Skepthical review: *Contrastive Learning of Merger Tree Embeddings for Likelihood-Free Cosmological Inference*

Summary

The paper presents a two-stage likelihood-free cosmological inference pipeline from dark-matter halo merger trees. A Graph Neural Network (GNN) is trained with a contrastive (NT-Xent) objective to embed each merger tree into a fixed-dimensional representation such that trees from the same cosmology (parameterized by Ω_m and σ_8) are close in embedding space while trees from different cosmologies are separated (Sec. 2.4, Sec. 3.4). These embeddings are then used as summary statistics in Sequential Neural Posterior Estimation (SNPE) with a normalizing-flow density estimator to infer posteriors over (Ω_m, σ_8) (Sec. 2.5, Sec. 3.5). A key claimed novelty is an "assembly-bias mitigation" data augmentation that resamples halo concentrations at fixed mass using an empirically fitted mass–concentration relation with scatter (Sec. 2.3, Sec. 3.3). On the reported dataset (1000 trees from 40 cosmologies; Sec. 2.1, Sec. 3.1), the method yields reasonably accurate and near-nominal uncertainty for Ω_m, but notably weaker and under-covered posteriors for σ_8 (Sec. 3.5–3.6). While the overall approach is timely and promising, the current manuscript is not yet fully convincing or reproducible: the simulation/data generation and sampling are under-specified, the augmentation's physical motivation and empirical benefit are not demonstrated, key baselines/ablations are missing, calibration is only partially assessed due to incomplete SBC (Sec. 2.6), and several methodological details (contrastive loss specification, batching, SNPE configuration) are too vague to audit or reproduce. Strengthening the experimental validation—especially regarding assembly-bias robustness, shortcut learning risks, and calibration—would substantially improve the paper's reliability and impact.

Strengths

- Timely and relevant goal: using merger-tree information for simulation-based cosmological inference, with explicit attention to assembly-bias concerns (Sec. 1).
- Clear high-level pipeline design (graph encoder → embedding → SNPE posterior) that could be reused for related SBI problems (Secs. 2.4–2.5).
- Methodological novelty in combining contrastive learning on merger-tree graphs with SNPE for posterior inference (Secs. 2.4–2.5).
- Useful exploratory analysis of engineered global merger-tree features and their correlations with Ω_m and σ_8, providing interpretive context and highlighting informative time-dependent concentration features (Sec. 3.2–3.2.3).
- Split-by-cosmology evaluation design (placing all trees from a cosmology into the same split) is directionally appropriate to reduce leakage across cosmologies (Sec. 2.1, Sec. 3.1).

- Reporting includes uncertainty-aware diagnostics (credible-interval coverage) in addition to point-estimate error metrics (Sec. 3.5–3.6).

Major issues

1. **Simulation suite, merger-tree construction, and cosmology sampling are under-specified, preventing assessment of realism/generalization and blocking reproducibility. Sec. 2.1 and Sec. 3.1 state "40 cosmologies, 25 trees each" across Ω_m and σ_8 ranges, but do not provide the simulation code/suite, box size, particle number, mass/force resolution, snapshot cadence/redshifts, halo finder, tree-building algorithm, halo selection criteria (mass thresholds and whether final-halo masses are fixed/narrow-binned), or how the 40 cosmologies were sampled (grid vs Latin hypercube vs random). It is also unclear whether the 25 trees per cosmology are statistically independent or share volume-specific artifacts.**

   Recommendation: Expand Sec. 2.1 (and cross-reference in Sec. 3.1) with a compact but complete dataset table: (i) simulation suite name and citation (or code + IC generator), (ii) box size, $N_{\text{particles}}$, particle mass, force softening, (iii) snapshot list/redshift range used for trees, (iv) halo finder and merger-tree builder (with versions/settings), (v) halo selection: final redshift, mass definition (e.g., M_{200c}), mass bin/range, and how trees are chosen (random? top-N by mass? per-environment?), and (vi) cosmology design: how 40 (Ω_m, σ_8) pairs are sampled and what prior bounds are. If trees come from shared volumes, describe how you ensure evaluation is not driven by volume/simulation-ID artifacts.

2. **The concentration-resampling augmentation is not convincingly motivated as "assembly-bias mitigation" and is not empirically validated; it may also create physically inconsistent/out-of-distribution trees. Sec. 2.3 and Sec. 3.3 rely on an M–C fit with extremely low R^2 (≈ 0.0036) and large scatter, yet the method is described as an assembly-bias correction. Resampling concentration independently conditional on mass (and ambiguously on scale factor) risks destroying temporal coherence along the main progenitor branch and correlations with other node properties (e.g., V_{max}), potentially altering cosmological signal and/or introducing artifacts. The paper does not show a with/without augmentation comparison for (i) inference metrics, (ii) sensitivity to assembly-bias proxies, or (iii) physical consistency checks.**

