Skepthical review: *Differentiable centisecond halo-model predictions in $-\Lambda CDM$ and beyond*

Summary

The manuscript presents **classy_szlite**, a pure-JAX halo-model pipeline for fast and differentiable predictions of the thermal SZ angular power spectrum, with emphasis on C_ℓ^{yy}. It combines CosmoPower ede-v2 neural emulators for cosmology-dependent inputs, FFTLog-based transforms (for $\sigma(R,z)$ and pressure-profile Fourier windows), and a JAX-native integration design that enables automatic differentiation and millisecond-level evaluations (reported as $\sim$ **5 ms** for fixed-cosmology evaluations and $\sim$ **20 ms** including cosmology setup; Sec. 2–3). The paper demonstrates (Sec. 5) gradient-based MAP estimation (L-BFGS(-B), Newton), and compares RW-MH (cobaya) versus NUTS (NumPyro) on an 8-bandpower tSZ dataset at two fixed cosmologies, reporting modern diagnostics (R-hat, ESS, divergences, E-BFMI) and validating gradients/Fisher matrices against finite differences. The work is timely and potentially impactful for inference, forecasting, and gradient-based methods in halo-model applications, but key aspects require clarification and stronger empirical support: the precise scope of "end-to-end differentiability" given non-JAX FFTLog components, the specification/reproducibility of the data likelihood and priors, validation against established reference pipelines (e.g. class_sz/CAMB/CLASS), and fair/fully documented benchmarking and sampler comparisons.

Strengths

- Clear motivation and positioning: explains why differentiability (not only speed) enables new inference workflows (Sec. 1, Sec. 6).
- Technically coherent architecture: JAX-first design with a practical "factory/closure" pattern to amortize cosmology setup and accelerate nuisance-only sampling (Sec. 3.3).
- Strong practical demonstration: MAP optimization + NUTS sampling with comprehensive diagnostics (R-hat, ESS, divergences, E-BFMI) and an informative RW-MH vs NUTS comparison (Sec. 5.1–5.6).
- Good numerical self-checks: gradients and Fisher matrices compared to finite differences with near machine-precision agreement (Sec. 5.6, Sec. 5.8).
- Open-source/reproducibility-oriented: code and emulator weights are made available, and the manuscript largely ties methodology to implementable components (Data Availability; Sec. 3, Sec. 6).
- Forward-looking discussion: outlines extensions to other tracers and differentiable inference paradigms (Fisher forecasts, VI, SBI), and comments on hardware backends (Sec. 6.1–6.3).

Major issues

1. **The manuscript repeatedly claims an "end-to-end", "fully differentiable", "JAX-traceable" pipeline, but Sec. 3.2–3.3 indicate reliance on non-JAX FFTLog / `mcfit.TophatVar` NumPy code paths that are "not safe to JIT-trace". As written, it is unclear which parts are actually inside the autodiff graph. This affects the scope of supported gradients—especially derivatives with respect to cosmological parameters if $\sigma(R, z)$ (and any downstream quantities depending on it) is computed outside JAX.**

   Recommendation: In Sec. 3.2–3.3, explicitly state what "differentiable" means in the public implementation. Add a compact "differentiability matrix" (table) listing which outputs (e.g. C_ℓ, bandpowers, likelihood) are differentiable with respect to which parameter blocks (cosmology vs pressure/profile vs nuisance) and whether gradients pass through: (i) emulator calls, (ii) $\sigma(R, z)$/FFTLog, (iii) HMF/bias, (iv) interpolation steps, (v) halo integrals. If $\sigma(R, z)$ is precomputed and treated as a constant inside the closure at fixed cosmology, say so and constrain the "end-to-end" language accordingly in Sec. 1–3 and Sec. 6. If cosmology gradients are intended, describe (even briefly) the concrete plan (JAX-native FFTLog, custom JVP/VJP, or alternative σ computation) and how this would affect timings.

2. **The worked-example likelihood is under-specified and therefore hard to reproduce/interpret (Sec. 5.1, Sec. 5.3–5.5). The source of the 8 bandpowers, their exact ℓ-binning/window functions, whether the data are real or synthetic, how the full 8×8 covariance was obtained (and its correlation structure), and the priors / parameterization for (P_0, β) are not fully stated. Because the posterior exhibits long tails, priors and parameter bounds materially affect results and sampler behavior.**

   Recommendation: Expand Sec. 5 with a precise likelihood specification: identify the dataset origin (or state it is synthetic), provide the 8 bandpowers with ℓ-bin edges/centers and uncertainties (table or appendix), define or reference the bandpower window functions used, and describe covariance estimation (analytic vs simulations; diagonal vs correlated; any conditioning). Write the Gaussian likelihood explicitly in terms of the data vector and covariance. State priors (forms and bounds) and whether parameters are sampled in linear/log space, and clarify whether MAP includes priors (Sec. 5.3). If the full numerical vectors/matrices live in the repository, explicitly point to file paths and ensure they match the manuscript.

