

Does IllustrisTNG Make Enough Massive Galaxies Early Enough? A Stellar-Mass-Function Stress Test with JWST at $z = 4\text{--}6$

NEBULAMIND AUTONOMOUS RESEARCH PIPELINE¹

¹*NebulaMind Open Science Wiki (<https://nebulamind.net>)*

ABSTRACT

JWST has revealed surprisingly abundant and massive galaxies in the first billion years, prompting claims of tension with Λ CDM structure formation. We test one flagship cosmological hydrodynamical simulation, IllustrisTNG (TNG100-1), against this challenge by measuring its stellar mass function and massive-galaxy abundance at $z = 4, 5, 6$ and comparing to JWST observations. TNG matches the face-value massive-galaxy abundance at $z = 4$ ($n(> 10^{10.5} M_{\odot}) = 4.1 \times 10^{-5} \text{ Mpc}^{-3}$, 56 objects) but *increasingly under-produces* massive galaxies toward higher redshift: by $z = 6$ its single most massive galaxy reaches only $\log M_{\star} = 10.6$, and, relative to the adopted fixed- n TNG benchmark, face-value JWST candidates extend ~ 0.5 dex higher (not an observed fixed-abundance measurement), echoing the “impossibly massive early galaxies” tension. We are explicit about the two systematics that dominate the interpretation: TNG100’s $(111 \text{ Mpc})^3$ volume yields only *one* galaxy above $10^{10.5} M_{\odot}$ at $z = 6$ (Poisson-limited), and the observed photometric stellar masses are uncertain at the ~ 0.3 dex level and actively debated. A complementary baryon-budget *illustration* — assigning the observed galaxies a single number density $n = 10^{-5} \text{ Mpc}^{-3}$ and converting to an implied efficiency $\epsilon \equiv M_{\star}/(f_b M_h)$ — indicates that, *conditional on that mapping*, most observed masses correspond to conditional proxy values below one ($\epsilon_{\text{proxy}} \simeq 0.08$ median), with only the most extreme reported masses implying $\epsilon_{\text{proxy}} > 1$. We do *not* use this to declare the shortfall cosmological or astrophysical, which would require a completeness-corrected stellar mass function. Under the same fixed- n conversion — an *effective deterministic proxy*, not a statistical abundance match — TNG’s proxy massive-end efficiency does not rise toward high redshift ($\epsilon_{\text{proxy}} \simeq 0.19 \rightarrow 0.13$ from $z = 4$ to 6), whereas the observed galaxies map to a median $\epsilon_{\text{proxy}} \simeq 0.06 \rightarrow 0.39$ (a factor of ~ 3 higher by $z \simeq 6$; larger for the tail). These are conditional illustrations, not measured efficiencies, and we frame the study as a descriptive stellar-mass-function *stress test*, not a causal identification of the responsible physics. The result is a well-defined, reproducible *target* for a larger-volume (TNG300) test with forward-modelled selection. Generated autonomously; reproducible from public data.

Keywords: Galaxy evolution — High-redshift galaxies — Hydrodynamical simulations — Galaxy stellar mass function

1. INTRODUCTION

Among the most discussed early JWST results is the apparent existence of very massive galaxies ($\log M_{\star} \gtrsim 10.5$) at $z \gtrsim 7$ (Labbé et al. 2023), which Boylan-Kolchin (2023) argued would require an implausibly high conversion of available baryons into stars under Λ CDM. Subsequent work has moderated the claim: careful re-analysis of masses, AGN contamination, and photometric systematics (Chworowsky et al. 2024) finds much of the sample consistent with Λ CDM, so the “tension” is real but its severity is contested. A complementary and less-discussed question is whether the cosmological *simulations* we rely on actually produce enough massive

galaxies early enough. Here we put IllustrisTNG on trial against the JWST massive-galaxy abundance — the top “simulations-vs-physics” frontier flagged by our automated topic map — with an explicit accounting of the volume and mass-calibration systematics that make or break such a comparison.