   Recommendation: Reframe the augmentation as an exploratory robustness test unless you can demonstrate a measurable benefit. Concretely: (i) clarify whether the M–C relation is fit globally or per scale-factor bin (resolve Sec. 2.3 vs Sec. 3.3 inconsistency), and report fit parameters + scatter per bin if applicable; (ii) add ablations: train/evaluate the full pipeline with and without augmentation, reporting Ω_m/σ_8

RMSE and coverage (Sec. 3.5–3.6); (iii) add robustness stratification: evaluate errors/coverage across quantiles of assembly-bias proxies identified in Sec. 3.2.3 (e.g., conc_scatter_final_snapshot, $\delta \log_{10} c_{\mathrm{sat-cen}}$, nummajor_mergers_mpb), comparing augmented vs non-augmented; (iv) add physical consistency checks pre/post augmentation (distributions of C, C–V_{max} relation, and correlation of concentration along MPB). If gains are not clear, tone down claims in the Abstract, Sec. 1, Sec. 3.6, and Sec. 4.

3. **Key baselines and ablations are missing, so the added value of (a) contrastive learning, (b) the GNN, and (c) augmentation cannot be isolated. Sec. 3.2 shows engineered global features strongly correlate with Ω_m (and somewhat with σ_8), but there is no inference baseline using these features. There is also no comparison to a supervised GNN trained to regress (Ω_m, σ_8) directly, nor to a contrastive variant better aligned with continuous labels. Without these baselines, it is unclear whether the proposed method materially improves over simpler summaries or whether σ_8 weakness is a representation-learning artifact.**

   Recommendation: Add a baseline/ablation block in Sec. 3.5 (or a dedicated subsection): (i) engineered-features → SNPE baseline (all 35 features as summaries; Sec. 3.2); (ii) a reduced physically motivated feature subset baseline (e.g., early-time MPB concentration features plus formation time and merger counts; Sec. 3.2.3); (iii) supervised GNN regression baseline (predict Ω_m, σ_8 with MSE; use penultimate embeddings as SNPE summaries, or directly compare regression performance); and (iv) augmentation ablation (on/off). Report the same RMSE and coverage metrics used for the main model, emphasizing σ_8.

4. **The contrastive setup uses discrete cosmology IDs as positives/negatives (Sec. 2.4) despite continuous targets (Ω_m, σ_8), and it may encourage shortcut learning via non-cosmological cues. Defining positives as "any other tree from the same cosmology" can let the model exploit simulation- or selection-specific artifacts constant within cosmology (e.g., tree size distributions, mass-range differences, resolution effects, volume/seed effects) rather than cosmology-dependent physics. The current formulation also does not enforce smoothness or metric structure with respect to continuous (Ω_m, σ_8), which may contribute to poor σ_8 performance (Sec. 3.5–3.6).**

   Recommendation: First, fully specify the positive/negative construction (Sec. 2.4): are positives only two augmentations of the same tree, or also different trees sharing the same cosmology? Then add diagnostics to detect shortcuts: (i) correlate embeddings with simple structural properties (node count, max depth, final mass) and report whether these predict cosmology; (ii) provide kNN cosmology classification accuracy in embedding space on held-out cosmologies to test generalization; (iii) if final masses vary, explicitly control for mass (e.g., narrow mass bin, reweighting, or conditioning

on mass). Finally, consider a continuous-label-aware objective: supervised contrastive loss weighted by distance in (Ω_m, σ_8), and/or an auxiliary regression head jointly trained with contrastive loss. Reassess σ_8 constraints after this change (Sec. 3.5).

5. **Inference calibration assessment is incomplete without SBC (Sec. 2.6), and current claims rely on limited coverage estimates that are insufficient to establish posterior validity—especially given σ_8 undercoverage and Ω_m overcoverage. Sec. 3.5–3.6 reports coverage on a small test set (trees from 6 cosmologies) and a single credible level (90%), but does not provide multi-level coverage curves, PIT-style diagnostics, or SBC rank histograms. It is also unclear how intervals are computed (central vs HPD) and how aggregation across trees/cosmologies is done.**

   Recommendation: Either complete SBC as proposed in Sec. 2.6 or substantially qualify calibration statements throughout (Sec. 3.5–3.6, Sec. 4.3–4.4). At minimum: (i) report coverage across multiple nominal levels (e.g., 50/68/90/95%) for each parameter; (ii) clarify interval type (HPD vs central) and aggregation (per-tree vs per-cosmology); (iii) add PIT histograms (or equivalent) for Ω_m and σ_8; (iv) if SBC remains blocked, explicitly describe the technical failure mode and what part of the pipeline prevents SBC.

