

[81] Anatoly Klypin, Andrey V. Kravtsov, Octavio Valenzuela, and Francisco Prada. "Where Are the Missing Galactic Satellites?" In: 522.1 (Sept. 1999), pp. 82–92. DOI: 10.1086/307643. arXiv: astro-ph/9901240 [astro-ph].

[82] Anatoly Klypin, Gustavo Yepes, Stefan Gottlöber, Francisco Prada, and Steffen Heß. "MultiDark simulations: the story of dark matter halo concentrations and density profiles." In: 457.4 (Apr. 2016), pp. 4340–4359. DOI: 10.1093/mnras/stw248. arXiv: 1411.4001 [astro-ph.CO].

[83] Doogesh Kodi Ramanah, Tom Charnock, Francisco Villaescusa-Navarro, and Benjamin D. Wandelt. "Super-resolution emulator of cosmological simulations using deep physical models." In: 495.4 (July 2020), pp. 4227–4236. DOI: 10.1093/mnras/staa1428. arXiv: 2001.05519 [astro-ph.CO].

[84] Andrey V. Kravtsov and Stefano Borgani. "Formation of Galaxy Clusters." In: 50 (Sept. 2012), pp. 353–409. DOI: 10.1146/annurev-astro-081811-125502. arXiv: 1205.5556 [astro-ph.CO].

[85] Andrey V. Kravtsov, Alexey Vikhlinin, and Daisuke Nagai. "A New Robust Low-Scatter X-Ray Mass Indicator for Clusters of Galaxies." In: 650.1 (Oct. 2006), pp. 128–136. DOI: 10.1086/506319. arXiv: astro-ph/0603205 [astro-ph].

[86] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." In: *Commun. ACM* 60.6 (May 2017), 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

[87] Ivo Labbé et al. "A population of red candidate massive galaxies 600 Myr after the Big Bang." In: 616.7956 (Apr. 2023), pp. 266–269. DOI: 10.1038/s41586-023-05786-2. arXiv: 2207.12446 [astro-ph.GA].

[88] Cedric Lacey and Shaun Cole. "Merger rates in hierarchical models of galaxy formation." In: 262.3 (June 1993), pp. 627–649. DOI: 10.1093/mnras/262.3.627.

[89] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. "Contrastive Representation Learning: A Framework and Review." In: *arXiv e-prints*, arXiv:2010.05113 (Oct. 2020), arXiv:2010.05113. DOI: 10.48550/arXiv.2010.05113. arXiv: 2010.05113 [cs.LG].

[90] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://doi.org/10.1038/nature14539.

[91] W. Lee, A. Pillepich, J. ZuHone, D. Nelson, M. J. Jee, D. Nagai, and K. Finner. "Radio relics in massive galaxy cluster mergers in the TNG-Cluster simulation." In: 686, A55 (June 2024), A55. DOI: 10.1051/0004-6361/202348194. arXiv: 2311.06340 [astro-ph.GA].

[92] Wonki Lee, M. James Jee, Kyle Finner, Kim HyeongHan, Ruta Kale, Hyein Yoon, William Forman, Ralph Kraft, Christine Jones, and Aeree Chung. "Discovery of a Double Radio Relic in ZwCl1447.2+2619: A Rare Testbed for Shock-acceleration Models with a Peculiar Surface-brightness Ratio." In: 924.1, 18 (Jan. 2022), p. 18. DOI: 10.3847/1538-4357/ac32c5. arXiv: 2109.00593 [astro-ph.CO].

[93] Katrin Lehle, Dylan Nelson, Annalisa Pillepich, Nhut Truong, and Eric Rohr. "The heart of galaxy clusters: Demographics and physical properties of cool-core and non-cool-core halos in the TNG-Cluster simulation." In: 687, A129 (July 2024), A129. DOI: 10.1051/0004-6361/202348609. arXiv: 2311.06333 [astro-ph.GA].

[94] S. Lloyd. "Least squares quantization in PCM." In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[95] J. M. Lotz et al. "The Frontier Fields: Survey Design and Initial Results." In: 837.1, 97 (Mar. 2017), p. 97. DOI: 10.3847/1538-4357/837/1/97. arXiv: 1605.06567 [astro-ph.GA].

