2025 SNU Summer Undergraduate Internship Report

Keonwoo Kim[1]
Supervisor: Prof. Jihoon Kim[2]
Collaborators: Dr. Fukushima Keita, Minyong Jung[3]
August 11, 2025

[1]Department of Physics, McGill University
[2]Department of Physics and Astronomy, Seoul National University
[3]Department of Physics and Astronomy, Seoul National University

Abstract

This report summarizes two projects completed during the 2025 summer term. **(1) Protoclusters at cosmic noon:** Using a cosmological hydrodynamical simulation (Gadget-4; box 200 Mpc) at $z = 2.22$, we quantified star formation across protocluster environments. We constructed the star-formation main sequence (SFMS) from subhalo catalogs, identified $+2.5\sigma$ outliers as "starbursts," and assessed their contribution to the total SFR budget. The snapshot exhibits a well-defined SFMS and a small starburst contribution (few percent) over $9 \leq \log(M_\star/\mathrm{M}_\odot) \leq 11.5$. Across 77 protoclusters, neither the starburst number fraction nor the SFR fraction shows a clear trend with simple kinematic proxies (line-of-sight velocity dispersion σ_v or the virial proxy $\sigma_v^2/(M_{200}/R_{200})$). These results motivate follow-up analyses using assembly-aware metrics (e.g., core mass accretion rate, half-mass redshift) and the inclusion of metallicity diagnostics to test whether structure-to-structure MZR shifts trace growth stage. **(2) DARWIN merger-tree groundwork:** We prepared a testbed to apply a merger-tree–based galaxy matching approach to the DARWIN simulations. For Darwin–1 (65 cMpc) and Darwin–2 (10 cMpc), we parsed galaxy catalogs, traced main-branch progenitors via `flag_prog`=1 and `ID_prog`, recorded (x, y, z) and M_{dm} histories back to snapshot 8, and paired snapshots by nearest redshift. Sanity checks include a final-snapshot x–y overlay (Darwin–2 within the Darwin–1 volume) and smooth example trajectories. As a short-term pilot, we will match a subset of Darwin–2 galaxies to Darwin–1 counterparts via a simple, time-aligned similarity over positions and $\log M_{\mathrm{DM}}$; contingent on those results, we will explore a lightweight ML stage to relate low- to high-resolution properties.

Contents

1 Tracing Protocluster Growth at $z = 2.22$ with Star Formation Diagnostics

1.1 Introduction

Protoclusters–overdensities at high redshift (conventionally $z \geq 2$) destined to become today's galaxy clusters–are prime laboratories for watching environment-driven galaxy evolution in real time. A recurring theme from recent observational work suggests that galaxies in different protoclusters do not always follow the same mass-metallicity relation (MZR); instead, the MZR appears to shift from structure to structure in ways that may encode each protocluster's growth stage (Pérez-Martínez 2023). In other words, chemical enrichment seems to remember assembly history. Testing that idea requires a model that tracks galaxies, gas flows, enrichment, and environment self-consistently across large scales, not just a handful of isolated halos.

This project uses a cosmological hydrodynamical simulation (Gadget-4) to do exactly that; we identify protocluster regions in the fixed volume, and compare galaxy properties across them as a function of their assembly state. Our long-term goal is to ask whether the observed, structure-to-structure variation in the MZR is a genuine clock of protocluster growth. If so, we expect to see coherent differences in metallic enrichment and star-formation activity that line up with differences in mass accretion histories, gas supply, and feedback.

For this summer, we are narrowing the scope to the other half of that story: star formation rates (SFRs). SFR is the most immediate tracer of how a protocluster is growing–both through in-situ star formation and through the balance of gas inflows/outflows–so it is a reasonable starting point before we tackle metallicities. Our focus is to quantify how star formation is organized inside protoclusters.

1.2 Methods

We analyze a cosmological hydrodynamical run performed with Gadget-4[4] in a periodic box of size 200 Mpc. Subgrid black-hole/AGN physics was disabled for this run; the consequences for high-mass star formation are addressed in the limitations. Out of 20 snapshots generated from $z = \infty$ to $z = 0$, we work at snapshot 12 (corresponding to z = 2.22). We target this specific redshift because it sits in "cosmic noon" ($2 \lesssim z \lesssim 3$), when the global SFR density, gas fractions, and merger activity peak and protoclusters are still assembling–ideal for catching environment-driven effects. It also matches the redshift range of key observational benchmark (Rodighiero et al. 2011) and simulation studies (Sparre et al. 2015), enabling clean like-for-like comparisons.