6. **Core methodological details are missing for both the contrastive GNN and SNPE, preventing reproduction and making it hard to diagnose σ_8 underperformance. For the GNN/contrastive training (Sec. 2.4, Sec. 3.4), the manuscript lacks the explicit NT-Xent formula and key hyperparameters (temperature τ, similarity function, embedding normalization), batch composition (cosmologies per batch, samples per cosmology), optimizer and LR schedule, dropout/norm layers, epochs/early stopping, and the exact augmentation multiplicity K actually used. For SNPE (Sec. 2.5, Sec. 3.5), prior specification, flow architecture details, rounds/simulations per round, training epochs, and evaluation protocol are insufficiently documented.**

   Recommendation: Add a concise hyperparameter/configuration table (main text or appendix) covering: (i) explicit NT-Xent loss definition, similarity metric, τ, embedding normalization, batch construction, and how many positives per anchor; (ii) GNN architecture details (layer types, activations, normalization, dropout), optimizer/LR schedule, batch size, epochs, early stopping; (iii) the exact augmentation operator and K used in the reported results; (iv) SNPE specifics: priors on Ω_m and σ_8, flow type and architecture (layers/hidden sizes/spline bins), number of rounds, training steps, and number of posterior samples used for metrics. Include random seeds and library versions if possible.

7. **Normalization leakage: figures and/or preprocessing appear to use normalization statistics computed over the full dataset (including validation/test), which constitutes information leakage and can bias both representation**

learning and reported analyses (Figures 1–3 and related text).

Recommendation: Recompute all scalers (means/stds) using training data only, apply the same transform to validation/test, and regenerate affected figures/metrics. State this explicitly in Sec. 2.1/Sec. 3.1. If results change, report the updated performance numbers in Sec. 3.5–3.6.

Minor issues

1. Ambiguities and internal inconsistencies in the augmentation description: Sec. 2.3 suggests scale-factor binning for the M–C fit, while Sec. 3.3 reports a single global fit; augmentation multiplicity K is inconsistent ($K = 2$–3 in Sec. 2.3–2.4 vs "one augmented copy" in Sec. 3.3–3.4, and "700 augmented for 700 original").

   Recommendation: Make the augmentation definition unambiguous and consistent across Sec. 2.3–2.4 and Sec. 3.3–3.4: specify whether the M–C relation is global or per-bin, specify the final K used for all reported experiments, and adjust any dataset size statements accordingly. If multiple variants were tried, summarize them as an ablation.

2. Potential mathematical inconsistency in augmentation bounds: the text mixes sampling $\log_{10}(C)$ with clipping bounds that appear to be in linear C units (Sec. 2.3). As written, the perturbation distribution is ill-defined and could materially alter augmentation strength.

   Recommendation: State explicitly whether clipping bounds apply to C or to $\log_{10}(C)$. If sampling in log space, sample $\log_{10}(C)$, exponentiate to C, then clip in C (or justify clipping in log space with appropriate bounds). Use consistent notation throughout.

3. Engineered feature definitions (Sec. 2.2, Sec. 3.2–3.2.3) are helpful but not fully specified for reproduction: missing scale-factor bin edges, MPB interpolation/nearest-snapshot rule at fixed a, exact formulae for mass-weighted means and formation_time_half_mass, and the major-merger definition (mass ratio and time window).

   Recommendation: Add a compact feature-definition table (main text or appendix) listing all 35 engineered features with precise formulas, bin edges, interpolation rules, thresholds, and units. This will make Sec. 3.2–3.2.3 fully actionable and supports the baseline experiments.

4. Inference evaluation could be more diagnostic: Sec. 3.5–3.6 focuses on overall RMSE (as fraction of prior range) and one coverage level, but does not show how bias/variance/coverage depend on true parameter values or on tree properties, limiting insight into when/why σ_8 fails.

Recommendation: Add granular plots: posterior-mean error vs true Ω_m and σ_8, CI width vs true parameters, and binned coverage vs true parameters. Optionally stratify by simple tree properties (final mass, node count, MPB depth) or key engineered features to localize failure modes.

5. Embedding-quality assessment in Sec. 3.4 relies largely on PCA/t-SNE visualizations, which are qualitative and sensitive to hyperparameters; the manuscript does not report t-SNE settings or provide quantitative embedding diagnostics.

   Recommendation: Report t-SNE hyperparameters (perplexity, learning rate, seed) and add quantitative measures (e.g., linear/MLP prediction of Ω_m, σ_8 from embeddings on validation; kNN classification by cosmology ID; or clustering metrics). Provide training/validation loss curves over epochs to justify convergence.

6. Several figures use internal variable names, lack units/definitions, or have ambiguous histogram normalization ("Frequency" vs "Density"), reducing interpretability and reproducibility (e.g., Figures 1, 5–10, 13–18).

   Recommendation: Relabel axes with human-readable names and units, define all plotted variables in captions (especially engineered features), and standardize histogram normalization. Where correlation claims are made in captions/text, include quantitative correlation values (Pearson/Spearman) or revise wording.