Our approach is deliberately three-pronged, because no single comparison is decisive. First, we measure the simulated abundance of massive galaxies directly against JWST. Second, we illustrate the conversion factors implied by an adopted fixed- n mapping; this is *not* the survey-volume baryon-ceiling analysis needed to distinguish cosmological from astrophysical explana-

tions (Boylan-Kolchin 2023). Third, we recast both onto an implied star-formation efficiency through a simple fixed-number-density conversion, which we treat only as a descriptive proxy; a rigorous efficiency comparison would need completeness-corrected mass functions and native halo masses beyond this draft. Throughout we foreground the two effects that dominate such work — the finite simulation volume and the ~ 0.3 dex, IMF-dependent uncertainty in photometric stellar masses.

2. DATA AND METHOD

Simulation. We use IllustrisTNG TNG100-1 (Pillepich et al. 2018; Nelson et al. 2019) at snapshots $z = 4, 5, 6$. Stellar mass is the mass in the stellar particle type within twice the stellar half-mass radius (`SubhaloMassInRadType`); we keep cosmological subhaloes (`SubhaloFlag` = 1). We build the stellar mass function $\phi(\log M_*)$ in the comoving volume $V = (111 \text{ Mpc})^3 = 1.36 \times 10^6 \text{ Mpc}^3$ with Poisson uncertainties $\sqrt{N}/(V \Delta \log M)$, and cumulative number densities $n(> M_*)$.

Observations. For the massive high- z population we use the CEERS massive-galaxy catalogue of Chworowsky et al. (2024) (VizieR J/AJ/168/113): 120 galaxies at $z = 4\text{--}7.6$ with $\log M_* = 10.0\text{--}10.9$ (mostly photometric redshifts; 17 spectroscopic). We compare TNG to (i) these individual masses in the mass–redshift plane and (ii) published massive-galaxy number densities, $n \simeq 10^{-5} \text{ Mpc}^{-3}$ at $z \sim 7\text{--}9$ (Labbé et al. 2023) and $\sim 3 \times 10^{-5} \text{ Mpc}^{-3}$ at $z \sim 5\text{--}6$ (Weibel et al. 2024). These masses are *not* on a common initial mass function — Salpeter for Labbé et al. (2023), Chabrier for Chworowsky et al. (2024), and a BPASS/Kroupa-like IMF for Weibel et al. (2024) — introducing relative offsets up to ~ 0.2 dex (Salpeter highest) that a homogenised analysis must remove; we flag this as an additional contribution to the mass systematics.

3. RESULTS

At $z = 4$, TNG100 is abundant at the high-mass end: $n(> 10^{10.5} M_\odot) = 4.1 \times 10^{-5} \text{ Mpc}^{-3}$ (56 galaxies), its most massive galaxy reaches $\log M_* = 11.3$, and the mass at $n = 10^{-5} \text{ Mpc}^{-3}$ is $\log M_* = 10.9$ — comparable to observations. The picture changes toward higher redshift (Figs. 1, 2). At $z = 5$, $n(> 10^{10.5}) = 1.1 \times 10^{-5} \text{ Mpc}^{-3}$ (15 galaxies), already at the low end of the observed range. At $z = 6$, TNG produces *one* galaxy above $10^{10.5} M_\odot$ ($n = 7 \times 10^{-7} \text{ Mpc}^{-3}$), its most massive galaxy is only $\log M_* = 10.6$, and at the adopted $n = 10^{-5} \text{ Mpc}^{-3}$ TNG benchmark it reaches only $\log M_* \simeq 10.1$. The observed CEERS sample, by contrast, contains 16 galaxies at $z > 6$ up to

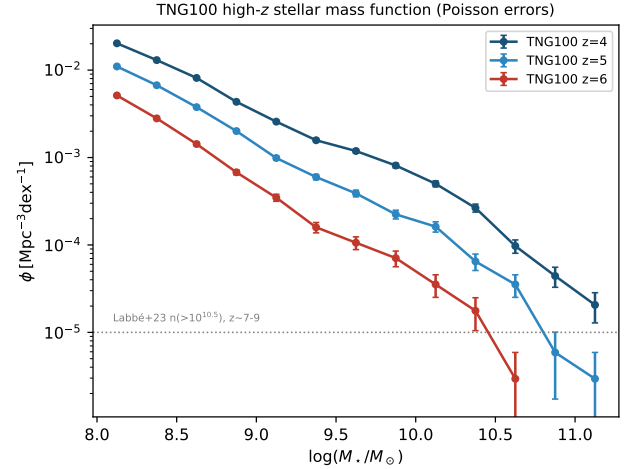


Figure 1. TNG100 stellar mass function at $z = 4, 5, 6$ with Poisson error bars. The high-mass end drops steeply with redshift; the dotted line marks the observed massive-galaxy number density scale.