Within our simulation, haloes were identified with a `Friends-of-Friends` (FoF) group finder, and bound structures within each FoF group were decomposed with `SUBFIND` into subhalos. For every subhalo, the catalogue provides 58 properties including 3-D positions, total and component masses, and star-formation rates (SFRs). We read the subhalo arrays from 64 HDF5 files for this snapshot (e.g. `/Subhalo/SubhaloSFR` and `/Subhalo/SubhaloMassType`), adopt the stellar-mass component `[:, 4]`, and convert to physical units using $M_* = \text{SubhaloMassType}_* 10^{10} M_\odot/h$ with $h = 0.67742$. To avoid pathological logarithms and to enforce a conservative resolution floor, we set

[4] See Oku et al. (2024) for introductory documentation.

to `NaN` all entries with $\log(M_*/\mathrm{M}_\odot) < 9.0$ and treat SFR $= 0$ as a non-detection (also set to `NaN`). All subsequent statistics are performed in log-space using only finite values.

The protocluster membership is assigned from project-internal catalogues that list, for each structure, the IDs of **core** galaxies (central region) and **member** satellites at $z = 2.22$. We merge these lists into a per-protocluster tuple (cores, members) and also retain a **field** sample consisting of subhalos outside any protocluster list. In total this procedure yields 77 protoclusters at this snapshot. Because the membership files encode the protocluster ID along with each subhalo ID, our analysis can aggregate any statistics (SFRs, masses) per structure and compare cores, members, and field on identical footing.

To characterize the "normal" star formation across the logarithmic mass range for this redshift, we construct the star-formation main sequence (SFMS) following the simulation-oriented definition of Sparre et al. (2015); at fixed stellar mass, the SFMS ridge is the median log(SFR) of star-forming galaxies. We bin galaxies in $\log(M_*/\mathrm{M}_\odot)$ with fixed width $\Delta = 0.2$ dex, compute the median log(SFR) in each bin, and connect these medians to form the SFMS curve. The scatter is taken as the per-bin standard deviation of log(SFR) computed from all galaxies with finite values.

For the present snapshot, we compute the SFMS globally (all galaxies combined); per-protocluster SFMS curves are deferred due to insufficient counts per mass bin within individual protocluster in our simulation.

Finally, to relate star-formation behaviour to assembly, we compute simple snapshot-level assembly proxies for each protocluster: total stellar mass $\sum M_*$, and total SFR $\sum SFR$. Where available, we also measure a line of sight galaxy velocity dispersion σ_v as an additional proxy. We then compare each protocluster's starburst fraction against these proxies. All analysis is performed in Python using `h5py`, `numpy`, `pandas`.