7. Dataset arithmetic inconsistency flagged in Sec. 3.1 narrative vs checks ($\mathbf{40 \times 25 = 1000}$ stated, but an executed check indicates a mismatch). Even if this is an artifact, it undermines trust in the data accounting.

   Recommendation: Re-audit and correct all dataset counts (cosmologies, trees per cosmology, total trees, train/val/test sizes) and ensure they are consistent across Sec. 2.1 and Sec. 3.1 (and any code-based checks). Include a small table summarizing split sizes.

8. References and positioning: some citations appear incomplete/inconsistent (e.g., dangling entries), and some tooling/framework mentions (e.g., FLORAH) are not clearly connected to what is actually used.

   Recommendation: Audit and fix the bibliography (authors/years/titles), remove or justify tangential citations, and ensure any mentioned frameworks are either used and documented or removed from the narrative.

9. Compute/resource reporting is absent, making it hard to judge scalability and practical applicability.

   Recommendation: Add a short paragraph (e.g., Sec. 4.4) reporting hardware, training time for the GNN and SNPE, and approximate dataset generation cost (if applicable).

Very minor issues

1. Typographical and formatting issues (duplicated sentences, broken line-wrapped citations, minor misspellings such as “unormized”, inconsistent spacing around symbols) reduce polish (Sec. 1, Sec. 2.2, Sec. 3.3–3.4, References).

   Recommendation: Perform a careful proofreading pass to remove duplicates, fix typos, repair citation formatting/line breaks, and standardize equation/punctuation spacing.

2. Section heading formatting inconsistencies (e.g., stray Markdown hash symbols in headings around Sec. 3.2–3.3) and inconsistent notation ($V_{\rm max}$ vs $V_{\rm max}$; snake_case feature names in plots) can confuse readers.

   Recommendation: Standardize heading styles in the final typeset version and unify notation across text and figures. Consider a short notation/abbreviation table for key variables and features.

3. Figure legibility issues (resolution, font sizes, jagged binning) and missing plot metadata (N, bin width, normalization) hinder readability and reproducibility.

   Recommendation: Export figures in vector/high-resolution formats, increase font sizes, use principled binning (e.g., Freedman–Diaconis), and include N/binning/normalization details in captions.

4. Logarithms of dimensional quantities are used without explicitly stating reference units (e.g., $\log_{10}$ mass, $\log_{10}$ $V_{\rm max}$), which is a minor but avoidable ambiguity.

   Recommendation: State the units before applying logs (e.g., $\log_{10}(M/M_\odot)$, $\log_{10}(V_{\rm max}/(\mathrm{km\ s^{-1}}))$) in Sec. 2.2 and in relevant figure captions.

Key statements and references

- ⚠ **Dark matter halo merger trees encode valuable information about the underlying cosmology and have been proposed as a rich data source for likelihood-free inference of cosmological parameters, but their use is complicated by non-linear relationships between tree structure, halo assembly histories, and cosmology, as well as by assembly bias, i.e. correlations between halo properties and formation history at fixed mass and cosmology (Benson et al., 2012; Jiang and van den Bosch, 2013; Ángel Chandro-Gómez et al., 2025).**
- *Reference(s):* Benson et al., 2012, Jiang and van den Bosch, 2013, Ángel Chandro-Gómez et al., 2025
- *Justification:* Benson et al., 2012 and Jiang and van den Bosch, 2013 support that merger trees depend on cosmology/dark matter physics (e.g., differences in mass and progenitor mass functions across cosmologies, and EPS-based trees enabling exploration across cosmological parameters). They also note complications/degeneracies

(e.g., similar mass accretion histories despite different cosmologies, algorithmic/systematic uncertainties). However, neither paper discusses likelihood-free inference nor assembly bias (correlations between halo properties and formation history at fixed mass), so those parts are not supported.

- ⚠ **The node-level features used to represent each merger tree—$\log_{10}$ (mass), $\log_{10}$ (concentration), $\log_{10}$ (V_{max}), and scale factor—as well as the edge_index connectivity, follow established practices for constructing and characterizing dark matter halo merger trees from N-body simulations (Parkinson et al., 2007; Jiang and van den Bosch, 2013; Benson et al., 2012).**

- *Reference(s):* Parkinson et al., 2007, Jiang and van den Bosch, 2013, Benson et al., 2012

- *Justification:* The cited works establish merger trees built around halo mass and time/redshift (e.g., conditional/progenitor mass functions, MAHs) and progenitor–descendant connectivity (Parkinson et al., 2007; Jiang and van den Bosch, 2013; Benson et al., 2012). However, they do not present $\log_{10}$ (concentration) or $\log_{10}$ (V_{max}) as standard node features; V_{max} is not discussed, and concentration appears only incidentally for mass-definition conversion (Benson et al., 2012). None of the papers describe an 'edge_index' data structure. Thus, mass/time and connectivity are consistent with practice, but concentration, V_{max}, and the specific edge_index representation are not supported by these references.

