# *Skepthical* review: *ST-based Component Separation of tSZ in the FLAMINGO Lensed Simulations*

## Summary

This manuscript evaluates scattering-transform (ST)–based component-separation strategies for recovering the thermal Sunyaev–Zel'dovich (tSZ) signal in FLAMINGO lensed simulation patches with realistic Simons Observatory (SO) + Planck noise. A six-frequency harmonic ILC is used as the main baseline under a careful, largely fair comparison protocol: explicit noise realisations are added to noiseless stacked simulations; beams are handled via per-patch effective ILC beams; and noise-debiased split-cross spectra are used for power-spectrum evaluation (Sec. 2.1–2.3, 4.1). Two ST-flavoured approaches are tested: FoCUS (Sec. 3.4), an ILC-anchored multi-frequency ST refinement aimed at suppressing CIB-like structure in cross-frequency residuals, and STsep (Sec. 3.3), a ScatCov/ST-optimisation approach.

Empirically, FoCUS provides only sub-percent improvements over ILC on the reported suite of Gaussian and non-Gaussian metrics, and is essentially indistinguishable from ILC in ST statistics and power spectra under the chosen configuration (Sec. 4.3, 4.6–4.7). By contrast, the strongest results come from a specific STsep configuration that is initialised from the six-frequency ILC output and stabilised by simulation-derived amplitude priors (mean/variance) plus a FLAMINGO-based contamination ensemble: in this hybrid ILC+ST setup, STsep substantially improves map-space errors and non-Gaussian/tail metrics on 20 noisy 5°×5° patches at 150 GHz (Sec. 4.6–4.7). A negative result is also clearly shown: an ILC-free, SED-difference–driven three-frequency STsep variant fails under realistic SO noise because the high-noise 217 GHz channel dominates (Sec. 4.4).

The paper's main added value is a careful realistic-noise evaluation of ST-based ideas beyond idealised/noiseless settings, and evidence that morphology-sensitive optimisation can meaningfully improve cluster-core/tail recovery when combined with ILC initialisation and simulation-calibrated amplitude constraints. The key limitations to address for a robust and appropriately scoped claim are: (i) clearly positioning the best-performing method as a hybrid ILC+ST pipeline (not an ILC-free replacement), (ii) quantifying how dependent the gains are on truth-derived priors/contamination ensembles and on hyperparameters, (iii) adding uncertainty/significance estimates across the finite patch sample, and (iv) strengthening the "bigger-picture" interpretation by connecting the improved map metrics to harmonic-space correlation/transfer-function behaviour and split-to-split determinism (Sec. 4.1, 4.6, 5–6).

## Strengths

- Careful and explicit experimental protocol for realistic evaluation: explicit addition of SO/Planck noise to noiseless stacked simulation products; clear split definitions; and use of split-cross spectra to avoid additive noise auto-bias (Sec. 2.1–2.3, 4.1, Eq. (14)).

- Fair beam policy is treated seriously (e.g., truth tSZ is beam-matched to each patch's ILC effective beam for map-level comparisons), improving interpretability of recovered spectra and map metrics (Sec. 2.3, 4.1, 5.1).
- Balanced reporting of positive and negative results: FoCUS yields only marginal gains; ILC-free three-frequency STsep fails at realistic noise; the successful configuration is clearly identified in Results (Sec. 4.3–4.4, 4.6, 6).
- Thorough multi-metric evaluation spanning Gaussian map metrics (r, RMS, MAE, power spectra) and non-Gaussian diagnostics (ScatCov distance, KS tests, tail statistics), which is appropriate for tSZ where rare peaks matter (Sec. 4.1–4.7).
- The ST anomaly map idea is potentially useful for localising under-recovered cluster cores that may not be obvious from global r or C$\ell$ comparisons alone (Sec. 3.5, Fig. 3).
- Clear high-level motivation: testing whether morphology-aware ST/ScatCov constraints can add value beyond second-order/linear estimators under realistic noise, which is timely for SO-era analyses (Sec. 1, 6).

## Major issues

1. The manuscript is conceptually ambiguous about what is "ILC-free" versus "ILC-anchored" in STsep, and the strongest reported gains come from a hybrid pipeline (ILC initialisation + simulation-trained amplitude priors + contamination ensemble), not from the ILC-free STsep described in parts of Sec. 3.3 and highlighted in Sec. 4.4. As written, some framing in Sec. 1 and Sec. 6 can be read as demonstrating an ILC-free replacement for ILC under realistic noise, which is contradicted by the three-frequency failure and by the reliance on ILC initialisation in Sec. 4.6.

   *Recommendation:* Make the distinction explicit and consistent throughout Sec. 3.3, Sec. 4.4, Sec. 4.6, Sec. 5, and Sec. 6: (i) a purely ILC-free, SED-difference–driven STsep variant that fails at realistic SO noise; versus (ii) the practical best-performing hybrid that is initialised from the six-frequency ILC map and stabilised by amplitude priors and a contamination ensemble. Update the Abstract/Sec. 1/Sec. 6 to describe the successful method explicitly as "ILC-initialised STsep" (or similar), and avoid calling that configuration "ILC-free." Also state unambiguously what s0 is in the $\lambda a\|s-s0\|^2$ term for the canonical six-frequency results (Sec. 3.3, Sec. 4.6), since different parts of the text suggest different initialisations.

2. The contribution of each ingredient in the best-performing STsep configuration is not identified (ST statistics vs ILC anchoring vs amplitude priors vs contamination ensemble). This limits interpretability ("why does it work?") and makes it hard to generalise or compare fairly to alternative post-processing/denoising approaches.