3. **Scientific/technical validation against established reference calculations is currently too thin given the central reliance on (i) CosmoPower ede-v2 emulators (Sec. 3.1) and (ii) a new numerical/integration implementation. There is no in-paper end-to-end comparison of C_ℓ^{yy} (1h/2h/total) against a**

CLASS/CAMB-based pipeline (e.g. class_sz) over representative cosmologies, nor a discussion of how emulator errors propagate to downstream halo-model observables and inferred parameters.

Recommendation: Add a validation subsection (Sec. 3 or Sec. 5) comparing classy_szlite to a reference pipeline (class_sz/CLASS/CAMB-based) for multiple cosmologies within the ede-v2 training domain and at least one profile setting. Show fractional differences versus ℓ (and ideally bandpower-level differences) for 1h, 2h, and total. Summarize emulator accuracy relevant for tSZ inputs (Sec. 3.1) and briefly discuss error propagation to C_ℓ^{yy} and to degeneracies (e.g. σ_8–P_0). Also state how points near emulator boundaries are handled.

4. **A central advertised use case is higher-dimensional joint cosmology+astrophysics inference (Sec. 1, Sec. 5.2, Sec. 6.1), yet the main demonstration is restricted to a 2-parameter (P_0, β) inference at fixed cosmology (Sec. 5). Scaling claims for NUTS and the differentiable approach at $d \gtrsim 6$–30 are extrapolated rather than empirically shown in this work.**

   Recommendation: Add at least one modest higher-dimensional example using the same pipeline and NUTS (e.g. include B and one additional GNFW shape parameter; or a small joint cosmology+profile run with $d \sim 6$–10, even with informative priors or a simplified/mock likelihood). Report wall time, gradient-evaluation counts, ESS, R-hat, divergences, and (ideally) ESS-per-gradient-evaluation. If infeasible, soften the language in Sec. 1, Sec. 5.2, Sec. 5.5, and Sec. 6.1 to clearly label higher-d performance as an expectation rather than a demonstrated result.

5. **Timing benchmarks and sampler-comparison methodology are not fully normalized or clearly documented (Sec. 2; Sec. 5.1–5.5; Sec. 6.2; Fig. 1). The RW-MH vs NUTS comparison mixes differing parallelism (e.g. RW-MH with "4 MPI walkers" vs NUTS chain execution assumptions), warmup/compilation inclusion is unclear, hardware details differ across locations (laptop CPU vs EPYC vs TPU), and the headline "$\sim 100\times$" statement appears numerically inconsistent in at least one place (Sec. 5.5).**

   Recommendation: For Sec. 5.1–5.5, state a clear protocol: (i) whether reported times include JAX/XLA compilation and warmup; (ii) whether NUTS chains are run sequentially or in parallel; (iii) for RW-MH, proposal tuning/adaptation, warmup, and whether "4 MPI walkers" implies 4 cores used concurrently (and whether times are wall-clock vs CPU-time). Report efficiency in units that factor out parallelism differences (e.g. ESS per forward-model evaluation, ESS per gradient evaluation) in addition to ESS/sec. Correct the inconsistent speedup arithmetic in Sec. 5.5 and reconcile the NUTS wall-time discrepancy between Table 1 and Fig. 4 caption by explicitly stating what differs (sample budget, hardware, warmup inclusion, parallelization). For

Fig. 1 / Sec. 6.2, standardize time units, document thread settings (OpenMP/JAX/XLA), and provide reproducible benchmark scripts and variability/error bars.

6. **Core mathematical definitions for the tSZ window function are not fully verifiable from the manuscript (Sec. 3.2; Sec. 4). In particular, $W_y(\ell, M, z)$ uses symbols that are not defined (e.g. ℓ_{500}, J_ℓ), the prefactor is dimensionally ambiguous, and consistency with Eq. (1)'s transform convention is difficult to audit.**

   Recommendation: Define ℓ_{500} explicitly (e.g. via $\ell_{500} = D_A(z)/r_{500}$ or equivalent) and provide an explicit definition of $J_\ell[\cdot]$ (including whether it contains the $r^2 dr$ measure and any projection factors). Re-check and state the dimensional consistency so that Compton-y is dimensionless, and ensure the W_y definition is demonstrably consistent with Eq. (1) under the substitutions $r = x r_{500}$ and $k = k_\ell$ (Sec. 4).

7. **The profile Fourier-transform integral is written to $r = \infty$ (Eq. (1), Sec. 3.2) while the inference explores outer slopes around $\beta \approx 2.7$ (Sec. 5.3–5.4), for which integrals like $\int r^2 P(r) dr$ can fail to converge without truncation/apodization. The manuscript does not state an $r_{\rm max}$, truncation scheme, or convergence/regularization strategy, so low-k/low-ℓ well-posedness is unclear from the PDF alone.**

   Recommendation: State explicitly whether the GNFW/pressure profile is truncated (and at what radius, e.g. multiple of r_{500} or $r_{\rm vir}$), apodized, or otherwise regularized before applying Eq. (1). Document the effective r-grid used in the implementation and clarify under what conditions on β the transform exists as $k \to 0$. If finite-r limits are used in code, reflect that in the mathematical description (Eq. (1) and surrounding text in Sec. 3.2).