[96] Lorenzo Lovisari, Stefano Ettori, Massimo Gaspari, and Paul A. Giles. "Scaling Properties of Galaxy Groups." In: *Universe* 7.5 (2021). ISSN: 2218-1997. DOI: 10.3390/universe7050139. URL: https://www.mdpi.com/2218-1997/7/5/139.

[97] L. B. Lucy. "A numerical approach to the testing of the fission hypothesis." In: 82 (Dec. 1977), pp. 1013–1024. DOI: 10.1086/112164.

[98] Maxim Markevitch and Alexey Vikhlinin. "Shocks and cold fronts in galaxy clusters." In: 443.1 (May 2007), pp. 1–53. DOI: 10.1016/j.physrep.2007.01.001. arXiv: astro-ph/0701821 [astro-ph].

[99] Kyoko Matsushita et al. "Suzaku Observation of the Metallicity Distribution in the Intracluster Medium of the Fornax Cluster." In: 59 (Jan. 2007), pp. 327–338. DOI: 10.1093/pasj/59.sp1.S327. arXiv: astro-ph/0609065 [astro-ph].

[100] B. J. Maughan, L. R. Jones, M. Pierre, S. Andreon, M. Birkinshaw, M. N. Bremer, F. Pacaud, T. J. Ponman, I. Valtchanov, and J. Willis. "Testing the galaxy cluster mass-observable relations at z = 1 with XMM-Newton and Chandra observations of XLSSJ022403.9-041328." In: 387.3 (July 2008), pp. 998–1006. DOI: 10.1111/j.1365-2966.2008.13313.x. arXiv: 0709.2300 [astro-ph].

[101] Sean L. McGee, Michael L. Balogh, Richard G. Bower, Andreea S. Font, and Ian G. McCarthy. "The accretion of galaxies into groups and clusters." In: 400.2 (Dec. 2009), pp. 937–950. DOI: 10.1111/j.1365-2966.2009.15507.x. arXiv: 0908.0750 [astro-ph.CO].

[102] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." In: *arXiv e-prints*, arXiv:1802.03426 (Feb. 2018), arXiv:1802.03426. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426 [stat.ML].

[103] Patrick McNamee and Zahra Nili Ahmadabadi. "Adaptive Extremum Seeking Control via the RMSprop Optimizer." In: *arXiv e-prints*, arXiv:2409.12290 (Sept. 2024), arXiv:2409.12290. DOI: 10.48550/arXiv.2409.12290. arXiv: 2409.12290 [math.OC].

[104] M. Meneghetti, E. Rasia, J. Merten, F. Bellagamba, S. Ettori, P. Mazzotta, K. Dolag, and S. Marri. "Weighing simulated galaxy clusters using lensing and X-ray." In: 514, A93 (May 2010), A93. DOI: 10.1051/0004-6361/200913222. arXiv: 0912.1343 [astro-ph.CO].

[105] J. Merten et al. "Creation of cosmic structure in the complex galaxy cluster merger Abell 2744." In: 417.1 (Oct. 2011), pp. 333–347. DOI: 10.1111/j.1365-2966.2011.19266.x. arXiv: 1103.2772 [astro-ph.CO].

[106] Chirag Modi, Yu Feng, and Uroš Seljak. "Cosmological reconstruction from galaxy light: neural network based light-matter connection." In: 2018.10, 028 (Oct. 2018), p. 028. DOI: 10.1088/1475-7516/2018/10/028. arXiv: 1805.02247 [astro-ph.CO].

[107] Ben Moore, Sebastiano Ghigna, Fabio Governato, George Lake, Thomas Quinn, Joachim Stadel, and Paolo Tozzi. "Dark Matter Substructure within Galactic Halos." In: 524.1 (Oct. 1999), pp. L19–L22. DOI: 10.1086/312287. arXiv: astro-ph/9907411 [astro-ph].