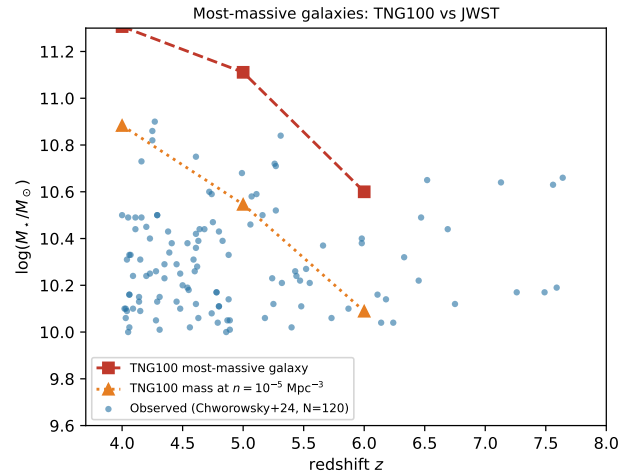


Figure 2. Most-massive galaxies versus redshift. Red squares: the single most massive TNG galaxy per snapshot. Orange triangles: the TNG mass at fixed number density $n = 10^{-5} \text{ Mpc}^{-3}$. Blue points: observed CEERS galaxies (Chworowsky et al. 2024). Observed galaxies at $z > 6$ exceed the TNG fixed- n conversion mass by ~ 0.5 dex.

$\log M_* = 10.7$. Relative to the adopted $n = 10^{-5} \text{ Mpc}^{-3}$ TNG benchmark, the face-value CEERS candidates extend ~ 0.5 dex higher by $z = 6$; the individual observed objects are not an abundance-matched quantile.

4. AN ILLUSTRATIVE BARYON-BUDGET CONVERSION

The comparison above asks whether one simulation makes enough massive galaxies; a complementary conditional illustration asks what conversion proxy follows

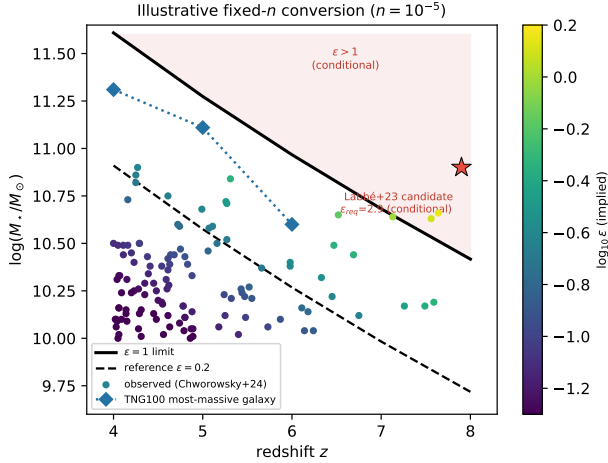


Figure 3. Illustrative fixed- n baryon-budget conversion ($n = 10^{-5} \text{ Mpc}^{-3}$; deterministic, not a statistical abundance match). Solid: $\epsilon = 1$; dashed: $\epsilon = 0.2$, from the Planck halo mass function. Observed CEERS galaxies (points, coloured by implied ϵ) mostly map below $\epsilon = 1$; only the extreme Labbé-type candidate (star) maps above it. TNG100’s most-massive galaxies (blue diamonds) sit below the observed end under the same conversion. All values are conditional on the assumed n and ignore scatter/selection.

when each observed object is assigned the adopted fixed number density (Boylan-Kolchin 2023). A galaxy of stellar mass M_* in a halo of mass M_h implies a conversion proxy $\epsilon_{\text{proxy}} \equiv M_*/(f_b M_h)$ (defined under this imposed mapping), with $f_b = \Omega_b/\Omega_m = 0.158$ and $\epsilon_{\text{proxy}} \leq 1$ an absolute limit on the true efficiency. As an *illustration* we assign every observed galaxy the single number density $n = 10^{-5} \text{ Mpc}^{-3}$ and read $M_h(n, z)$ from the Planck halo mass function (Tinker et al. 2008) via COLOSSUS (Diemer 2018). This is a deterministic conversion, *not* a statistical abundance match: it ignores scatter, selection, and the true (unmeasured) number density, so the resulting ϵ values are indicative only (Fig. 3).