   *Recommendation:* Add an ablation study (ideally a compact table in Sec. 4.6 or an Appendix) on the same 20 patches at 150 GHz, reporting r, RMS (or $\sigma e/\sigma t$), D_ST, KS distance, and tail-recovery statistics for: (a) ILC baseline; (b) STsep initialised

from ILC with priors (current best); (c) STsep initialised from ILC without mean/variance priors (or with weakened priors); (d) STsep initialised from SED-difference with priors; (e) STsep initialised from noise/zero with priors. If feasible, include a run where the ST loss is removed but priors/proximity remain (to quantify what portion is due to "ST morphology" versus amplitude anchoring and proximity regularisation).

3. STsep's performance depends critically on truth-derived amplitude priors ($\mu$, $V$) and on a FLAMINGO-based contamination ensemble (Sec. 3.3, 4.5–4.6), but robustness to prior/ensemble mis-specification is not quantified. This is a central "bigger-picture" risk for transfer to real data, where tSZ/foreground statistics and noise mismatch the training suite.

   *Recommendation:* Add a robustness/mismatch experiment (Sec. 4.6, Sec. 5.2, or Appendix): perturb $\mu$ *and* $V$ by $\pm$(10–50)% (and/or rescale contamination-ensemble amplitudes) and quantify changes in r, σe/σt, D_ST, KS, and tail recovery. If variants/sub-volumes exist, estimate priors on one subset and apply to another to emulate mismatch. Use the results to revise Sec. 6 to state clearly that current gains are conditional on reasonably well-matched simulation-assisted priors, and outline how such priors might be specified in practice (e.g., suites of imperfect foreground/tSZ simulations, cross-checks with external data).

4. Claims of "beating ILC and FoCUS on every metric" rely primarily on means over 20 patches without rigorous uncertainty quantification or hypothesis testing, despite visible patch-to-patch scatter (Sec. 4.1–4.2, 4.6–4.7). FoCUS-vs-ILC differences are explicitly within scatter; the same standard should be applied to STsep-vs-ILC comparisons, especially for modest absolute changes (e.g., r from ≈0.14 to ≈0.17).

   *Recommendation:* Augment Sec. 4.6–4.7 and the relevant figures/tables (e.g., Figs. 4, 9, 11; Table 4 if present) with uncertainty estimates across patches: mean±1σ, standard errors, and/or bootstrap confidence intervals for each key scalar metric. For direct comparisons, report the fraction of patches where STsep improves over ILC (paired comparison) and include a paired test (paired t-test or Wilcoxon signed-rank) for the principal metrics. Revise wording in the Abstract/Sec. 1/Sec. 6 to match the quantified statistical strength (significant vs modest).

5. The power-spectrum discussion indicates split-cross power remaining above beam-matched truth (factors ~2–5) due to residual contaminants common to splits, but the manuscript's main headline metrics are map-space (r, RMS, KS, tails). Without harmonic-space correlation/transfer-function diagnostics, it is hard to reconcile "better maps" with "excess power," and to interpret implications for typical tSZ science analyses (Sec. 4.1, Fig. 4, Sec. 5.1).

   *Recommendation:* Add harmonic-space recovery diagnostics alongside split-cross power: (i) a binned multipole-dependent correlation coefficient, e.g. ρ ℓ = Cℓ^{m×t}/sqrt(Cℓ^{m×m}Cℓ^{t×t}); and/or (ii) a transfer function estimate Tℓ =

$C\ell\hat{}\{m\times t\}/C\ell\hat{}\{t\times t\}$. These separate "extra residual power" from "true tSZ recovery." Also quantify split-to-split determinism/noise sensitivity by reporting Var(m_splitA − m_splitB) (or an equivalent) for each method (ILC, FoCUS, STsep), to support statements that STsep outputs are more deterministic across splits.

6. Beam handling and unit consistency are not fully well-defined in key multi-frequency equations and constraints. As written, Eq. (12) and SED-difference initialisations imply direct subtraction across bands without explicit beam matching, despite significantly different beams (Table 1). Similarly, constraints (9)–(10) compare $a_\nu \hat{y} + c_\nu$ to $d_\nu$ without explicit beam operators, and the unit/meaning of the optimisation variable s is ambiguous (Compton-y vs μK_CMB at 150 GHz), which also interacts with the ILC SED constraint normalisation ambiguity in Eq. (6) (Sec. 2.3, Sec. 3.1–3.4).

   *Recommendation:* Make the forward model explicit with beam operators (e.g., $d_\nu = B_\nu \star (a_\nu\, y + \ldots) + n_\nu$), then state clearly which objects are beam-equalised and at what stage (pre-smoothing all channels vs embedding $B_\nu$ in the constraints). Resolve the ILC SED normalisation ambiguity by explicitly defining the SED vector used in Eq. (6) (e.g., a := a_tSZ/a_tSZ(150) so $w\hat{}T\, a = 1$, or else use $w\hat{}T\, a = a_{150}$ and adjust the closed form accordingly). Finally, state explicitly whether s denotes (a) $\hat{y}$ (dimensionless), (b) the 150 GHz tSZ temperature in μK_CMB, or (c) a 150-normalised amplitude, and adjust Eq. (12) and the $a_\nu$ notation to be dimensionally consistent. Include an explicit statement of beam convention for STsep comparisons (why truth is smoothed to B_eff rather than the 150 GHz beam), and provide the distribution of B_eff (e.g., FWHM across patches) in an Appendix to interpret high-ℓ behaviour.

7. STsep optimisation stability and hyperparameter dependence are acknowledged (including "catastrophic divergence" without priors) but not systematically quantified, limiting reproducibility and transfer to other noise regimes (Sec. 3.3, Sec. 4.6).

