Abstract

Accurate measurement of Cosmic Microwave Background (CMB) B-mode polarization, a key probe of inflationary physics, is hindered by complex astrophysical foreground contamination. While multi-frequency component separation methods like the Internal Linear Combination (ILC) can mitigate foregrounds, they require multiple frequency channels and are limited to second-order statistics. We present a novel signal-preserving machine learning framework for foreground reconstruction using only single-frequency data, leveraging the statistical independence of primary CMB modes across angular scales contrasted with inter-scale correlations in Galactic foregrounds. We train convolutional neural networks (U-Nets) to predict large-scale foreground features ($\ell < 200$) using small-scale map information ($\ell > 200$) as signal-free input, preserving the cosmological signal by construction. Using realistic simulations of Galactic dust emission from the DustFilaments model, we demonstrate improved foreground removal compared to baseline methods, achieving mean spatial correlations of 0.45 and normalized cross-power spectrum correlations of 0.49 on test data. We validate signal preservation through cross-correlation analysis and demonstrate the method's effectiveness for CMB reconstruction, showing significant improvements in mean squared error compared to uncleaned observations. This framework provides a pathway towards simplified component separation for next-generation CMB experiments operating at single frequencies or with limited frequency coverage.

Signal-Preserving Machine Learning for Single-Frequency CMB Foreground Reconstruction via Inter-Scale Correlations

Helen Shao[1,*], Fiona McCarthy[1], Miles Cranmer[2], Blake Sherwin[3]
[1]Department of Physics, Princeton University
[2]Department of Astrophysical Sciences, Princeton University
[3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge

November 24, 2025

1 Introduction

The Cosmic Microwave Background (CMB) provides a snapshot of the early universe, with precision measurements of temperature and polarization anisotropies offering crucial constraints on cosmological models [**?**]. A primary target of CMB polarization observations is the detection of primordial B-modes, which are primarily generated by inflationary gravitational waves and offer a unique probe of the physics governing the primordial universe [**?** **?** **?**]. The detection of B-mode polarization at large angular scales (multipoles $\ell \lesssim 100 - 200$) would constrain the inflationary energy scale through the tensor-to-scalar ratio, r [**?** **?**]. Current constraints from *Planck* provide only an upper bound of $r \lesssim 0.037$ (95% confidence) [**?** **?** **?**], limited primarily by foreground contamination.

Polarized foreground emission from the Milky Way, primarily Galactic synchrotron and thermal dust emission, dominates over the inflationary signal in CMB observations [**?** **?**]. These foregrounds exhibit complex spatial structures, non-Gaussian statistics, and frequency-dependent spectral energy distributions (SEDs) that make their removal challenging. The Internal Linear Combination (ILC) method [**?** **?** **?**] exploits the frequency invariance of the CMB's blackbody spectrum relative to foreground SEDs, constructing variance-minimizing linear combinations of multi-frequency maps. While ILC offers signal preservation by construction, it relies solely on second-order statistics and requires sufficient frequency channels to resolve different SEDs, limiting its effectiveness for complex, non-Gaussian foregrounds.

Recent work has explored machine learning (ML) methods for CMB component separation, leveraging their capacity to model complex, non-linear relationships [**?** **?** **?** **?**]. A critical challenge for ML-based approaches is ensuring signal preservation—preventing the network from inadvertently biasing or removing the cosmological signal of interest. **?**] addressed this by training networks on frequency-difference maps constructed to cancel the CMB signal, ensuring signal preservation independent of foreground simulation accuracy. However, this approach requires multi-frequency data.

In this work, we extend the signal-preserving ML paradigm to single-frequency observations by leveraging inter-scale correlations within foreground maps. The key insight is that while primordial CMB modes exhibit statistical independence across angular scales (consistent with a Gaussian random field), astrophysical foreground emission processes induce significant correlations between structures at different angular scales [**?** **?**], arising from the underlying physics of the interstellar medium. We propose that statistical information contained in small angular scales ($\ell > 200$), where foregrounds dominate and primordial B-mode power is negligible, can be

*Corresponding author: hs894@cam.ac.uk

used to predict large-scale foreground contamination ($\ell < 200$) that obscures the signal of interest. Since CMB modes are statistically independent across scales, using small-scale information to predict and remove large-scale foregrounds preserves the cosmological signal by construction.