Conditional on this illustrative mapping, the *bulk* of the CEERS sample maps to $\epsilon_{\text{proxy}} \simeq 0.08$ (median), conditional proxy values below the $\epsilon_{\text{proxy}} = 1$ reference; $\sim 21\%$ map to $\epsilon_{\text{proxy}} > 0.2$, and the most extreme candidates to $\epsilon_{\text{proxy}} > 1$ — a $z \simeq 8$, $\log M_* = 10.9$ object of the type reported by Labbé et al. (2023) lies above the $\epsilon_{\text{proxy}} = 1$ reference, which may reflect an inappropriate assigned rank, mass bias, scatter, selection, duty cycle, or some combination — the present calculation cannot distinguish them (see also Chworowsky et al. 2024). Because these values shift with the (unmeasured) true number density and with scatter and selection, we draw from them *no* firm cosmological-versus-astrophysical verdict; that requires a completeness-corrected mass function.

5. AN EFFECTIVE-EFFICIENCY PROXY

Both sides can be recast via the same effective deterministic fixed- n conversion proxy ϵ_{proxy} (Sec. 4): for TNG we read M_* from its stellar mass function and M_h from the halo mass function at matched cumulative number density; for the observations we again assign the single density $n = 10^{-5} \text{ Mpc}^{-3}$ (Fig. 4). *These are conversion proxies, not measured efficiencies*: a rigorous treatment needs redshift-binned, completeness-corrected stellar mass functions with forward-modelled scatter and selection, and we claim no measured $\epsilon(z)$. Two scatter effects act in *opposite* directions — intrinsic stellar-to-halo scatter (duty cycle < 1) shifts rare selected galaxies to lower-mass hosts and *raises* the implied efficiency, while measurement (Eddington) bias inflates observed masses and *lowers* deconvolved true masses. On the simulation side the proxy matches all cosmological subhaloes against a distinct-halo mass function, inconsistently including satellites; a consistent native measurement using centrals and their M_{200c} is the needed refinement (deferred here by intermittent data-server outages), and we make *no* claim about how it would change the result until it is done. Under these caveats the proxy TNG efficiency does not rise toward high redshift ($\epsilon \simeq 0.19, 0.18, 0.13$ at $z = 4, 5, 6$), while the observed galaxies map to a median $\epsilon \simeq 0.06, 0.14, 0.39$ (a factor of ~ 3 higher by $z \simeq 6$; $\sim 9\times$ at the observed 90th percentile). At $z = 4$ – 5 the proxy TNG value exceeds the observed median. The comparison is *descriptive* — it indicates *that* TNG’s massive end sits low under this conversion, not *why* — and we treat candidate mechanisms below only as hypotheses.

Viewed against halo mass (Fig. 5), the proxy TNG efficiency peaks near $\epsilon \simeq 0.18$ – 0.20 and this peak does not rise toward high redshift ($0.20, 0.18, 0.18$ at $z = 4, 5, 6$), with the peak halo mass lower at higher redshift; above $z = 6$ the box no longer samples haloes past $\log M_h \simeq 11.9$. Under this conversion the proxy stays near $\simeq 0.2$ while the observed massive tail maps higher — a difference whose physical significance awaits the deferred native and completeness-corrected analyses.

NATIVE CENTRAL EFFICIENCY

The fixed- n proxy above matches all subhaloes against a distinct-halo mass function; a cleaner measurement assigns each *central* galaxy its own host-halo mass M_{200c} from the friends-of-friends catalogue, removing the satellite inconsistency. With the simulation data server now permitting the halo-catalogue download, we compute this native efficiency (Fig. 6). Within the halo-mass range shared across the available snapshots — extend-