8. **Key numerical settings controlling accuracy and runtime are not systematically documented (Sec. 3.1–3.3; Sec. 5; Sec. 6.2). Important details (mass/redshift grid ranges and resolution, FFTLog configuration, interpolation schemes, ℓ sampling, integration rules, and default configuration used to reproduce figures/timings) are scattered or implicit, which limits strict reproducibility and makes it harder to judge accuracy/runtime trade-offs.**

   Recommendation: Add a dedicated "Numerical configuration" subsection (e.g. Sec. 3.4 or a preamble to Sec. 5) that lists: (i) z and M grid definitions (ranges, spacing, sizes); (ii) FFTLog settings for $\sigma(R, z)$ and pressure transforms (grid sizes, bias/padding, extrapolation); (iii) ℓ arrays and bandpower windowing; (iv) integration strategy (vectorization, quadrature, any adaptivity); (v) runtime scaling with grid sizes. Point to a single default config file / parameter dictionary in the repository that reproduces the paper's figures and timings.

Minor issues

1. The general "tracer-agnostic" halo-model template claim (Sec. 4; Sec. 6.1–6.3) is currently more conceptual than demonstrated, since only tSZ C_ℓ^{yy} is implemented and shown. Extensions to kSZ^2, CIB, galaxy–lensing, cluster counts typically require additional ingredients (HOD/selection functions, mass–observable relations, redshift kernels) beyond merely swapping W_T.

   Recommendation: Either (i) temper the generality claims in the Abstract/Sec. 4/Sec. 6 to clearly state that the current public focus is tSZ C_ℓ^{yy}, or (ii) add one concrete non-tSZ worked example (e.g. a simple $y \times g$ cross-spectrum) including the explicit kernel/window and timing. In Sec. 6.3, briefly enumerate the additional modeling components required per tracer and what is realistically planned for near-term releases.

2. Physical interpretation of the worked-example GNFW parameter shifts and degeneracies is easy to overread as astrophysical tension (Sec. 5.3–5.4), even though many relevant quantities are fixed (e.g. hydrostatic bias B, other GNFW parameters, cosmology).

   Recommendation: Add a concise caveat in Sec. 5.3–5.4 listing what is fixed (B, c_{500}, γ, α, cosmology, etc.) and how these assumptions could shift inferred (P_0, β). Optionally include a small sensitivity test (vary B or swap profile model) or point directly to repository scripts enabling such tests, emphasizing that Sec. 5 is primarily methodological.

3. Notation inconsistency in Sec. 5: the heading refers to "tSZ $C_\ell^{\gamma\gamma}$ bandpowers" while the observable elsewhere is C_ℓ^{yy}. This can be confusing (shear γ vs Compton-y).

   Recommendation: Standardize Sec. 5 and relevant figure captions to C_ℓ^{yy} (or explicitly define γ if it is intended to denote y).

4. ESS / integrated autocorrelation time definitions are internally inconsistent (Sec. 5.1): ESS is written as $N/(1 + 2\,\tau_{\rm int})$ but later reasoning uses ESS $\approx N/\tau_{\rm int}$, which yields different numbers for the same $\tau_{\rm int}$.

   Recommendation: Define $\tau_{\rm int}$ precisely (including whether it includes the lag-0 term and whether it is per-chain or pooled) and use one consistent ESS relation throughout; update the text so the reported ESS values match the stated convention.

5. Units conventions are not fully explicit: Sec. 4 gives $dV/(d\Omega dz) = \chi^2/H(z)$ while elsewhere c is explicit (e.g. $m_e c^2$ in W_y). Without a statement (e.g. $c = 1$ in the cosmological part, or H in {\rm Mpc}^{-1}), dimensional consistency is harder to audit.

   Recommendation: Add a brief units convention statement near Sec. 4 (and/or in a notation table), clarifying whether $c = 1$ is assumed for χ, H or how emulator outputs encode units.

6. Sampler-performance conclusions rely heavily on a mean-based "accuracy" metric for one parameter (Fig. 10; Sec. 5.5). Given non-Gaussian/heavy-tailed posteriors, mean error alone may not capture posterior agreement robustly.

   Recommendation: Complement the mean-based metric with at least one additional posterior-distance/summary metric (e.g. error on both (P_0, β) mean and covariance; KS/Wasserstein distance for 1D marginals; or bandpower-predictive checks). Also report ESS/sec (or ESS per eval) for both parameters, not only P_0.

7. Interpolation and potential non-smoothness: if the implementation uses operations like `searchsorted` for piecewise interpolation on grids, gradients can become non-smooth and may affect HMC/NUTS stability in some regimes (Sec. 3.3; implied by JAX implementation choices).