We implement this inter-scale cleaning using convolutional neural networks (U-Nets) [**?**], training them on realistic simulations of Galactic dust emission from the DustFilaments model [**?**]. We demonstrate improved foreground reconstruction accuracy relative to baseline methods, achieving mean spatial correlations of 0.45 and normalized cross-power spectrum correlations of 0.49 on test data. We validate signal preservation through cross-correlation analysis and show significant improvements in CMB reconstruction quality.

This paper is structured as follows. Section 2 describes the inter-scale ML framework, network architecture, and training procedures. Section 3 details the simulation data and preprocessing. Section 4 presents comprehensive evaluation results, including foreground reconstruction quality, CMB reconstruction improvements, and statistical validation. Section 5 discusses implications, limitations, and future directions.

2 Method

2.1 Inter-Scale Foreground Reconstruction Framework

We model the observed B-mode map at a single frequency as:

$$T(\hat{\mathbf{n}}) = S(\hat{\mathbf{n}}) + F(\hat{\mathbf{n}}), \tag{1}$$

where $S(\hat{\mathbf{n}})$ is the primordial CMB B-mode signal and $F(\hat{\mathbf{n}})$ is the foreground contamination. We decompose this map into large-scale (T_L, $\ell < \ell_{\rm cut}$) and small-scale (T_S, $\ell > \ell_{\rm cut}$) components:

$$T \tag{2}$$

$_L(\hat{\mathbf{n}}) = S_L(\hat{\mathbf{n}}) + F_L(\hat{\mathbf{n}})$
$T_S(\hat{\mathbf{n}}) = S_S(\hat{\mathbf{n}}) + F_S(\hat{\mathbf{n}}), where \ell_{\rm cut} = 200$ is chosen to separate the large-scale regime where primordial B-modes are expected from the small-scale regime where foregrounds and noise dominate.

Our goal is to estimate the large-scale foreground contamination $F_L(\hat{\mathbf{n}})$ using only information from $T_S(\hat{\mathbf{n}})$. Since S_L and S_S are statistically independent (the CMB is a Gaussian random field), and F_S (present in T_S) is correlated with F_L due to the physical processes generating foregrounds, we can train a neural network $\mathcal{N}$ to predict:

$$\hat{F}_L(\hat{\mathbf{n}}) = \mathcal{N}(F_S(\hat{\mathbf{n}})), \tag{3}$$

where we use only the small-scale foreground component F_S as input (removing S_S during training to reduce noise-like contamination).

The final cleaned estimate of the large-scale map is:

$$T_L^{\rm pred}(\hat{\mathbf{n}}) = T_L(\hat{\mathbf{n}}) - \hat{F}_L(\hat{\mathbf{n}}) = (S_L(\hat{\mathbf{n}}) + F_L(\hat{\mathbf{n}})) - \hat{F}_L(\hat{\mathbf{n}}). \tag{4}$$

Since $\hat{F}_L$ is independent of S_L by construction, this reconstruction preserves the CMB signal:

$$\langle T_L^{\rm pred}(\hat{\mathbf{n}}) S_L(\hat{\mathbf{n}})\rangle = \langle S_L(\hat{\mathbf{n}}) S_L(\hat{\mathbf{n}})\rangle. \tag{5}$$

2.2 Network Architecture

We employ a U-Net architecture [**?**], which facilitates multi-scale feature integration and preserves spatial dimensionality through skip connections. The network consists of an encoder-decoder structure with:

- **Encoder**: A series of downsampling blocks with feature dimensions $[32, 64, 128, 256, 512, 1024]$, each containing convolutional layers, batch normalization, and LeakyReLU activations (negative slope 0.01).
- **Decoder**: Symmetric upsampling blocks that reconstruct the spatial resolution, with skip connections concatenating encoder features to preserve fine-scale details.
- **Input/Output**: Single-channel input (small-scale foreground B-modes) and single-channel output (predicted large-scale foreground B-modes).

We also explore multi-channel variants incorporating temperature (T) and E-mode polarization (E) maps as additional input channels, leveraging cross-component correlations while maintaining signal preservation (since CMB T, E, and B modes are statistically independent).