[108] Ben Moore, Neal Katz, George Lake, Alan Dressler, and Augustus Oemler. "Galaxy harassment and the evolution of clusters of galaxies." In: 379.6566 (Feb. 1996), pp. 613–616. DOI: 10.1038/379613a0. arXiv: astro-ph/9510034 [astro-ph].

[109] John S. Mulchaey. "X-ray Properties of Groups of Galaxies." In: 38 (Jan. 2000), pp. 289–335. DOI: 10.1146/annurev.astro.38.1.289. arXiv: astro-ph/0009379 [astro-ph].

[110] Kirpal Nandra et al. "The Hot and Energetic Universe: A White Paper presenting the science theme motivating the Athena+ mission." In: *arXiv e-prints*, arXiv:1306.2307 (June 2013), arXiv:1306.2307. DOI: 10.48550/arXiv.1306.2307. arXiv: 1306.2307 [astro-ph.HE].

[111] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. "Simulations of X-ray clusters." In: 275.3 (Aug. 1995), pp. 720–740. DOI: 10.1093/mnras/275.3.720. arXiv: astro-ph/9408069 [astro-ph].

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

[113] Dylan Nelson, Annalisa Pillepich, Mohammadreza Ayromlou, Wonki Lee, Katrin Lehle, Eric Rohr, and Nhut Truong. "Introducing the TNG-Cluster simulation: Overview and the physical properties of the gaseous intracluster medium." In: 686, A157 (June 2024), A157. DOI: 10.1051/0004-6361/202348608. arXiv: 2311.06338 [astro-ph.GA].

[114] Dylan Nelson et al. "The IllustrisTNG simulations: public data release." In: *Computational Astrophysics and Cosmology* 6.1, 2 (May 2019), p. 2. DOI: 10.1186/s40668-019-0028-x. arXiv: 1812.05609 [astro-ph.GA].

[115] Mehdi Noroozi and Paolo Favaro. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles." In: *arXiv e-prints*, arXiv:1603.09246 (Mar. 2016), arXiv:1603.09246. DOI: 10.48550/arXiv.1603.09246. arXiv: 1603.09246 [cs.CV].

[116] Michelle Ntampaka et al. "The Role of Machine Learning in the Next Decade of Cosmology." In: 51.3, 14 (May 2019), p. 14. DOI: 10.48550/arXiv.1902.10159. arXiv: 1902.10159 [astro-ph.IM].

[117] Roderik A. Overzier. "The realm of the galaxy protoclusters. A review." In: 24.1, 14 (Nov. 2016), p. 14. DOI: 10.1007/s00159-016-0100-3. arXiv: 1610.05201 [astro-ph.GA].

[118] Rüdiger Pakmor, Volker Springel, Andreas Bauer, Philip Mocz, Diego J. Munoz, Sebastian T. Ohlmann, Kevin Schaal, and Chenchong Zhu. "Improving the convergence properties of the moving-mesh code AREPO." In: 455.1 (Jan. 2016), pp. 1134–1143. DOI: 10.1093/mnras/stv2380. arXiv: 1503.00562 [astro-ph.GA].

[119] Rüdiger Pakmor et al. "Magnetic field formation in the Milky Way like disc galaxies of the Auriga project." In: 469.3 (Aug. 2017), pp. 3185–3199. DOI: 10.1093/mnras/stx1074. arXiv: 1701.07028 [astro-ph.GA].

[120] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. "Normalizing Flows for Probabilistic Modeling and Inference." In: *arXiv e-prints*, arXiv:1912.02762 (Dec. 2019), arXiv:1912.02762. DOI: 10.48550/arXiv.1912.02762. arXiv: 1912.02762 [stat.ML].

[121] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *arXiv e-prints*, arXiv:1912.01703 (Dec. 2019), arXiv:1912.01703. DOI: 10.48550/arXiv.1912.01703. arXiv: 1912.01703 [cs.LG].

[122] P. J. E. Peebles and J. T. Yu. "Primeval Adiabatic Perturbation in an Expanding Universe." In: 162 (Dec. 1970), p. 815. DOI: 10.1086/150713.