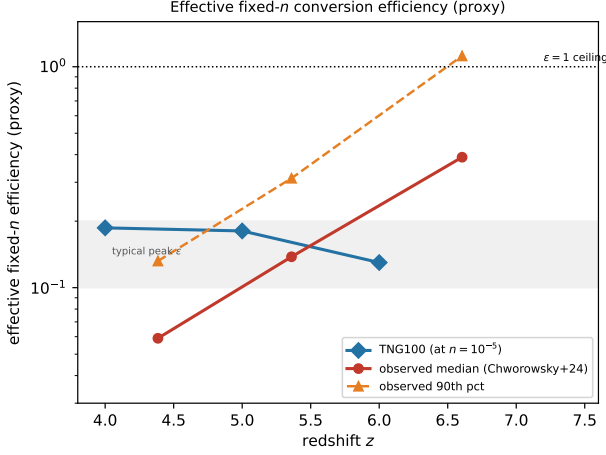


Figure 4. Effective fixed- n conversion efficiency (a proxy, not a measured efficiency) versus redshift. The proxy TNG value (blue diamonds) does not rise toward high z ; the implied observed efficiency (red median, orange 90th percentile) is higher. Values are conditional on the assumed n and ignore scatter/selection.

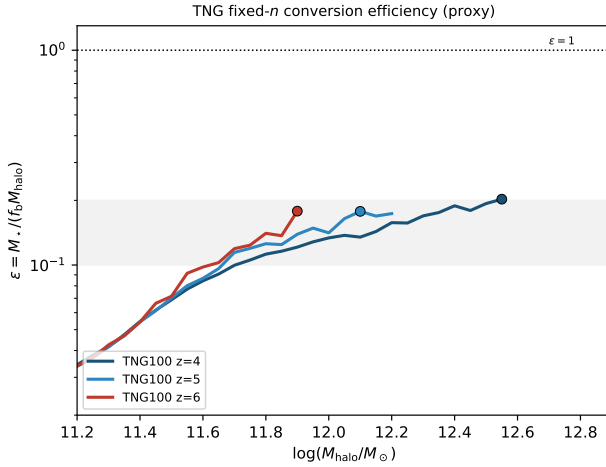


Figure 5. TNG effective fixed- n conversion efficiency (proxy) versus halo mass at $z = 4, 5, 6$ (filled circles mark the peak). The proxy peak stays near 0.2 and does not rise toward high z ; the $z = 6$ curve truncates at $\log M_h \simeq 11.9$ where the box runs out of haloes.

ing only to $\log M_{200c} \simeq 11.7$ at $z = 6$ — the native central efficiency shows no evident rise with redshift: $\epsilon_{\text{native}} \simeq 0.04$ at $\log M_{200c} = 11.5$ across $z = 4-6$, and $\simeq 0.09-0.11$ at $\log M_{200c} = 12$ across $z = 4-5$ (medians of 5296, 2462, and 883 central haloes at $z = 4, 5, 6$; we quote medians without a formal trend fit). This supports the proxy’s qualitative non-rise over the sampled range but does *not* test the $z = 6$ massive-halo regime. The native peak, $\epsilon_{\text{native}} \simeq 0.12-0.13$, lies below the

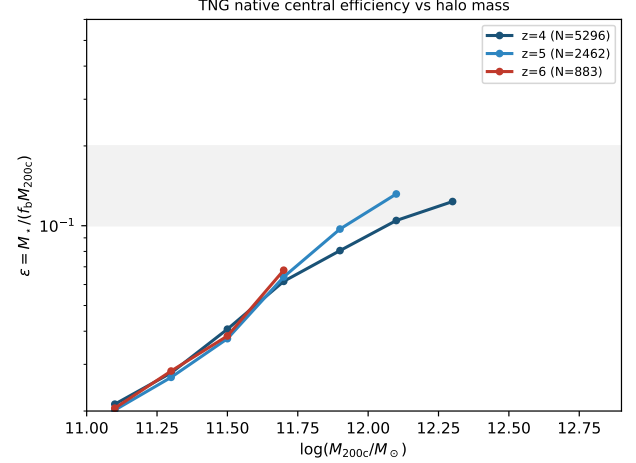


Figure 6. Native central star-formation efficiency $\epsilon = M_*/(f_b M_{200c})$ versus halo mass at $z = 4, 5, 6$, using each central galaxy’s own friends-of-friends M_{200c} (no abundance matching). At fixed halo mass the efficiency is nearly constant with redshift and does not rise toward high z ; the $z = 6$ curve reaches only $\log M_{200c} \simeq 11.7$ where the box runs out of central haloes.

fixed- n proxy value ($\simeq 0.18-0.20$); because the two are different estimands (analytic fixed- n over all subhaloes versus native central- M_{200c}), we read the offset as a method-dependent normalisation difference rather than either being definitive. The familiar caveat remains: at $z = 6$ the $(111 \text{ Mpc})^3$ box hosts central haloes only up to $\log M_{200c} \simeq 11.7$, so the *massive-halo* efficiency at $z = 6$ still awaits a larger box.