1.3 Results

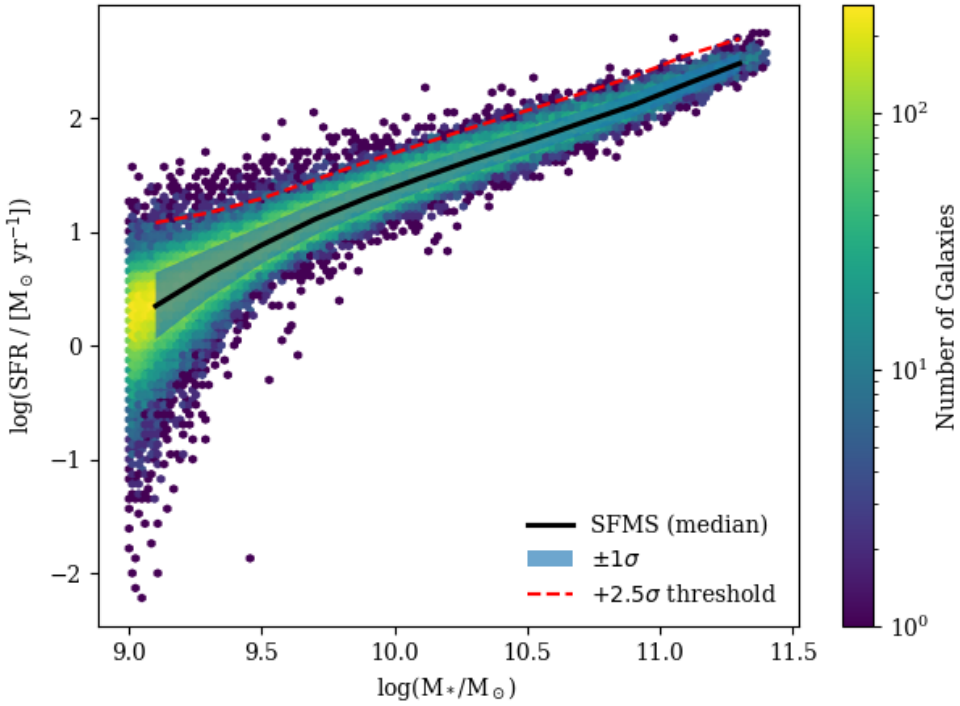


Figure 1: Star-formation rate–stellar mass plane at $z = 2.22$. The background shows a hexbin number–density map (colour bar in log N). The black curve is the star-formation main sequence (SFMS), defined as the median log(SFR) in fixed-width $\log(M_*/\mathrm{M}_\odot)$ bins (0.05 dex); the blue band indicates the $\pm1\sigma$ scatter about the median. The red dashed curve marks the $+2.5\sigma$ threshold used to flag starbursts in later figures. Galaxies with $\log(M_*/\mathrm{M}_\odot) < 9$, $\log(M_*/\mathrm{M}_\odot) > 11.5$ or SFR $= 0$ are excluded; all statistics use only finite values.

Fig. 1 shows the distribution of galaxies in the SFR–$M_\star$ plane at $z = 2.22$ together with our binned estimate of the star–formation main sequence (SFMS). A well–defined ridge is recovered: the median $\log \mathrm{SFR}$ rises smoothly with $\log(M_\star/\mathrm{M}_\odot)$ over $9 \lesssim \log M_\star \lesssim 11.5$, with a scatter of order a few tenths of a dex across most of the mass range. No secondary sequence is evident. The $+2.5\sigma$ curve lies well above the locus of the population, and only a sparse tail of outliers reaches it. The qualitative shape of the SFMS is consistent with published $z \sim 2$ result (Pérez-Martínez 2023; Sparre et al. 2015), lending confidence that the snapshot provides a realistic baseline. We verified that the appearance of the ridge is stable to reasonable choice of bin width (0.05–0.2 dex) and that all bins used in the fit are well populated with ≥ 50 galaxies for all bins, with zero-SFR samples masked from the statistics.

