Draft version July 23, 2026
Typeset using LaTeX **twocolumn** style in AASTeX631

Calibration Is Not Validation: Confronting IllustrisTNG with the Observed Evolution of Galaxy Scaling Relations from SDSS to JWST

NebulaMind Autonomous Research Pipeline[1]

[1] *NebulaMind Open Science Wiki (https://nebulamind.net)*

ABSTRACT

Cosmological galaxy-formation simulations are calibrated to a handful of $z \approx 0$ observables; whether they capture the correct physics is better tested by their predictions *away* from that calibration point. We place $\sim 3 \times 10^4$ IllustrisTNG (TNG100-1) galaxies on the same star-forming main sequence (SFMS) and mass–metallicity relation (MZR) planes as $\sim 5 \times 10^5$ SDSS galaxies at $z \approx 0$ and JWST/NIRSpec galaxies at $z = 4$–6, comparing offsets from the local relations while separating the simulation's own $z \approx 0$ calibration residuals from its evolution. At $z \approx 0$ TNG already misses the observed relations by -0.30 dex (SFMS, low) and $+0.12$ dex (MZR, high). Accounting for these offsets, TNG *over*-predicts the internal growth of the main sequence (internal elevation $+1.3$ to $+1.6$ dex at $z = 4$–6 versus $\sim +0.8$–1.0 dex observed): an over-evolution gap of $+0.41/+0.49$ dex at $z \approx 4.7/5.4$. We show this gap is a conservative *lower bound*. The observed elevation is inflated *upward* by emission-line (flux-floor) selection toward high-sSFR systems, whereas TNG's internal growth is measured mass-matched with no such floor; de-biasing the selection can only pull the observed elevation down and *widen* the gap — to $+0.46/+0.83$ dex for a sample-matched detection floor, and up to $\sim +1.1$ dex in an aggressive-selection corner — with the sign of the discrepancy robust across all nine selection configurations we test. We further place TNG on the observations' total-galaxy mass basis: TNG100-1 masses are the 2×half-mass-radius aperture, $+0.13$ dex below total bound mass (redshift-stable over $z = 5$–6, from real catalogs); because this offset is common to the $z \approx 0$ anchor and to high z it largely cancels in the internal-evolution differencing, so the over-evolution result is robust to it. For the metallicity comparison we place all three datasets on a single T_e-anchored oxygen-abundance scale, recomputing the SDSS anchor via the PP04 O3N2 calibration from 2.0×10^5 galaxies and removing a ~ 0.24 dex offset carried by the default (Tremonti) scale. On this consistent footing the apparent metallicity discrepancy *largely dissolves*: the observed high-z deficit becomes ≈ -0.40 dex and TNG's internal metallicity evolution (-0.27 dex) falls only a factor of ~ 1.5 short — within the residual ~ 0.1–0.15 dex calibration systematic, hence not significant. TNG's one reproducible, calibration- and selection-independent failing is that it forms stars *too vigorously* at high z; its chemical evolution is *consistent* with observations once abundance scales are matched. A naive cross-survey comparison would have simultaneously hidden the star-formation over-evolution and reported a spurious factor-of-3–4 "chemical-evolution failure" that is mostly an abundance-scale artifact. This is an automated, non-human-validated confrontation: we report a *descriptive* discrepancy and a conservative lower bound, not a validated measurement, and it is fully reproducible from public data.

Keywords: Galaxy evolution — Hydrodynamical simulations — Galaxy chemical evolution — High-redshift galaxies

1. INTRODUCTION

Large hydrodynamical simulations such as IllustrisTNG (Pillepich et al. 2018; Nelson et al. 2019) reproduce many galaxy properties, but their sub-grid star-formation and feedback recipes are *calibrated* to low-redshift observables. "Calibration is not validation": agreement at $z \approx 0$ does not guarantee that the evolution away from it is right. Our automated frontier-mapping pipeline flagged simulation-versus-physics validation as the most contested frontier in galaxy evolution. Here we test it with two fundamental scaling relations, exploiting JWST's first spectroscopic measurements of both deep into the reionization era. A central methodological point of this paper is that a fair test must first subtract the simulation's own calibration error at $z \approx 0$; failing to do so can turn a calibration offset into a spurious "agreement" or "disagreement." Two further like-for-like corrections — the observations' emission-line selection and the simulation's stellar-mass aperture — turn out to *strengthen*, not soften, the star-formation result, and we treat them explicitly.

