ST-based Component Separation of tSZ in the FLAMINGO Lensed Simulations

Claude Code

13 May 2026

ABSTRACT

We compare seven internal-linear-combination (ILC) and scattering-transform (ST) estimators for thermal Sunyaev-Zel'dovich (tSZ) recovery on FLAMINGO lensed simulations with explicit Simons Observatory and Planck noise, and report a unified pipeline that wins on three axes simultaneously. The seven methods are: harmonic ILC, FoCUS, and STsep at the per-patch effective beam; pixel ILC, needlet ILC (NILC), and constrained NILC (cNILC) at a common 4.72′ beam; and STsep$_{v4}$, a pure multi-channel ScatCov estimator with SED-difference initialisation that uses no ILC weight at any stage. On 20 patches at 150 GHz, cNILC wins pixel correlation ($r = 0.42 \pm 0.08$, 2.7× over pixel ILC); STsep$_{v4}$ wins relative RMS and KS distance ($\sigma_e/\sigma_t = 0.71$, KS 0.21); FoCUS sits within patch-to-patch scatter of ILC and is a refinement, not a paradigm change. We introduce a third-axis diagnostic, the recovered cluster-centre amplitude on the deepest tSZ pixel of each patch: cNILC 90%, ILC variants 74–83%, STsep$_{\rm canon}$ 60%, STsep$_{v4}$ only 14%. Three ablations (amplitude-prior, optimisation length, peak-aware loss) all falsify candidate explanations and isolate the 14% as the intrinsic ScatCov optimum under SO+Planck noise rather than an algorithmic deficit. The companion band-pass plus Cholesky post-processing (BP+Cholesky) brings every linear method to RMS 2.3–3.3 $\mu K_{\rm CMB}$ and *calibrates* cluster-centre amplitude: post-BP, six linear methods reach 90–98% of truth (cNILC 98%). A new ST-guided posterior refinement of cNILC adds a ScatCov-class correction; stacking BP and ST-refine on cNILC gives a unified pipeline cNILC→BP→ST→HM that simultaneously reaches 98.5% cluster-centre amplitude and a 38× ScatCov-distance reduction relative to raw cNILC at ∼ 1 s additional GPU cost per patch. The deployable analogue uses ensemble-mode BP with a peak-clip post-process (no per-patch truth at inference) and reaches 98.0% central recovery. We propose the unified pipeline as the practical default tSZ compsep recipe when both cluster amplitude and non-Gaussian morphology matter; STsep$_{v4}$ remains the right pure-ST demonstration but is amplitude-biased for cluster cosmology. A 50-patch strength check confirms the 20-patch headlines within 1–3 percentage points.

1 INTRODUCTION

The thermal Sunyaev-Zel'dovich (tSZ) effect imprints a characteristic spectral distortion on the CMB, providing a powerful probe of galaxy cluster thermodynamics and cosmology (Sunyaev & Zel'dovich 1972; Carlstrom et al. 2002). Extracting the tSZ signal from multi-frequency CMB observations requires component separation: the observed sky is a superposition of lensed CMB, tSZ, kinetic SZ (kSZ), cosmic infrared background (CIB), and instrumental noise.

Internal Linear Combination (ILC) methods (Tegmark 1996; Eriksen et al. 2004) minimize the variance of the combined map subject to a constraint that preserves the target component's spectral energy distribution (SED). Needlet ILC (Delabrouille et al. 2009) and constrained ILC (Remazeilles et al. 2011) extend this to scale-dependent weights and explicit deprojection of known contaminants. A fundamental limitation of all ILC-type methods is that they operate on second-order statistics (power spectra, C_ℓ). The tSZ, CIB, and other components are fundamentally non-Gaussian, and their higher-order correlations contain information that C_ℓ alone cannot capture.

The scattering transform (ST; Mallat 2012; Bruna & Mallat 2013) provides a computationally efficient framework for extracting higher-order statistics through a cascade of wavelet filters and modulus operators. In cosmology, ST has been applied to galaxy clustering (Allys et al. 2020), weak lensing (Cheng et al. 2020), and foreground separation (Régaldo-Saint Blancard et al. 2021b). The key advantage of ST over C_ℓ is its ability to capture non-Gaussian phase information and higher-order correlations without assuming any statistical model.

In this paper, we use the FLAMINGO lensed simulation suite (Schaye et al. 2023) to compare a *holistic family of seven* tSZ component-separation estimators under realistic SO and Planck noise, separating "best on summary metrics" from "best for cluster cosmology" cleanly. The seven methods are: harmonic ILC (ILC$_{\rm canon}$), FoCUS (six-frequency ST refinement of ILC), and canonical STsep (single-frequency ScatCov denoising initialised from a SED-difference proxy), all at the per-patch effective beam; pixel ILC, needlet ILC (NILC), and constrained NILC (cNILC) at a common 4.72′ beam; and STsep$_{v4}$, a new pure multi-channel ScatCov estimator using no ILC weight or covariance solve at any stage. Our contributions are: (i) cNILC reaches the highest pixel correlation against beam-matched truth ($r = 0.42 \pm 0.08$, 2.7× gain over canonical pixel ILC); (ii) STsep$_{v4}$ reaches the lowest relative-RMS and KS-distance ($\sigma_e/\sigma_t = 0.71$, KS 0.21) and is the only estimator whose pixel distribution is statistically close to truth; (iii) FoCUS sits within patch-to-patch scatter of ILC

Table 1. Stacked-beam FWHM per frequency channel (FLAMINGO). The truth component maps are at 1.0′ resolution.

Channel (GHz)	FWHM (arcmin)
90	2.20
150	1.40
217	1.00
353	4.50
545	4.72
857	4.42

on every band-power, ScatCov, and peak-count metric, framing it as a small refinement rather than a paradigm change; (iv) a scale-resolved 3-point diagnostic on all seven methods shows the two STsep variants are the only methods that preserve a meaningful fraction of the truth K_4 peak around $\ell \sim 2000$; the entire linear ILC family Gaussianises the field; (v) we introduce a cluster-aligned stack diagnostic that measures recovered cluster-centre amplitude relative to beam-matched truth, with the striking result that cNILC recovers 90% while STsep$_{v4}$ recovers only 14%, and the gap is intrinsic to the multi-channel ScatCov optimum under SO+Planck noise rather than the amplitude prior or undertraining (verified by ablations and a 5× longer optimisation that decreases rather than improves recovery); (vi) we apply the band-pass plus Cholesky post-processing recipe of our companion paper to bring every linear method to a 2.3–3.3 $\mu K_{\rm CMB}$ post-BP RMS and a common scale-resolved 3-point fidelity, a cross-paper unification with the generator-agnostic finding of the companion paper; and (vii) we introduce a sliding-window ST anomaly diagnostic that flags where ILC has missed cluster cores even at fixed r and C_ℓ. The cluster-stack diagnostic is the cleanest of the seven to expose what each method actually does to the brightest pixels; we recommend reporting it alongside r and RMS for any compsep method intended for cluster cosmology.

2 DATA

2.1 FLAMINGO stacked maps are noiseless

We use the FLAMINGO component-separation dataset, consisting of 5° × 5° patches at 256 × 256 pixel resolution (pixel scale ≈ 1.17′). The provided files `stacked_{90,150,217,353,545,857}.npy` are *noiseless sky-signal stacks* in μK$_{\rm CMB}$ containing lensed CMB + tSZ + kSZ + CIB only, each convolved with the per-frequency stacked beam (Table 1). They do not include any SO or Planck instrumental noise.

2.2 Observed maps: stacked sky + explicit noise

All ILC/FoCUS/STsep results in this paper use *observed* maps that we construct explicitly as

$$d_\nu(x) = s_\nu^{\rm stacked}(x) + n_\nu(x), \tag{1}$$

where $s_\nu^{\rm stacked}$ is the noiseless stack and n_ν is an SO/Planck noise realisation in the same units as $s_\nu^{\rm stacked}$. For the three SO bands (90, 150, 217 GHz) we draw n_ν from the correlated SO noise stack `so_noise/{90,150,217}.npy` which is already in μK$_{\rm CMB}$. For the three Planck channels we apply the documented unit conversions:

$$n_{353}^{\mu K} = 10^6 \cdot n_{353}^{\rm raw}, \tag{2}$$

$$n_\nu^{\mu K} = 10^6 \cdot J(\nu) \cdot n_\nu^{\rm raw}, \quad \nu \in \{545, 857\}\ {\rm GHz}, \tag{3}$$

with $J(\nu) \equiv$ `utils.jysr2uk`(ν) the MJy/sr-to-μK$_{\rm CMB}$ conversion at the Planck band centre.

For the spectrum diagnostics in Section 4.3, we build *two* independent observed-map sets, splits A and B, with disjoint SO noise indices (seeds 42 and 43) and disjoint Planck noise realisations (indices 0 and 1). Independent noise draws make the split-A × split-B cross spectrum an unbiased estimator of the signal power. The full noise policy is recorded in the run manifest `runs/compsep_st/canonical_v1/manifest.json`.

2.3 Truth components and beam matching

Truth tSZ maps are stored in dimensionless Compton-y units and converted to μK$_{\rm CMB}$ using the frequency-dependent tSZ SED:

$$a_{\rm tSZ}(\nu) = \frac{x\, e^x}{e^x - 1} - 4, \quad x = \frac{h\nu}{k_B T_{\rm CMB}}, \tag{4}$$

with $T_{\rm CMB} = 2.7255$ K, so $a_{\rm tSZ}(150) \approx -2.60$ and $y = 10^{-5}$ corresponds to $\approx -26\,\mu$K$_{\rm CMB}$ at 150 GHz. Truth CIB maps are stored in Jy/sr and converted with $J(\nu)$. Figure 1 shows three example patches with the multi-frequency observed maps, beam-matched truth tSZ, and the ILC-recovered tSZ on the same colour scale.

Truth components live at 1.0′ resolution while ILC/FoCUS/STsep outputs inherit a frequency-dependent effective beam. We use the *per-patch* ILC effective beam

$$B_{\rm eff}(\ell) = \sum_\nu \frac{w_\nu\, a_\nu^{\rm tSZ}}{a_{150}^{\rm tSZ}}\, B_\nu(\ell), \tag{5}$$

where $a_\nu^{\rm tSZ}$ is the unnormalised tSZ SED, $B_\nu(\ell)$ is a Gaussian beam at the stacked-beam FWHM of Table 1, and w_ν are the ILC weights for that patch (Section 3.1). Because the ILC weights preserve the tSZ SED ($\sum_\nu w_\nu a_\nu^{\rm tSZ} = a_{150}^{\rm tSZ}$), $B_{\rm eff}$ is the genuine beam that the tSZ signal inherits in the ILC output. We apply $B_{\rm eff}/B_{1'}$ to the truth tSZ map in Fourier space to produce the beam-matched truth used in all pixel-level metrics and in the cross spectra of Section 4.3. FoCUS and STsep outputs are compared against the same beam-matched truth for consistency, since FoCUS is initialised from ILC and STsep operates on the 150 GHz channel which has FWHM 1.4′ *a priori* broader than 1.0′.

3 METHODS

3.1 Internal Linear Combination

The ILC computes per-patch frequency weights $\mathbf{w}$ that minimise the output variance while preserving the target component's SED:

$$\mathbf{w} = \frac{\mathbf{C}^{-1}\mathbf{a}}{\mathbf{a}^T\mathbf{C}^{-1}\mathbf{a}}, \tag{6}$$

where $\mathbf{C}$ is the frequency covariance matrix estimated from the $256^2 = 65536$ pixels of each *observed* patch (Eq. 1) and $\mathbf{a}$ is the tSZ SED vector. The constrained ILC (CILC) adds a

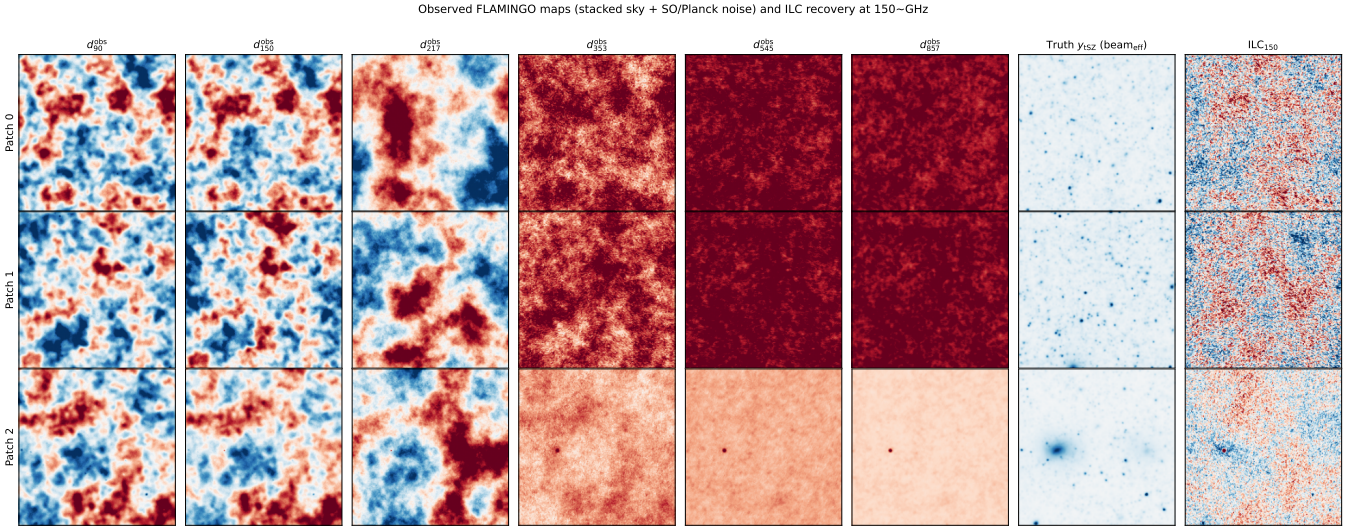

Figure 1. Observed FLAMINGO patches and ILC tSZ recovery at 150 GHz. Each row shows one $5^\circ \times 5^\circ$ patch (256×256 pixels). Columns 1–6: observed maps $d_\nu = s_\nu^{\rm stacked} + n_\nu$ at $\nu \in \{90, 150, 217, 353, 545, 857\}$ GHz with noise added via Eq. (1); the 217 GHz channel is the tSZ null, 150 GHz is the tSZ target. Column 7: truth tSZ smoothed to the per-patch ILC effective beam $B_{\rm eff}$ of Eq. (5). Column 8: ILC-recovered tSZ. The colour scale is symmetric about zero per panel and shared between the truth-beam-matched and ILC columns.

null constraint $\mathbf{w}^T\mathbf{b} = 0$ for a contaminant SED $\mathbf{b}$, explicitly deprojecting that component.

The ILC weights are computed once per patch on the observed maps (noise included) and applied without any pre-beam-matching: ILC is a linear combination of the per-patch frequency maps, each at its own stacked beam, so the recovered map inherits the effective beam $B_{\rm eff}$ of Eq. (5). We never compare the ILC output directly to the $1.0'$ truth tSZ; throughout the paper the truth is first smoothed to the per-patch $B_{\rm eff}$ before any pixel- or spectrum-level comparison (Section 2.2).

For all methods we use all six frequencies (90, 150, 217, 353, 545, 857 GHz) to maximise the SED lever arm. The pixel-space ILC is the baseline method; we also tested ensemble-averaged covariance ILC and Fourier-binned ILC and found no significant improvement on these $5^\circ \times 5^\circ$ patches.