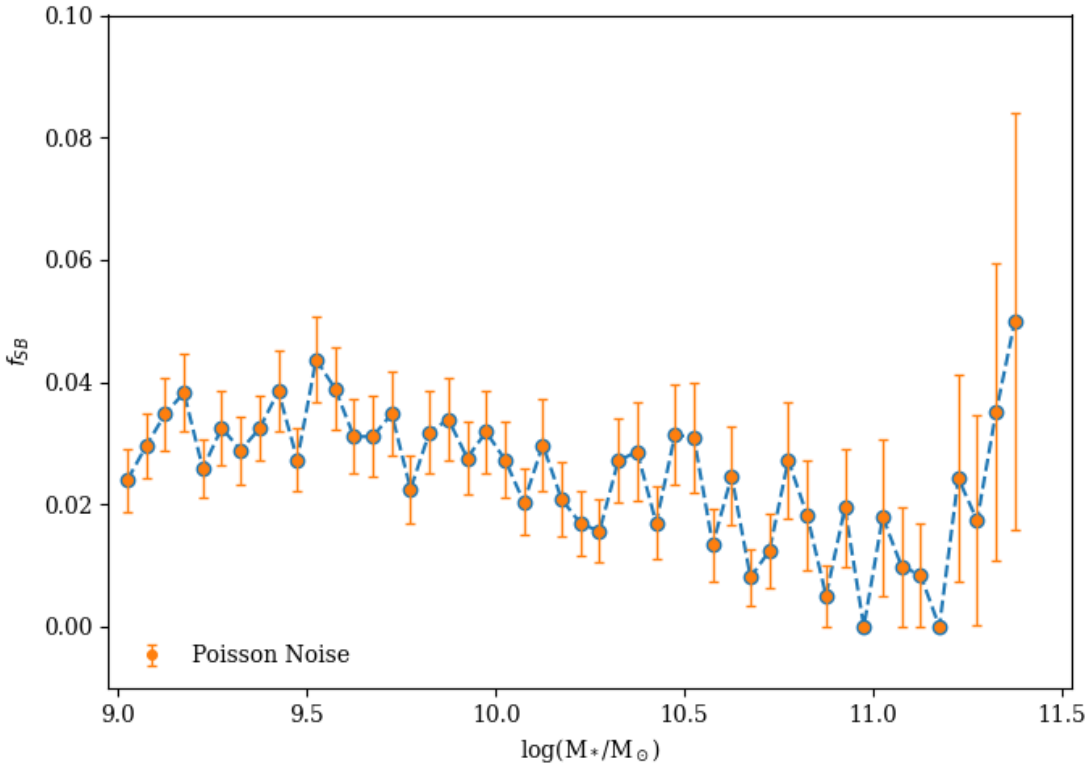
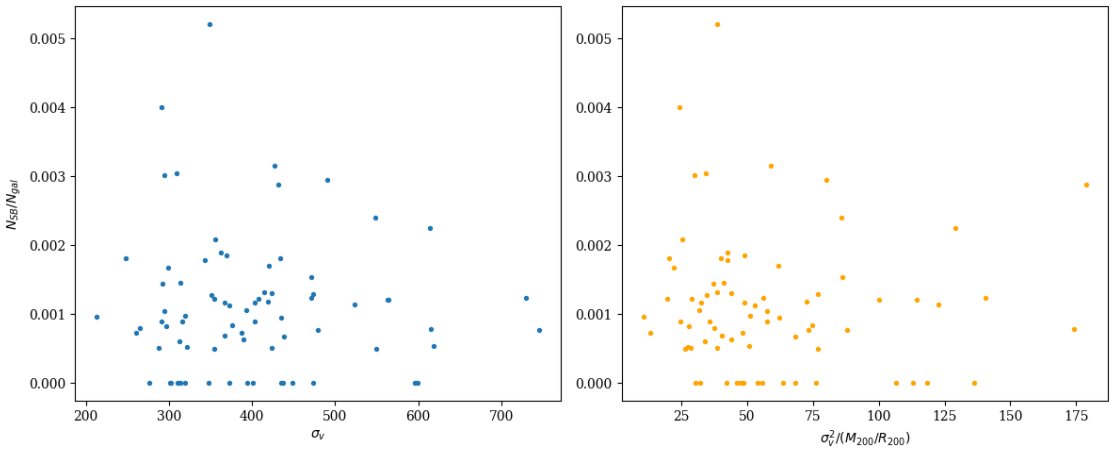
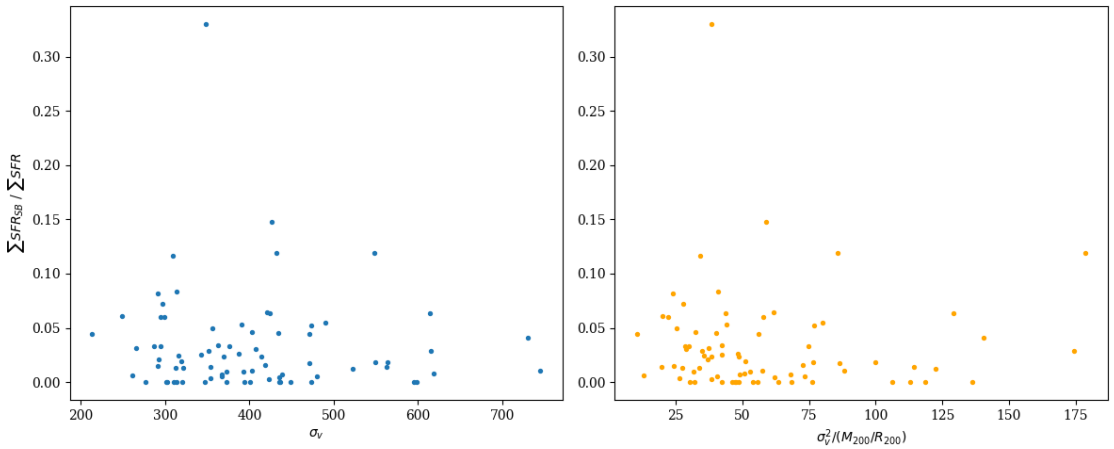


Figure 2: Fraction of the total SFR contributed by starburst outliers as a function of stellar mass at $z = 2.22$. Points (blue) show, in 0.05-dex $\log(M_\star/\mathrm{M}_\odot)$ bins, $f_{\mathrm{SB}} \equiv \sum \mathrm{SFR}_{(\Delta>2.5\sigma)} / \sum \mathrm{SFR}_{\mathrm{all}}$, where $\Delta = \log(\mathrm{SFR}) - \mathrm{SFMS}(M_\star)$ and the SFMS and σ are those from Fig. 1. Error bars (orange) are Poisson-only uncertainties for weighted sums, $\mathrm{Var}(\sum w_i) = \sum w_i^2$ with $w_i = \mathrm{SFR}_i$. All bins contain > 50 galaxies; the upturn at the highest masses carries large uncertainties due to low counts of massive systems.