   *Recommendation:* Provide a compact sensitivity analysis (Sec. 4.6 or Appendix): vary $\lambda_c$, $\lambda_a$, learning rate, number of steps, and (optionally) N_ens batching, and report how RMS ratio, D_ST, KS, and tail recovery respond. Highlight stable ranges and failure modes. If full scans are too costly, include a small set of representative alternate configurations (e.g., ×0.5 and ×2 for key weights, shorter/longer runs) to demonstrate that conclusions do not hinge on a narrowly tuned setting.

8. FoCUS is positioned as a methodological contribution (Sec. 3.4) but its negative/marginal result is under-diagnosed: there is limited exploration of λ, ST statistic choices, or frequency-pair selection, and it is unclear whether FoCUS is inherently weak under realistic noise or simply under-tuned/under-specified.

   *Recommendation:* Either (a) add a compact FoCUS characterisation (Sec. 4.7): scan λ over $\sim 10^{-4}$–1, report the update size $\|s\_FoCUS - s\_ILC\| / \|s\_ILC\|$ and changes in r/RMS/D_ST/KS/tails, and try at least one alternative residual choice (e.g., combining Δ90,217 and Δ150,217 or another pair motivated by CIB/tSZ contrast); or (b) ex-

plicitly reframe FoCUS as an exploratory proof-of-concept/negative result, shorten Sec. 3.4 accordingly, and tone down claims in Sec. 1 and Sec. 6 to match the demonstrated utility.

9. The ST anomaly diagnostic is highlighted as a key contribution (Sec. 3.5, Fig. 3, Sec. 6), but it is currently qualitative and partially supervised (learned direction \hat d from training tiles). Its practical meaning, calibration, and robustness to noise/beam changes are not quantified, making it hard to evaluate beyond visualisation.

   *Recommendation:* Add a quantitative evaluation in Sec. 5.3 or Appendix: define ground-truth "cluster" masks (halo catalogue or truth |y| thresholds) and compute ROC/PR curves and AUC for anomaly-score detection of cluster regions, comparing anomaly maps derived from ILC vs STsep outputs. Clarify whether the diagnostic uses the same training/priors as STsep and discuss any circularity. If such analysis is infeasible, explicitly downgrade claims in Sec. 3.5 and Sec. 6 to describe the diagnostic as exploratory/qualitative and defer calibration to future work.

10. Several implementation and reproducibility details remain implicit (ST configuration, contamination ensemble construction/batching, spectra details), which may prevent independent re-implementation without access to code (Sec. 2.2–2.3, 3.1–3.3, 4.1).

    *Recommendation:* Add an "implementation checklist" (Appendix is fine) specifying: the full ST/ScatCov configuration (orders S1–S4 used, scales/orientations, any orientation averaging, normalisation/self-normalisation); which coefficients enter $\Phi$ in Eq. (11) vs FoCUS Eq. (13) vs D_ST; contamination ensemble selection and normalisation across frequencies (including whether noise is added and how); batching over N_ens during optimisation (batch size, whether ensemble moments are recomputed per step); details of map apodisation/pixel window in power spectra; and explicit $\ell$-bin edges/centres for "24 log-spaced bins over $500 \leq \ell \leq 6000$." If code will be released, provide a repository URL and version/commit; otherwise provide enough detail to replicate results.

## Minor issues

1. Definition of the ScatCov/$\Phi$ feature vector and which coefficients are used is spread across Sec. 3.2–3.3 and is not fully explicit; likewise the relationship between $\Phi(\cdot)$ and $\mathrm{ST}(\cdot)$ notation in Eq. (13) is unclear.

   *Recommendation:* In Sec. 3.2, give a compact explicit definition of $\Phi$ as used in this paper (orders, scales, orientations, any averaging, normalisation/standardisation). Then reference that same $\Phi$ in Sec. 3.3 and Sec. 3.4 when defining losses (Eqs. (11), (13)). If $\mathrm{ST}(\cdot)$ in Eq. (13) is identical to $\Phi(\cdot)$, use one symbol consistently; otherwise define $\mathrm{ST}(\cdot)$ and how it differs.

2. Some key ILC-baseline contextual claims are qualitative (e.g., "r≈0.14 saturates the correlation budget" or that more sophisticated ILC variants show "no significant improvement"), making it harder to judge baseline strength and headroom (Sec. 3.1, 4.6, 5.1).

   *Recommendation:* Provide a short quantitative baseline comparison: report performance deltas for any tested ILC variants (ensemble-averaged, Fourier-binned) and/or include one additional standard baseline if feasible (e.g., a constrained ILC/NILC-like variant). Alternatively, give a simple upper-bound/benchmark (e.g., Wiener/Fisher estimate) to contextualise achievable r and σe/σt on these patches under the stated noise/foreground model.

3. The three-frequency STsep negative result (Sec. 4.4) is described mainly via r/RMS on only 6 patches, with incomplete specification of priors/hyperparameters and limited ST-native diagnostic reporting.

   *Recommendation:* In Sec. 4.4, explicitly state which priors ($\mu$, $V$) and weights (λa, λc) are used; describe any alternative initialisations attempted (e.g., 3-frequency ILC seed, zero seed) and outcomes; and, if feasible, report at least D_ST/KS/tail metrics even in failure mode. Justify the N=6 choice and comment on stability if more patches were used.

4. Figure presentation is often not stand-alone quantitative: missing numeric colorbars/units in key map figures (e.g., Fig. 1), inconsistent scaling across panels, and limited depiction of residuals.

   *Recommendation:* Add numeric colorbars with explicit units (Compton-y vs μK_CMB), standardise color scales within comparable panels, and include residual panels (method − truth_beam) with matched scales and summary numbers. Add scale indicators/pixel size where helpful, and embed clear row/column labels in the figure rather than relying on caption text.

5. Many plots summarising patch ensembles do not show uncertainty/variability clearly (e.g., mean curves without bands), and axes/captions sometimes omit units/normalisations or sample sizes (Figs. 2, 4–6, 9–10).