   Recommendation: Briefly document the interpolation scheme(s) used for emulator outputs and grids, and comment on whether any non-smooth steps exist in the computational graph. If relevant, note tested stability (divergences) and/or consider smoother interpolation alternatives for parameters that move grid indices.

8. Figure set (esp. Fig. 1 and figures in Sec. 5) contains several presentation/reproducibility issues flagged by the structured review: mixed/inconsistent time units (Fig. 1), unclear inclusion/exclusion of compilation/warmup, incomplete axis/legend definitions, and potential accessibility issues (small fonts, color choices).

   Recommendation: Standardize units (prefer ms), ensure captions are self-contained (hardware + protocol + sample sizes), add error bars/variability where appropriate, and adopt colorblind-safe palettes with redundant encodings (markers/linestyles).

9. Emulator interface documentation (Sec. 3.1) mixes standard cosmological parameters with emulator-specific conventions/symbols (e.g. $N_{\rm ut}$; outputs like $\log_{10}(k^3 P_k)$), which may hinder external use.

   Recommendation: Add a small table in Sec. 3.1 listing each emulator, its inputs, outputs, units/conventions, and grid definitions, and define any nonstandard symbols at first use.

10. Tracer bias notation ambiguity: Eq. (3) uses $b_T(M, z)$ while elsewhere a single halo bias model is implied (Sec. 3.3; Tinker-10).

    Recommendation: Clarify whether $b_T \equiv b_{\rm halo}(M, z)$ for all tracers in this paper or whether tracer-specific effective biases are intended in the general template.

11. Table 1 contains at least one ambiguous numeric format ("12.3/6") and could be clearer about whether entries are χ^2/dof or reduced χ^2, and which timings include warmup/compilation.

Recommendation: Label columns unambiguously (separate χ^2 and dof, or state explicitly χ^2/dof) and add footnotes specifying timing definitions (warmup, compilation, parallelism assumptions).

Very minor issues

1. Typographical/naming inconsistencies: "classy_szLite" vs "classy_szlite", "cl_yyFACTORY" vs "cl_yyFactory" vs "cl_yy_factory", "c1_yy" vs "cl_yy", "TSZ" vs "tSZ", inconsistent capitalization of hardware names (Sec. 2; Sec. 3.3; Sec. 5; Sec. 6.2).

   Recommendation: Do a final standardization pass across text, figures, and code references: choose one spelling for the package and key API functions, standardize "tSZ" and C_ℓ^{yy}, and fix small typos (e.g. "c1_yy" → "cl_yy").

2. Bibliography issues: duplicated or inconsistent Bolliet et al. 2023a/2023b entries (same title/arXiv), and several incomplete "in prep." references (References).

   Recommendation: Disambiguate or merge duplicate references; where possible replace "in prep." with arXiv IDs or add clarifying notes, and fill missing journal/publication metadata.

3. Some long, dense sentences and potentially confusing notation choices (e.g. using Z as a Z-score metric, which can be confused with Bayesian evidence Z; Sec. 1; Sec. 5.5).

   Recommendation: Split a few long sentences for readability and explicitly define the Z-score symbol (and state it is not the evidence) where used.

4. Figure readability: some fonts/line widths are small for print; legends sometimes rely on captions; some panels lack labels.

   Recommendation: Increase font sizes/line widths, add panel labels, and include concise in-figure legends so figures remain interpretable when viewed standalone or in grayscale.

Key statements and references

- • **The ede-v2 CosmoPower emulators used in classy_szlite, originally trained and validated for the ACT DR6 extended-cosmology analysis, reproduce standard ΛCDM predictions at the default early-dark-energy parameter value $f_{\rm EDE} = 10^{-3}$, with emulator outputs agreeing with CAMB-based references to well under 0.1σ on ACT DR6 parameter constraints.**
- *Reference(s):* Spurio Mancini et al. 2022, Bolliet et al. 2023a, Bolliet et al. 2023b
- • **The variance $\sigma(R, z)$ entering the Tinker halo mass function and bias, and the Fourier transform $\tilde{u}(k|M, z)$ of the GNFW pressure profile, are computed with FFTLog as implemented in the mcfit library, using the**

TophatVar transform for $\sigma(R,z)$ and mcfit.SphericalBessel for $\tilde{u}(k|M,z)$, with the $P(k)$ emulator's log-uniform k-grid chosen to satisfy FFTLog's requirements.