- ✖ **Global merger-tree features such as total mass in scale-factor bins, mass-weighted mean concentration and V_{max}, main progenitor branch properties, formation time, number of major mergers, assembly-bias proxies, and structural metrics are motivated by previous work showing that these quantities trace halo growth and assembly histories and can correlate with cosmological parameters and assembly bias (Bansal et al., 2023; Burgarella et al., 2022; Pu et al., 2025).**

- *Reference(s):* Bansal et al., 2023, Burgarella et al., 2022, Pu et al., 2025

- *Justification:* Bansal et al. (2023) use merger trees to study SMBH–halo mass evolution but do not discuss global merger-tree features like concentration or V_{max}, nor correlations with cosmological parameters or assembly bias. Burgarella et al. (2022) analyzes galaxy SEDs and emission lines, unrelated to halo merger-tree metrics or assembly bias. Pu et al. (2025) examines progenitor diversity and radial contributions in simulated stellar halos, focusing on assembly histories and progenitor counts, but not on the listed global features or correlations with cosmological parameters or assembly bias. Therefore, the claimed motivation and correlations are not supported by these papers.

- ✔ **The assembly-bias–mitigating data augmentation strategy is grounded in empirical and theoretical studies of the halo mass–concentration (M–C) relation, which is typically modeled as a relation of the form $\log_{10}(\text{concentration}) = \text{A} \cdot \log_{10}(\text{mass}) + \text{B}$ with a non-negligible intrinsic scatter, as inferred from observations and simulations of dark matter halos (Biviano et al., 2017; Gilman et al., 2019; Gu et al., 2022).**
- *Reference(s):* Biviano et al., 2017, Gilman et al., 2019, Gu et al., 2022
- *Justification:* Supported. The cited works treat the halo mass–concentration relation as a power-law in log space with scatter, based on observations and simulations. Biviano et al., 2017 fit a log-linear relation for clusters (e.g., Eq. 13: $\log c_{200} = (1.0 \pm 1.4) - (0.03 \pm 0.09) \cdot \log M_{200}$) and find a lognormal scatter (~0.22 dex; consistent with simulated intrinsic scatter ~0.2–0.3). Gilman et al., 2019 constrain $c(M)$ observationally via strong lensing with a power-law parameterization (Eq. 4) and include a non-zero scatter of 0.1 dex in concentration. Gu et al., 2022 fit a log-linear c–M relation (Eq. 5: $\log c_{200} = \beta + \alpha \cdot \log M_{200}$) across $10^{11.6}$–$10^{14.1}$ $M_\odot$ and compare with simulations, noting observational scatter. Thus, the typical modeling form and non-negligible scatter are grounded in empirical and theoretical studies as referenced.
- ✖ **The contrastive learning setup employs the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss, a supervised contrastive objective in which trees from the same cosmology (including their augmentations) form positive pairs and trees from different cosmologies form negative pairs, following recent applications of NT-Xent-style contrastive learning in astrophysical and cosmological contexts (Gondhalekar et al., 2024; Wilkinson et al., 2025; Perez et al., 2025).**
- *Reference(s):* Gondhalekar et al., 2024, Wilkinson et al., 2025, Perez et al., 2025
- *Justification:* Gondhalekar et al., 2024 and Perez et al., 2025 use SimCLR with the NT-Xent loss in an unsupervised setup where positives are two augmentations of the same image and all other images are negatives; they do not use supervised pairing or any cosmology/'trees' labels. Wilkinson et al., 2025 includes both unsupervised SimCLR (NT-Xent) and a supervised contrastive loss, but not for cosmology or trees. None of the cited works describe a supervised contrastive scheme where positives are from the same cosmology and negatives from different cosmologies.
- ✔ **Sequential Neural Posterior Estimation (SNPE), implemented via the sbi Python package, is used as the likelihood-free inference engine to learn a neural density estimator (e.g., a normalizing flow) that approximates the posterior $p(\text{cosmology} \mid \text{embedding})$, building on prior work that has successfully applied SNPE to cosmological and astrophysical inference problems (Zhang et al., 2023; Erickson et al., 2024; Kosiba et al., 2024).**
- *Reference(s):* Zhang et al., 2023, Erickson et al., 2024, Kosiba et al., 2024