   *Recommendation:* For ensemble curves, overlay mean±1σ (or bootstrap CI bands) and clearly label axes with units/normalisations. In captions, state N_patches and whether curves are means/medians, and annotate reference lines or baselines where small differences matter (optionally add difference panels/insets when curves overlap).

6. Some metric definitions are slightly ambiguous: e.g., whether D_ST uses $\|\cdot\|2$ or $\|\cdot\|2\hat{}2$, and how z-scores/tail thresholds are defined and normalised when computing "cluster pixel" recovery (Sec. 4.2, 4.7, 5.3).

*Recommendation:* Define D_ST unambiguously as either ‖ΔΦ‖2/‖Φtruth‖2 or ‖ΔΦ‖2^2/‖Φtruth‖2^2 and use consistently. For tail metrics, define z (mean/variance computed from which map/region), and state explicitly whether truth and recovered maps are standardised using the same reference statistics before thresholding.

## Very minor issues

1. Typographical/formatting inconsistencies appear throughout: split words (e.g. “reach ing”), HTML entities (>, <), inconsistent “SO/Planck” vs “SQ/Planck,” inconsistent unit spacing, and minor typos (e.g. “three_freq.cannual”) (Sec. 1, Sec. 4.1, Fig. 1, Sec. 4.4).

   *Recommendation:* Perform a final proofreading/LaTeX cleanup pass: replace HTML entities with proper math symbols, standardise SO/Planck naming, fix typos and split words, and enforce consistent unit formatting (e.g., 150\,\mathrm{GHz}).

2. Minor notation/cross-reference inconsistencies: e.g., D_ℓ^{m†} mentioned without definition; inconsistent split-cross spectrum notation across sections; inconsistent Fig./Figure and Sec./Section styling (Sec. 4.1, Sec. 5.1–5.3).

   *Recommendation:* Add brief first-use definitions for all symbols, standardise split-cross notation across Sec. 4–5 (or add a notation recap at the start of Sec. 4), and use a uniform cross-reference style throughout.

3. Some numeric/procedural statements are hard to verify without additional specifics (e.g., “24 log-spaced bins over 500≤ℓ≤6000” without bin edges; FoCUS coefficient difference (a90−a217)≈1.67 without listing a90 and a217).

   *Recommendation:* Provide explicit ℓ-bin edges/centres (table or caption) and list the av values (or a reference table) used to compute reported SED differences so readers can reproduce simple arithmetic checks.

4. A few references are incomplete or inconsistently formatted (e.g., missing journal/arXiv details, stray punctuation) (References section).

   *Recommendation:* Ensure all references have complete, consistently formatted bibliographic fields (journal/arXiv IDs, year/volume/pages where applicable) and that in-text citations match the bibliography entries.

## Key statements and references

- ⚠ **Internal Linear Combination (ILC) methods minimize the variance of a combined multi-frequency map subject to preserving a target component’s spectral energy distribution, with extensions such as Needlet ILC and constrained ILC introducing scale-dependent weights and explicit deprojection of contaminant SEDs, but all ILC-type methods fundamentally operate**

**only on second-order statistics (power spectra) and therefore cannot exploit the non-Gaussian higher-order correlations characteristic of tSZ and CIB signals.**

- *Reference(s):* Tegmark M., 1996, Eriksen H. K., et al., 2004, Remazeilles M., Delabrouille J., Cardoso J.-F., 2011
- *Justification:* Remazeilles, Delabrouille, Cardoso, 2011 explicitly defines ILC as a variance-minimizing linear combination with unit response to the target SED (Eq. 3), introduces needlet (scale/pixel-localized) ILC (Sec. 2.3), and describes constrained/multidimensional ILC with explicit deprojection of CMB and SZ (Sec. 3.5), all using covariance matrices R (second-order). However, the papers do not state that ILC methods cannot exploit higher-order (non-Gaussian) statistics or discuss CIB non-Gaussianity; this limitation is not directly supported. Eriksen H. K., et al., 2004 is unrelated to ILC.

- ⚠ **The scattering transform (ST), originally developed as a cascade of wavelet filters and modulus operators, has been applied in cosmology to galaxy clustering, weak lensing, and foreground separation, and its key advantage over power-spectrum-based methods is the ability to capture non-Gaussian phase information and higher-order correlations without assuming a specific statistical model for the field.**

- *Reference(s):* Mallat, 2012, Bruna and Mallat, 2013, Cheng S., Ting Y.-S., Ménard B., Bruna J., 2020
- *Justification:* Bruna and Mallat, 2013 defines the scattering transform as a cascade of wavelet convolutions and modulus/averaging operators and shows it captures higher-order moments beyond the power spectrum, enabling discrimination of non-Gaussian structures (Abstract; Secs. 2.3, 3.2; Fig. 5). However, this paper does not discuss cosmology applications such as galaxy clustering, weak lensing, or foreground separation. Thus, only part of the statement is supported.

- ✖ **The STsep separator implemented here follows the formalism introduced for dust polarisation and dust–CIB separation, in which an estimate of the target signal is obtained by enforcing that (i) mock data built from the estimate plus a contamination ensemble match the observed data in scattering-transform space, and (ii) the observed data minus the estimate match the contamination ensemble in the same space, thereby pinning down the target component without recourse to ILC weights.**

- *Reference(s):* Régaldo-Saint Blancard et al., 2021b, Régaldo-Saint Blancard et al., 2021a, Auclair et al., 2024
- *Justification:* Neither Régaldo-Saint Blancard et al., 2021b nor Auclair et al., 2024 describe an STsep scattering-transform-based component separation. Régaldo-Saint Blancard et al., 2021b presents MCMC fitting of a turbulence model using power spectra and structure functions, not scattering-transform constraints or component

separation without ILC weights. Auclair et al., 2024 focuses on PTA gravitational-wave backgrounds and related astrophysical/cosmological implications, with no discussion of dust polarisation/CIB separation or ST-based matching. Thus the statement is not supported by the attached papers.