[123] Ying-jie Peng et al. "Mass and Environment as Drivers of Galaxy Evolution in SDSS and zCOSMOS and the Origin of the Schechter Function." In: 721.1 (Sept. 2010), pp. 193–221. DOI: 10.1088/0004-637X/721/1/193. arXiv: 1003.4747 [astro-ph.CO].

[124] S. Perlmutter et al. "Measurements of Ω and Λ from 42 High-Redshift Supernovae." In: 517.2 (June 1999), pp. 565–586. DOI: 10.1086/307221. arXiv: astro-ph/9812133 [astro-ph].

[125] C. E. Petrillo et al. "Finding strong gravitational lenses in the Kilo Degree Survey with Convolutional Neural Networks." In: 472.1 (Nov. 2017), pp. 1129–1150. DOI: 10.1093/mnras/stx2052. arXiv: 1702.07675 [astro-ph.GA].

[126] C. Pfrommer, R. Pakmor, K. Schaal, C. M. Simpson, and V. Springel. "Simulating cosmic ray physics on a moving mesh." In: 465.4 (Mar. 2017), pp. 4500–4529. DOI: 10.1093/mnras/stw2941. arXiv: 1604.07399 [astro-ph.GA].

[127] Annalisa Pillepich et al. "Simulating galaxy formation with the IllustrisTNG model." In: 473.3 (Jan. 2018), pp. 4077–4106. DOI: 10.1093/mnras/stx2656. arXiv: 1703.02970 [astro-ph.GA].

[128] Anders Pinzke, S. Peng Oh, and Christoph Pfrommer. "Giant radio relics in galaxy clusters: reacceleration of fossil relativistic electrons?" In: 435.2 (Oct. 2013), pp. 1061–1082. DOI: 10.1093/mnras/stt1308. arXiv: 1301.5644 [astro-ph.CO].

[129] Planck Collaboration et al. "Planck 2015 results. XIII. Cosmological parameters." In: 594, A13 (Sept. 2016), A13. DOI: 10.1051/0004-6361/201525830. arXiv: 1502.01589 [astro-ph.CO].

[130] Planck Collaboration et al. "Planck 2018 results. VI. Cosmological parameters." In: 641, A6 (Sept. 2020), A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209 [astro-ph.CO].

[131] Bianca M. Poggianti et al. "GASP. I. Gas Stripping Phenomena in Galaxies with MUSE." In: 844.1, 48 (July 2017), p. 48. DOI: 10.3847/1538-4357/aa78ed. arXiv: 1704.05086 [astro-ph.GA].

[132] G. W. Pratt, M. Arnaud, A. Biviano, D. Eckert, S. Ettori, D. Nagai, N. Okabe, and T. H. Reiprich. "The Galaxy Cluster Mass Scale and Its Impact on Cosmological Constraints from the Cluster Population." In: 215.2, 25 (Feb. 2019), p. 25. DOI: 10.1007/s11214-019-0591-0. arXiv: 1902.10837 [astro-ph.CO].

[133] G. W. Pratt, J. H. Croston, M. Arnaud, and H. Böhringer. "Galaxy cluster X-ray luminosity scaling relations from a representative local sample (REXCESS)." In: 498.2 (May 2009), pp. 361–378. DOI: 10.1051/0004-6361/200810994. arXiv: 0809.3784 [astro-ph].

[134] William H. Press and Paul Schechter. "Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation." In: 187 (Feb. 1974), pp. 425–438. DOI: 10.1086/152650.

[135] K. Rajpurohit et al. "Deep VLA Observations of the Cluster 1RXS J0603.3+4214 in the Frequency Range of 1-2 GHz." In: 852.2, 65 (Jan. 2018), p. 65. DOI: 10.3847/1538-4357/aa9f13. arXiv: 1712.01327 [astro-ph.GA].

[136] E. Rasia, S. Borgani, S. Ettori, P. Mazzotta, and M. Meneghetti. "On the Discrepancy between Theoretical and X-Ray Concentration-Mass Relations for Galaxy Clusters." In: 776.1, 39 (Oct. 2013), p. 39. DOI: 10.1088/0004-637X/776/1/39. arXiv: 1301.7476 [astro-ph.CO].