- *Justification:* Kosiba et al., 2024 explicitly use Sequential Neural Posterior Estimation with the sbi Python package for likelihood-free inference, training a neural density estimator (an MDN) to approximate the posterior $p(\Omega_m, \sigma_8 \mid \textbf{compressed XOD embedding})$ (Section 3; Fig. 6). Erickson et al., 2024 apply NPE/SNPE to strong-lens mass modeling, learning approximate posteriors from images via a CNN and sequential proposals (Sections 3.1–3.2), demonstrating successful astrophysical application on real HST data. Zhang et al., 2023 introduces the nbi framework for NPE/SNPE in astronomy, employing normalizing flows (MAF) as neural density estimators and discussing sbi as a related toolkit (Sections 2–3). Together, these support the claim that SNPE (via sbi) is used to learn a neural density estimator approximating $p(\textbf{cosmology} \mid \textbf{embedding})$, building on prior SNPE successes in cosmology and astrophysics.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The manuscript is primarily methodological/ML-focused and contains limited explicit mathematics. The main explicit analytic elements are (i) feature normalization, (ii) Pearson correlation usage, (iii) a log–log linear mass–concentration relation with a stated scatter, and (iv) a probabilistic description of concentration resampling for data augmentation. The central learning objectives (NT-Xent contrastive loss and SNPE posterior learning) are described verbally without explicit equations, limiting symbolic verification.

Checked items

1. ✔ **Node-feature tensor definition** (Sec. 2.1, p.2)
    - **Claim:** Each merger tree has node features $\boldsymbol{x}$ of shape $(N_{\text{nodes}}, 4)$ consisting of $\log_{10}$(mass), $\log_{10}$(concentration), $\log_{10}(V_{\text{max}})$, and scale factor a.
    - **Checks:** symbol/definition consistency, dimensional sanity
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Each node corresponds to a halo at some snapshot/scale factor., $\log_{10}$ is applied to mass, concentration, and V_{max} consistently.
    - **Notes:** The four features are listed consistently across Methods and Results. Strict dimensional consistency for logs would require explicit reference units, but within-paper usage is consistent.
2. ⚠ **Target tensor y and cosmology identity for positives** (Sec. 2.1 and Sec. 2.4, p.2–p.3)

- **Claim:** Graph-level targets y contain (Ω_m, σ_8), and positive pairs in contrastive learning are trees with the same original y (same cosmology).
- **Checks:** symbol/definition consistency, logical consistency
- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
- **Assumptions/inputs:** There are 40 discrete cosmologies, each repeated across multiple trees., Equality of cosmology is well-defined in training (likely via a cosmology ID).
- **Notes:** The paper says features and targets are normalized globally. If y is normalized to floats, exact equality comparisons for 'same y' are not robust unless a discrete cosmology label is used. The manuscript does not explicitly define the equivalence relation (ID vs float equality).

3. ✔ **Normalization formula (z-scoring)** (Sec. 2.1, p.2 and Sec. 3.1, p.4)

    - **Claim:** Each feature/parameter is normalized by subtracting its global mean and dividing by its global standard deviation.
    - **Checks:** algebraic form, definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Means/standard deviations are computed over the intended dataset scope.
    - **Notes:** The described transformation is algebraically standard and consistent with later statements that normalized quantities have approximately zero mean and unit variance.

4. ✔ **Pearson correlation usage** (Sec. 3.2, p.5–p.6)

    - **Claim:** Pearson correlations are computed between unnormalized engineered features and unnormalized cosmological parameters (Ω_m, σ_8).
    - **Checks:** definition consistency, sanity constraints
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Pearson r is computed in the usual centered-covariance normalized-by-std manner.
    - **Notes:** No explicit formula is given, but nothing in the text contradicts standard Pearson correlation properties (e.g., r in $[-1, 1]$).

5. ✔ **Mass–concentration relation functional form** (Sec. 2.3, p.3 and Sec. 3.3, p.10)

    - **Claim:** A log–log linear relation is fitted: $\log_{10}(C) = A \log_{10}(M) + B$ (example fitted values given later).
    - **Checks:** algebraic form, notation consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** C denotes concentration (positive)., M denotes mass (positive).

- **Notes:** The relation is syntactically consistent and matches the earlier ' $A \log_{10}(\text{mass}) + \text{B}$ ' template.

6. ✔ **Reported fitted relation and scatter symbols** (Sec. 3.3, p.10)
    - **Claim:** The fit is $\log_{10}(C) = -0.0304 \times \log_{10}(M) + 1.071$ and scatter is $\sigma_{\log C} = 0.3603$ (std dev around the median relation).
    - **Checks:** symbol consistency, distribution parameter coherence
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** $\sigma_{\log C}$ is a standard deviation in log10 space., Scatter is computed around the fitted line.
    - **Notes:** The scatter symbol $\sigma_{\log C}$ is consistently referenced as being in $\log_{10}(C)$. The paper does not define whether scatter is conditional on mass bins or global; that affects meaning but not internal algebra.
7. ✔ **Augmentation sampling distribution definition** (Sec. 2.3, p.3 and Sec. 3.3, p.10)
    - **Claim:** For each node, a new $\log_{10}$(concentration) is sampled from a Gaussian centered at the predicted median $\log_{10}(C)$ with std dev equal to observed scatter $\sigma_{\log C}$.
    - **Checks:** probabilistic definition consistency
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Sampling is performed in log10 space., Gaussian is over $\log_{10}(C)$.
    - **Notes:** As a distributional statement in log space, this is coherent given the fitted log–log relation and $\sigma_{\log C}$.
8. ✖ **Clipping bounds for resampled concentrations** (Sec. 3.3, p.10 (also related to Sec. 2.3, p.3))
    - **Claim:** Resampled concentrations are clipped to physically plausible bounds observed in the original dataset (0.0002 to 3.767).
    - **Checks:** notation/unit consistency, distribution support consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** The model samples $\log_{10}(C)$ but bounds are stated without specifying whether they apply to C or $\log_{10}(C)$.
    - **Notes:** The text explicitly says $\log_{10}$(Concentration) is resampled, but the clipping bounds (0.0002, 3.767) strongly suggest linear C bounds. If applied to $\log_{10}(C)$, these imply $\log_{10}(C) \geq 0.0002$ (i.e., $C \geq\sim 1$) and allow $\log_{10}(C)$ up to 3.767 (i.e., C up to thousands), which is inconsistent with typical 'physically plausible' wording and with the earlier use of $\log_{10}$(con-