- ✖ **Previous ScatCov-based separation work on noiseless or effectively noise-free data suggested that a spectral-energy-distribution (SED) difference combined with ST consistency could outperform ILC at three frequencies, but when realistic SO noise levels are included at 90/150/217 GHz the 217 GHz noise dominates the SED-difference initialisation and the multi-frequency STsep cannot recover the truth tSZ signal, demonstrating that realistic instrumental noise is a quantitative gate for ILC-free, limited-frequency ScatCov separation.**

- *Reference(s):* Régaldo-Saint Blancard et al., 2021a, Auclair et al., 2024, Tsouros et al., 2026

- *Justification:* Tsouros et al., 2026 applies scattering-transform statistics to single-frequency (353 GHz) polarized dust separation, comparing mainly to GNILC maps, and does not discuss SED-difference initializations, ILC performance at three frequencies, Simons Observatory noise at 90/150/217 GHz, or tSZ recovery. Auclair et al., 2024 concerns nanohertz gravitational waves and is unrelated. Hence the claim about multi-frequency ScatCov separation vs ILC and SO noise effects is not supported by these papers.

- ✖ **The STsep formalism adapted from earlier work, when combined with an ILC initialisation and target-mean/target-variance priors estimated from a disjoint truth-tSZ training set, yields a single-frequency 150 GHz separator that outperforms six-frequency ILC and FoCUS on FLAMINGO under realistic SO+Planck noise on all tested metrics, including pixel correlation, pixel RMS, ScatCov distance, KS distance, and extreme negative-tail cluster-pixel recovery.**

- *Reference(s):* Régaldo-Saint Blancard et al., 2021a, Auclair et al., 2024, Tsouros et al., 2026

- *Justification:* Tsouros et al., 2026 presents a scattering-transform based separation of polarized dust at 353 GHz from Planck data, comparing mainly to GNILC and using power/cross-spectra; it does not describe an STsep formalism with ILC initialization, tSZ-focused priors, single-frequency 150 GHz separation, FLAMINGO tests, SO+Planck noise, or the listed metrics (pixel correlation, RMS, ScatCov, KS, negative-tail recovery). Auclair et al., 2024 concerns pulsar-timing gravitational-wave backgrounds and is unrelated. Hence the claim is not supported by the attached papers.

- ⚠ **Earlier applications of scattering transforms to component separation and dust emission analysis demonstrated that ST-based statistics can separate polarized dust emission and distinguish dust from CIB structures,**

**motivating the present adaptation of these techniques to tSZ recovery in the FLAMINGO simulations.**

- *Reference(s):* Régaldo-Saint Blancard et al., 2021a, Régaldo-Saint Blancard et al., 2021b, Tsouros et al., 2026
- *Justification:* Tsouros et al., 2026 states that prior ST-based component-separation work separated polarized dust emission from instrumental noise in Planck data and that ST can distinguish Galactic dust from CIB structures (citing earlier applications). However, neither Tsouros et al., 2026 nor Régaldo-Saint Blancard et al., 2021b mention any adaptation to tSZ recovery or FLAMINGO simulations. Thus, only the first part of the statement is supported.

## Mathematical consistency audit

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

**Maths relevance:** substantial

The paper's core methods (ILC weighting, effective beam propagation, split-cross spectra, and ScatCov-based optimisation objectives) are mathematical and depend on consistent definitions of the tSZ SED vector, units (Compton-y vs µKCMB), and the treatment of per-frequency beams in multi-frequency combinations and constraints. Several central equations are individually plausible, but key parts are not internally checkable due to missing or ambiguous normalisation and beam-operator definitions.

### Checked items

1. ✔ **Observed map construction** (Eq. (1), Sec. 2.2, p.2)
    - **Claim:** Observed maps are noiseless stacked sky plus an explicit noise realisation: dν(x) = s_stacked_ν(x) + nν(x).
    - **Checks:** symbol consistency, units/dimensional consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** s_stacked_ν and nν are in the same units (µKCMB)., Noise is additive in the map domain.
    - **Notes:** As a definition, Eq. (1) is consistent; later statements about unit conversions aim to ensure unit alignment.
2. ✔ **tSZ SED definition** (Eq. (4), Sec. 2.3, p.2)
    - **Claim:** Defines atSZ(ν) = x e^x/(e^x−1) − 4 with x = hν/(kB TCMB).
    - **Checks:** definition consistency, symbol definition completeness
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Non-relativistic tSZ frequency dependence in thermodynamic temperature units., TCMB is constant and specified.

- **Notes:** All symbols in x are defined; atSZ is explicitly dimensionless and is later used as a scaling factor.

3. ✔ **ILC effective beam expression** (Eq. (5), Sec. 2.3, p.2)
    - **Claim:** The tSZ component in the ILC output inherits an effective beam Beff(ℓ) = Σν wν (aνtSZ/a150tSZ) Bν(ℓ).
    - **Checks:** algebra between definitions, symbol consistency
    - **Verdict:** PASS; confidence: medium; impact: critical
    - **Assumptions/inputs:** Each frequency map contains a tSZ contribution proportional to aνtSZ and convolved by Bν., ILC output is ŝ = Σν wν dν with no pre-beam-matching., The ILC constraint preserves the target tSZ response at 150 GHz (see text).
    - **Notes:** Algebra is correct if the constraint is Σν wν aνtSZ = a150tSZ. However, this hinges on how a is normalised/defined in Eq. (6); see the ILC-weight item.