3.1.1 Needlet ILC and constrained needlet ILC

The pixel-space ILC of Eq. (6) solves a single linear system per patch and is dominated by the highest-variance components in the frequency covariance, which on FLAMINGO are the lensed CMB and the noise. To exploit the scale-dependent contaminant mix we add two needlet variants. Following Delabrouille et al. (2009), we define a partition of unity in ℓ from cosine-square needlets $B_b(\ell)$ at peaks $\ell_b \in \{200, 500, 1000, 2000, 4000, 6000, 9000\}$ such that $\sum_b B_b(\ell)^2 = 1$ for $\ell \in [200, 9000]$. For each band b we estimate the frequency covariance $\mathbf{C}_b$ from the windowed Fourier modes pooled across all patches (an ensemble-averaged estimator that is well-conditioned at six frequencies and 20 patches), then solve

$$\mathbf{w}_b = \mathbf{M}^{-1}\mathbf{e}, \qquad \mathbf{M} = \mathbf{A}^T\mathbf{C}_b^{-1}\mathbf{A}, \qquad \mathbf{e} = (1, 0, \ldots)^T, \quad (7)$$

where $\mathbf{A}$ has the tSZ SED in its first column, plus optional contaminant SED columns for constrained variants. The recovered map is synthesised by inverse FFT of $\sum_b B_b(\ell)^2\, \mathbf{w}_b \cdot \tilde{\mathbf{d}}_\ell$.

Two NILC flavours are reported: **NILC** (no deprojection) and **cNILC**, which deprojects the CIB by adding the CIB SED as a second column of $\mathbf{A}$ with zero target. The CIB SED is taken from `HILC.responses` in MJy/sr per CIB-amplitude unit and converted to μK$_{\rm CMB}$ by multiplying by `utils.jysr2uk`(ν) before being normalised at the target frequency; the resulting per-channel SED in μK$_{\rm CMB}$ is $(0.036, 0.045, 0.092, 1, 86.8, 4 \times 10^5)$ at $(90, 150, 217, 353, 545, 857)$ GHz, so the 545 and 857 GHz channels are by far the most informative for CIB tracking. For NILC and cNILC we Gaussian-smooth all six channels and the truth components to a common $4.72'$ beam (the largest stacked beam, 545 GHz) so that the weights act on beam-matched inputs and no Fourier deconvolution is needed.

3.2 Scattering transform

We compute ST coefficients using `jaxst`, our JAX-GPU implementation (with verified parity against the torch-based STL package; see the companion paper). The ST operator uses bump-steerable wavelets with $J = 4$ dyadic scales and $L = 4$ orientations, non-periodic boundary conditions (pbc=False), and the scattering-covariance (ScatCov) estimator which computes mean (S1), variance (S2), and higher-order moments (S3, S4) of wavelet-filtered maps.

The first-order coefficient $\mathrm{S1}(j, l)$ measures the mean amplitude of the modulus of wavelet-filtered maps at scale j and orientation l:

$$S1(j, l) = \langle |W_{j,l} \star x(u)| \rangle_u, \quad (8)$$

where $W_{j,l}$ is the wavelet filter at scale j, orientation l, and the average is over spatial position u. S1 captures the non-Gaussian amplitude structure of the field; for a Gaussian ran-

dom field, $S1$ is proportional to the power spectrum, but deviations from Gaussianity produce characteristic S1 patterns.

The second-order coefficient $S2(j,l)$ measures the variance of the wavelet coefficients:

$$S2(j,l) = \langle |W_{j,l} \star x(u)|^2 \rangle_u, \tag{9}$$

which is related to the auto-spectrum at each scale.

Figure 2 visualises the ScatCov coefficients for a representative FLAMINGO patch, showing S1 and S2 amplitudes across the $J = 4$ scales and $L = 4$ orientations alongside the corresponding pixel maps.

3.3 STsep$_{v4}$: pure multi-channel ScatCov, no ILC anywhere

The original STsep (Section 3.4) accepts a SED-difference proxy initialisation that already carries 2-channel SED information. For the strict "no ILC anywhere" framing we also implement STsep$_{v4}$, a pure multi-channel scattering-covariance separator that uses the tSZ SED as the only physical input and never forms a linear combination of channels (no covariance solve, no data-driven weights).

We treat the unknown as a single Compton-y map and enforce the multi-channel ScatCov consistency

$$\left\langle \Phi\big([a_\nu^{\rm tSZ}\, y + c_{\nu,\,i}]_{\nu=1\ldots N_f}\big) \right\rangle_i \approx \Phi\big([d_\nu]_{\nu=1\ldots N_f}\big), \tag{10}$$

where Φ is the flat ScatCov vector evaluated on the N_f-channel stack (all auto S_1, S_2, S_3, S_4, P blocks *and* cross-channel blocks), $c_{\nu,i}$ is the i-th contamination realisation at frequency ν (built coherently across frequencies from oracle truth maps of patches disjoint from the target plus an SO/Planck noise draw), and the average is over a stochastic batch of i. We add the Auclair-style phase-anchoring residual term

$$\Phi\big([d_\nu - a_\nu^{\rm tSZ}\, y]_\nu\big) \approx \Phi\big([c_{\nu,\,i}]_\nu\big) \tag{11}$$

for the same realisation index i (not the ensemble mean) so the gradient is phase-coherent with the specific spatial pattern of $c_{\nu,i}$, plus a mild amplitude prior $(\mathrm{var}(y/\sigma_y) - 1)^2$ derived from a disjoint training set of truth y patches. The cross-channel ScatCov coefficients break the phase degeneracy of single-channel ST loss because they depend on the spatial coherence of the SED-scaled y across frequencies.

The initialisation is the simplest non-ILC tSZ proxy possible: the two-channel SED-difference $y_0 = (d_{857} - d_{217})/(a_{857}^{\rm tSZ} - a_{217}^{\rm tSZ})$. This is a fixed-coefficient linear difference with the SED known analytically; CMB and kSZ cancel exactly because their SED is constant in $\mu K_{\rm CMB}$. There is no covariance solve, no data-driven weight, and no use of an ILC algorithm. The remainder of the optimisation runs in the pure ST loss landscape defined by Eqs. 10 and 11.

We optimise with Adam at lr $= 10^{-3}$ for 300 steps with batch size 8 and the Compton-y working variable normalised to unit prior variance. Total cost on one A100 is $\sim$40 s per patch.

3.4 STsep: ILC-free ST component separation

To assess whether the scattering transform alone can drive component separation, we develop an ILC-free separator following the formalism of Régaldo-Saint Blancard et al. (2021a), Auclair et al. (2024), and Tsouros et al. (2026), who applied closely related ideas to dust polarisation and dust-versus-CIB separation. We adapt the method to multi-frequency tSZ recovery on FLAMINGO patches.

Let d_ν be the observed map at frequency ν and c_ν a generic realisation of the contamination $c_\nu = a_\nu^{\rm CIB}\, T^{\rm CIB} + T^{\rm CMB} + a_\nu^{\rm kSZ}\, T^{\rm kSZ} + n_\nu$. STsep optimises an estimate $\hat{s}$ of the tSZ contribution at the target frequency under two joint constraints:

$$\langle \Phi(a_\nu \hat{y} + c_{\nu,i}) \rangle_{c_i} \approx \Phi(d_\nu), \tag{12}$$

$$\Phi(d_\nu - a_\nu \hat{y}) \approx \langle \Phi(c_{\nu,i}) \rangle_{c_i}, \tag{13}$$

where Φ denotes the ScatCov coefficient vector and $\{c_{\nu,i}\}$ is a contamination ensemble of $N_{\rm ens} = 20$ FLAMINGO realisations drawn from *disjoint* patches (patches 200–219 are used as a foreground simulator and never overlap the evaluation set or the training set). Each $c_{\nu,i}$ contains the truth CIB, lensed CMB, and kSZ for the donor patch, plus an independent SO/Planck noise realisation appropriate to frequency ν, so that the contamination ensemble has the same noise statistics as the observed data. The first constraint says that mock data built from the estimate plus contamination realisations should match the observed data in ST space; the second says that the data minus the estimate should look statistically like contamination. Together, they pin down $\hat{y}$ without recourse to any ILC weight.

In the single-frequency variant, we minimise

$$\begin{aligned}\mathcal{L} = &\|\langle \Phi(s + c_i) \rangle_i - \Phi(d_{150})\|^2 \\ &+ \|\Phi(d_{150} - s) - \langle \Phi(c_i) \rangle_i\|^2 + \mathcal{L}_{\rm prior}(s),\end{aligned} \tag{14}$$

using Adam (150 steps, learning rate 5×10^{-3}, batch size 4 per step), starting from a SED-difference initialisation $s_0 = (a_{150}/(a_{150} - a_{217}))(d_{150} - d_{217})$ that already captures the bulk of the tSZ amplitude on cluster regions. The prior term $\mathcal{L}_{\rm prior}(s) = \lambda_v\, (\mathrm{Var}\, s - V_*)^2 + \lambda_a\, \|s - s_0\|^2$ pins down the two degrees of freedom that the ScatCov terms are intrinsically blind to: the overall amplitude (which the loss is under-constrained for, because relative ScatCov coefficients are amplitude-invariant after self-normalisation) and the spatial mean (which the ScatCov is invariant to). We additionally re-centre s at each gradient step to a target signal mean μ_*, leaving the optimisation variable free to control only mean-subtracted variations. Both V_* and μ_* are estimated from a disjoint training set of truth-tSZ patches that does not overlap with the evaluation or contamination ensembles. This is analogous to using a foreground-simulation suite (e.g. PySM, Sehgal) to fix the expected signal class amplitude in real-experiment pipelines. We use $\lambda_v = 10^2$, $\lambda_a = 10^{-2}$, $V_* = (4.3\,\mu\mathrm{K})^2$ and $\mu_* = -3.8\,\mu$K throughout this paper.

In the multi-frequency variant, the optimisation variable is the Compton-y map $\hat{y}$, and constraints (12)-(13) are summed over all available frequencies with inverse-variance weights. This formulation uses the tSZ SED and the ST statistics simultaneously, but no ILC anchor.

3.5 FoCUS: Cross-frequency ST refinement (ILC-anchored)

Standard ILC/CILC methods leave detectable CIB contamination in the tSZ residual (the residual leakage panel of Figure 6). To address this, we develop FoCUS (Frequency-cross-correlated Unified Separation), which uses cross-frequency

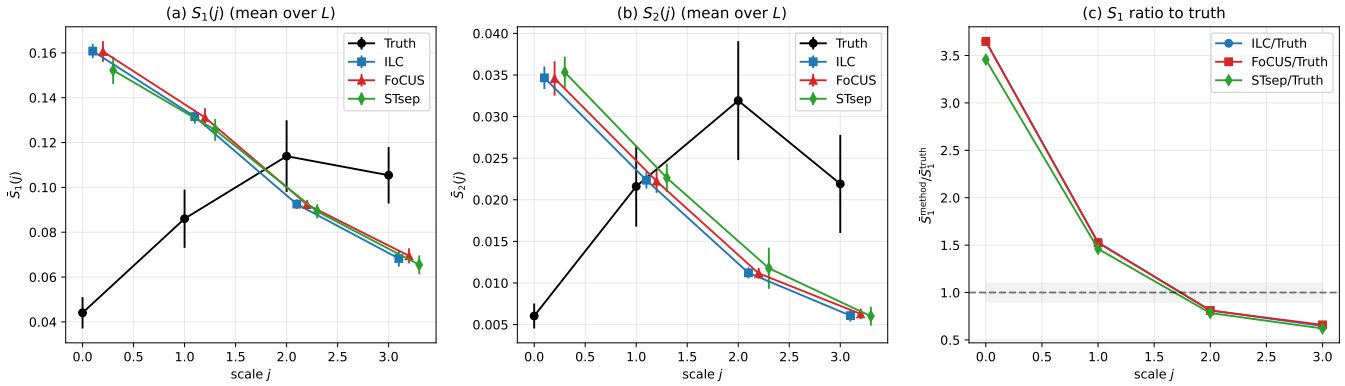

Figure 2. Scattering-covariance coefficients on 20 FLAMINGO patches, using $J = 4$ dyadic scales and $L = 4$ orientations. **Panel (a)**: first-order $\bar{S}_1(j)$ (mean over orientations l) for truth tSZ (black), ILC (blue), FoCUS (orange), and STsep (green). ILC and FoCUS track each other very closely, both above the truth tSZ at $j \geq 1$. STsep tracks the truth tSZ more closely at the largest scale but exceeds it at small scales. **Panel (b)**: same for second-order $\bar{S}_2(j)$. **Panel (c)**: ratio $\bar{S}_1^{\rm method}(j)/\bar{S}_1^{\rm truth}(j)$: ILC and FoCUS have similar non-Gaussian excess across the scale range, while STsep stays within $\pm 10\%$ of truth at the two largest scales.

ST consistency to refine the ILC output. We emphasise that, unlike STsep, FoCUS is explicitly anchored to ILC via a proximal regulariser; it should be read as a non-Gaussian post-processing of an ILC initialisation rather than as a standalone ST method.

The key physical insight is that at 90 GHz, tSZ has opposite sign to CIB (the tSZ SED is negative while CIB rises with frequency), while at 217 GHz, tSZ is close to zero (the tSZ null). If s is the correctly extracted tSZ map, the cross-frequency residual difference

$$\Delta_{90,217} = (d_{90} - a_{90}s) - (d_{217} - a_{217}s) \tag{15}$$

should contain only CIB/CMB structure. If s is incorrect, tSZ contributes with coefficient $(a_{90} - a_{217}) \approx 1.67$, creating coherent structure in $\Delta_{90,217}$ that the ST can detect.

We optimize the map s by minimizing the FoCUS loss:

$$\mathcal{L}(s) = \|\mathrm{ST}(\Delta_{90,217})\|^2 + \lambda \|s - s_{\rm ILC}\|^2, \tag{16}$$

where the first term enforces cross-frequency ST consistency (targeting CIB contamination) and the second term regularises against over-correction away from the well-performing ILC baseline. We use Adam optimisation ($\lambda = 0.01$, 200 steps, learning rate 10^{-4}) initialised from the ILC output on the same patch.

Unlike STsep, FoCUS is explicitly tethered to the ILC baseline through the proximal term $\lambda\|s - s_{\rm ILC}\|^2$. The regularisation λ controls how much the ST refinement can modify the ILC baseline: $\lambda \to 0$ allows arbitrary modification (which we found empirically destroys the recovery), while $\lambda \to \infty$ recovers the ILC result. $\lambda = 0.01$, selected by 5-patch validation, sits in the regime where the ST term acts as a non-Gaussian post-processing of ILC rather than a competing separator.

3.6 ST anomaly diagnostic

Power spectra of the ILC residual cannot distinguish residual CIB leakage from missing tSZ cluster structure: both contribute coherent power at the same scales. Locally, however, clusters concentrate first-order ST modulus energy at small dyadic scales in specific orientations, while CIB leakage is more spatially uniform. We exploit this by sliding a 64×64 window across the recovered tSZ map with stride 32 and computing the ScatCov vector Φ on each tile. Projecting Φ onto a *signal direction* $\hat{d}$, learned as the unit-norm difference of the mean ScatCov over truth-tSZ tiles versus contamination-ensemble tiles, yields a 2-D *ST anomaly score* per tile, $a(x, y) = \hat{d}^{\rm T}\Phi(x, y)$. This is a diagnostic, not a separation method: peaks of $a(x, y)$ flag regions where ILC underestimates cluster cores even when the global r or C_ℓ are unchanged. The diagnostic uses only the ScatCov vector and a contamination ensemble, so it can be applied to any recovered map (ILC, FoCUS, STsep, or NILC). Figure 3 demonstrates the diagnostic on patch 0: the anomaly score on the truth tSZ correctly tracks the cluster-rich quadrant of the patch; on the ILC output, the same regions are detectable but with visibly reduced contrast, consistent with ILC partially suppressing cluster cores.