centration) as the stored feature. The manuscript must clarify the scale of the bounds and the order of operations (sample in log, exponentiate, clip, then log again vs clip in log space).

9. ✖ **Number of augmented copies per tree** (Sec. 2.3, p.3 vs Sec. 3.3–3.4, p.10)
    - **Claim:** Methods: K augmented copies per tree (e.g., $K = 2$–3). Results: one augmented copy per tree; training set size reported as 700 original + 700 augmented.
    - **Checks:** definition consistency, count/constraint consistency (symbolic)
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** Training set contains 700 original trees (as stated in Sec. 3.1).
    - **Notes:** The augmentation multiplicity parameter K is inconsistent between sections. The reported dataset size (1400 trees) implies $K = 1$, contradicting the earlier '$K = 2$–3' description.

10. ✖ **Scale-factor conditional M–C relation vs global fit** (Sec. 2.3, p.3 vs Sec. 3.3, p.10)
    - **Claim:** Methods describe using an M–C relation and scatter derived for a node's scale-factor bin; Results describe a single fit using all halos across masses and epochs.
    - **Checks:** conditional definition consistency
    - **Verdict:** FAIL; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Scale-factor bins exist and are used elsewhere for engineered features.
    - **Notes:** These are different mathematical objects: $p(\log_{10} C | \log_{10} M, \mathrm{bin(a)})$ vs $p(\log_{10} C | \log_{10} M)$ pooled over a. The text should be aligned to one definition because it changes the augmentation distribution and thus the contrastive invariances being trained.

11. ⚠ **NT-Xent / supervised contrastive loss definition** (Sec. 2.4, p.3)
    - **Claim:** The model uses NT-Xent (temperature-scaled cross-entropy) to pull together same-cosmology embeddings and push apart different-cosmology embeddings.
    - **Checks:** derivation/verifiability, symbol definition completeness
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** A specific loss variant is implemented (standard NT-Xent vs supervised contrastive).
    - **Notes:** No explicit mathematical form is provided (no equation for the numerator/denominator, similarity measure, normalization, or temperature). Without the loss definition, symbolic verification of the paper's central optimization objective is not possible.

12. ✔ **Posterior notation for SNPE** (Sec. 2.5, p.4)
    - **Claim:** SNPE approximates the posterior $p(\mathbf{cosmology} \mid \mathbf{embedding})$ for (Ω_m, σ_8) given embeddings.
    - **Checks:** notation consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Cosmology refers to the 2D parameter vector (Ω_m, σ_8).
    - **Notes:** The conditional notation is consistent with later statements about sampling from the inferred posterior. No further equations are given to audit.

Limitations

- The provided PDF text contains almost no explicit equations beyond the log–log mass–concentration fit; key objectives (NT-Xent loss, SNPE training objective, normalizing flow density) are described verbally without mathematical specification, preventing full symbolic verification.
- No explicit definitions are provided for several engineered global features (exact binning rules, mass-weighting formulas, branching-factor definition), so those quantities cannot be audited for algebraic correctness—only for consistency of description.
- This audit does not evaluate numerical values, reported performance metrics, plots, or empirical claims; it only checks internal symbolic/analytic consistency of definitions and stated formulas.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

11 of 12 automated numeric consistency checks passed. One key internal-consistency check related to the stated dataset size vs (cosmologies × trees per cosmology) failed, while split counts, augmentation totals, RMSE-from-MSE conversions, parameter-range spans, and RMSE-as-percent-of-range statements were numerically consistent within stated tolerances.

Checked items

1. ✖ **C1_dataset_trees_per_cosmology** (Page 4, Sec. 3.1 (Dataset Characteristics))
    - **Claim:** "These trees originate from 40 unique cosmological parameter pairs (Ω_m, σ_8), with 25 trees per cosmology" and earlier "dataset of 1000 merger trees from 40 unique cosmologies."
    - **Checks:** parts_to_total
    - **Verdict:** FAIL
    - **Notes:** Exact integer consistency expected.