4. ⚠ **ILC weight formula vs stated constraint** (Eq. (6) and surrounding text, Sec. 3.1, p.2)
    - **Claim:** Weights minimizing variance with SED preservation are w = C−1 a / (aT C−1 a), and the weights preserve tSZ SED as Σν wν aνtSZ = a150tSZ.
    - **Checks:** derivation logic (implied), constraint/normalisation consistency, symbol definition consistency
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** Standard constrained quadratic minimization under linear response constraint., a is the tSZ SED vector.
    - **Notes:** Eq. (6) corresponds to a unity-response constraint wT a = 1 (with that same a). The text asserts w preserves response equal to a150tSZ, which requires either a to be normalised by a150 or a different constraint/derivation. The paper does not explicitly state which convention is used, so internal consistency cannot be confirmed.

5. ✔ **Scattering first-order coefficient** (Eq. (7), Sec. 3.2, p.3)
    - **Claim:** S1(j,l) is the spatial mean of the modulus of wavelet-filtered maps.
    - **Checks:** definition consistency, notation consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Wavelet convolution Wj,l ⋆ x(u) is well-defined on the patch with stated boundary conditions.
    - **Notes:** Equation is a standard definitional form; symbols are defined locally.

6. ✔ **Scattering second-order coefficient** (Eq. (8), Sec. 3.2, p.3)
    - **Claim:** S2(j,l) is the spatial mean of squared modulus of wavelet coefficients.

- **Checks:** definition consistency, notation consistency
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** Same conventions as Eq. (7).
- **Notes:** Equation is consistent with a variance/energy-like moment of wavelet coefficients.

7. ⚠ **STsep constraints (multi-frequency)** (Eqs. (9)–(10), Sec. 3.3, p.3)
   - **Claim:** An estimate $\hat{y}$ is constrained so that ST statistics of ($a_\nu \hat{y}$ + contamination) match data, and data minus $a_\nu \hat{y}$ matches contamination in ST space.
   - **Checks:** symbol/definition consistency, units/dimensional consistency, missing operator/omitted steps
   - **Verdict:** UNCERTAIN; confidence: high; impact: critical
   - **Assumptions/inputs:** $\Phi$ maps a field to a ScatCov coefficient vector., Contamination ensemble $\{c_{\nu,i}\}$ represents non-tSZ components plus noise., All compared quantities are in the same space (including beam response).
   - **Notes:** As written, comparisons $a_\nu \hat{y} + c_{\nu,i}$ versus $d_\nu$ ignore per-frequency beam operators, despite different stacked beams across $\nu$. The constraints are not well-defined unless $\hat{y}$ is implicitly beam-convolved per $\nu$ or all maps are pre-matched to a common beam, neither of which is stated here.

8. ✔ **STsep single-frequency loss** (Eq. (11), Sec. 3.3, p.3)
   - **Claim:** Minimises squared differences of ST coefficients between (s+ci) and data, and between (data−s) and contamination mean, plus a prior term.
   - **Checks:** algebra/structure of objective, units consistency, notation consistency
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Single-frequency at 150 GHz; s, d150, ci are all maps at the same beam and units., Ensemble averaging $\langle \cdot \rangle_i$ is over contamination realisations.
   - **Notes:** Formally consistent as an objective function in one frequency channel. It avoids cross-frequency beam issues provided all operands live at the 150 GHz map resolution/beam.

9. ⚠ **SED-difference initialisation** (Definition of s0 in Sec. 3.3 and Sec. 4.4, p.3 and p.5–6)
   - **Claim:** Initialises $s0 = (a150/(a150 - a217))(d150 - d217)$ as a tSZ proxy.
   - **Checks:** algebraic derivation (implied), assumption consistency, beam/operator consistency
   - **Verdict:** UNCERTAIN; confidence: high; impact: critical

- **Assumptions/inputs:** Model dν = aν y + common contamination (CMB/kSZ) + noise., a217 approximately nulls tSZ (or is known)., d150 and d217 are directly subtractable.
- **Notes:** The algebra is correct under the stated simplified mixture model with identical beams/transfer functions for both channels. However, Table 1 indicates different beams at 150 and 217 GHz; the paper also states no pre-beam-matching is applied. Without an explicit beam-matching step (or including beams in the formula), the subtraction is not strictly consistent.

10. ⚠ **Prior term and claimed flat directions** (Lprior(s) definition, Sec. 3.3, p.4)
    - **Claim:** Adds Lprior(s) = λv(Var s − V*)^2 + λa‖s − s0‖^2 and re-centres s to target mean* $\mu$ because ScatCov is invariant to mean and (after self-normalisation) amplitude.
    - **Checks:** logic consistency, missing derivation/implementation dependence
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** The ScatCov implementation used is mean-invariant and largely amplitude-invariant under its internal normalisations.
    - **Notes:** The form of the prior is mathematically consistent, but the stated invariances of Φ depend on specific ScatCov normalisation details not shown in the paper; verification would require explicit definition of Φ and its normalisation.

11. ⚠ **FoCUS residual difference definition** (Eq. (12), Sec. 3.4, p.4)
    - **Claim:** Defines Δ90,217 = (d90 − a90 s) − (d217 − a217 s) to isolate non-tSZ residual structure.
    - **Checks:** units/dimensional consistency, beam/operator consistency, symbol meaning consistency
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** s represents the same underlying tSZ field used to predict contributions at both frequencies via aν., Subtractions are performed between like-beam, like-unit maps.
    - **Notes:** Without a clear statement of whether s is Compton-y or a 150 GHz μK map (and whether aν is absolute SED or relative scaling), dν − aν s is not verifiably unit-consistent. Additionally, d90 and d217 have different beams; predicting and subtracting aν s from each channel generally requires applying the channel beam to s (or pre-matching beams), which is not shown.