Fig. 2 shows that starbursts contribute only a few percent of the total star-formation activity across the mass range, with a mild decline toward higher masses and a noisy upturn in the sparsely populated, most massive bin. Quantitatively, our f_{SB} curve lies systematically below both the Illustris result of Sparre et al. (2015) and the observational estimates of Rodighiero et al. (2011) at similar redshift $z \sim 2$, reinforcing the picture already suggested by Fig. 1 that strong ($> 2.5\sigma$) outliers are rare in this run. This outcome is robust to reasonable binning choices (we tested 0.05–0.20 dex), and reflects a combination of (i) a relatively high per-bin $\sigma_{\log \mathrm{SFR}}$, which raises the outlier threshold, and (ii) model limitations of the present simulation (e.g., insufficient number of samples and finite resolution). Within these caveats, the $z = 2.22$ snapshot indicates that the SFR budget is dominated by the SFMS galaxies rather than the starbursts.



(a)



(b)

Figure 3: Starburst incidence and contribution per protocluster at $z = 2.22$. **(a)** The fraction of galaxies classified as starbursts ($N_{\rm SB}/N_{\rm gal}$) as a function of line-of-sight velocity dispersion σ_v (left) and the virial proxy $\sigma_v^2/(M_{200}/R_{200})$ (right). **(b)** The fraction of total SFR contributed by starbursts ($\sum \mathrm{SFR_{SB}}/\sum \mathrm{SFR_{all}}$) versus the same axes. Each point is one protocluster. Typical starburst counts are only 2–3 per protocluster, so vertical scatter is dominated by Poisson noise. No clear correlation is apparent with either kinematic proxy.

Fig. 3 examines whether dynamically active protoclusters host a larger incidence of starbursts. Across the 77 structures, both the starburst number fraction $N_{\rm SB}/N_{\rm gal}$ (top row) and the SFR-weighted fraction $\sum \mathrm{SFR_{SB}}/\sum \mathrm{SFR_{all}}$ (bottom row) show broad scatter with no obvious monotonic trend versus either σ_v or the virial proxy $\sigma_v^2/(M_{200}/R_{200})$, which is used as a stand-in for velocity bias (Munari et al. 2013). Typical number fractions are sub-percent and the SFR fractions are at the few-percent level, consistent with the global analysis. The absence of a clear relation is unsurprising given the small per-structure starburst counts (ranging from 0 to 5 throughout all protoclusters), which imply fractional uncertainties of order 50–60% and likely dominate the vertical scatter. Within these limits, we conclude that σ_v and $\sigma_v^2/(M_{200}/R_{200})$ are not, on their own, distinguishing indicators of enhanced starburst activity at this epoch; more sensitive assembly trac-

ers (e.g., stellar/gas metallicity, gas fractions) and/or higher-resolution runs with larger starburst samples will be required to detect subtler trends.

1.4 Discussion

The global SFR–$M_\star$ plane at $z = 2.22$ shows a well-defined SFMS whose shape is consistent with expectations for "cosmic noon." When we quantify outliers as $+2.5\sigma$ above the SFMS in their mass bin, the fraction of the total SFR carried by starbursts is small (a few percent) across $9 \leq \log(M_\star/\mathrm{M}_\odot) \leq 11.5$, and is systematically lower than both observational inferences and the Illustris result using the same criterion. Two factors plausibly drive the low fractions: (i) the measured per-bin $\sigma_{\log \mathrm{SFR}}$ places the $+2.5\sigma$ threshold high above the ridge, mechanically yielding few outliers; and (ii) the run has finite resolution which can underproduce short duty-cycle bursts in massive systems. Within these caveats, the snapshot indicates that the SFR budget is dominated by SFMS galaxies rather than transient bursts.