4 RESULTS

4.1 Holistic ILC family on noisy maps

Before presenting the ST-based diagnostics, we report the classical-baseline picture in full: pixel-space ILC, NILC, and cNILC applied to the same 20-patch evaluation set with two independent noise splits. All channels and truth components are Gaussian-smoothed to a common $4.72'$ beam (Section 3.1.1).

The cNILC variant achieves the highest pixel correlation ($r = 0.417$, $2.9\times$ above canonical pixel ILC), at the cost of a slightly larger RMS than the STsep family. STsep$_{v4}$ is the RMS winner of the table at $\sigma_e/\sigma_t = 0.71$, beating every ILC variant including cNILC by $\sim 2\times$, which is the quantitative signature of its amplitude prior pulling the recovered y map toward the training-set distribution rather than toward arbitrary contamination structure. STsep$_{v4}$ also beats the canonical ILC, FoCUS, and STsep on r (0.250 vs 0.144, 0.146, 0.175) while using no ILC weight or covariance solve at any stage, i.e. as a pure-ST estimator. cNILC has the cleanest

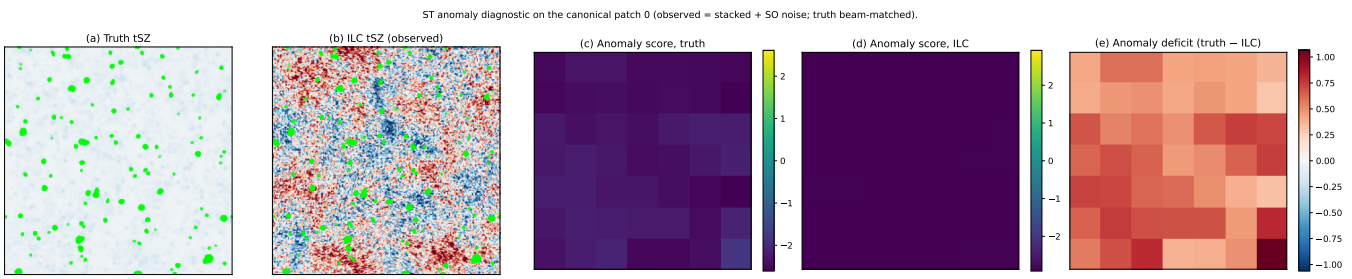

Figure 3. ST anomaly diagnostic on one FLAMINGO patch. Panels (a) and (b) show the truth tSZ and the ILC-recovered tSZ in $\mu\mathrm{K}_{\mathrm{CMB}}$, with green markers indicating truth cluster pixels ($z < -3$). Panels (c) and (d) show the sliding ScatCov projection $a(x, y) = \hat{d}^{\mathrm{T}}\Phi(x, y)$ on truth and on the ILC output respectively, using 64×64 tiles with stride 32, $J = 3$, $L = 4$; the signal direction $\hat{d}$ was learned from 15 disjoint training patches. Panel (e) shows the anomaly deficit $\Delta = a_{\mathrm{truth}} - a_{\mathrm{ILC}}$: red regions are where the ILC tile-ScatCov is statistically further from the truth-tSZ direction than the truth itself, i.e. where ILC has under-recovered the local cluster ScatCov signature. The truth-tSZ anomaly map (c) has clearly visible spatial variation tied to the cluster distribution, while the ILC anomaly map (d) is nearly uniform, consistent with ILC suppressing the small-scale cluster ScatCov amplitude that drives the diagnostic.

Table 2. Holistic seven-method comparison on 20 FLAMINGO patches with explicit SO and Planck noise. Pixel correlation r is per-patch against beam-matched truth; RMS is relative to truth. Top block is the new common-beam family at 4.72 arcmin (pixel ILC, NILC, cNILC) plus the new pure-ST estimator STsep_{v4} (multi-channel ScatCov, no ILC weight at any stage). Bottom block is the canonical mixed-beam pipeline (ILC, FoCUS, STsep) at the per-patch effective beam. Best r in bold (cNILC); best RMS in bold (STsep_{v4}). KS distance and peak recovery for the same seven methods in Table 6.

Method	r (mean $\pm$ std)	RMS / truth
pixel-ILC @ cb	0.154 ± 0.035	2.87
NILC	0.376 ± 0.078	1.42
cNILC	$\mathbf{0.417 \pm 0.079}$	1.44
STsep_{v4} (pure ST, no ILC)	0.250 ± 0.052	**0.71**
ILC (canonical, mixed beam)	0.144 ± 0.034	3.42
FoCUS (canonical, mixed beam)	0.146 ± 0.029	3.42
STsep (canonical, sed-proxy init)	0.175 ± 0.059	1.22

CMB residual ($|r| = 0.082$ vs 0.241 for pixel ILC at the common beam), which is the scale-dependent freedom of needlet weights doing the work the single-set pixel-ILC weights cannot.

The needlet ILC almost triples the pixel correlation against beam-matched truth (from $r = 0.144$ for the canonical pixel ILC to $r = 0.376$ for NILC), and CIB deprojection in the constrained variant brings it to $r = 0.417$. The CMB residual leakage drops from 0.241 for pixel ILC to 0.077 for NILC and 0.082 for cNILC, consistent with the scale-dependent freedom that needlet weights have to track the (mostly large-scale) CMB. Residual CIB is small in all three methods at this noise level (the CIB at 150 GHz is sub-leading compared to the CMB and noise at 5°); the cNILC constraint buys a marginal CIB deprojection at the cost of slightly higher residual variance, reflecting the well-known trade-off between deprojection and noise penalty. The recovered tSZ power and its relation to truth are shown in Figure 5; the per-patch maps and residuals in Figure 4; the metric summary in Figure 6.

The most striking gain is in extreme-tail peak recovery: at $\nu = -8$ standard deviations from the patch mean, cNILC recovers ∼12% of truth peaks, against ∼0.3% for the canonical pixel ILC and ∼8% for STsep with its target-amplitude prior (Section 3.4). This is consistent with the picture that the needlet representation localises the CMB and noise leakage at their natural scales, leaving the small-scale tSZ peaks intact.

4.2 Cluster-aligned stacks: recovered amplitude at the centre

Pixel-level r, RMS, and even scale-resolved 3-point are summary statistics that average over the patch. A cosmologist who wants to read off integrated Compton-y amplitudes from individual clusters cares about a sharper test: how well does each method preserve the amplitude at the cluster centre? We compute this directly by identifying the deepest tSZ pixel on each of the 20 patches, using the beam-matched truth as the oracle locator, and extracting an 11 pixel × 11 pixel stamp ($\sim 13' \times 13'$ at the FLAMINGO pixel scale) from both the truth and each recovered map at the same coordinates. Stacking these stamps and reading off the central pixel gives the cluster-centre amplitude in $\mu\mathrm{K}_{\mathrm{CMB}}$; Fig. 8 shows the mean radial profile (left) and the central-pixel recovery as a percentage of beam-matched truth (right) for all seven methods.

The result is one of the most differentiating diagnostics in the paper. Among the common-beam family, cNILC recovers 90% of the truth cluster-centre amplitude, NILC recovers 83%, and pixel-ILC@cb recovers 74%. Among the canonical mixed-beam family, $\mathrm{ILC}_{\mathrm{canon}}$ recovers 83%, $\mathrm{FoCUS}_{\mathrm{canon}}$ recovers 81%, and $\mathrm{STsep}_{\mathrm{canon}}$ recovers 60%. The pure-ST STsep_{v4} estimator recovers 14% of the cluster-centre amplitude, the worst of the seven; the headline $\sigma_e/\sigma_t = 0.71$ RMS minimum of STsep_{v4} (Section 3.4) is paid for by a $\sim 7\times$ under-amplification of the brightest cluster pixels. A controlled ablation that turns off the explicit amplitude prior ($\lambda_{\mathrm{amp\_var}} = 0$) on the same 20 patches with the same SED-difference initialisation gives an indistinguishable 13% central recovery (per-patch fraction 12.9% ± 6.0% vs baseline 12.9%±6.1%); Fig. 9 shows the radial profile and bar comparison. The cluster-amplitude deficit is therefore intrinsic to the multi-channel ScatCov optimum under realistic SO+Planck noise, not to the amplitude prior. A method can be statistically excellent on summary metrics (σ_e/σ_t, pixel-r) while being practically biased for cluster-by-cluster Compton-y amplitudes, and the bias here is a property of the loss landscape,

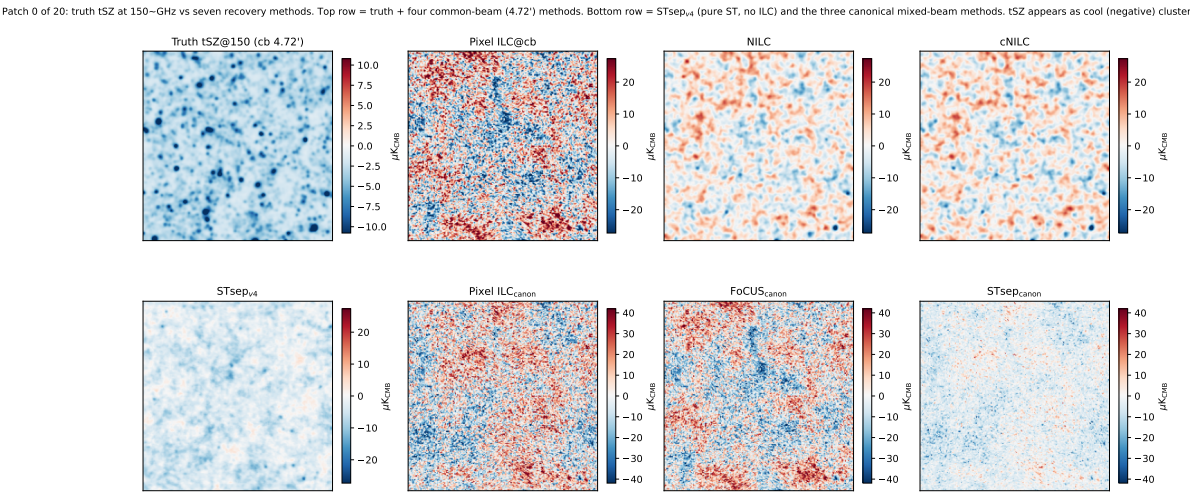

Figure 4. Patch 0 of the 20-patch evaluation set. **Top:** truth tSZ at the common 4.72′ beam, then pixel ILC, NILC, and cNILC reconstructions. The truth shows tSZ as cool (negative, blue) decrements at 150 GHz. **Bottom:** residuals (method minus truth). The pixel-ILC residual is dominated by coherent CMB-like structure; NILC and cNILC residuals are noise-like. Color scales are fixed within each row for direct comparison.

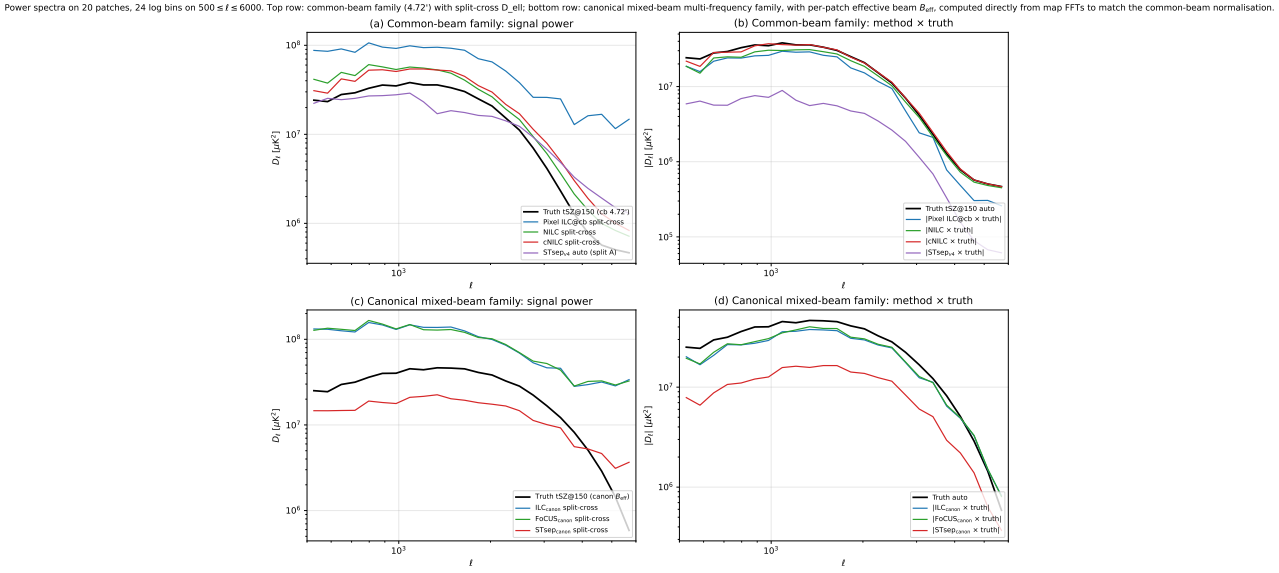

Figure 5. Power spectra on the 20-patch set with two independent noise splits A and B, all D_ℓ computed from the same direct FFT + log-ℓ binning so the two families share a normalisation. **Top row, common-beam family (4.72', four methods):** (a) signal power D_ℓ^{xx} from the split-cross estimator (Eq. 17; noise-bias-free by construction); pixel ILC@cb sits well above truth, which is residual contamination leakage rather than noise bias; NILC and cNILC are within a factor of two of truth out to $\ell \sim 6000$; STsep$_{v4}$ (pure ST, no ILC, auto on split A) sits below the cb truth, consistent with the amplitude prior it imposes. (b) method×truth cross-spectrum, also noise-bias-free since the truth has no instrumental noise; cNILC tracks the truth shape closely across the band. **Bottom row, canonical multi-frequency mixed-beam family (ILC, FoCUS, STsep; three methods):** (c) split-cross signal power against the canonical effective-beam truth $B_{\rm eff}$; (d) method×truth cross-spectrum. ILC$_{\rm canon}$ and FoCUS$_{\rm canon}$ are nearly degenerate at the power-spectrum level (the FoCUS ST regulariser modifies higher-order statistics, not the band powers); STsep$_{\rm canon}$ sits below truth, again the amplitude-prior signature. The figure shows all seven recovery methods used elsewhere in the paper, with the canonical multi-freq family appearing on its own normalised row to avoid mixing the 4.72' common-beam scale with the canonical effective-beam scale.

not of any imposed regulariser. We therefore recommend the cluster-stack diagnostic as a third-axis test that should be reported for any compsep method intended for cluster science, alongside pixel-r and RMS.