[137] Adam G. Riess et al. "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant." In: 116.3 (Sept. 1998), pp. 1009–1038. DOI: 10.1086/300499. arXiv: astro-ph/9805201 [astro-ph].

[138] Adam G. Riess et al. "A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s^{-1} Mpc^{-1} Uncertainty from the Hubble Space Telescope and the SHoES Team." In: 934.1, L7 (July 2022), p. L7. DOI: 10.3847/2041-8213/ac5c5b. arXiv: 2112.04510 [astro-ph.CO].

[139] Kenneth Rines, Margaret J. Geller, Antonaldo Diaferio, and Michael J. Kurtz. "Measuring the Ultimate Halo Mass of Galaxy Clusters: Redshifts and Mass Profiles from the Hectospec Cluster Survey (HeCS)." In: 767.1, 15 (Apr. 2013), p. 15. DOI: 10.1088/0004-637X/767/1/15. arXiv: 1209.3786 [astro-ph.CO].

[140] Vera C. Rubin and W. Kent Ford Jr. "Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions." In: 159 (Feb. 1970), p. 379. DOI: 10.1086/150317.

[141] George B. Rybicki and Alan P. Lightman. *Radiative processes in astrophysics*. 1979.

[142] Craig L. Sarazin. *X-ray emission from clusters of galaxies*. 1988.

[143] Kevin Schaal and Volker Springel. "Shock finding on a moving mesh - I. Shock statistics in non-radiative cosmological simulations." In: 446.4 (Feb. 2015), pp. 3992–4007. DOI: 10.1093/mnras/stu2386. arXiv: 1407.4117 [astro-ph.CO].

[144] Joop Schaye et al. "The EAGLE project: simulating the evolution and assembly of galaxies and their environments." In: 446.1 (Jan. 2015), pp. 521–554. DOI: 10.1093/mnras/stu2058. arXiv: 1407.7040 [astro-ph.GA].

[145] Federico Sembolini, Gustavo Yepes, Marco De Petris, Stefan Gottlöber, Luca Lamagna, and Barbara Comis. "The MUSIC of galaxy clusters - I. Baryon properties and scaling relations of the thermal Sunyaev-Zel'dovich effect." In: 429.1 (Feb. 2013), pp. 323–343. DOI: 10.1093/mnras/sts339. arXiv: 1207.4438 [astro-ph.CO].

[146] Federico Sembolini et al. "nIFTy galaxy cluster simulations – II. Radiative models." In: *Monthly Notices of the Royal Astronomical Society* 459.3 (Apr. 2016), pp. 2973–2991. ISSN: 0035-8711. DOI: 10.1093/mnras/stw800. eprint: https://academic.oup.com/mnras/article-pdf/459/3/2973/8107345/stw800.pdf. URL: https://doi.org/10.1093/mnras/stw800.

[147] Christopher J. Shallue and Andrew Vanderburg. "Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90." In: 155.2, 94 (Feb. 2018), p. 94. DOI: 10.3847/1538-3881/aa9e09. arXiv: 1712.05044 [astro-ph.EP].

[148] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer." In: *arXiv e-prints*, arXiv:1701.06538 (Jan. 2017), arXiv:1701.06538. DOI: 10.48550/arXiv.1701.06538. arXiv: 1701.06538 [cs.LG].

[149] Ravi K. Sheth and Giuseppe Tormen. "Large-scale bias and the peak background split." In: 308.1 (Sept. 1999), pp. 119–126. DOI: 10.1046/j.1365-8711.1999.02692.x. arXiv: astro-ph/9901122 [astro-ph].

[150] T. W. Shimwell et al. "The LOFAR Two-metre Sky Survey. I. Survey description and preliminary data release." In: 598, A104 (Feb. 2017), A104. DOI: 10.1051/0004-6361/201629313. arXiv: 1611.02700 [astro-ph.IM].

[151] Debora Sijacki, Volker Springel, Tiziana Di Matteo, and Lars Hernquist. "A unified model for AGN feedback in cosmological simulations of structure formation." In: 380.3 (Sept. 2007), pp. 877–900. DOI: 10.1111/j.1365-2966.2007.12153.x. arXiv: 0705.2238 [astro-ph].