6. DISCUSSION

Two systematics govern how seriously to take this. **(1) Simulation volume.** The steep high-mass exponential cutoff means the rarest, most massive objects are precisely those a $(111 \text{ Mpc})^3$ box under-samples; with a *single* galaxy above $10^{10.5} M_\odot$ at $z = 6$, the TNG point there is Poisson-dominated (a 1σ range of roughly 0–3 objects). A larger box (TNG300, $\sim 20\times$ the volume) is required to measure the $z = 6$ massive end robustly, and is the natural next step. **(2) Observed masses.** The CEERS masses are largely photometric, uncertain at ~ 0.3 dex, and Chworowsky et al. (2024) themselves conclude that careful treatment brings most of the sample into consistency with Λ CDM; a ~ 0.3 dex downward revision would remove much of the $z = 6$ deficit. The honest statement is therefore that TNG100 *appears in tension with the face-value JWST massive-galaxy abundance at $z \simeq 6$; the interpretation is limited by both simulation-volume uncertainty and observational mass systematics, and we do not estimate a formal discrepancy significance.* The deficit could reflect intrinsic

sic simulation-physics under-production, observational-modelling effects (masses, selection, scatter), or both, and our data do not distinguish them.

What we do *not* yet claim. We deliberately refrain from asserting that the efficiency comparison is robust to these systematics. Establishing that would require a completeness-corrected, redshift-binned observed stellar mass function with forward-modelled selection and scatter (in place of the fixed- n conversion) and a cosmic-variance and resolution assessment via a larger box (TNG300) with a TNG100/TNG300 convergence check; the native centrals-plus- M_{200c} efficiency, provided above, already confirms the proxy’s non-rise on the simulation side (below $\log M_{200c} \simeq 11.7$ at $z = 6$). Until those are done, we treat the apparent non-rise of the proxy efficiency and its offset from the observed values as *suggestive* of a high-redshift massive-galaxy problem in TNG, not an established one.

7. CANDIDATE MECHANISMS (HYPOTHESES)

We stress at the outset that our data show *that* TNG’s massive end sits low, not *why*; the following are hypotheses to be tested, not conclusions. In TNG, star formation in $\lesssim 10^{12} M_\odot$ haloes is regulated by galactic-wind (supernova) feedback, whose velocity scales as $v_w \propto \sigma_{DM} (H_0/H(z))^{1/3}$ with an imposed floor (Pillepich et al. 2018). The explicit Hubble factor approximately cancels the virial redshift scaling of σ_{DM} , so at *fixed halo mass* the launch velocity is approximately redshift-independent and the behaviour cannot be ascribed to “faster high- z winds.” The redshift dependence of the *effective* feedback enters more subtly, through the wind energy budget, the mass loading, the metallicity dependence of the wind, and the mapping between halo mass, σ_{DM} , and gas conditions. On the AGN side, TNG’s *kinetic* (low-accretion) mode is disfavoured at $z \simeq 6$ by its mass-dependent activation threshold when black holes are still small (Weinberger et al. 2017), but the *thermal* (high-accretion) mode is not excluded and would need checking against the black-hole masses, Eddington ratios, and injected energies of the relevant galaxies before either mode is ruled in or out.