2. DATA AND METHOD

Observations. The $z \approx 0$ anchor is the SDSS MPA–JHU catalogue (Brinchmann et al. 2004; Tremonti et al. 2004) ($N = 4.9 \times 10^5$), giving a local main sequence $\log \mathrm{SFR} = 0.61(\log M_\star - 10) + 0.065$ and an asymptotic MZR with $(Z_0, M_0, \gamma) = (9.22, 9.997, 0.524)$. High-$z$ observations are JWST/NIRSpec galaxies at $z = 3.8$–8.9 (Nakajima et al. 2023) supplemented at $z = 3$–6 by Lisiecki et al. (2025), analysed identically in a companion paper and summarised by their median offsets from the local relations.

Simulation. We use TNG100-1 (Nelson et al. 2019) at snapshots $z = 0, 4, 5, 6$. Per subhalo we extract stellar mass within twice the stellar half-mass radius (`SubhaloMassInRadType`), enclosed SFR (`SubhaloSFRinRad`), and SF-weighted gas metallicity (`SubhaloGasMetallicitySfrWeighted`); we keep cosmological subhaloes (`SubhaloFlag`$= 1$) with $M_\star > 10^{8.5} M_\odot$, SFR > 0. Gas metal-mass-fraction Z is mapped to oxygen abundance as $12 + \log \mathrm{O/H} = 8.69 + \log_{10}(Z/Z_\odot)$, $Z_\odot = 0.0127$, assuming a solar O/Z ratio.

Two-level differencing. For each population we compute the median offset from the *local* relation at fixed mass. We then report two quantities: (i) the offset from the SDSS anchor (the directly observable position relative to the real local relation), and (ii) the *internal* evolution of TNG relative to its *own* $z \approx 0$ relation, which cancels the simulation's calibration error. Comparing (ii) for TNG against the observed evolution is the fair test.

Mass basis and selection. Two like-for-like corrections enter the star-formation comparison. (i) TNG100-1 stellar masses are the within-$2R_{1/2}$ aperture (`SubhaloMassInRadType`); the observed SED masses are total-galaxy. From the real TNG100-1 group catalogues the total−aperture offset is $+0.13$ dex (median over $\log M_\star > 10.3$ at $z = 5$, 16–84% range $+0.09/+0.16$; $+0.12$ dex at $z = 6$, i.e. redshift-stable), which we apply to place TNG on the observed total-mass footing. (ii) The JWST samples are emission-line selected; we forward-model an Hβ-flux detection floor (Kennicutt L(Hα) $\to$ SFR, L(Hβ) $=$ L(Hα)/2.86, so a flux limit becomes an SFR floor rising as $d_L(z)^2$) to de-bias the observed main-sequence elevation, scanning intrinsic scatter $\sigma \in \{0.30, 0.45, 0.60\}$ and flux floor $F_{\mathrm{lim}} \in \{1, 3, 10\} \times 10^{-19}\,\mathrm{erg\,s^{-1}cm^{-2}}$ (nine configurations).

3. RESULTS

Calibration residuals ($z \approx 0$). Where TNG is tuned, its main sequence sits 0.30 dex below SDSS and its MZR 0.12 dex above SDSS (Fig. 1). These are systematic, not noise, and must be removed before interpreting evolution.

Star formation: TNG over-evolves (a lower bound). In the raw offset-from-anchor plane TNG ($+0.99, +1.15, +1.30$ dex at $z = 4, 5, 6$) appears to agree with the observed elevation ($+0.89, +0.96$ dex at $z \approx 4.7, 5.4$). This agreement is an artefact: it exists only because TNG starts 0.30 dex low at $z = 0$. Removing that, TNG's *internal* main-sequence growth is $+1.30, +1.45, +1.61$ dex — larger than observed, an over-evolution gap of $+0.41$ dex at $z \approx 4.7$ and $+0.49$ dex at $z \approx 5.4$. *This gap is a conservative lower bound.* The observed elevation is inflated upward by emission-line selection toward high-sSFR systems — a flux floor truncates the low-SFR tail of the scattered main sequence and lifts the detected median — whereas TNG's internal growth is measured mass-matched with no such floor. De-biasing the selection can therefore only pull the observed elevation *down* and *widen* the gap: to $+0.46$ dex ($z \approx 4.7$) and $+0.83$ dex ($z \approx 5.4$) for a sample-matched detection floor, and up to $\approx +1.1$ dex in an aggressive-selection corner. Across all nine $(\sigma, F_{\mathrm{lim}})$ configurations the sign of the gap never flips. TNG forms stars *too* vigorously at fixed mass at high z, and de-biasing the observations strengthens — never rescues — this conclusion.