- *Reference(s):* Hamilton, 2000, Talman, 1978, Li, 2019

- • **classy_szlite adopts the Tinker 2008 halo mass function with the redshift-dependent parameters given in their Table 2 and the Tinker 2010 linear bias expressed in terms of the peak height $\nu = \delta_c/\sigma(M,z)$ at $\Delta_{\rm crit} = 500$, thereby matching the halo-population modelling used in class_sz for direct inter-code comparisons.**

- *Reference(s):* Tinker et al., 2008, Tinker et al., 2010

- • **In the worked tSZ bandpower example, a NumPyro implementation of the no-U-turn sampler (NUTS) using exact JAX gradients achieves an effective sample size of ≈ 1400 with $\hat{R} \leq 1.003$ and zero divergences in 4×2000 post-warmup samples, whereas a cobaya random-walk Metropolis chain tuned to the same posterior requires ≈ 14–17 minutes to reach $n_{\rm eff} \approx 1900$, demonstrating that NUTS is about $100\times$ faster wall-for-wall at matched accuracy on this 2D problem, consistent with theoretical scaling arguments for RW-MH and HMC.**

- *Reference(s):* Hoffman & Gelman 2014, Phan et al., 2019, Torrado & Lewis, 2021

- • **The ede-v2 CosmoPower emulator suite used by classy_szlite covers ΛCDM, m_ν–ΛCDM, wCDM, $N_{\rm eff}$–ΛCDM, and early-dark-energy combinations thereof, and has been shown to achieve $\lesssim 0.1\sigma$ accuracy relative to CAMB-based calculations for ΛCDM and extended cosmologies in recent ACT DR6 and DESI DR2 analyses, making it sufficient for essentially all current CMB and large-scale-structure survey targets without resorting to a full Boltzmann solver.**

- *Reference(s):* Spurio Mancini et al. 2022, Bolliet et al. 2023a, Calabrese et al. 2025

Mathematical consistency audit

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

Maths relevance: light

The PDF contains a small number of central analytic expressions: a spherical-Bessel Fourier transform for the pressure profile (Eq. (1)), a generic 1-halo/2-halo Limber-template for angular power spectra (Eqs. (2)–(3)), a stated form for the tSZ window function W_y (inline), and a Fisher-matrix expression (Eq. (4)). Most other content is computational/algorithmic. The core

halo-model template is structurally consistent, but key definitions needed to verify the tSZ window normalization and profile-transform well-posedness are missing or ambiguous, and there is an internal inconsistency in the ESS/$\tau_{\rm int}$ formula statements.

Checked items

1. ✔ **Alpha-beta sigmoid activation definition** (Sec. 3.1, p.2)
    - **Claim:** Defines the hidden-layer activation as $h_{\rm out} = (\beta + \sigma(\alpha z)(1 - \beta))z$.
    - **Checks:** algebra, notation consistency, sanity/limiting cases
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** σ denotes the logistic sigmoid, α and β are scalar parameters per hidden layer (or broadcastable to z)
    - **Notes:** Expression is algebraically well-formed. Limiting cases behave sensibly: $\beta = 1$ gives $h_{\rm out} = z$; $\beta = 0$ gives $h_{\rm out} = \sigma(\alpha z)z$.
2. ⚠ **Spherical Fourier transform of pressure profile** (Eq. (1), Sec. 3.2, p.3)
    - **Claim:** Defines $\tilde{u}(k|M, z) = 4\pi \int_0^\infty r^2 \left[\frac{\sin(kr)}{kr}\right] P\left(\frac{r}{r_{500}(M,z)}\right) dr$.
    - **Checks:** algebra, dimensional/units, existence/convergence, definition consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: critical
    - **Assumptions/inputs:** This is intended as a 3D spherical-Bessel (j_0) transform with the author's chosen Fourier convention, P is the 3D radial pressure profile (or proportional to it)
    - **Notes:** The form matches a standard $4\pi \int r^2 j_0(kr) f(r) dr$ convention, but the written upper limit ∞ raises a well-posedness concern when later-sampled outer slopes are around $\beta \approx 2.7$: the $k \to 0$ limit involves $\int r^2 P(r) dr$, which would require $\beta > 3$ (or an explicit truncation/regularization) to converge. The PDF does not state truncation/apodization, so existence at low k /low ℓ cannot be verified.
3. ✔ **Dimensionless lookup variable for profile transform** (Sec. 3.2, p.3)
    - **Claim:** Stores $\tilde{u}$ as a 1-D lookup over $s = kr_{500}/c_{500}$ for the Arnaud-10 profile.
    - **Checks:** dimensional/units, definition consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** c_{500} is the GNFW concentration-like parameter relating $r_s = r_{500}/c_{500}$, Shape dependence is intended to be captured as a function of kr_s
    - **Notes:** s is dimensionless if k is inverse-length. Using $r_s = r_{500}/c_{500}$ is consistent with tabulating transforms for profiles expressed in terms of r/r_s.
4. ✔ **1-halo Limber template for angular cross-power** (Eq. (2), Sec. 4, p.3)

- **Claim:** $C_\ell^{XY,\mathrm{1h}} = \int dz \, \frac{dV}{d\Omega dz} \int d\ln M \, \frac{dn}{d\ln M} W_X(\ell, M, z) W_Y(\ell, M, z)$.
- **Checks:** algebra, notation consistency, dimensional/units (structural)
- **Verdict:** PASS; confidence: high; impact: moderate
- **Assumptions/inputs:** W_T are the projected Fourier-space window functions appropriate for the defined tracer fields, $dn/d\ln M$ corresponds to the same mass definition used in W_T
- **Notes:** Structure is internally consistent: the use of $d\ln M$ with $dn/d\ln M$ is consistent, and separating tracer dependence into W_T is coherent.