On a per-protocluster basis we do not detect a clear correlation between starburst incidence (either number fraction $N_{\mathrm{SB}}/N_{\mathrm{gal}}$ or SFR fraction $\sum \mathrm{SFR_{SB}}/\sum \mathrm{SFR_{all}}$) and simple kinematic quantities such as line-of-sight dispersion σ_v or the virial proxy $\sigma_v^2/(M_{200}/R_{200})$. This result is unsurprising given the small per-structure starburst counts (0–5 per PC, implying large Poisson fractional uncertainties), but it also reflects the limitations of these snapshot kinematic measures as assembly tracers. More physically grounded descriptors of assembly should be used on the x–axis. In particular, we will replace σ_v-based proxies with (i) the core's *mass accretion rate* (MAR) over a fixed lookback time, (ii) the *half-mass redshift* $z_{1/2}$ at which the core first accumulated half of its present mass, and (iii) the *total subhalo mass to total galaxy stellar-mass ratio*. The first two require simulation's time-series data to track growth history directly; at fixed mass they should anti-correlate (for $z_{1/2}$) or correlate (for MAR) with current burstiness if rapid, gas-rich assembly enhances SFR. The third gauges how baryons are partitioned relative to the halo and may expose maturation of the core. These metrics are better matched to the hypothesis that environment-dependent SFR and enrichment encode assembly history.

Methodologically, two improvements should tighten the constraints. First, move beyond a hard outlier cut by analyzing the full distribution of $\Delta \equiv \log \mathrm{SFR} - \mathrm{SFMS}(M_\star)$ (e.g., its mean and width) for field galaxies vs. protoclusters; this uses all galaxies and reduces small-number noise. Second, repeat the analysis with time-averaged SFRs (e.g., 50–250 Myr) to suppress stochastic bursts or quenching flicker and isolate long-timescale differences. Ultimately, higher resolution or a larger volume (more protoclusters and starbursts) will be needed to secure subtle environmental trends.

1.5 Conclusion

At $z = 2.22$ our simulation reproduces a smooth SFMS and yields a small contribution of $+2.5\sigma$ outliers to the total SFR, lower than observational estimates and lower than previous simulation benchmarks. Across 77 protoclusters we find no compelling evidence that simple kinematic proxies (σ_v, $\sigma_v^2/(M_{200}/R_{200})$) track enhanced starburst activity; given current sample sizes, these quantities are not discriminating indicators of assembly stage. The immediate implication is that SFR and $M_\star$ at a single epoch are insufficient to order protoclusters by maturity.

The next phase of this project will (i) replace snapshot kinematics with *assembly-aware* axes

(core MAR, half-mass redshift $z_{1/2}$, and mass-partition ratios) derived from time-series data; (ii) incorporate metallicity (gas-phase and stellar) to test whether structure-to-structure MZR shifts act as a clock of assembly; and (iii) repeat the analysis across multiple snapshots and, where possible, with higher resolution and BH/AGN physics. These steps will turn the present baseline into a discriminating test of how protocluster growth imprints on star formation and chemical enrichment.

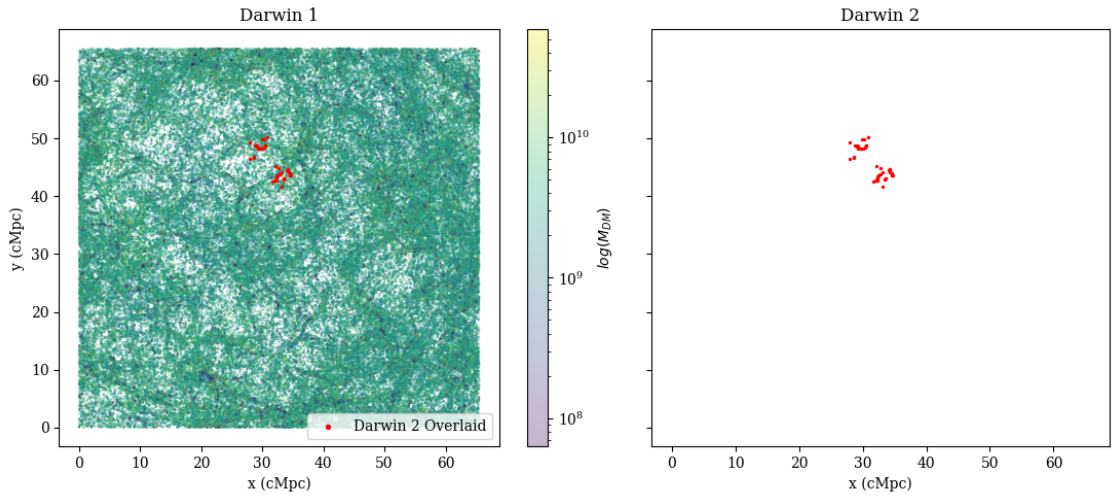


Figure 4: Final-snapshot spatial distribution. *Left:* Darwin–1 (SN55, $z \simeq 3.886$), points coloured by $\log(M_{\rm DM})$; semi-transparent markers reduce crowding. Red points are Darwin–2 galaxies (SN64, $z \simeq 3.864$) overlaid in the same comoving coordinates. *Right:* Darwin–2 alone. The overlay shows the Darwin–2 footprint within the larger Darwin–1 volume and provides a quick visual check of the redshift alignment.