[152] Joseph Silk. "Cosmic Black-Body Radiation and Galaxy Formation." In: 151 (Feb. 1968), p. 459. DOI: 10.1086/149449.

[153] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *International Conference on Learning Representations (ICLR)* (2015). arXiv: 1409.1556. URL: https://arxiv.org/abs/1409.1556.

[154] Randall K. Smith, Nancy S. Brickhouse, Duane A. Liedahl, and John C. Raymond. "Collisional Plasma Models with APEC/APED: Emission-Line Diagnostics of Hydrogen-like and Helium-like Ions." In: 556.2 (Aug. 2001), pp. L91–L95. DOI: 10.1086/322992. arXiv: astro-ph/0106478 [astro-ph].

[155] Volker Springel. "E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh." In: 401.2 (Jan. 2010), pp. 791–851. DOI: 10.1111/j.1365-2966.2009.15715.x. arXiv: 0901.4107 [astro-ph.CO].

[156] Volker Springel, Carlos S. Frenk, and Simon D. M. White. "The large-scale structure of the Universe." In: 440.7088 (Apr. 2006), pp. 1137–1144. DOI: 10.1038/nature04805. arXiv: astro-ph/0604561 [astro-ph].

[157] Volker Springel and Lars Hernquist. "Cosmological smoothed particle hydrodynamics simulations: a hybrid multiphase model for star formation." In: 339.2 (Feb. 2003), pp. 289–311. DOI: 10.1046/j.1365-8711.2003.06206.x. arXiv: astro-ph/0206393 [astro-ph].

[158] Volker Springel, Simon D. M. White, Giuseppe Tormen, and Guinevere Kauffmann. "Populating a cluster of galaxies - I. Results at z=0." In: 328.3 (Dec. 2001), pp. 726–750. DOI: 10.1046/j.1365-8711.2001.04912.x. arXiv: astro-ph/0012055 [astro-ph].

[159] Volker Springel et al. "Simulations of the formation, evolution and clustering of galaxies and quasars." In: 435.7042 (June 2005), pp. 629–636. DOI: 10.1038/nature03597. arXiv: astro-ph/0504097 [astro-ph].

[160] R. A. Sunyaev and Ya. B. Zeldovich. "Small-Scale Fluctuations of Relic Radiation." In: 7.1 (Apr. 1970), pp. 3–19. DOI: 10.1007/BF00653471.

[161] R. A. Sunyaev and Ya. B. Zeldovich. "The Observations of Relic Radiation as a Test of the Nature of X-Ray Radiation from the Clusters of Galaxies." In: *Comments on Astrophysics and Space Physics* 4 (Nov. 1972), p. 173.

[162] R. Teyssier. "Cosmological hydrodynamics with adaptive mesh refinement. A new high resolution code called RAMSES." In: 385 (Apr. 2002), pp. 337–364. DOI: 10.1051/0004-6361:20011817. arXiv: astro-ph/0111367 [astro-ph].

[163] Jeremy Tinker, Andrey V. Kravtsov, Anatoly Klypin, Kevork Abazajian, Michael Warren, Gustavo Yepes, Stefan Gottlöber, and Daniel E. Holz. "Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality." In: 688.2 (Dec. 2008), pp. 709–728. DOI: 10.1086/591439. arXiv: 0803.2706 [astro-ph].

[164] Matteo Viel, George D. Becker, James S. Bolton, and Martin G. Haehnelt. "Warm dark matter as a solution to the small scale crisis: New constraints from high redshift Lyman-α forest data." In: 88.4, 043502 (Aug. 2013), p. 043502. DOI: 10.1103/PhysRevD.88.043502. arXiv: 1306.2314 [astro-ph.CO].

[165] A. Vikhlinin, A. Kravtsov, W. Forman, C. Jones, M. Markevitch, S. S. Murray, and L. Van Speybroeck. "Chandra Sample of Nearby Relaxed Galaxy Clusters: Mass, Gas Fraction, and Mass-Temperature Relation." In: 640.2 (Apr. 2006), pp. 691–709. DOI: 10.1086/500288. arXiv: astro-ph/0507092 [astro-ph].