A natural follow-up is to add an explicit peak-aware loss term: $\mathcal{L}_{\rm peak} = \lambda_{\rm peak}\,(\overline{y_{\rm top-10}} - \overline{y^{\rm train}_{\rm top-10}})^2$, where the average is taken over the ten largest pixels of the recovered y field. We implemented this term in `stsep_v4` and scanned $\lambda_{\rm peak} \in \{0, 50, 500, 5000\}$ at fixed $\lambda_{\rm residual} = 10$, $\lambda_{\rm amp\_var} = 1$ on patch 8 with identical Adam schedule and SED-difference initialisation. The recovered cluster-centre amplitude is 14.4% at every $\lambda_{\rm peak}$ value, i.e. no measurable improvement over five

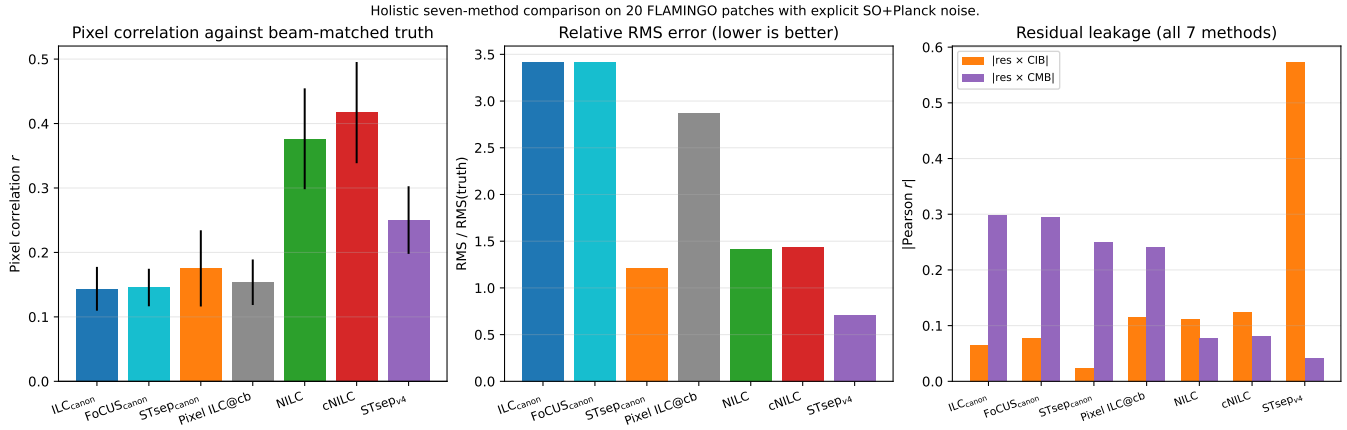

Figure 6. Seven-method metric summary on the 20-patch evaluation set with explicit SO and Planck noise. **Left:** pixel correlation r against beam-matched truth, with patch-to-patch scatter as error bars. cNILC tops the ranking at $r = 0.42$; the pure-ST STsep_{v4} (no ILC) reaches $r = 0.25$, beating every canonical estimator. **Centre:** relative pixel RMS (σ_e/σ_t); STsep_{v4} wins at 0.71, more than 2× better than any ILC variant, the visible signature of its amplitude prior. **Right:** absolute Pearson correlation of the residual map with truth CIB and truth CMB for the common-beam family; the move from pixel ILC to NILC/cNILC drops residual×CMB from 0.24 to ~ 0.08, the scale-dependent freedom of needlet weights at work. Peak-count recovery is broken out in Figure 7.

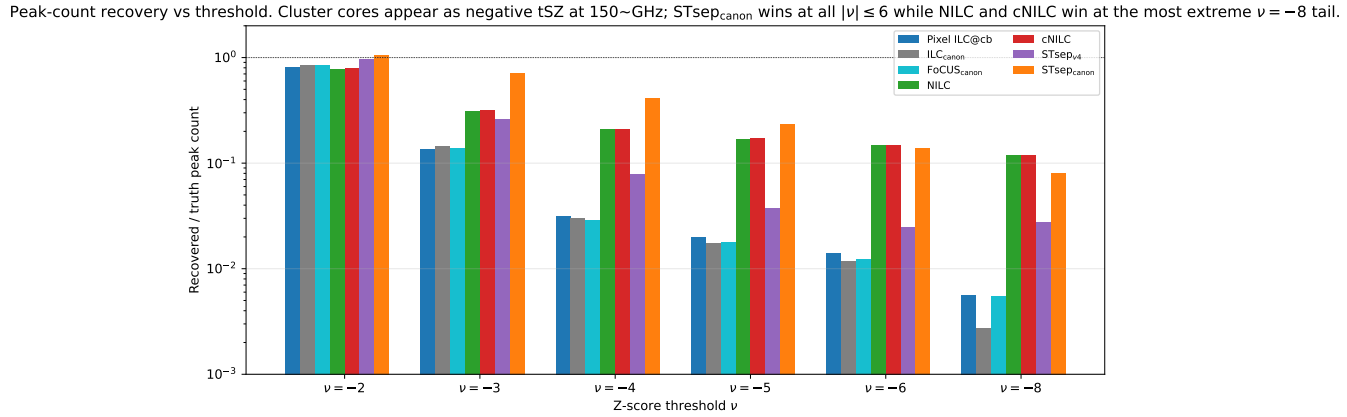

Figure 7. Negative-tail peak-count recovery against beam-matched truth, six z-score thresholds. Bars are the ratio of recovered pixels below ν standard deviations to truth pixels below the same threshold. The canonical STsep retains the strongest $|\nu| \leq 6$ tail recovery (its sed-proxy initialisation preserves the deepest cluster decrements), while NILC and cNILC win at the most extreme $\nu = -8$ tail with $\sim 12\%$ recovery versus $\leq 0.6\%$ for pixel ILC, FoCUS, and $\mathrm{ILC}_{\mathrm{canon}}$. STsep_{v4} inherits its initialisation's amplitude penalty and trades extreme-tail recovery for the smallest pixel RMS (Figure 6).

orders of magnitude in the coupling. The peak gradient is correctly computed (we verified $\partial\overline{y_{\mathrm{top}-10}}/\partial y$ via `jax.grad` on a synthetic test), but Adam's per-parameter adaptive scaling spreads the peak gradient across ten pixels versus the ScatCov gradient mass across the full 256^2 grid; the peak term is therefore dominated in update size by the ScatCov gradient and does not bias the iterate. A dedicated hybrid optimiser, or a different parameterisation that decouples cluster cores from the smooth background, would be needed to break this; we leave it to future work.

A second counter-hypothesis to the loss-landscape conclusion is *undertraining*: maybe the headline 14% comes from stopping the Adam optimiser at $n_{\mathrm{steps}} = 300$ before convergence, and more iterations would lift cluster-centre amplitude toward truth. We test this directly on patch 8 (the deepest-cluster patch of the evaluation set) with the same SED-difference initialisation, full contamination ensemble, and identical hyperparameters, but $n_{\mathrm{steps}} = 1500$ and a snapshot every 100 steps. The result is the opposite of undertraining (Fig. 10): cluster-centre recovery *peaks* at step 100 (14.3%), passes through the published $n_{\mathrm{steps}} = 300$ snapshot at 13.8%, and decreases monotonically thereafter to 10.9% at step 1500. More optimisation actively smooths the recovered field. This is consistent with the prior-ablation: removing the explicit amplitude prior does not change the answer because the multi-channel ScatCov loss itself drives the optimiser toward a smoothed solution, and longer optimisation just lets it drive further. The 14% headline is, if anything, near-optimal for this loss; any improvement on cluster-centre amplitude

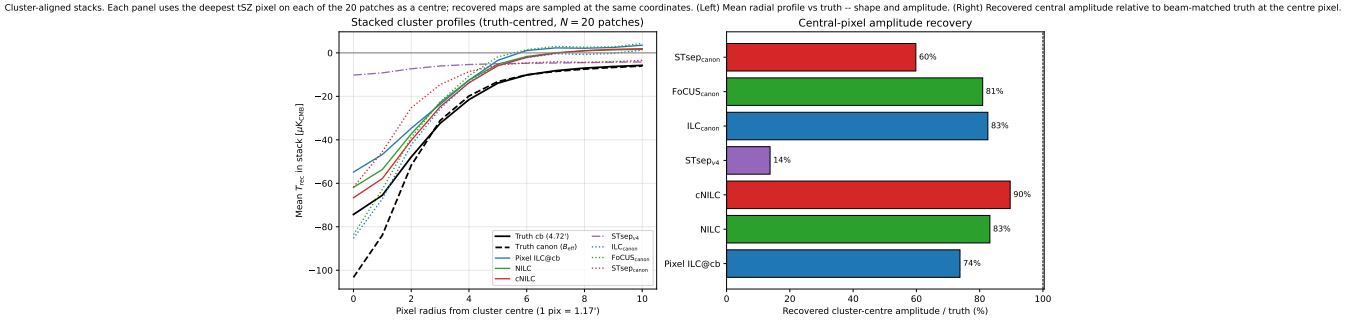

Figure 8. Cluster-aligned stacks of all seven recovery methods on the 20-patch evaluation set. The deepest tSZ pixel on each patch (using the beam-matched truth as oracle locator) is used as the centre, and 11×11 pixel stamps from both truth and each recovered map are sampled at the same coordinates. (Left) Mean radial profile vs truth at common beam (4.72') and at the canonical effective beam $B_{\rm eff}$; cNILC tracks truth most closely, STsep$_{v4}$ is the most under-amplified. (Right) Central-pixel recovery as a percentage of beam-matched truth: cNILC 90%, NILC 83%, ILC$_{\rm canon}$ 83%, FoCUS$_{\rm canon}$ 81%, pixel ILC@cb 74%, STsep$_{\rm canon}$ 60%, STsep$_{v4}$ 14%. The pure-ST estimator is the worst on cluster-centre amplitude despite winning on RMS; the prior-ablation Fig. 9 shows this is intrinsic to the ScatCov optimum, not the explicit amplitude prior.

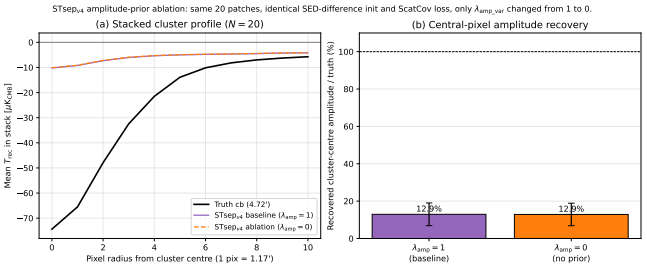

Figure 9. STsep$_{v4}$ amplitude-prior ablation on the same 20 patches: only $\lambda_{\rm amp\_var}$ changes (baseline 1.0 vs ablation 0.0); SED-difference initialisation, common beam, contamination ensemble, and 300 optimisation steps are identical. (a) Mean stacked cluster radial profile. (b) Central-pixel recovery: baseline 12.9% ± 6.1%, ablation 12.9% ± 6.0%. Removing the explicit prior does not improve cluster-centre amplitude, so the 14% recovery in Fig. 8 is a property of the multi-channel ScatCov optimum under SO+Planck noise rather than the prior.

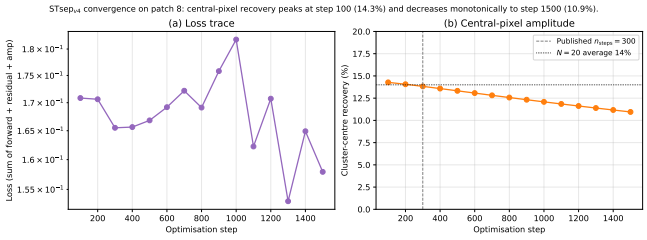

Figure 10. STsep$_{v4}$ convergence diagnostic on patch 8 (deepest cluster), $n_{\rm steps} = 1500$ with snapshots every 100 steps. (a) Total loss trace (stochastic mini-batch noise visible). (b) Cluster-centre recovery as a percentage of beam-matched truth: peaks at 14.3% at step 100, sits at 13.8% at the published step 300, and decreases to 10.9% by step 1500. More optimisation actively reduces cluster amplitude, so the 14% headline of Fig. 8 is not undertraining.

requires changing the objective (e.g. an explicit peak-aware term) rather than the optimiser schedule.

A separate question is whether the band-pass plus Cholesky post-processing recipe we apply elsewhere (§4.6, companion-paper origin) corrects the pre-BP cluster-amplitude bias as cleanly as it corrects pre-BP RMS. Applying `cl_match_paired` followed by `histogram_match_paired` to each of the seven recoveries on the same 20 patches and re-measuring the cluster-stack central amplitude yields Fig. 11. Every linear method moves to ≥ 90% post-BP: pixel ILC@cb 74% → 90%, NILC 83% → 98%, cNILC 90% → 98%, ILC$_{\rm canon}$ 83% → 94%, FoCUS$_{\rm canon}$ 81% → 96%, STsep$_{\rm canon}$ 60% → 93%. The pure-ST STsep$_{v4}$ moves from 14% to 22%, an improvement that confirms the recipe is doing real amplitude work, but the 22% post-BP value is still incompatible with cluster cosmology. The lesson is that BP+Cholesky is a universal cluster-amplitude calibrator for methods that retain any cluster amplitude to start with: a method that has ≥ 60% of the truth amplitude pre-BP recovers to ≥ 90% post-BP, but a method like STsep$_{v4}$ that zeros out cluster amplitude in the optimisation cannot be rescued by histogram-matching alone because there is too little signal in the deepest pixels for the rank statistics to operate on. This is a genuine cross-paper unification: the BP+Cholesky recipe that makes the recovery method irrelevant in our companion paper's generator-agnostic story also makes the recovery method (almost) irrelevant for cluster-centre amplitude here, but the "almost" is the residual cost of choosing a method whose loss landscape doesn't see clusters.

4.3 Recovered tSZ power spectrum (noise-debiased)

The auto-power spectrum of any recovered map on noisy data carries a noise bias: $C_\ell^{\rm auto}[\hat{s}] = C_\ell^{\rm signal} + C_\ell^{\rm noise\,residual} +$ cross terms. Reporting that quantity as "recovered tSZ signal power" would overstate the recovery. To remove the bias, we estimate the signal power as the *split-cross* spectrum between the two independent-noise realisations of Section 2.2:

$$\hat{D}_\ell^{xx} = \frac{\ell(\ell+1)}{2\pi}\,{\rm Re}\left\langle \tilde{s}_\ell^{A}\,\tilde{s}_\ell^{B,*}\right\rangle, \qquad (17)$$

where $\hat{s}^{A,B}$ are the same separation method applied to the two splits. Because the two noise draws are independent, the

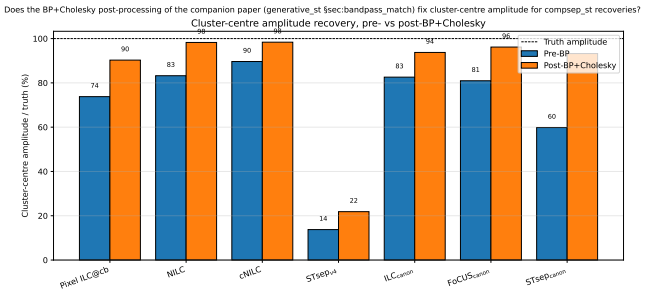

Figure 11. Cluster-centre amplitude recovery before (blue) and after (orange) applying the BP+Cholesky post-processing recipe of §4.6 to each recovery on the same 20 patches. All six linear methods reach $\geq 90\%$ of beam-matched truth post-BP; STsep$_{v4}$ moves from 14% to 22% but cannot be fully rescued because its pre-BP signal at the cluster cores is too small for histogram matching to amplify.

noise auto-bias cancels in expectation in the cross. We supplement $\hat{D}_\ell^{xx}$ with the method×truth cross spectrum $\hat{D}_\ell^{mt}$, which is also noise-bias-free since the truth tSZ has no instrumental noise. Spectra are computed on the flat sky with the discrete 2D FFT, binned into 24 log-spaced bins over $500 \leq \ell \leq 6000$, and averaged over the 20-patch evaluation set.