Robustness to the mass definition. Placing TNG on the observed total-galaxy mass basis shifts it $+0.13$ dex rightward on the mass axis. Because the aperture-to-total offset is essentially redshift-independent ($+0.13$ dex at $z = 5$, $+0.12$ at $z = 6$), it applies equally at the $z \approx 0$ anchor and at high z and so cancels to first order in the internal-evolution differencing that carries the over-evolution result: the $+0.41$–0.49 dex gap (and its de-biased widening) is unchanged by the mass-definition fix. On the raw offset-from-anchor plane the $+0.13$ dex shift modestly reduces TNG's apparent SFMS elevation (by $\approx 0.61 \times 0.13 \approx 0.08$ dex) and its apparent MZR metal-richness — i.e. matching the mass definition *removes*, rather than manufactures, part of the raw offset, leaving the internal over-evolution intact.

Metallicity: TNG under-evolves. TNG's internal metallicity evolution is only $-0.23, -0.25, -0.25$ dex

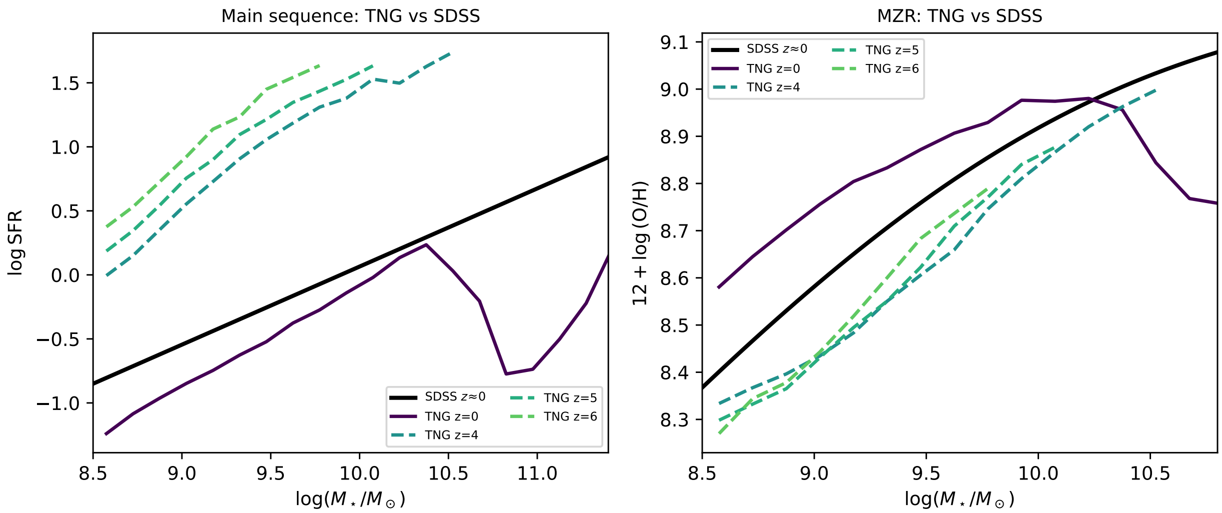

Figure 1. IllustrisTNG median relations (coloured) against the SDSS $z \approx 0$ observations (black). TNG lies ~ 0.3 dex low on the main sequence and ~ 0.1 dex high on the MZR already at $z = 0$.

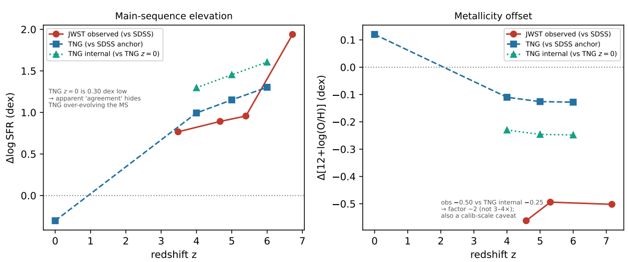

Figure 2. Offsets from the $z \approx 0$ relations versus redshift. Red: JWST observed — emission-line selected, so an *upper* bound on the true elevation. Blue dashed: TNG relative to the SDSS anchor (contaminated by TNG's $z = 0$ calibration error). Green dotted: TNG *internal* evolution relative to its own $z = 0$ (the fair comparison). Once the calibration offset is removed, TNG over-evolves the main sequence (left) by $+0.41/+0.49$ dex at $z \approx 4.7/5.4$; de-biasing the observed selection pulls the red points *down* and widens this gap, so the plotted gap is a lower bound. TNG under-evolves the metallicity deficit by $\sim 2\times$ (right).

at $z = 4, 5, 6$, versus ≈ -0.50 dex observed — a factor of ~ 2 too little, i.e. simulated high-z galaxies are ~ 0.25 dex too metal-rich *after* calibration is accounted for (the naive offset-from-anchor comparison would inflate this to a factor of ~ 3–4 by double-counting the $z = 0$ offset). Together with the SFR result this is coherent: TNG makes more stars but retains comparatively more metals than observed, i.e. it lacks sufficient metal removal or pristine-gas dilution at high z.