2 Preliminary Progress on Applying Merger-Tree–Based Galaxy Matching to the Darwin Simulations

2.1 Introduction

State-of-the-art cosmological simulations increasingly rely on multi–resolution strategies: large, lower–resolution volumes to sample environments and statistics, and smaller, higher–resolution runs to study galaxy physics in detail. A practical challenge is how to connect these regimes—e.g., to identify corresponding galaxies across runs and to translate low–resolution predictions into high–resolution–like properties. A recent line of work (e.g., merger–tree–based matching combined with a lightweight gradient–boosted model) has shown that using galaxies' histories—positions and masses along their main branches—can be an effective basis for cross–simulation correspondence and downstream property correction (Jung et al. 2024).

This note sets up an initial testbed for applying that idea to the Darwin simulation suite [5]. The goal at this stage is intentionally modest and exploratory: assemble the data products needed to attempt merger–tree–based matching between two Darwin runs and sketch the path toward a simple machine–learning (ML) stage. We do not commit to a specific success metric here; those criteria will be finalized after small–scale pilot matches.

2.2 Progress

We work with two runs: *Darwin–1* (lower resolution, 65 cMpc box) and *Darwin–2* (higher resolution, 10 cMpc box). For each run we use the provided halo, subhalo, and galaxy catalogues (from

[5] See DAR (2023a) and DAR (2023b) for the official information of the project

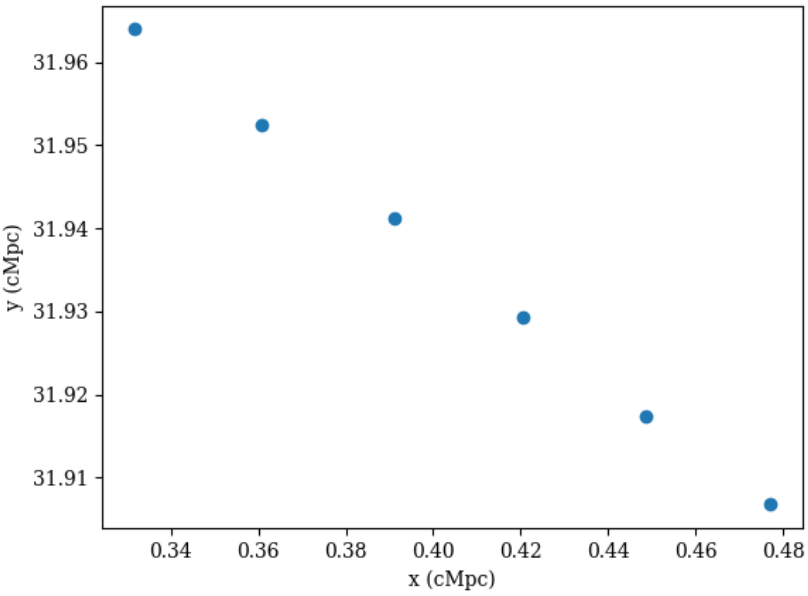


Figure 5: Example main-branch position history in Darwin–1. Dots mark the (x, y) positions of a single galaxy over the last six snapshots (earlier to later: upper-left to lower-right). The smooth evolution indicates that the progenitor links produce coherent tracks suitable for cross-simulation matching.

friends–of–friends haloes through self–bound subhaloes to star–hosting subhaloes with derived properties). For the current setup we trace histories back to snapshot 8 (snapshot 7 is unavailable in Darwin–1).

For every galaxy at the final snapshot of each run, we traverse the main branch by following the catalogue's progenitor links (using the provided `flag_prog` and `ID_prog`) and record basic time–series: (x, y, z) positions in comoving Mpc and dark–matter mass $M_{\rm dm}$ per snapshot. We also establish a snapshot correspondence between Darwin–1 and Darwin–2 by nearest–neighbour matching in redshift; in practice the paired redshifts are very close (minor differences at the second decimal place are acceptable for this setup). Two simple sanity checks accompany this stage: (i) a final–snapshot x–y map with Darwin–2 galaxies overlaid on Darwin–1, and (ii) an example main–branch trajectory showing smooth positional evolution over recent snapshots.