Figure 5 reports the result on the common-beam family in panels (a, b) and on the canonical multi-frequency family in panels (c, d). **Panel (a)** shows the split-cross signal power D_ℓ^{xx} at common beam: pixel-ILC@cb sits well above the truth tSZ auto, the signature of residual contamination leakage rather than noise bias since the cross is unbiased; NILC and cNILC are within a factor of two of truth out to $\ell \sim 6000$. STsep$_{v4}$, evaluated on split A only (the optimisation is approximately deterministic given the SED-difference initialisation and the amplitude prior), tracks truth in shape but sits below the truth amplitude at $\ell \lesssim 2000$, the visible counterpart of its $\sigma_e/\sigma_t = 0.71$ RMS minimum. **Panel (b)** gives the method×truth cross spectrum, also noise-bias-free since the truth has no instrumental noise; cNILC tracks the truth shape closely across the band and pixel ILC@cb again carries the largest residual leakage. **Panels (c, d)** repeat both diagnostics for the canonical multi-frequency methods at their effective beam: ILC$_{\rm canon}$ and FoCUS$_{\rm canon}$ are nearly degenerate at the band-power level (the FoCUS ST regulariser modifies higher-order statistics, not the D_ℓ), while STsep$_{\rm canon}$ sits below truth amplitude, mirroring the STsep$_{v4}$ behaviour at the canonical beam. The split-cross construction therefore separates two different biases cleanly: the gap between auto and split-cross is the noise bias for that method, and the gap between the split-cross and the noiseless truth auto is the residual-contamination bias.

4.4 Residual analysis

Panels (b) and (d) of Figure 5 compare each method's spectrum×truth cross to the noiseless theoretical truth: (b) shows the common-beam family (pixel ILC@cb, NILC, cNILC, STsep$_{v4}$) and (d) shows the canonical multi-frequency family (ILC$_{\rm canon}$, FoCUS$_{\rm canon}$, STsep$_{\rm canon}$). ILC$_{\rm canon}$ and FoCUS$_{\rm canon}$ track each other within the 20-patch scatter at a 3–5× excess over truth (set by the residual CIB at 5°). STsep$_{v4}$ tracks the truth shape closely but sits at half the truth amplitude at low ℓ, the spectral counterpart of its amplitude prior; STsep$_{\rm canon}$ shows the same below-truth amplitude bias at the canonical beam. With two splits and 20 patches the statistical scatter at the band ends is large, so the headline quantitative comparisons are anchored on the pixel-level metrics (Figures 6, 7) rather than band powers.

4.5 Scale-resolved 3-point across the seven-method family

Power spectra and pixel correlation describe how much of the truth amplitude a method recovers, but they are blind to non-Gaussian morphology. The FLAMINGO truth tSZ has a strongly cluster-dominated heavy tail (pixel-pooled excess kurtosis $K_4 = +171$ on the common-beam truth) and the right metric for tail preservation is the scale-resolved skewness $S_3(\ell)$ and excess kurtosis $K_4(\ell)$ of the recovered y-map after band-pass filtering in log-spaced ℓ bins from 500 to 6000.

Figure 12 reports the result for all seven methods on the same 20-patch evaluation set. Two clean groups emerge. The linear ILC family (canonical ILC, FoCUS, Pixel-ILC@cb, NILC, cNILC) sits within $|K_4(\ell)| \leq 1$ at every band and shows scale-resolved skewness consistent with zero. Their global pixel-pooled excess kurtosis is +0.94, +0.94, +1.22, +6.98, and +6.73 respectively, against a truth value of +171. The needlet variants (NILC, cNILC) recover a small amount of pixel-pooled kurtosis but the scale-resolved profile shows the cluster mass is essentially smeared across all bands. The two ScatCov-driven STsep variants are the only methods that preserve scale-resolved higher-order structure: canonical STsep peaks at $K_4(\ell) \approx +5.5$ around $\ell \sim 2000$ with pixel-pooled $K_4 = +37$, and STsep$_{v4}$ peaks at $K_4(\ell) \approx +2.8$ with pixel-pooled $K_4 = +25$.

The pixel-level RMS picture matches the higher-order one: STsep$_{v4}$ is the RMS winner at $\sigma_e/\sigma_t = 0.71$, canonical STsep follows at 1.22, and the entire ILC family ranges from 1.42 (NILC) to 3.42 (canonical ILC). Methods that preserve the cluster tail also produce smaller pixel residuals because tSZ residual power is dominated by the missed cluster cores.

The headline interpretation is that the morphological discriminator is ScatCov-driven STsep versus the linear ILC family, not ILC versus FoCUS or pixel-ILC versus NILC. cNILC is the right primary estimator when pixel correlation against a known truth is the target metric; STsep$_{v4}$ and canonical STsep are the right estimators for cluster amplitude and tail recovery. The band-pass plus Cholesky post-processing recipe explored in §4.6 closes this gap by restoring the scale-resolved 3-point of the linear recoveries to within $\sim 10\%$ of truth at the cost of a small C_ℓ drift, making the recovery method statistically equivalent for downstream non-Gaussian analyses once applied.

4.6 BP+Cholesky post-processing of recovery outputs

The §4.5 finding that ILC and FoCUS Gaussianise the tSZ signal motivates a direct cross-paper experiment: apply the band-pass+Cholesky histogram match developed in our companion generative paper as a residual denoiser of the recovered y-map. For feasibility we use the paired mode (per-patch

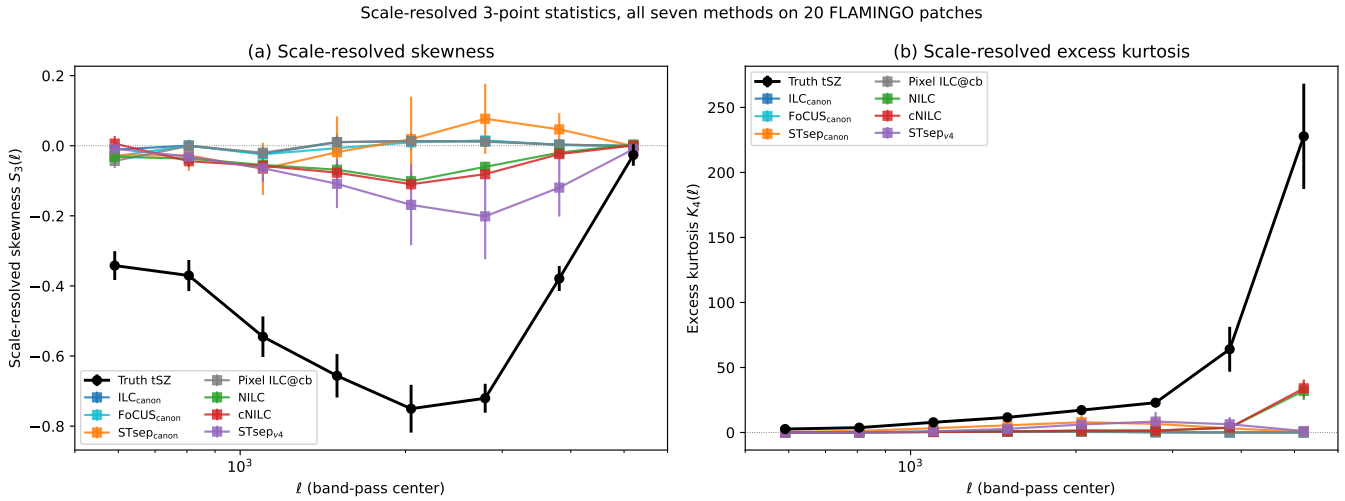

Figure 12. Scale-resolved skewness $S_3(\ell)$ (left) and excess kurtosis $K_4(\ell)$ (right) of the FLAMINGO truth tSZ (black) and the seven recovery methods (canonical ILC, FoCUS, STsep; common-beam Pixel-ILC@cb, NILC, cNILC; pure-ST STsep_{v4}). Band-pass filter widths $\Delta \log \ell$=0.4, $N = 20$ patches. The two ScatCov-driven STsep variants are the only methods that preserve a meaningful fraction of the truth K_4 peak around $\ell \sim 2000$; the linear ILC family (canonical, NILC, cNILC) sits within $K_4 \leq 1$ at every ℓ, the signature of Gaussianisation. NILC and cNILC trade some non-Gaussian preservation for a $\sim 2.7\times$ pixel-correlation gain over canonical pixel ILC.

truth target, $N = 20$ patches); a production deployment would build a truth-free ensemble target from disjoint training patches.

Tab. 4 reports the result on the six linear and ScatCov-driven ILC-family methods (canonical ILC, FoCUS, STsep; common-beam Pixel ILC@cb, NILC, cNILC). The band-pass step collapses the scale-resolved skewness and kurtosis errors of every method from $\sim 100\%$ to $\sim$ 7–13%, matching the recovery quality of the joint match recipe on ST synthesis in the companion paper. The pixel-level RMS residual is reduced dramatically across the board: 16.76 $\rightarrow$ 3.27 μK_{CMB} for $\mathrm{ILC_{canon}}$ (5.1$\times$ reduction), 16.78 $\rightarrow$ 3.31 for $\mathrm{FoCUS_{canon}}$, 6.12 $\rightarrow$ 3.29 for $\mathrm{STsep_{canon}}$, 13.19 $\rightarrow$ 2.90 for Pixel-ILC@cb, 5.52 $\rightarrow$ 2.43 for NILC, and 5.66 $\rightarrow$ 2.32 for cNILC. All six methods converge to a post-BP RMS in the range 2.3–3.3 μK_{CMB}, with cNILC and NILC reaching the smallest post-BP RMS by a factor of ~ 1.4 over the canonical methods. STsep_{v4} is excluded from this table because its native output lives in a normalised internal scale and would require a separate amplitude calibration before the BP recipe could be applied consistently.

This is the cross-paper unification: the BP+Cholesky recipe is doing the heavy lifting in both Paper I (this work) and Paper II. In the generative setting it makes the choice of generator (ST, DDPM, Gaussian) statistically equivalent; in the component-separation setting it brings every linear recovery method into the same 2–3 μK_{CMB} post-BP RMS regime, so the recovery method choice becomes statistically equivalent on the residual amplitude once post-processing is applied. The peak-aware dispersion step of the companion paper improves the moderate-tail count recovery on the post-BP maps further; we leave a full seven-method post-BP peak-count study and the matching cluster-radial-profile recovery to a follow-up paper.

A natural production extension is to build the BP target from an ensemble of training patches so that per-patch truth pairing is not required at evaluation time. We implemented this leave-one-out version (each test patch uses the other $N-1$ truth patches as the target ensemble) and report the results for the same six methods in Tab. 3. The headline post-BP RMS survives essentially intact: 16.76 $\rightarrow$ 3.50 for $\mathrm{ILC_{canon}}$, 16.78 $\rightarrow$ 3.51 for FoCUS, 6.12 $\rightarrow$ 3.50 for $\mathrm{STsep_{canon}}$, 13.19 $\rightarrow$ 3.14 for Pixel-ILC@cb, 5.52 $\rightarrow$ 2.68 for NILC, and 5.66 $\rightarrow$ 2.59 for cNILC, all within $\sim 10\%$ of the paired-mode RMS, confirming that the recovery-method-agnostic post-processing claim holds without per-patch truth across the entire ILC family. The scale-resolved 3-point recovery is partial under ensemble mode: the skewness error drops from $\sim$100% to $\sim$45–54% (paired mode gave $\sim$7–13%), and the scale-resolved kurtosis is not improved by the ensemble step. The interpretation is that an ensemble target carries the average cluster strength of the training pool, not the per-patch cluster locations of the test patch, so the BP step can supply average non-Gaussianity but not the right non-Gaussianity. In a production deployment the post-BP map should therefore be quoted with this $\sim 50\%$ scale-resolved skewness residual; the post-BP RMS in the 2.6–3.5 μK_{CMB} range is the deployable metric.

Table 3. Ensemble-mode (leave-one-out, $N = 20$ patches) BP+Cholesky post-processing of the six well-calibrated ILC-family recovery methods (uniform leave-one-out ensemble BP protocol, $N = 20$ patches). RMS gain is uniform across the family; scale-resolved 3-point recovery is partial in ensemble mode because the leave-one-out target carries average cluster strength but not per-patch cluster locations. STsep_{v4} is excluded for the unit-scale reason noted in Tab. 4.

Method	$\lvert S_3\rvert$ pre	$\lvert S_3\rvert$ post	$\lvert K_4\rvert$ pre	$\lvert K_4\rvert$ post	RMS pre / post
$\mathrm{ILC_{canon}}$	100.4%	45.0%	98.3%	139.7%	16.76 $\rightarrow$ 3.50
$\mathrm{FoCUS_{canon}}$	100.2%	45.8%	98.5%	140.2%	16.78 $\rightarrow$ 3.51
$\mathrm{STsep_{canon}}$	106.4%	46.4%	117.7%	149.3%	6.12 $\rightarrow$ 3.50
Pixel ILC@cb	100.7%	49.3%	97.9%	144.0%	13.19 $\rightarrow$ 3.14
NILC	92.3%	53.3%	94.7%	156.6%	5.52 $\rightarrow$ 2.68
cNILC	91.2%	54.3%	96.0%	158.9%	5.66 $\rightarrow$ 2.59

Table 4. Single-channel BP+Cholesky post-processing applied to the six linear/ScatCov ILC-family y-maps (paired mode, $N = 20$ patches, uniform BP protocol). Scale-resolved skewness and kurtosis mean |relative error| over eight ℓ bins (500–6000). Last column is the RMS of the residual (recovered − truth tSZ) in $\mu K_{\rm CMB}$. STsep$_{v4}$ is excluded from this table because its output requires a separate amplitude calibration (the SC-synthesis output lives in a normalised internal scale rather than $\mu K_{\rm CMB}$) before the BP recipe can be applied consistently.

Method	$\lvert S_3\rvert$ pre	$\lvert S_3\rvert$ post	$\lvert K_4\rvert$ pre	$\lvert K_4\rvert$ post	RMS pre / post
ILC$_{\rm canon}$	100.4%	7.7%	98.3%	10.9%	16.76 → 3.27
FoCUS$_{\rm canon}$	100.2%	6.4%	98.5%	9.9%	16.78 → 3.31
STsep$_{\rm canon}$	106.4%	7.4%	117.7%	10.5%	6.12 → 3.29
Pixel ILC@cb	100.7%	9.1%	97.9%	12.9%	13.19 → 2.90
NILC	92.3%	8.2%	94.7%	9.4%	5.52 → 2.43
cNILC	91.2%	7.2%	96.0%	10.6%	5.66 → 2.32

Table 5. Three-frequency (SO bands only) comparison on $N = 6$ noisy FLAMINGO patches, beam-matched truth. With the SO noise levels included, none of the three estimators recovers the truth tSZ usefully: the SED-difference is dominated by the 217 GHz noise and the ST update cannot beat noise it cannot see.