[166] A. Vikhlinin et al. "Chandra Cluster Cosmology Project III: Cosmological Parameter Constraints." In: 692.2 (Feb. 2009), pp. 1060–1074. DOI: 10.1088/0004-637X/692/2/1060. arXiv: 0812.2720 [astro-ph].

[167] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Greg Snyder, Dylan Nelson, and Lars Hernquist. "Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe." In: 444.2 (Oct. 2014), pp. 1518–1547. DOI: 10.1093/mnras/stu1536. arXiv: 1405.2921 [astro-ph.CO].

[168] Rachel Ward, Xiaoxia Wu, and Leon Bottou. "AdaGrad stepsizes: Sharp convergence over nonconvex landscapes." In: *arXiv e-prints*, arXiv:1806.01811 (June 2018), arXiv:1806.01811. DOI: 10.48550/arXiv.1806.01811. arXiv: 1806.01811 [stat.ML].

[169] Rainer Weinberger et al. "Simulating galaxy formation with black hole driven thermal and kinetic feedback." In: 465.3 (Mar. 2017), pp. 3291–3308. DOI: 10.1093/mnras/stw2944. arXiv: 1607.03486 [astro-ph.GA].

[170] N. Werner, A. Finoguenov, J. S. Kaastra, A. Simionescu, J. P. Dietrich, J. Vink, and H. Böhringer. "Detection of hot gas in the filament connecting the clusters of galaxies Abell 222 and Abell 223." In: 482.3 (May 2008), pp. L29–L33. DOI: 10.1051/0004-6361:200809599. arXiv: 0803.2525 [astro-ph].

[171] S. D. M. White and M. J. Rees. "Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering." In: 183 (May 1978), pp. 341–358. DOI: 10.1093/mnras/183.3.341.

[172] Robert P. C. Wiersma, Joop Schaye, Tom Theuns, Claudio Dalla Vecchia, and Luca Tornatore. "Chemical enrichment in cosmological, smoothed particle hydrodynamics simulations." In: 399.2 (Oct. 2009), pp. 574–600. DOI: 10.1111/j.1365-2966.2009.15331.x. arXiv: 0902.1535 [astro-ph.CO].

[173] Idit Zehavi et al. "Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity." In: 736.1, 59 (July 2011), p. 59. DOI: 10.1088/0004-637X/736/1/59. arXiv: 1005.2413 [astro-ph.CO].

[174] Ya B. Zel'dovich. "Special Issue: the Cosmological Constant and the Theory of Elementary Particles." In: *Soviet Physics Uspekhi* 11.3 (Mar. 1968), pp. 381–393. DOI: 10.1070/PU1968v011n03ABEH003927.

[175] Ya. B. Zel'dovich. "Gravitational instability: An approximate theory for large density perturbations." In: 5 (Mar. 1970), pp. 84–89.

[176] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization." In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

[177] F. Zwicky. "Die Rotverschiebung von extragalaktischen Nebeln." In: *Helvetica Physica Acta* 6 (Jan. 1933), pp. 110–127.

[178] lightly.ai. *Lightly: A computer vision framework for self-supervised learning*. https://github.com/lightly-ai/lightly. Version To appear. Accessed: 2024-07-25. 2024.

[179] R. J. van Weeren, F. de Gasperin, H. Akamatsu, M. Brüggen, L. Feretti, H. Kang, A. Stroe, and F. Zandanel. "Diffuse Radio Emission from Galaxy Clusters." In: 215.1, 16 (Feb. 2019), p. 16. DOI: 10.1007/s11214-019-0584-z. arXiv: 1901.04496 [astro-ph.HE].

[180] Reinout J. van Weeren, Huub J. A. Röttgering, Marcus Brüggen, and Matthias Hoeft. "Particle Acceleration on Megaparsec Scales in a Merging Galaxy Cluster." In: *Science* 330.6002 (Oct. 2010), p. 347. DOI: 10.1126/science.1194293. arXiv: 1010.4306 [astro-ph.CO].