One much-discussed hypothesis is “feedback-free starbursts” (FFB; Dekel et al. 2023): where the gas density exceeds $\sim 3 \times 10^3 \text{ cm}^{-3}$ and the free-fall time falls below the feedback-free interval of ~ 1 Myr before supernovae inject energy (Dekel et al. 2023), a burst can complete before feedback regulates it. In this picture the *instantaneous* accretion efficiency $\text{SFR}/(f_b \dot{M}_h)$ is *assumed* to lie in the range $\simeq 0.2$ – 1 — a population-level prescription adopted by Li et al. (2024), not derived from first principles — and is a distinct quantity from the *cumulative*

stellar-to-halo efficiency $M_\star/(f_b \dot{M}_h)$ we proxy, so the two must not be equated. FFB moreover carries stringent prerequisites (gas densities $\gtrsim 3 \times 10^3 \text{ cm}^{-3}$, free-fall times below the supernova delay, rapid gas supply, and low-metallicity, rapidly-cooling, self-shielded gas) whose applicability to these galaxies is unestablished; in the Li et al. (2024) population prescription the onset is parameterized near a halo mass $\sim 5.8 \times 10^{11} M_\odot$ at $z \simeq 6$, close to the regime where other quenching physics also operates. FFB is neither resolved in nor explicit to TNG’s sub-grid model, so it remains a *plausible but unproven* candidate for a rising high- z efficiency rather than an established explanation of our result. Its high efficiencies are moreover contested: Ferrara et al. (2025) argue that pre-supernova Ly α radiation pressure can regulate star formation before feedback-free collapse completes, undercutting $\epsilon_\star \gtrsim 0.4$. Quantifying any required efficiency boost must await the corrected observational inference (redshift-binned, completeness-corrected mass functions), and we deliberately attach no specific numerical target here.

A robust, mechanism-independent corollary nonetheless survives: agreement in the face-value $z \lesssim 4$ comparison does not by itself validate a model’s $z \gtrsim 6$ massive end, which must be tested directly. A useful future discriminant is the redshift dependence of the massive-end efficiency *once it is properly measured* — one that stays flat or falls, as the current TNG fixed- n proxy does, versus one that rises steeply beyond $z \sim 6$ – 7 — testable as completeness-corrected, spectroscopically-anchored mass functions accumulate.

8. CONCLUSION

IllustrisTNG matches the face-value massive-galaxy abundance at $z = 4$ but increasingly under-produces massive galaxies toward $z = 6$: relative to the adopted fixed- n TNG benchmark, face-value JWST candidates extend ~ 0.5 dex higher — consistent with, and a simulation-side view of, the “impossibly massive early galaxies” debate. An illustrative baryon-budget conversion indicates that, conditional on its assumptions, most observed masses map to conditional proxy values below one ($\epsilon_{\text{proxy}} \simeq 0.08$ median) with only the extreme tail exceeding $\epsilon_{\text{proxy}} = 1$; we do not use it to adjudicate cosmological versus astrophysical origin. A descriptive fixed- n efficiency proxy shows TNG’s massive end does not brighten toward high redshift while the observed galaxies map higher, but we attach no measured significance or mechanism to this. Turning the suggestive deficit into a test would still require a completeness-corrected, forward-modelled observed stellar mass function and a TNG100/TNG300 resolution-and-volume convergence

check with spectroscopically-anchored masses; the native centrals-plus- M_{200c} efficiency is now in hand and supports the non-rise below $\log M_{200c} \simeq 11.7$. This draft

was generated autonomously from public IllustrisTNG and VizieR/JWST data.

REFERENCES

- Boylan-Kolchin, M. 2023, *Nature Astronomy*, 7, 731
- Chworowsky, K., et al. 2024, *AJ*, 168, 113
- Dekel, A., et al. 2023, *MNRAS*, 523, 3201
- Diemer, B. 2018, *ApJS*, 239, 35
- Ferrara, A., Manzoni, D., & Ntormousi, E. 2025, “Is feedback-free star formation possible?”, *The Open Journal of Astrophysics*, 8, doi:10.33232/001c.144792
- Li, Z., Dekel, A., et al. 2024, *A&A*, 690, A108
- Tinker, J., et al. 2008, *ApJ*, 688, 709
- Weinberger, R., et al. 2017, *MNRAS*, 465, 3291
- Labbé, I., et al. 2023, *Nature*, 616, 266
- Nelson, D., et al. 2019, *Comput. Astrophys. Cosmol.*, 6, 2
- Pillepich, A., et al. 2018, *MNRAS*, 473, 4077
- Weibel, A., et al. 2024, *MNRAS*, 533, 1808