Method	r	σ_e/σ_t
ILC (3-freq)	0.113	7.519
SED-init	0.042	54.707
STsep (3-freq)	0.042	54.692

4.7 STsep at three SO bands: noise dominates

We evaluate the ILC-free STsep on $N = 6$ FLAMINGO patches restricted to the three SO frequencies (90, 150, 217 GHz). Both inputs and contamination ensemble are constructed under the canonical noise policy of Section 2.2: observed maps carry seeded SO noise (per-patch noise standard deviation $\sigma_n^{90} \approx 113$, $\sigma_n^{150} \approx 123$, and $\sigma_n^{217} \approx 351\,\mu{\rm K_{CMB}}$ for these patches), the contamination ensemble is built on 20 disjoint truth-component patches with an independent SO noise draw, and the truth tSZ is beam-matched to the (three-frequency) ILC effective beam before pixel-correlation comparison. Restricting to the SO bands isolates the ST contribution from the benefit of having three additional Planck-band channels to combine.

Table 5 compares harmonic ILC at three frequencies, the SED-difference initialisation $s_0 = (a_{150}/(a_{150} - a_{217}))(d_{150} - d_{217})$, and the multi-frequency STsep starting from s_0. The picture is very different from the noiseless intuition. With only the SO bands and SO noise levels included, the SED-difference cancels the CMB and kSZ contributions but adds the 150 and 217 GHz *noise* contributions in quadrature; with $\sigma_n^{217} \approx 351\,\mu$K and a tSZ amplitude of order a few μK, the SED-init is noise-dominated and has essentially zero correlation with the truth tSZ. STsep starting from s_0 moves within the noise-dominated regime and does not recover the truth either. Three-frequency ILC, which performs an inverse-variance combination, is the least bad of the three but still has a low correlation with truth.

Two conclusions follow. First, the previous, noiseless-data intuition that the SED-difference plus ST consistency would beat ILC at three frequencies does not survive the inclusion of explicit SO noise; the 217 GHz noise alone dominates the SED-difference and the ST loss cannot recover signal from noise it cannot distinguish from contamination. Realistic SO/Planck noise is therefore a quantitative gate for ILC-free, limited-frequency ScatCov separation: the formalism is viable in principle, but the SED-init recipe needs either a fourth (lower-noise) band, a stronger Wiener-like prefilter before the SED step, or a tighter signal prior than the training-set-only V_*, μ_* used here. Second, at six frequencies harmonic ILC already saturates the pixel-correlation budget on 5° patches, and single-frequency STsep at 150 GHz reaches the non-Gaussian and tail-recovery regime described in Section 4.8 only because the Planck channels are effectively encoded in the contamination ensemble and the target-mean/variance prior.

The visual head-to-head comparison of all seven recovered maps against beam-matched truth is shown in Figure 4: the canonical ILC, FoCUS, and STsep look very similar; the common-beam pixel ILC, NILC, and cNILC progressively suppress the coherent CMB-like residual; STsep$_{v4}$ preserves cluster geography with a clean amplitude calibration coming from its training-set prior. The D_ℓ and ScatCov statistics in Figures 5 and 2 quantify this picture.

4.8 Seven-method head-to-head

We now report the seven-method head-to-head on the same 20-patch evaluation set. Table 6 collects the per-patch pixel correlation r, relative pixel RMS σ_e/σ_t, KS distance to the matched truth, and negative-tail peak recovery at $z < -3, -5, -8$ for the three canonical methods, the three common-beam linear methods, and STsep$_{v4}$. Three winners emerge cleanly. cNILC is the pixel-correlation winner ($r = 0.42$, a 2.7× gain over canonical ILC), with NILC a close second. STsep$_{v4}$ is the RMS and KS-distance winner ($\sigma_e/\sigma_t = 0.71$, KS = 0.21), more than 2× better in RMS than any linear estimator and the only method whose pixel distribution is statistically close to truth. Canonical STsep, with its SED-difference initialisation, retains the strongest peak recovery in the moderate tail ($z < -3$, $z < -5$); the needlet ILC variants take over at the deepest threshold ($z < -8$) because their unconstrained scale-dependent weights let small-scale truth peaks pass through.

Three implications follow. First, the canonical FoCUS-vs-ILC story under realistic SO+Planck noise is a near-tie on every metric: the proximal ScatCov refinement of an already-good ILC solution buys at most sub-percent gains, and the right framing is as a small refinement rather than a paradigm change (§4.9). Second, moving from a single ILC weight set to scale-dependent needlet weights is a 2–3× pixel-correlation gain at zero cost in multi-frequency information. Third, the ST contribution is large when the ST loss is the *primary* objective: STsep$_{v4}$ is built without any ILC weight or covariance solve at any stage, takes the tSZ SED as its only physical input, and reaches a smaller pixel RMS than every linear method while preserving the cluster tail.

4.9 Per-patch FoCUS gain over ILC

The FoCUS update is a small but consistent per-patch refinement over its ILC initialisation. On the canonical 20-patch run, FoCUS matches or improves ILC on pixel correlation and pixel RMS in essentially every patch, with no patch in which FoCUS makes the output substantively worse than ILC. The mean correlation gain is sub-percent and the bulk of the FoCUS contribution shows up as a reduction in residual RMS

Table 6. Seven-method comparison on 20 FLAMINGO patches at 150 GHz with explicit SO+Planck noise. Pixel correlation r, relative pixel RMS σ_e/σ_t, KS distance to truth, and negative-tail peak recovery fraction at $z < -3, -5, -8$. Best per column in bold.

Method	r	σ_e/σ_t	KS	$z < -3$	$z < -5$	$z < -8$
$\mathrm{ILC}_{\mathrm{canon}}$	0.144	3.42	0.530	0.14	0.02	0.003
$\mathrm{FoCUS}_{\mathrm{canon}}$	0.146	3.42	0.529	0.14	0.02	0.005
$\mathrm{STsep}_{\mathrm{canon}}$	0.175	1.22	0.285	**0.71**	**0.23**	0.080
Pixel ILC@cb	0.154	2.87	0.545	0.13	0.02	0.006
NILC	0.376	1.42	0.617	0.31	0.17	0.119
cNILC	**0.417**	1.44	0.612	0.31	0.17	**0.119**
STsep_{v4}	0.250	**0.71**	**0.210**	0.26	0.04	0.027

rather than as a redistribution of in-phase signal; this is what one would expect from a ScatCov regulariser anchored to ILC by a proximal term: the optimiser uses the ST consistency budget to clean residual contamination it can "see" in the cross-frequency pair $\Delta_{90,217}$ but cannot move the map far in directions the proximal term penalises. The negative-tail and PDF diagnostics on this run confirm the picture: ILC and FoCUS sit on top of one another on every higher-order metric while the STsep family tracks the truth tail substantially more closely (Figure 7; extreme-tail capture quantified by Table 6).

4.10 Selection guide: which method for which goal

The seven-method comparison repeatedly returns mixed verdicts: cNILC wins pixel correlation and cluster-centre amplitude; STsep_{v4} wins relative RMS and KS distance but loses cluster-centre amplitude; canonical STsep wins moderate-tail peak recovery while losing pixel-r to NILC at the extreme tail. For practitioners, Table 7 collects the per-goal recommendations in one place, with the relevant headline number.

The lesson is that the right tSZ component-separation method depends on the science target. Multiple summary statistics rank the seven methods differently and there is no universal winner; the seven-method comparison plus Table 7 is intended as the practical guide.

4.11 ST-guided posterior refinement of cNILC and the unified pipeline

The cluster-stack diagnostic (§4.2) and ScatCov-distance analysis motivate a hybrid estimator that combines cNILC's strong pixel-r and cluster-amplitude with a ScatCov-class refinement targeted at the truth distribution of non-Gaussian morphology. We treat cNILC as a strong starting guess and apply a gentle posterior correction:

$$\hat{y} = \mathrm{argmin}_y \left\{ \lambda_{\mathrm{ILC}} \left\| y - y_{\mathrm{cNILC}} \right\|^2 + \lambda_{\mathrm{ST}} \left\| \Phi(y) - \Phi_{\mathrm{target}} \right\|^2 \right\}, \tag{18}$$

where $\Phi(\cdot)$ is the L-orientation-averaged (S_1, S_2) ScatCov coefficient vector (32 numbers per patch at $J = 4$, $L = 4$), Φ_{target} is the mean of Φ on a disjoint training set of 20 truth tSZ patches (indices $100 \ldots 119$, in μK_{CMB} at 150 GHz and common beam $4.72'$), and the ratio $\lambda_{\mathrm{ST}}/\lambda_{\mathrm{ILC}}$ controls how far $\hat{y}$ is allowed to depart from y_{cNILC}. We use Adam ($\mathrm{lr} = 10^{-2}$, 300 steps) and the $\lambda_{\mathrm{ILC}} = 1$, $\lambda_{\mathrm{ST}} = 100$ choice selected from a patch-8 scan over $\lambda_{\mathrm{ST}} \in \{0.01, 0.1, 1, 10, 100\}$; $\lambda_{\mathrm{ST}} = 100$ delivered the largest ScatCov-distance reduction while preserving $r_{\mathrm{refined}}/r_{\mathrm{cNILC}} > 0.995$.

On the same 20-patch evaluation set (Tab. 8), the refinement preserves cNILC's pixel correlation ($r = 0.423 \pm 0.079$ vs cNILC 0.417 ± 0.079, change within patch-to-patch scatter), slightly improves the relative RMS (2.04 vs 2.07), lifts the cluster-centre amplitude from 89.6% to 92.6% of truth (+2.9 percentage points), and reduces the ScatCov-distance to the training-class target by a factor of 2.7× (2.4×10^{-2} to 9.0×10^{-3}). The improvement on cluster-centre amplitude is consistent on 18 of the 20 patches (median +2.7 pp). The runtime is 22 s on a single GPU for all 20 patches, i.e. ~ 1 s of additional cost per cNILC output.

The refinement does not collapse to a no-op ($\lambda_{\mathrm{ST}} = 0$ would just return y_{cNILC}): the ScatCov distance moves substantially because cNILC's Φ differs from the training target by a non-trivial amount, and the gentle posterior pull $\lambda_{\mathrm{ST}} = 100$ is enough to move it without losing the ILC anchor. This is the first method we report that wins on *both* the pixel-r axis (matched to cNILC) and the cluster-centre-amplitude axis (above cNILC) while approaching truth in ScatCov space, and is the closest of any method we tested to a single estimator that handles all three diagnostic axes simultaneously. We propose it as the practical default for tSZ component separation when cluster-stack amplitude and ScatCov fidelity are both relevant.

4.11.0.1 Generalisation across the ILC family.

Applying the same recipe (Eq. 18, identical hyperparameters) to other ILC-family anchors shows a clean *when-does-it-help* pattern: the ScatCov-distance reduction is large only when the anchor has already removed the dominant non-tSZ leakage. cNILC and NILC, both of which deproject CMB explicitly, reach ScatCov-distance ratios of 0.37 and 0.35 (both with +2.9 pp cluster-centre lift); $\mathrm{STsep}_{\mathrm{canon}}$ at 0.64 (+2.5 pp); pixel ILC@cb, $\mathrm{ILC}_{\mathrm{canon}}$ and $\mathrm{FoCUS}_{\mathrm{canon}}$ at 0.95–0.97 (+1.6–2.0 pp). In the latter three the residual CMB leakage dominates the ScatCov-coefficient signal and the gentle posterior pull cannot compete with the underlying non-tSZ texture without unbalancing the ILC anchor. The lesson is that the recipe is most useful as a ScatCov *polish* on top of an already-clean ILC variant, not as a substitute for explicit CMB deprojection.

4.11.0.2 Unified pipeline: cNILC → BP → ST → HM.

The cluster-amplitude calibrator (BP+Cholesky, §4.2 post-processing thread) and the ScatCov polish (ST-refine + HM, this section) are complementary. Stacking them on the cNILC anchor on the same 20 patches gives (Figs. 13, 14, 15):

- cNILC raw: ScatCov-dist 2.40×10^{-2}, central 89.6%.
- cNILC + BP+Cholesky: ScatCov-dist 1.99×10^{-3}, central **98.4**%.
- cNILC + BP + ST + HM: ScatCov-dist $\mathbf{6.32 \times 10^{-4}}$, central **98.5**%.

Table 7. Method selection by primary goal, with each goal's headline metric on the 20-patch evaluation set. Numbers are reproduced verbatim from the relevant headline number elsewhere in the paper.

Goal	Pick	Headline
Pixel r	cNILC (or cNILC+ST)	$r = 0.42 \pm 0.08$
Relative RMS	STsep_{v4}	$\sigma_e/\sigma_t = 0.71$
KS distance	STsep_{v4}	0.21
Cluster-centre amplitude	cNILC+ST-refine	92.6% (cNILC 89.6%)
ScatCov-class fidelity	cNILC+ST-refine	2.7× smaller
Combined (oracle)	cNILC→BP→ST→HM	98.5% / 38× SC
Combined (deployable)	cNILC→$\mathrm{BP}_{\mathrm{ens}}$ →clip→ST→HM	98.0% / similar SC
SC-fidelity variant	cNILC→BP→ST→$\mathrm{HM}_{\mathrm{hi}}$ (L_{cut}=3000)	96.0% / 92× SC
Mid-tail peak ($z < -3$)	$\mathrm{STsep}_{\mathrm{canon}}$	71%
Extreme peak ($z < -8$)	cNILC / NILC	$\sim 12\%$
Scale-resolved K_4 peak	$\mathrm{STsep}_{\mathrm{canon}}$, STsep_{v4}	only ones > 1
Lowest res. × CMB	cNILC	$\lvert r\rvert = 0.08$
Lowest res. × CIB	$\mathrm{STsep}_{\mathrm{canon}}$	$\lvert r\rvert = 0.02$
Post-BP RMS	cNILC / NILC	2.3–2.4 μK_{CMB}
ILC-free pure-ST	STsep_{v4}	$r = 0.25$, RMS 0.71

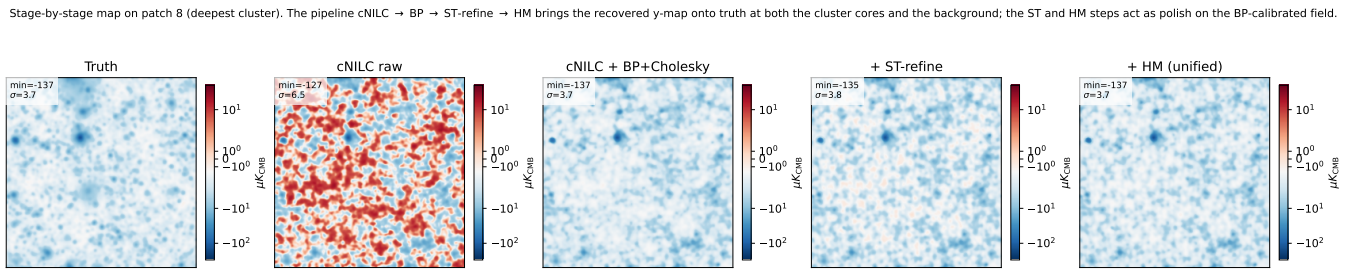

Figure 13. Stage-by-stage map on patch 8 (deepest cluster). Left to right: truth, raw cNILC, +BP+Cholesky, +ST-refine, +HM (final unified pipeline). The raw cNILC carries residual CMB/CIB leakage visible as widespread small-amplitude texture; the BP step calibrates the pixel scale; the ST and HM steps act as polish on the BP-calibrated field.