12. ⚠ **FoCUS loss definition** (Eq. (13), Sec. 3.4, p.4)
    - **Claim:** Minimises L(s) = ‖ST(Δ90,217)‖^2 + λ‖s − sILC‖^2.
    - **Checks:** objective well-posedness, notation consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate

- **Assumptions/inputs:** ST( · ) returns a coefficient vector and ‖ · ‖ is an L2 norm in that coefficient space., s and sILC are in the same pixel domain and units.
- **Notes:** As an optimisation objective it is structurally consistent, but because Δ90,217 in Eq. (12) is itself uncertain (units/beams), the loss inherits that ambiguity. Also, ST vs Φ notation is not reconciled.

13. ✔ **ST anomaly score definition** (Sec. 3.5, p.4)
    - **Claim:** Defines anomaly score per tile as a(x,y) = d̂^T Φ(x,y), with d̂ a unit-norm difference of mean ScatCov vectors between truth-tSZ tiles and contamination tiles.
    - **Checks:** linear-algebra consistency, definition clarity
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Φ(x,y) is a vector; d̂ is a vector of same dimension.
    - **Notes:** Inner-product definition is consistent; learning d̂ as a normalised mean-difference direction is well-defined.

14. ✔ **Split-cross spectrum estimator** (Eq. (14), Sec. 4.1, p.5)
    - **Claim:** Uses Dℓ^{xx} = ℓ(ℓ+1)/(2π) Re⟨ s̃^A_ℓ s̃^{B,*}_ℓ ⟩ as a noise-bias-free estimator when split noises are independent.
    - **Checks:** expectation/linearity sanity check, notation consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Noise in splits A and B is independent and additive., Fourier binning average ⟨ · ⟩ is over modes in a bin.
    - **Notes:** Under independence, E[nA nB*]=0 so additive noise auto-bias cancels in expectation. The estimator form is consistent with the flat-sky Dℓ convention stated.

15. ⚠ **Normalised ScatCov distance DST** (Sec. 4.6 and Sec. 6, p.6 and p.12)
    - **Claim:** Defines DST = ‖Φmethod − Φtruth‖2/‖Φtruth‖2 as an ST-native error metric.
    - **Checks:** dimensionless ratio check, notation clarity
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** Φ vectors are comparable (same normalisation, same coefficient ordering).
    - **Notes:** The ratio is dimensionless if the same norm is used in numerator and denominator, but the paper does not clarify whether ‖ · ‖2 denotes a norm or squared norm; this affects interpretation but not basic dimensional consistency.

**Limitations**

- Several key mathematical checks depend on implementation-specific details not defined in the text (exact ScatCov normalisation/invariances; precise definition of $\mathrm{ST}(\cdot)$ vs $\Phi(\cdot)$).
- Beam handling across frequencies is central to verifying multi-frequency formulas; the paper states no pre-beam-matching for ILC, but does not explicitly specify the corresponding beam treatment for STsep (multi-frequency) and FoCUS, preventing full symbolic verification.
- The audit does not validate any numerical approximations (e.g., a217 ≈ 0, approximate SED values) or empirical claims; it only assesses symbolic consistency.

## Numerical results audit

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

Of 16 numeric candidates checked, 14 PASS and 2 are UNCERTAIN (not enough supporting numbers to recompute/compare). No FAIL results were found. Key cross-references (Table 3 vs Figure 7(b); Table 4 vs narrative/abstract rounding) and multiple derived computations (percent reduction, ratios, pixel scale, delta-r, percent-from-ratios) were consistent within stated tolerances.

**Checked items**

1. ✔ **C1** (p1 Abstract)
    - **Claim:** Pixel RMS ratio improvement claim: STsep has pixel RMS $\sigma e/\sigma t = 1.22$ vs ILC's 3.42, described as "64% lower than ILC's 3.42".
    - **Checks:** percent_reduction_from_baseline
    - **Verdict:** PASS
    - **Notes:** Computed reduction = 64.3275%, consistent with claimed 64% within 1 percentage point rounding tolerance.
2. ✔ **C2** (p1 Abstract)
    - **Claim:** Extreme-tail improvement claim: "30× at $z < -8$, 14× at $z < -5$". Table 4 later lists ILC vs STsep recovery fractions.
    - **Checks:** ratio_claim_check
    - **Verdict:** PASS
    - **Notes:** Computed ratios: $z<-8$: 0.08/0.003=26.67 vs claimed 30× (within 20% rel tol); $z<-5$: 0.231/0.017=13.59 vs claimed 14× (within 20% rel tol).
3. ✔ **C3** (p2 §2.1)
    - **Claim:** Patch pixel count consistency: "256 × 256 pixel resolution" and later "256^2 = 65536 pixels".

- **Checks:** integer_product
- **Verdict:** PASS
- **Notes:** 256×256 equals 65536 exactly.

4. ✔ **C4** (p2 §2.1)
   - **Claim:** Pixel scale consistency: "5° × 5° patches at 256 × 256 pixel resolution (pixel scale ≈ 1.17′)".
   - **Checks:** unit_conversion_and_division
   - **Verdict:** PASS
   - **Notes:** Computed pixel scale = (5×60)/256 = 1.171875 arcmin, consistent with ≈1.17′.

5. ✔ **C5** (p2 Table 1)
   - **Claim:** Table 1 FWHM values: verify internal monotonic/expected comparisons mentioned elsewhere (e.g., "150 GHz channel has FWHM 1.4′").
   - **Checks:** cross_reference_value_match
   - **Verdict:** PASS
   - **Notes:** Text FWHM(150)=1.4′ matches Table 1; also 1.4′ > 1.0′ (217 GHz) as expected.