5. ✔ **2-halo Limber template and tracer integrals** (Eq. (3), Sec. 4, p.3)

- **Claim:** $C_\ell^{XY,\mathrm{2h}} = \int dz \, \frac{dV}{d\Omega dz} P_{\mathrm{lin}}(k_\ell, z) \prod_{T\in X,Y} I_T(\ell, z)$, with $I_T \equiv \int d\ln M \, \frac{dn}{d\ln M} \, b_T(M, z) W_T(\ell, M, z)$.
- **Checks:** algebra, notation consistency, structural sanity
- **Verdict:** PASS; confidence: medium; impact: moderate
- **Assumptions/inputs:** Linear bias factorization is intended (two-halo term proportional to P_{lin} times products of biased tracer weights), b_T is either the halo bias or an effective tracer bias
- **Notes:** The product form is algebraically equivalent to $C \propto I_X I_Y$. Minor clarity gap: b_T notation suggests tracer-specific bias, while elsewhere a single halo bias model is described.

6. ⚠ **Limber mapping and comoving volume element** (Sec. 4, p.3)

- **Claim:** Uses $k_\ell = (\ell + 1/2)/\chi(z)$ and $\frac{dV}{d\Omega dz} = \chi^2/H(z)$.
- **Checks:** dimensional/units, definition consistency
- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
- **Assumptions/inputs:** χ is comoving distance and H is the Hubble rate, Either $c = 1$ is assumed in cosmological distances or H is expressed in inverse-length units
- **Notes:** Dimensionally, $dV/d\Omega dz$ typically requires consistency between units of χ and H; the PDF does not state whether $c = 1$ is assumed for these terms, while it keeps c explicitly elsewhere (e.g., $m_e c^2$). A units convention is needed to confirm consistency.

7. ⚠ **tSZ window function expression** (Inline after Eq. (3), Sec. 4, p.3)

- **Claim:** States $W_y(\ell, M, z) = (\sigma_T/m_e c^2)\,(4\pi r_{500}^3/\ell_{500}^2) J_\ell[P_e(x|M, z)]$, with J_ℓ the spherical-Bessel projection at k_ℓ and $x = r/r_{500}$.
- **Checks:** dimensional/units, definition consistency, compatibility with Eq. (1)
- **Verdict:** UNCERTAIN; confidence: medium; impact: critical

- **Assumptions/inputs:** ℓ_{500} is a characteristic angular multipole scale associated with r_{500} and $D_A(z)$, J_ℓ is a specific integral operator acting on the radial profile
- **Notes:** Key quantities are undefined (ℓ_{500}, the precise definition of J_ℓ), preventing verification of the r_{500} and D_A dependence and whether this expression is consistent with Eq. (1) plus the 3D→2D projection. As written, the prefactor's r_{500} power is ambiguous without knowing what J_ℓ includes.

8. ✔ **Fisher matrix for Gaussian likelihood with fixed covariance** (Eq. (4), Sec. 5.8, p.5)
    - **Claim:** $F_{ij}(\theta) = (\partial_i \mu)^T \Sigma^{-1} (\partial_j \mu)$, with $\mu(\theta)$ the model bandpower vector.
    - **Checks:** algebra, notation consistency, assumption consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Likelihood is Gaussian in the data vector with parameter-independent covariance Σ
    - **Notes:** Correct under the stated assumption of fixed Σ. The transpose/inner-product structure is consistent for μ as a vector.
9. ✖ **ESS and integrated autocorrelation time relations** (Sec. 5.1, p.4)
    - **Claim:** States ESS $= N/(1 + 2\tau_{\rm int})$ and later states $\tau_{\rm int} \approx 8$ is in agreement with ESS $\approx N/\tau_{\rm int}$.
    - **Checks:** algebra, definition consistency
    - **Verdict:** FAIL; confidence: high; impact: minor
    - **Assumptions/inputs:** N is the total number of draws, $\tau_{\rm int}$ is the integrated autocorrelation time
    - **Notes:** The two ESS relations are inconsistent unless $\tau_{\rm int}$ is defined differently in each statement. The text uses both as if they were simultaneously valid for the same $\tau_{\rm int}$, which is a mathematical/definition inconsistency that should be corrected by fixing the $\tau_{\rm int}$ convention and corresponding ESS formula.

Limitations

- Audit is limited to the PDF text provided; several key definitions (e.g., ℓ_{500} and J_ℓ) are not given in the PDF, preventing verification of the tSZ window normalization.
- The paper frequently references implementation details (code modules, libraries) without fully specifying the analytic conventions (Fourier conventions, truncation of profiles, unit system), which limits purely symbolic verification.
- No appendices or step-by-step derivations are included for the halo-model/tSZ-specific normalizations; where intermediate definitions are missing, items are marked UNCERTAIN rather than inferred.