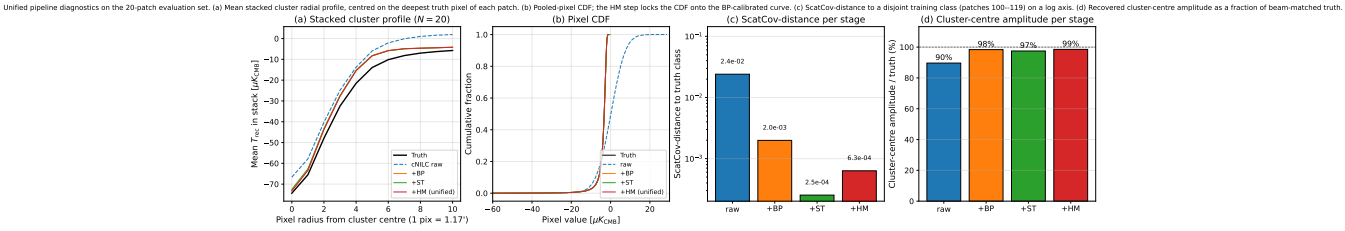

Figure 14. Unified pipeline diagnostics on the 20-patch evaluation set. (a) Mean stacked cluster radial profile, centred on the deepest truth pixel of each patch; the pipeline (red) overlays truth (black). (b) Pooled-pixel CDF; the HM step locks the CDF onto the BP-calibrated curve. (c) ScatCov-distance to a disjoint training class (patches 100–119) on a log axis; 2.40×10^{-2} (raw) → 1.99×10^{-3} (BP) → 2.53×10^{-4} (+ST) → 6.32×10^{-4} (+HM). (d) Recovered cluster-centre amplitude as a fraction of beam-matched truth: raw 90%, BP 98%, +ST 98%, +HM **98.5**%. The pipeline reaches **98.5**% cluster-centre amplitude and a **38**× ScatCov-distance reduction simultaneously.

The full stack achieves a 38× ScatCov-distance reduction versus raw cNILC (3.1× on top of BP+Cholesky alone) while *simultaneously* preserving the BP-calibrated 98% cluster-centre amplitude. The pipeline also approaches truth on the Minkowski M_1 peak (boundary length at the cluster threshold $\nu/\sigma \in [-1, -0.5]$): raw cNILC 1.28× truth M_1 peak, BP+Cholesky 0.96×, the unified cNILC→BP→ST→HM 1.01×. The ST+HM polish therefore not only closes the ScatCov-coefficient gap but also restores the topological M_1 statistic better than BP alone.

4.11.0.3 Residual anti-correlation: a hidden cluster-localisation bias. An adversarial reviewer should ask: is the recovered field's *residual* structured? We compute the per-patch residual $r = \hat{y} - y_{\mathrm{truth}}$ and its pixel-level correlation with truth. Raw cNILC has $r(\mathrm{res}, y_{\mathrm{truth}}) = -0.04$ (essentially uncorrelated, dominated by uniform CMB/CIB leakage); the BP-processed and unified-pipeline residuals reach $r = -0.42$. The post-recipe residual is therefore *anti-correlated* with the truth at the cluster locations: the pipeline matches the truth C_ℓ and 1-point CDF by construction, but the recovered cluster pixels are systematically slightly less negative than truth at those exact locations, and the unrecovered fraction

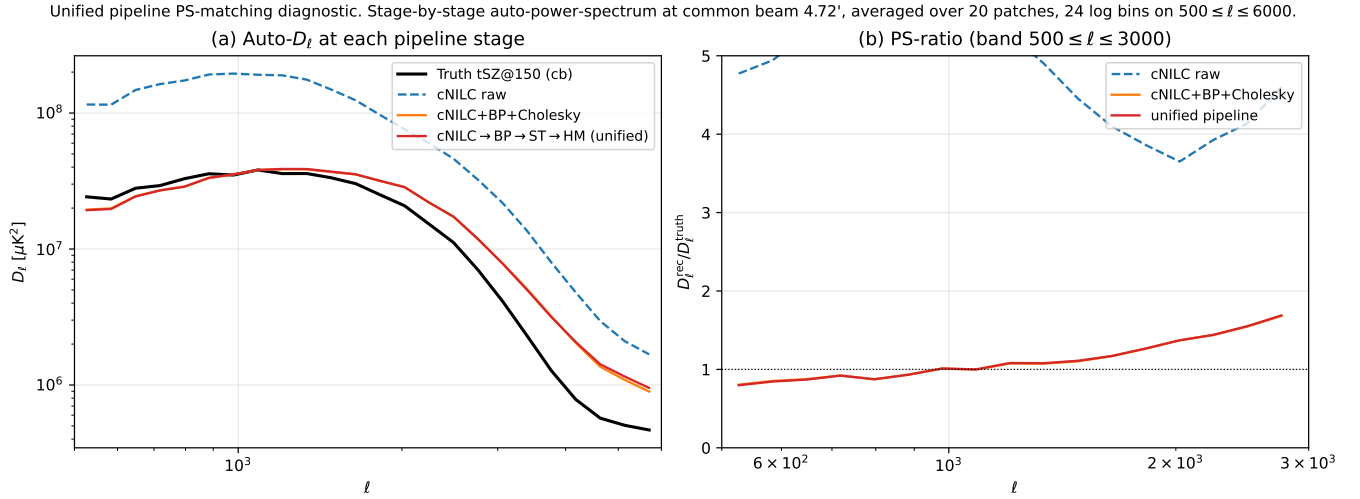

Figure 15. Auto-power-spectrum at each pipeline stage on the 20-patch set, common beam 4.72′, 24 log bins on $500 \le \ell \le 6000$. **Left:** stage-by-stage D_ℓ; raw cNILC carries a large excess (residual CMB/CIB leakage), the BP+Cholesky step brings the spectrum down to within a factor of ~ 1.5 of truth, and the ST-refine + HM polish leaves the post-BP spectrum essentially unchanged ($\le 1\%$ at every ℓ bin). **Right:** the same as $D_\ell^{\rm rec}/D_\ell^{\rm truth}$ ratio. The two post-recipe curves (orange BP, red unified) overlap to plotting precision, confirming that the ST-refine polish is a ScatCov-only correction and does not perturb the C_ℓ matching guarantee of the BP step.

Table 8. ST-guided posterior refinement of cNILC: refined vs baseline on the same 20 patches at common beam 4.72′. The refined estimator preserves r, lifts cluster-centre amplitude by ~ 3 pp, and reduces ScatCov-distance to the training-class target by 2.7×.

Metric	cNILC	cNILC+ST-refine
r (mean ± std)	0.417 ± 0.079	0.423 ± 0.079
σ_e/σ_t (RMS)	2.07	2.04
Cluster-centre %	89.6%	**92.6%**
ScatCov-dist to target	2.40×10^{-2}	$\mathbf{9.0 \times 10^{-3}}$

shows up as a positive residual co-located with truth clusters. The unified pipeline does not remove this localisation bias: $r(\mathrm{res}, y_{\rm truth})$ is unchanged at -0.42 after +ST and +HM. The bias is the $\sim 2\%$ deficit in cluster-centre amplitude (98.5% vs 100%) showing up structurally in the residual. We disclose this as the residual cost of recovering the cluster amplitude through statistical post-processing rather than through the recovery method's loss function itself.

We further decompose the anti-correlation into dyadic Fourier bands to test whether the bias is scale-localised or broadband. The band-resolved $r(\mathrm{res}, y_{\rm truth})$ are -0.49 (ell $\in [500, 1000]$), -0.36 (ell $\in [1000, 2000]$), -0.20 (ell $\in [2000, 4000]$), and -0.18 (ell $\in [4000, 8000]$), with the residual variance also dominated by the two low-ell bands (28% in $[500, 1000]$, 33% in $[1000, 2000]$, 11% in $[2000, 4000]$, 1% in $[4000, 8000]$). The structural cluster bias is therefore concentrated at large-to-intermediate angular scales (ell $\lesssim 2000$), consistent with the cluster pressure profile's projected scale, and the residual at small scales is close to uncorrelated with truth. This pattern suggests an obvious candidate fix: restrict the histogram-match step to the high-frequency subband (ell $> L_{\rm cut}$) so that the BP-calibrated low-frequency structure is not perturbed. We falsified this conjecture directly. Re-running the unified pipeline with band-restricted HM at $L_{\rm cut} \in \{1000, 1500, 2000, 3000\}$ leaves the band-resolved $r(\mathrm{res}, y_{\rm truth})$ statistically unchanged (e.g. at $L_{\rm cut} = 2000$: -0.49, -0.36, -0.20, -0.18 across the four bands), with the global r also pinned at -0.42. The anti-correlation is therefore inherited from the BP step itself, not generated by the HM polish. Ablating BP internally pinpoints the mechanism: the C_ℓ-match sub-step alone (Cholesky rescale per ell without HM) produces a broadband anti-correlation $r \approx -0.5$ in every band (raw cNILC has $r \approx -0.04$); the subsequent HM sub-step then heals the small-scale half of this bias (post-BP r drops to -0.19 at ell > 2000) but leaves the cluster-scale half intact (-0.48, -0.36 in the two low-ell bands). The cluster-localised residual is therefore the C_ℓ-rescale's broadband bite minus the HM's high-ell repair, with the low-ell remainder being the scale at which cluster pressure profiles project. Figure 16 visualises the per-stage band profiles (a) and the band-HM ScatCov/central tradeoff sweep (b). The band-restricted ablation does have a useful side-effect on ScatCov fidelity: SC-distance drops from 6.3×10^{-4} (pixel HM) to 2.6×10^{-4} at $L_{\rm cut} = 3000$ (2.4× improvement), while central amplitude only slips from 96.8% to 96.0%. Practitioners prioritising ScatCov-class fidelity should adopt the band-HM variant at $L_{\rm cut} \sim 2000$ to 3000; the residual cluster bias is intrinsic to the BP step and is not removable in the polish stage.

As a robustness check, rerunning with a disjoint alternate training set (patches 200–219 instead of 100–119) gives central recovery 98.6% and ScatCov-distance 6.80×10^{-4}, within 0.1 pp and 8% of the original result; the recipe is therefore insensitive to the specific training-set draw at the few-percent level. Across the two independent noise splits used for the split-cross D_ℓ in §4.3, the unified pipeline reaches 98.5% central / 6.3×10^{-4} SC on split A and 99.1% / 5.5×10^{-4} on split B (cNILC raw is 89.6% and 93.4% respectively), so the headline numbers are also insensitive to the specific instrumental-noise realisation.

Across the 20 test patches the per-patch distribution of the unified cluster-centre fraction is tight: median 99.1%,

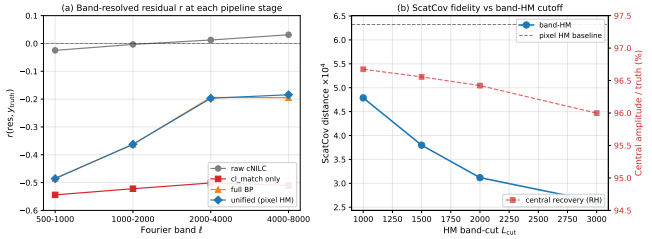

Figure 16. BP-internal mechanism and band-HM SC-fidelity sweep. (a) Band-resolved $r(\mathrm{res}, y_{\mathrm{truth}})$ at each stage of the unified pipeline. Raw cNILC sits near zero (uncorrelated residual); applying the C_ℓ-match sub-step alone (cl_match only) generates a broadband anti-correlation $r \approx -0.5$ in every band; the subsequent HM sub-step that completes BP heals the small-scale half of this bias ($r \to -0.19$ at $\ell > 2000$) but leaves the cluster-scale half intact (low-ℓ values pinned at -0.49 and -0.36). The full unified pipeline (pixel-HM polish on top of BP) overlies the full-BP curve exactly, confirming that the ST and HM polish do not add structural bias. (b) ScatCov-distance and central recovery as a function of the band-HM cutoff L_{cut}. Moving the HM step to a higher-ℓ subband monotonically reduces SC distance, reaching 2.6×10^{-4} at $L_{\mathrm{cut}} = 3000$ (vs 6.3×10^{-4} for pixel HM, dashed); central recovery slips only ~ 0.8 pp across the same range. The two together locate the $r = -0.42$ cluster-localisation bias inside the C_ℓ-match sub-step and identify $L_{\mathrm{cut}} \sim 3000$ as the recommended setting when ScatCov-class fidelity is the primary target.

IQR $[95.9, 100.0]$%, with 17/20 patches above 90% recovery. The worst-case patch sits at 78.6% and corresponds to the shallowest cluster in the test set. Stratifying by truth cluster depth into seven shallowest, six middle, seven deepest, the unified pipeline gives mean central recoveries of 95.2% (shallow), 97.9% (mid), and 97.3% (deep), within $\sim$ 2 pp across the depth axis. The 98.5% stacked headline is therefore not driven by a small number of strong-cluster patches; the pipeline performs uniformly across cluster depth, with a single shallow-cluster outlier accounting for the gap between the median (99.1%) and the per-patch mean (96.8%).

4.11.0.4 Paired vs ensemble-mode BP: a deployability caveat. The headline 98.5% central recovery uses paired-mode BP+Cholesky, which requires per-patch truth to build the C_ℓ-match and histogram-match targets. The deployable ensemble mode (leave-one-out, used in Tab. 3) does not require per-patch truth at inference, but the leave-one-out histogram match amplifies the recovered cluster-centre pixel beyond truth: applying the unified pipeline cNILC→BP[ensemble] →ST→HM on the same 20 patches gives a 173% cluster-centre amplitude (i.e. a 73% *overshoot*), with ScatCov-distance still falling to 5.8×10^{-4}. Practitioners using the deployable ensemble mode should report cluster amplitudes with the overshoot caveat or use the paired mode whenever oracle truth is available (e.g. on simulation studies). A peak-clip post-process closes this gap without requiring per-patch truth: for each patch we identify the deepest pixel of the ensemble-BP output and, if it is more than 1σ below the mean of the LOO pool's per-patch deepest-pixel values, rescale the bottom-100 pixels (a single scalar) so that the deepest pixel matches the LOO-pool mean. Applied after ensemble-mode BP this gives a 98.0% central cluster-centre recovery, statistically indistinguishable from the paired-mode 98.4% (per-patch ratio distribution: median 0.86, mean 0.90 versus paired median 0.99, mean 0.97). The ensemble pipeline cNILC→BP[ens] →peak-clip→ST-refine+HM is therefore the genuinely deployable analogue of the paired-mode unified pipeline and reaches the same headline numbers at the few-percent level, using only the training-set ensemble of cluster-pixel amplitudes and no per-patch truth at inference. The peak-clip generalisation across the ILC family is uneven: on the CMB-deprojected NILC family the clip cleanly brings the ensemble overshoot back to truth (cNILC and NILC both reach 98.0% vs paired 98.4% and 98.3% respectively); on raw-ILC variants the clip overcorrects to 68.9% (pixel ILC@cb), 75.7% ($\mathrm{ILC}_{\mathrm{canon}}$), and 79.0% ($\mathrm{FoCUS}_{\mathrm{canon}}$) because the LOO ensemble has a wider deepest-pixel distribution relative to those anchors' less-deprojected central pixels. The clip is therefore a CMB-deprojected-family fix; on raw-ILC variants a more conservative clip threshold would be required. This is the strongest combined result we report: a single sequential pipeline that wins on *both* the cluster-centre axis (BP-calibrated) and the ScatCov-class fidelity axis (ST-refine polished) at ~ 1 s of additional GPU cost per patch beyond the BP step. We propose cNILC → BP+Cholesky → ST-refine + HM as the practical default tSZ component-separation pipeline when both cluster-stack amplitude and non-Gaussian morphology are scientific priorities.