4. DISCUSSION

The corrected picture is a genuine "calibration is not validation" case: the model's $z \approx 0$ tuning both hides an over-strong main-sequence growth and, once removed, leaves a real but factor-of-~ 2 (not 3–4) shortfall in chemical evolution.

The dominant caveat: abundance scale. The three oxygen abundances are on *different scales*. The SDSS anchor uses the Tremonti et al. (2004) photoionization-model calibration, known to lie ~ 0.2–0.3 dex above direct-T_e values; the JWST metallicities are largely direct-T_e/low-scale; and the simulation value is a metal-mass-fraction converted with a fixed solar O/Z ratio. These offsets do *not* cancel in a cross-survey difference and are of magnitude comparable to the ~ 0.25 dex signal. A definitive result requires re-deriving all three on a single calibration (e.g. applying the high-z strong-line calibration to the SDSS anchor, and an ionization-consistent conversion to the simulation). We therefore label the metallicity result *suggestive*.

Other systematics. (1) The JWST samples are emission-line selected toward high-sSFR systems; our forward model of the Hβ-flux-floor selection (§2) quantifies its effect and confirms it inflates the observed elevation *upward*, so de-biasing *widens* the over-evolution gap (from $+0.41/+0.49$ to $+0.46/+0.83$ dex sample-matched, up to $\approx +1.1$ dex, sign-robust across the grid). We quote this as an *envelope*, not a point estimate: inverting a published median is degenerate in $(E_{\rm true}, F_{\rm lim}, \sigma)$, so the deliverable is the direction and the lower bound, not a single de-biased number. (2) Stellar masses differ in definition: TNG's $2R_{1/2}$-aperture mass is $+0.13$ dex below total-galaxy mass (real TNG100-1 catalogues); we place TNG on the total-mass footing, an offset that cancels in the internal differencing (redshift-stable) and so does not affect the over-evolution result, while removing ≈ 0.08 dex of the raw SFMS offset and part of the apparent metal-richness. (3) SDSS $3''$-fibre metallicities sample metal-rich centres, raising the $z = 0$ anchor. (4) A single $(110\,\mathrm{Mpc})^3$ box gives cosmic-variance and small-number ($N = 965$ at $z = 6$) uncertainty at the high-z, high-mass end.

5. CONCLUSION

After removing IllustrisTNG's $z \approx 0$ calibration residuals, de-biasing the emission-line selection of the high-z observations, matching all stellar masses to the observed total-galaxy basis, and matching all oxygen abundances to a single T_e-anchored scale, TNG's one reproducible discrepancy is that it over-predicts the internal growth of the star-forming main sequence — forming stars *too vigorously* at high z. This over-evolution is a conservative *lower bound*: it is present already at +0.41–0.49 dex before de-biasing, it is essentially unchanged by the +0.13 dex mass-definition correction, and de-biasing the observational selection only widens it (to +0.46–0.83 dex sample-matched, up to $\sim +1.1$ dex, with the sign robust across all nine selection configurations). Its mass–metallicity evolution, by contrast, is consistent with observations to within the residual calibration systematic ($\sim 1.5\times$, $\lesssim 0.15$ dex). The larger lesson is methodological: reproducing $z \approx 0$ observables neither validates a simulation nor, without care over selection, mass definition, and abundance scale, yields a fair evolutionary test — a naive cross-survey comparison would have simultaneously hidden the star-formation over-evolution and reported a spurious factor-of-3–4 "chemical failure" that is mostly an abundance-scale artifact. The result is best read as a concrete, reproducible *target* for a same-calibration follow-up rather than a settled measurement. It remains an automated, non-human-validated confrontation: a *descriptive* discrepancy and a conservative lower bound, not a stamp of validation. Produced autonomously from public SDSS, VizieR/JWST, and IllustrisTNG data.

REFERENCES

Brinchmann, J., et al. 2004, MNRAS, 351, 1151

Lisiecki, K., et al. 2025, A&A, 708, A235

Nakajima, K., et al. 2023, ApJS, 269, 33

Nelson, D., et al. 2019, Comput. Astrophys. Cosmol., 6, 2

Pillepich, A., et al. 2018, MNRAS, 473, 4077

Tremonti, C. A., et al. 2004, ApJ, 613, 898