Numerical results audit

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

Of 18 audited numeric items: 12 PASS, 2 FAIL, and 4 UNCERTAIN. The main failures are (i) a claimed $\sim 100\times$ wall-time speedup that computes to $\sim 9.36\times$ from the stated times, and (ii) a cross-location inconsistency in reported NUTS wall time (200 s vs ~ 40 s). Several other checks are descriptive or heuristic and cannot be strictly verified from the provided numerals alone.

Checked items

1. ⚠ **C1** (Page 1, Abstract)
    - **Claim:** "gradient-based optimisation reaches the MAP in fewer than ~ 40 forward-and-gradient evaluations (~ 0.4 s wall)"
    - **Checks:** wall_time_from_eval_count
    - **Verdict:** UNCERTAIN
    - **Notes:** Implied time per evaluation is $0.4/40 = 0.01$ s, but this is a heuristic with no directly comparable per-eval number reported in the checked inputs.
2. ✔ **C2** (Page 3, Section 3.3)
    - **Claim:** "Reverse-mode autodiff through the closure ... costs ~ 17 ms ... The closure performs only the halo-model integration ... is ~ 5 ms warm."
    - **Checks:** ratio_check
    - **Verdict:** PASS
    - **Notes:** Computed ratio $17/5 = 3.4$, consistent with "$\sim 3\times$" within the stated tolerance.
3. ✔ **C3** (Page 3, Section 3.3)
    - **Claim:** "The full pipeline including a fresh cosmology costs ~ 20 ms, with the emulator forward pass contributing ~ 2–3 ms and the halo-model integration the remainder."
    - **Checks:** parts_vs_total_range
    - **Verdict:** PASS
    - **Notes:** Remainder is $20 - 3 = 17$ ms to $20 - 2 = 18$ ms; positive and within $[0, \text{total}]$.
4. ✔ **C4** (Page 4, Section 5.1)
    - **Claim:** "ESS = N/(1+2 \tau_{\rm int}) \in [466, 504] ... The integrated autocorrelation time is $\tau_{\rm int} \approx 8$ ... in agreement with ESS $\approx N/\tau_{\rm int}$ (Figure 9)." for 4 chains × 1000 samples
    - **Checks:** ess_from_N_and_tau

- **Verdict:** PASS
- **Notes:** With $N = 4000$ and $\tau_{\text{int}} = 8$, $N/\tau_{\text{int}} = 500$ lies within $[466, 504]$. Also computed $N/(1 + 2\tau_{\text{int}}) = 4000/17 \approx 235.29$ as an alternative formula mentioned, which does not match the stated range.

5. ✖ **C5** (Page 5, Section 5.5)
   - **Claim:** "NUTS reaches $|Z| < 0.1\,\sigma$ at ~ 11 s, whereas the cobaya RW-MH chain needs ~ 103 s ... roughly a $\sim 100\times$ wall-for-wall advantage"
   - **Checks:** speedup_ratio
   - **Verdict:** FAIL
   - **Notes:** Computed speedup $103/11 \approx 9.36$, not consistent with $\sim 100\times$.

6. ✔ **C6** (Page 5, Section 5.5)
   - **Claim:** "The asymptotic ESS-accumulation rates are ~ 10 ESS/s (NUTS) vs ~ 2.3 ESS/s (cobaya RW-MH), a factor of ~ 4"
   - **Checks:** ratio_check
   - **Verdict:** PASS
   - **Notes:** Computed $10/2.3 \approx 4.35$, consistent with "~ 4" within tolerance.

7. ✔ **C7** (Page 5, Section 5.5)
   - **Claim:** "... because the autocorrelation length of the RW-MH chain ($\tau_{\text{int}} \sim 20$ ...) is much longer than the NUTS chain's ($\tau_{\text{int}} \approx 8$)."
   - **Checks:** autocorr_ratio
   - **Verdict:** PASS
   - **Notes:** Computed ratio $20/8 = 2.5$.

8. ✔ **C8** (Page 6, Table 1 caption)
   - **Claim:** "χ^2_{bf} is quoted at 6 degrees of freedom (8 bandpowers minus 2 fitted parameters)."
   - **Checks:** degrees_of_freedom_subtraction
   - **Verdict:** PASS
   - **Notes:** $8 - 2 = 6$ matches reported dof.

9. ✔ **C9** (Page 6, Table 1 (L-BFGS-B rows))
   - **Claim:** Bestfit shows "12.3/6" and caption states 6 dof; confirm that the table's "12.3/6" corresponds to $\chi^2 = 12.3$ with dof=6 (not a computed ratio).
   - **Checks:** format_consistency_check
   - **Verdict:** PASS

- **Notes:** Parsed numerator/denominator consistent; reduced χ^2 would be $\mathbf{12.3/6 = 2.05}$, but the intended display meaning cannot be verified from arithmetic alone.