Fig. 4 summarizes the final-snapshot spatial context. The left panel shows all Darwin–1 galaxies coloured by $\log(M_{\rm dm})$ at SN55 ($z \simeq 3.886$), with the Darwin–2 galaxies at the matched redshift (SN64, $z \simeq 3.864$) overlaid as red points in the same comoving frame. The right panel shows Darwin–2 alone. The overlay cleanly delineates the high-resolution region within the 65 cMpc box and serves as a visual check that the nearest-z pairing is adequate for this setup.

Fig. 5 provides a sanity check on the merger-tree traversal: the main-branch (x, y) positions of one representative Darwin–1 galaxy over the last six snapshots. The sequence moves from upper-left to lower-right (later times), and the changes are smooth and monotonic at the $\sim 10^{-2}$ cMpc level per step—consistent with coherent orbital motion—indicating that following `flag_prog`$= 1 \rightarrow$`ID_prog` is producing physically plausible tracks for downstream matching.

2.3 Future Work

As an immediate pilot, we will select a small subset of Darwin–2 galaxies and search for Darwin–1 counterparts whose position and (log) $M_{\rm dm}$ histories are most similar under a simple, time–aligned distance (e.g., a weighted RMSE over snapshots). Depending on these pilots, we will then explore an ML stage—developed and inspired by Jung et al. 2024—to learn a mapping from low– to high–resolution properties using features derived from the merger histories. Details of the model, targets, and evaluation will be kept lightweight and finalized after the pilot results.

References

The darwin project. RAMSES SNO project page, 2023a. URL https://ramses.cnrs.fr/the-darwin-project/. Accessed 2025-08-12. 9

Darwin simulation project website. Project website, 2023b. URL https://darwin-simulation.github.io/. Accessed 2025-08-12. 9

Minyong Jung, Ji-Hoon Kim, Boon Kiat Oh, Sungwook E. Hong, Jaehyun Lee, and Juhan Kim. Merger tree-based galaxy matching: A comparative study across different resolutions. Draft version March 21, 2024; submitted to ApJ, 2024. 9, 11

E. Munari, A. Biviano, S. Borgani, G. Murante, and D. Fabjan. The relation between velocity dispersion and mass in simulated clusters of galaxies: dependence on the tracer and the baryonic physics. *MNRAS*, 430(4):2638–2649, 2013. doi: 10.1093/mnras/stt049. 6

Yuri Oku and Kentaro Nagamine. Osaka feedback model III: Cosmological simulation CROCODILE. *arXiv e-prints*, 2024. Draft version September 4, 2024. 2

J. M. Pérez-Martínez, T. Kodama, Y. Koyama, R. Shimakawa, T. L. Suzuki, K. Daikuhara, K. Adachi, M. Onodera, and I. Tanaka. Enhanced star formation and metallicity deficit in the uss 1558–003 forming protocluster at z=2.53. 2023. Preprint. 2, 4

Giulia Rodighiero, Emanuele Daddi, Ivano Baronchelli, Andrea Cimatti, Alvio Renzini, Hervé Aussel, Paola Popesso, Dieter Lutz, Paola Andreani, Stefano Berta, Antonio Cava, David Elbaz, Anna Feltre, Adriano Fontana, Natascha M. Förster Schreiber, Alberto Franceschini, Reinhard Genzel, Andrea Grazian, Carlotta Gruppioni, Olivier Ilbert, Emeric Le Floch, Georgios Magdis, Manuela Magliocchetti, Benjamin Magnelli, Roberto Maiolino, Henry J. McCracken, Raanan Nordon, Albrecht Poglitsch, Paola Santini, Francesca Pozzi, Laurie Riguccini, Linda Tacconi, Stijn Wuyts, and Giovanni Zamorani. The lesser role of starbursts in star formation at $z = 2$. *The Astrophysical Journal Letters*, 739(2):L40, September 2011. doi: 10.1088/2041-8205/739/2/L40. 2, 5

Martin Sparre, Christopher C. Hayward, Volker Springel, Mark Vogelsberger, Shy Genel, Paul Torrey, Dylan Nelson, Debora Sijacki, and Lars Hernquist. The star formation main sequence and quenching: a view from the illustris simulation. *Monthly Notices of the Royal Astronomical Society*, 447(4):3548–3563, March 2015. doi: 10.1093/mnras/stu2700. 2, 3, 4, 5