2. ✔ **C2_split_counts_sum** (Page 4-5, Sec. 3.1 (Normalization and Dataset Splitting))
    - **Claim:** "The resulting split comprises 700 trees (28 cosmologies) for training, 150 trees (6 cosmologies) for validation, and 150 trees (6 cosmologies) for testing."
    - **Checks:** parts_to_total
    - **Verdict:** PASS
    - **Notes:** Exact integer sum expected.
3. ✔ **C3_split_cosmologies_sum** (Page 4-5, Sec. 3.1 (Dataset Characteristics and split description))
    - **Claim:** Split by unique cosmologies: "28 cosmologies for training, 6 for validation, and 6 for testing" out of "40 unique cosmologies."
    - **Checks:** parts_to_total
    - **Verdict:** PASS
    - **Notes:** Exact integer sum expected.
4. ✔ **C4_trees_per_cosmology_within_splits** (Page 4-5, Sec. 3.1 (Dataset Characteristics and split description))
    - **Claim:** Given "25 trees per cosmology", the split counts should match: training 28 cosmologies→700 trees; validation 6→150; test 6→150.
    - **Checks:** derived_count_check
    - **Verdict:** PASS
    - **Notes:** Exact integer consistency expected.
5. ✔ **C5_augmented_training_set_size** (Page 10, Sec. 3.4 (Contrastive Embedding Learning))
    - **Claim:** "The training utilizes the augmented training set (1400 trees: 700 original + 700 augmented)."
    - **Checks:** parts_to_total
    - **Verdict:** PASS
    - **Notes:** Exact integer sum expected.
6. ✔ **C6_rmse_from_mse_omega_m** (Page 12, Sec. 3.5 (Quantitative Performance Metrics))
    - **Claim:** "MSE … Ω_m: 0.00069 … Taking the square root, the RMSE for Ω_m is approximately 0.026."
    - **Checks:** sqrt_consistency
    - **Verdict:** PASS
    - **Notes:** Reported as approximate; allow modest rounding.

7. ✔ **C7__rmse__from__mse__sigma8** (Page 12, Sec. 3.5 (Quantitative Performance Metrics))
    - **Claim:** "MSE … σ_8: 0.006412 … For σ_8, the RMSE is approximately 0.080."
    - **Checks:** sqrt_consistency
    - **Verdict:** PASS
    - **Notes:** Reported as approximate; should match closely.
8. ✔ **C8__omega__m__range__span** (Page 4, Sec. 3.1 + Page 12, Sec. 3.5)
    - **Claim:** Ω_m varies from 0.103 to 0.4734 (span 0.3704).
    - **Checks:** range_span
    - **Verdict:** PASS
    - **Notes:** Exact subtraction to 4 decimals should match.
9. ✔ **C9__sigma8__range__span** (Page 4, Sec. 3.1 + Page 12, Sec. 3.5)
    - **Claim:** σ_8 varies from 0.603 to 0.9918 (span 0.3888).
    - **Checks:** range_span
    - **Verdict:** PASS
    - **Notes:** Exact subtraction to 4 decimals should match.
10. ✔ **C10__rmse__fraction__of__range__omega__m** (Page 12, Sec. 3.5)
    - **Claim:** "RMSE for Ω_m … about 7.0% of the parameter range." using RMSE $\approx \mathbf{0.026}$ and range span 0.3704.
    - **Checks:** percentage_of_range
    - **Verdict:** PASS
    - **Notes:** 'About' suggests rounding; allow a few tenths of a percent.
11. ✔ **C11__rmse__fraction__of__range__sigma8** (Page 12, Sec. 3.5)
    - **Claim:** "RMSE … about 20.6% of its range." using RMSE$\approx \mathbf{0.080}$ and σ_8 span 0.3888.
    - **Checks:** percentage_of_range
    - **Verdict:** PASS
    - **Notes:** Allow for rounding of RMSE and percent.
12. ✔ **C12__mc__relation__scatter__vs__clip__bounds** (Page 10, Sec. 3.3 (Data Augmentation))
    - **Claim:** Resampled $\mathbf{log}_{10}$(Concentration) uses $\sigma_{\log C} = \mathbf{0.3603}$, then is clipped to bounds 0.0002 to 3.767; check bounds order and positivity.
    - **Checks:** sanity_check_bounds
    - **Verdict:** PASS
    - **Notes:** Pure logical check.

Limitations

- Audit based only on provided parsed PDF text; no underlying simulation/tree data, engineered-feature matrices, posterior samples, or training logs are available to recompute many reported statistics.
- Plot-derived numerical verification (e.g., from histogram/heatmap axes or colorbars) is excluded because it would require reading pixels/graphics rather than explicit numbers in text.