10. ⚠ **C10** (Page 6, Table 1 (RW-MH baseline))
    - **Claim:** Baseline RW-MH: "$n_{\text{eff}} \approx \mathbf{1900}$ ($\sim \mathbf{5300}$ accepted steps with acceptance $\sim \mathbf{13\%}$)"
    - **Checks:** accepted_vs_total_steps
    - **Verdict:** UNCERTAIN
    - **Notes:** Implied total proposals $\approx \mathbf{5300/0.13 \approx 40,769.23}$, but no explicit total was provided to confirm.
11. ✔ **C11** (Page 6, Table 1)
    - **Claim:** Check consistency between NUTS ESS $\sim \mathbf{1400}$ and wall 200 s with implied ESS rate; compare to claimed ~10 ESS/s rate.
    - **Checks:** rate_from_total
    - **Verdict:** PASS
    - **Notes:** Implied rate $\mathbf{1400/200 = 7}$ ESS/s, within the loose tolerance of the claimed ~10 ESS/s.
12. ✔ **C12** (Page 5, Section 5.5)
    - **Claim:** Gold-standard chain: "500 warmup + 4000 samples × 4 chains ... ESS $\sim \mathbf{1400}$"
    - **Checks:** ess_upper_bound_check
    - **Verdict:** PASS
    - **Notes:** Total post-warmup draws $= \mathbf{4000 \times 4 = 16,000}$; ESS=1400 is below this bound.
13. ✔ **C13** (Page 1 Abstract; Page 2 Section 2; Page 2 Figure 1 caption)
    - **Claim:** Cumulative acceleration claim: "from $\sim \mathbf{30}$ s ... to $\sim \mathbf{5\ ms}$" and "$\sim \mathbf{6000}\times$ acceleration"
    - **Checks:** speedup_factor
    - **Verdict:** PASS
    - **Notes:** $\mathbf{30/0.005 = 6000}$ exactly.
14. ✔ **C14** (Page 2, Section 2 (item iv))
    - **Claim:** "The $\sim \mathbf{40}\times$ gain over the previous generation ..." comparing $\sim \mathbf{200\ ms}$ to $\sim \mathbf{5\ ms}$ (fixed-cosmology closure) or to $\sim \mathbf{20\ ms}$ (full pipeline).
    - **Checks:** speedup_factor
    - **Verdict:** PASS
    - **Notes:** $\mathbf{200/5 = 40}$ matches the claimed $\sim \mathbf{40}\times$ gain; $\mathbf{200/20 = 10}$ does not, indicating the claim aligns with the fixed-cosmology timing.

15. ✔ **C15** (Page 6, Figure 4 caption)
    - **Claim:** "Wall time per cosmology: ~ 0.4 s L-BFGS-B + ~ 40 s NUTS (8000 samples $\times$ 4 chains)."
    - **Checks:** samples_count_multiplication
    - **Verdict:** PASS
    - **Notes:** $8000 \times 4 = 32,000$ total draws.
16. ✖ **C16** (Page 6, Table 1 vs Page 6, Figure 4 caption)
    - **Claim:** NUTS wall time: Table 1 lists 200 s, while Figure 4 caption states $\sim$ 40 s NUTS (8000 samples $\times$ 4 chains).
    - **Checks:** cross_reference_consistency
    - **Verdict:** FAIL
    - **Notes:** Computed ratio $200/40 = 5$; discrepancy requires contextual reconciliation (e.g., different budgets/settings).
17. ⚠ **C17** (Page 6, Table 1 caption)
    - **Claim:** "publication-grade budget (500 warmup + 4000 samples, R-hat ≤ 1.003, ESS ~ 1400)" and Table 1 NUTS wall is 200 s; compute total post-warmup draws and compare to ESS.
    - **Checks:** ess_fraction
    - **Verdict:** UNCERTAIN
    - **Notes:** Descriptive recomputation: total post-warmup draws = $4000 \times 4 = 16,000$; ESS fraction = $1400/16,000 = 0.0875$. No explicit fraction claim was provided to verify.
18. ✔ **C18** (Page 3, Section 3.1)
    - **Claim:** P_k emulator grid: "1000 points spanning $k \in [5 \times 10^{-4}, 10]$ Mpc^{-1}, extrapolate to $k_{\min} = 10^{-4}$ Mpc^{-1}."
    - **Checks:** range_order_check
    - **Verdict:** PASS
    - **Notes:** Ordering holds: $1 \times 10^{-4} < 5 \times 10^{-4} < 10$, and $n_{\text{points}} = 1000$ is positive.

Limitations

- Only parsed text and embedded figure/table text from the provided PDF pages were used; no external data, code, or repositories were accessed.
- No values were extracted from plotted curves or points in figures (pixel-based extraction disallowed); only textual numerals were audited.
- Many performance, convergence, and accuracy claims depend on runtime logs, chains, datasets, or implementation details not contained in the PDF; these are listed as unverified.