6. ✔ **C6** (p2 §2.3 Eq. (4))
   - **Claim:** tSZ SED scaling example: "atSZ(150) ≈ −2.60 and y = 10−5 corresponds to ≈ −26 µKCMB at 150 GHz."
   - **Checks:** scalar_multiplication_and_power_of_ten
   - **Verdict:** PASS
   - **Notes:** Computed (−2.60)×(1e−5)×(1e6 µK/K)=−26.0 µK, matching the stated example.

7. ✔ **C7** (p4 §3.3)
   - **Claim:** Prior variance parameter: "V∗ = (4.3 µK)^2".
   - **Checks:** square_value
   - **Verdict:** PASS
   - **Notes:** Computed 4.3^2 = 18.49 µK^2 for reference; no separate numeric V* value was provided to compare against.

8. ✔ **C8** (p4 §3.3 and p3 §3.3)
   - **Claim:** Training/ensemble sizes: contamination ensemble "Nens = 20" and optimisation hyperparameters "150 steps, batch size 4 per step"; compute number of batch items processed.
   - **Checks:** simple_count_multiplication
   - **Verdict:** PASS

- **Notes:** Computed total batch items = 150×4 = 600, consistent with the implied count.

9. ⚠ **C9** (p4 §3.4)
   - **Claim:** FoCUS coefficient difference: "(a90 − a217) ≈ 1.67". Verify arithmetic if a90 and a217 are provided elsewhere; if not, this is only checkable as a stand-alone constant match if repeated.
   - **Checks:** repeated_constant_match
   - **Verdict:** UNCERTAIN
   - **Notes:** Cannot recompute without a90 and a217; and only one occurrence is available here, so repetition consistency cannot be assessed.
10. ⚠ **C10** (p5 §4.1 Eq. (14))
    - **Claim:** Spectrum binning statement: "binned into 24 log-spaced bins over $500 \le \ell \le 6000$". Check that bin count and endpoints are consistent with any later reported bin edges if present (none in text).
    - **Checks:** parameter_consistency_across_mentions
    - **Verdict:** UNCERTAIN
    - **Notes:** No explicit bin edges/centers provided to compare against a generated 24-bin logspace between 500 and 6000 (which would imply 25 edges).
11. ✔ **C11** (p6 Table 3 and p8 Figure 7(b))
    - **Claim:** Three-frequency mean pixel correlations: Table 3 lists r=0.113 (ILC 3-freq), r=0.042 (SED-init), r=0.042 (STsep 3-freq), matching Figure 7(b) labels 0.113, 0.042, 0.042.
    - **Checks:** cross_reference_value_match
    - **Verdict:** PASS
    - **Notes:** All three table values exactly match the corresponding Figure 7(b) numeric annotations in the provided inputs.
12. ✔ **C12** (p6 §4.4)
    - **Claim:** SED-init and STsep (3-freq) RMS ratios: Table 3 lists $\sigma e/\sigma t$ = 54.707 vs 54.692; verify that the claim "converges to the same pixel correlation within numerical precision" is consistent at least for r (exactly equal) and that RMS are close.
    - **Checks:** difference_within_tolerance
    - **Verdict:** PASS
    - **Notes:** Pixel correlations match exactly (0.042 vs 0.042). RMS ratios differ by 0.015, which is small and within the stated <0.02 closeness criterion.
13. ✔ **C13** (p5 Table 2 and p5 §4.3)

- **Claim:** S1 correlations: Table 2 gives residual vs CIB S1 = 0.970 and vs CMB S1 = 0.831, identical for ILC and FoCUS; check equality across methods and alignment with text statements (r=0.97 vs 0.83).
- **Checks:** table_internal_consistency
- **Verdict:** PASS
- **Notes:** ILC equals FoCUS for both correlations (0.970 and 0.831), and 0.970 $>$ 0.831 as stated.

14. ✔ **C14** (p5 §4.3)
    - **Claim:** Residual S1 ratio range statement: "ranges from 1.15 at j=0 to 1.36 at j=2", implying 15–36% more amplitude; check percent conversion.
    - **Checks:** percent_from_ratio
    - **Verdict:** PASS
    - **Notes:** (1.15−1)×100=15% and (1.36−1)×100=36%, matching the stated 15–36% range.
15. ✔ **C15** (p5 §4.3)
    - **Claim:** Δr computation: "ILC residual's r = 0.970 is only Δr = 0.013 above the pure-tSZ baseline [r = 0.958]."
    - **Checks:** difference
    - **Verdict:** PASS
    - **Notes:** Computed Δr = 0.970−0.958 = 0.012, which is consistent with the reported 0.013 under rounding tolerance.
16. ✔ **C16** (p8 Table 4 vs p7 §4.6 text)
    - **Claim:** Table 4 means vs narrative: verify text-reported metrics match Table 4 (e.g., r: 0.144→0.175; σe/σt: 3.42→1.22; DST: 0.780→0.733; KS: 0.216→0.192).
    - **Checks:** cross_reference_value_match
    - **Verdict:** PASS
    - **Notes:** Table 4 σe/σt values (3.416, 1.215) round to the narrative 2-decimal values (3.42, 1.22) exactly; other listed Table 4 values match as given in the provided inputs.

**Limitations**

- Only the provided parsed text/images from the PDF were used; no external constants, code, or datasets were accessed.
- No checks requiring extraction of numerical values from plotted curves or image pixels were included.
- Several claims are qualitative (e.g., 'factor ~few', 'about 50%') without explicit numbers; only statements with explicit numerics were selected as candidates.

- Some internal consistency checks are limited to rounding-level verification because the paper reports rounded summary statistics (e.g., 1.22 vs 1.215).
- Some statements cannot be recomputed from the provided inputs because they depend on external functions/constants or additional underlying values not given (e.g., utils.jysr2uk(ν); physical constants; a90 and a217; explicit bin edges/centers; data/noise realisations and statistical assumptions).