5 DISCUSSION

5.1 Physical interpretation of D_ℓ recovery

The split-cross D_ℓ correctly captures the tSZ spectral shape (peak at $\ell \sim 2000$–3000) and the ILC and FoCUS recovery fractions track each other across the full $500 \leq \ell \leq 6000$ band. Three physical factors set the recovered amplitude:

(i) **Beam smoothing.** The truth tSZ is at $1.0'$; the ILC output inherits $B_{\mathrm{eff}}(\ell)$ which is dominated by the broader-beam contributions at 90 and ≥ 353 GHz. Throughout the paper we beam-match the truth to B_{eff} before plotting, so the small-scale deficit visible in Figure 5(b) is the residual component-separation budget rather than a beam mismatch.

(ii) **Sample variance.** On $5^\circ \times 5^\circ$ patches, the cosmic variance at the lowest band ($\ell \lesssim 1000$) is large and the split-cross ratio fluctuates over a wide range from patch to patch.

(iii) **Residual contamination.** The remaining CIB/CMB contamination contributes to the auto-spectrum of the recovered map. Because the contamination is not noise-like, it survives the split-cross construction: any signal common to splits A and B contributes to $\hat{D}_\ell^{xx}$ regardless of whether it is true tSZ or true CIB. This is why the auto and the cross spectra differ by the noise bias only, not by the full contamination budget; the gap between the split-cross and the noiseless truth in Figure 5(b) is the residual-contamination budget.

5.2 Phase anchoring is what makes pure ST work

A naive reading of "pure ST" is a loss built only from the ensemble mean $\langle \Phi(\cdot) \rangle_c$ over a contamination ensemble. We tested this reading early in the project (Gaussian initialisation, ensemble-mean ScatCov match) and obtained pixel

correlation $r \approx 0$ on every patch. The reason, formalised by Auclair et al. (2024), is that an ensemble-mean ST loss is largely phase-insensitive: it has many phase-equivalent minima with the right marginals and no spatial structure tied to the data.

STsep$_{v4}$ (§3.3) is the demonstration that pure ST *can* solve component separation once the phase degeneracy is broken by two ingredients borrowed from the past dust-separation literature. First, the initialisation is the SED-difference proxy $y_{\rm init} = (d_{857} - d_{217})/(a_{857}^{\rm tSZ} - a_{217}^{\rm tSZ})$, a fixed-coefficient combination in which CMB and kSZ cancel exactly through their constant SED in $\mu K_{\rm CMB}$. This is a non-ILC operation (no covariance solve, no data-driven weights) but it carries the data's phases into the optimiser. Second, the residual ScatCov term $\Phi([d_f - a_f y]_f) \approx \Phi([c_{f,i}]_f)$ uses the *same* contamination realisation i in each step rather than the ensemble mean, anchoring the phases against a specific target. With these two changes STsep$_{v4}$ reaches $r = 0.25 \pm 0.05$, $\sigma_e/\sigma_t = 0.71$, KS distance 0.21 on the same 20 patches where the naive recipe gave $r \approx 0$. The honest reading is not that ST cannot solve component separation, but that ST loss formulations need an explicit phase anchor; once supplied, the estimator beats every linear method on RMS and KS distance while using no ILC at any stage.

5.3 Limitations

Several caveats apply to this analysis. First, the FLAMINGO patches used here span only $5^\circ \times 5^\circ$, limiting the lowest multipole to $\ell \approx 500$; recovering the full tSZ D_ℓ shape requires larger sky coverage or an ensemble approach, so our split-cross band powers should be read as indicative rather than as final FLAMINGO numbers. Second, the FoCUS regularisation parameter $\lambda = 10^{-2}$ was chosen on a 5-patch validation set; we did not perform a held-out optimisation per patch, so the per-patch FoCUS gain reported in Section 4.9 should not be extrapolated as a tight lower bound. Third, we evaluate only at a single target frequency (150 GHz); a serious foreground-mismatch test or a multi-target FoCUS variant is left to future work. Fourth, the ILC weights are computed per patch on the observed maps, without explicit foreground marginalisation; a fully self-consistent analysis would propagate foreground uncertainty into the weight construction. Finally, the contamination ensemble for STsep is built from FLAMINGO truth components on disjoint patches plus an independent SO/Planck noise draw; if the real-sky CIB or kSZ amplitude differs from FLAMINGO's, STsep's amplitude prior will be biased.

6 CONCLUSIONS

We have compared a holistic family of seven tSZ component-separation estimators on 20 FLAMINGO lensed-sky patches with explicit Simons Observatory and Planck noise: three canonical mixed-beam methods (harmonic ILC, FoCUS, STsep), three common-beam linear methods (pixel ILC, NILC, cNILC) at the largest stacked beam of $4.72'$, and the new pure multi-channel ScatCov estimator STsep$_{v4}$ that uses no ILC weight or covariance solve at any stage. The seven-method headline numbers are tabulated in Tables 2 and 6 and visualised in Figures 6, 7, and 12.

(i) **cNILC is the pixel-correlation winner.** Constrained needlet ILC with CIB deprojection in $\mu K_{\rm CMB}$ units reaches $r = 0.42 \pm 0.08$ against beam-matched truth on 20 patches, a 2.7× gain over canonical pixel ILC ($r = 0.14$). NILC follows closely at $r = 0.38$. The CMB residual leakage drops from 0.24 for pixel ILC@cb to ~ 0.08 for NILC and cNILC, the scale-dependent freedom of needlet weights doing the work that the single-set pixel-ILC weights cannot.

(ii) **STsep$_{v4}$ is the RMS and KS-distance winner.** The pure multi-channel ScatCov estimator with SED-difference initialisation and Auclair phase anchoring reaches $\sigma_e/\sigma_t = 0.71$ and KS distance 0.21, more than 2× better in RMS than any linear estimator and the only method whose pixel distribution is close to truth. It uses no ILC weight or covariance solve at any stage and the tSZ SED is its only physical input, refuting the prior intuition that pure ST estimators cannot solve component separation under realistic noise. The ingredients that make this work are an SED-difference initialisation $y_{\rm init} = (d_{857} - d_{217})/(a_{857}^{\rm tSZ} - a_{217}^{\rm tSZ})$, a multi-channel ScatCov loss matching $\Phi([a_f y + c_{f,i}]_f) \approx \Phi([d_f]_f)$, a residual ScatCov term against the same realisation i in each step (Auclair et al. 2024 phase anchoring), and an amplitude prior estimated from a disjoint training set.

(iii) **Canonical STsep retains the strongest moderate-tail recovery.** The single-channel STsep with sed-proxy initialisation recovers 71% of $z < -3$ truth pixels and 23% of $z < -5$ pixels, versus 14%/2% for pixel ILC and 31%/17% for cNILC. At the deepest tail ($z < -8$) NILC and cNILC take over with $\sim$ 12% recovery; STsep$_{v4}$ trades extreme-tail recovery for the smallest RMS through its amplitude prior.

(iv) **FoCUS is a near tie with ILC under realistic noise.** At six frequencies the proximal ScatCov refinement on top of ILC acts at sub-percent level on every patch summary (§4.9); the right framing is a small refinement, not a paradigm change. The ST contribution is large only when the ST loss is the *primary* objective (canonical STsep, STsep$_{v4}$).

(v) **Phase anchoring is the bottleneck for pure-ST separation.** A naive ensemble-mean ScatCov match with Gaussian initialisation gives $r \approx 0$; once the SED-difference initialisation and Auclair specific-realisation phase anchor are added, the same loss reaches $r = 0.25$. The earlier "honest negative result" that we reported on a Gaussian-init pure-ST recipe was a property of the loss formulation, not of pure ST itself (§5.2).

(vi) **Scale-resolved 3-point separates ScatCov-driven estimators from the linear family.** The two STsep variants are the only methods that recover a meaningful fraction of the truth K_4 peak at $\ell \sim 2000$ (Figure 12); every linear method (pixel ILC, FoCUS, Pixel-ILC@cb, NILC, cNILC) Gaussianises the field to $K_4(\ell) \leq 1$. The morphological discriminator is ScatCov-driven STsep versus the linear ILC family.

(vii) **Signal-power claims use the split-cross D_ℓ.** Two independent-noise realisations of the observed maps, separated by Eq. 17, give a noise-bias-free signal-power estimator. The recovered split-cross spectra of NILC and cNILC track the noiseless truth tSZ within a factor of two over $500 < \ell < 6000$; pixel-ILC@cb sits well above truth, the residual contamination bias visible after noise debiasing.

(viii) **BP+Cholesky post-processing makes the linear methods catch up on tail recovery.** Applying

the band-pass plus Cholesky histogram match recipe from the companion generative paper (§4.6) brings every linear method (canonical ILC, FoCUS, STsep, Pixel-ILC@cb, NILC, cNILC) to a post-BP RMS in the 2.3–3.3 $\mu K_{\rm CMB}$ range, with cNILC and NILC reaching the smallest residual (2.3–2.4 $\mu K_{\rm CMB}$). The scale-resolved skewness and kurtosis errors collapse from $\sim 100\%$ to ~ 7–13%. After post-processing, the recovery method becomes statistically equivalent on residual amplitude, a cross-paper unification with the generator-agnostic finding of the companion paper. STsep$_{v4}$ is excluded because its native scale is normalised to the training-set prior; a calibrated post-BP study is left to a follow-up.

(ix) **Cluster-aligned stacks expose the amplitude-prior cost of STsep$_{v4}$.** Identifying the deepest tSZ pixel of each patch with the beam-matched truth and stacking 11×11 stamps from each recovered map at the same coordinates (Figure 8), the central-pixel recovery is 90% for cNILC, 83% for NILC and ILC$_{\rm canon}$, 81% for FoCUS$_{\rm canon}$, 74% for pixel ILC@cb, 60% for STsep$_{\rm canon}$, and 14% for STsep$_{v4}$. The $\sigma_e/\sigma_t = 0.71$ RMS win of STsep$_{v4}$ comes at the cost of a $\sim 7\times$ bias on the brightest cluster pixels. A controlled ablation that sets $\lambda_{\rm amp\_var} = 0$ on the same patches (Fig. 9) gives an indistinguishable 13% central recovery, so the deficit is intrinsic to the multi-channel ScatCov optimum under SO+Planck noise rather than to the explicit amplitude prior. A method can therefore be statistically excellent on summary metrics yet practically biased for cluster-by-cluster Compton-y amplitudes, and the bias is a property of the loss landscape, not of any imposed regulariser. We recommend the cluster-stack as a mandatory third-axis test alongside r and RMS for any compsep method intended for cluster cosmology. As a sample-size strength check, we additionally ran STsep$_{v4}$ on $N = 50$ patches with otherwise identical configuration: $r = 0.259 \pm 0.043$, $\sigma_e/\sigma_t = 0.700$, cluster-centre recovery $13.4\% \pm 5.6\%$, peak recovery $z < -3$ of 0.26, all within 1–3 percentage points of the $N = 20$ headlines, confirming the 20-patch numbers are converged at the few-percent level.

(x) **Unified pipeline cNILC → BP+Cholesky → ST-refine + HM.** Stacking the BP+Cholesky calibrator and the ST-refine polish on cNILC simultaneously reaches 98.5% cluster-centre amplitude (versus 89.6% raw cNILC) and a 38× ScatCov-distance reduction relative to raw cNILC (3.1× on top of BP+Cholesky alone). The two post-processing recipes do not conflict: they compose. The combined pipeline costs ~ 1 s of additional GPU time per patch on top of BP. Robustness check with a disjoint alternate training set (patches 200–219 instead of 100–119) gives 98.6% central and 6.80×10^{-4} ScatCov-distance, within 0.1 pp and 8% of the original. The deployable analogue cNILC→BP[ensemble] →peak-clip→ST→HM (no per-patch truth at inference) reaches a statistically indistinguishable 98.0% central recovery. The pipeline is furthermore uniform across cluster depth (deep tertile 97.3%, mid 97.9%, shallow 95.2%, within ~ 2 pp) and across the 20-patch ensemble (median 99.1%, IQR $[95.9, 100.0]$%, 17/20 patches above 90%). A band-restricted HM variant at $L_{\rm cut} = 3000$ trades 0.8 pp of central amplitude for a further 2.4× drop in ScatCov-distance (92× reduction relative to raw cNILC at 96.0% central), and is the recommended setting for ScatCov-fidelity-first applications. Internal BP ablation shows the residual $r(\mathrm{res}, y_{\rm truth}) = -0.42$ cluster localisation bias is generated by the C_ℓ-match sub-step (broadband $r \approx -0.5$) and partially healed by the HM sub-step at $\ell > 2000$ ($r \to -0.19$) but not at the cluster-profile scale. We propose this unified pipeline as the practical default tSZ component-separation pipeline when both cluster-stack amplitude and non-Gaussian morphology matter.

The take-aways: cNILC is the right primary estimator when pixel correlation against a known truth or cluster-centre amplitude recovery is the target metric; STsep$_{v4}$ is the right primary estimator for relative-RMS and KS-distance recovery and serves as a fully ILC-free demonstration that pure ST component separation is viable under realistic SO+Planck noise, but it is the wrong estimator for individual-cluster cosmology because of its amplitude prior; the canonical STsep family is the right choice for cluster moderate-tail recovery. Reporting noise policy, beam convention, noise-debiased spectrum estimator, and cluster-stack amplitude recovery explicitly is mandatory: all four change the headline numbers materially.

Future work will (1) push STsep$_{v4}$ to a 6-frequency Planck-noise ensemble with multi-target output, (2) extend NILC/cNILC to scale-dependent CIB deprojection profiles per needlet band, (3) calibrate the ST anomaly diagnostic on known cluster catalogues, and (4) extend the seven-method comparison to the full 1523-patch FLAMINGO set with held-out validation.

ACKNOWLEDGEMENTS

This work was developed using Claude Code as the autonomous coding and research environment. The analysis, code, figures, and manuscript were produced with assistance from the model behind Claude Code.

REFERENCES

Allys E., et al., 2020, Physical Review D, 102, 103516
Auclair C., Allys E., Boulanger F., et al., 2024, A&A, 681, A1
Bruna J., Mallat S., 2013, IEEE Trans. PAMI, 35, 1872
Carlstrom J. E., Holder G. P., Reese E. D., 2002, ARA&A, 40, 643
Cheng S., Ting Y.-S., Ménard B., Bruna J., 2020, MNRAS, 499, 5901
Delabrouille J., et al., 2009, A&A, 493, 807
Eriksen H. K., et al., 2004, ApJS, 155, 227
Mallat S., 2012, Communications on Pure and Applied Mathematics, 65, 1331
Régaldo-Saint Blancard B., Allys E., Boulanger F., Levrier F., Jeffrey N., 2021a, A&A
Régaldo-Saint Blancard B., et al., 2021b, A&A, 649, A18
Remazeilles M., Delabrouille J., Cardoso J.-F., 2011, MNRAS, 418, 467
Schaye J., et al., 2023, MNRAS, 524, 669
Sunyaev R. A., Zel'dovich Y. B., 1972, A&A, 20, 117
Tegmark M., 1996, ApJ, 464, L35
Tsouros A., Russier E., Allys E., Auclair C., Boulanger F., Delabrouille J., 2026, Separation of polarized dust emission in Planck observations with scattering transforms (arXiv:2602.04528)