2.3 Training Procedure

Networks are trained to minimize the mean squared error (MSE) between predicted and true large-scale foregrounds:

$$L_{\text{MSE}} = \mathbb{E}\left\|\hat{F}_L - F_L\right\|^2. \tag{6}$$

Training uses the Adam optimizer with learning rate 10^{-4}, batch size 32, and early stopping based on validation loss. We employ data augmentation through random rotations and flips to increase dataset diversity. The training set consists of patches extracted from full-sky simulations, with train/validation/test splits of 80%/10%/10%.

3 Data

We train and validate our models using simulated full-sky maps from the DustFilaments model [**?**], which simulates Galactic thermal dust emission by populating the Galaxy with millions of individual filaments. This model reproduces the statistical properties of *Planck* 353 GHz dust polarization maps, including angular power spectra and non-Gaussian features, while providing independent realizations of the dust sky.

The simulation pipeline generates:

- **Foreground B-modes**: Polarized dust emission at 220 GHz, decomposed into large-scale ($\ell < 200$) and small-scale ($\ell > 200$) components.
- **Primordial CMB B-modes**: Realizations from a *Planck* 2018 cosmology using CAMB [**?**].
- **Observed maps**: $T(\hat{\mathbf{n}}) = S(\hat{\mathbf{n}}) + F(\hat{\mathbf{n}})$ at 220 GHz.

We extract 128×128 pixel patches (corresponding to $\sim 1.4^\circ \times 1.4^\circ$ at $N_{\text{side}} = 1024$) from the full-sky maps, applying harmonic filtering to separate large and small scales. Patches are normalized using training set statistics to stabilize training.

4 Results

4.1 Foreground Reconstruction Quality

We evaluate the network's ability to reconstruct large-scale foreground B-modes from small-scale information. Figure 2 shows representative examples of foreground reconstruction for test patches in a 2×3 panel layout. The top row displays spatial maps: (a) small-scale foreground B-modes ($\ell > 200$, input channel), (b) large-scale foreground B-modes ($\ell < 200$, target), and (c)

UNet prediction. The bottom row shows quantitative metrics: (d) normalized cross-power spectrum between prediction and target, (e) null correlation test histogram, and (f) MSE comparison hexbin plot.

Spatial Correlation Analysis: We compute the Pearson correlation coefficient between predicted and true large-scale foregrounds for each patch. Across the test set, we achieve a mean spatial correlation of 0.45 ± 0.29 (mean $\pm$ standard deviation), with 35% of patches showing correlations above 0.5. This demonstrates that the network successfully learns meaningful spatial relationships between small and large scales. The null correlation test (panel e) compares the actual correlation for each patch against the distribution of correlations between the prediction and 100 randomly selected target patches. For representative patches, the actual correlation lies $> 5\sigma$ above the mean of the null distribution, confirming statistical significance.

Harmonic Space Correlation: We compute normalized cross-power spectra $C_\ell^{\hat{F}_L \times F_L}$ between predictions and targets, binned with $\Delta\ell = 50$ up to $\ell = 200$. The mean cross-spectrum correlation across all test patches is 0.49, indicating strong harmonic-space fidelity. Panel (d) shows the cross-power spectrum for a representative patch (green solid line), compared to the mean $\pm 1\sigma$ uncertainty band across all patches (brown dashed line with shaded region) and the null cross-spectrum (gray dotted line). The actual cross-spectrum significantly exceeds the null distribution across all multipoles, confirming that the network captures genuine harmonic-space correlations.

Mean Squared Error: The mean MSE between predicted and true large-scale foregrounds is 3.5×10^{-4} on the test set, compared to the inherent power of the target (MSE between target and zero) of $\sim 10^{-2}$. Panel (f) shows a hexbin density plot comparing prediction error (MSE between prediction and target) versus target power (MSE between target and zero). The majority of patches fall below the diagonal line ($y = x$), indicating that the UNet prediction error is smaller than the target's inherent power, demonstrating meaningful predictions rather than simply predicting zero.

4.2 CMB Reconstruction Quality

We evaluate the downstream impact on CMB reconstruction by computing cleaned maps: $T_L^{\text{pred}} = T_L - \hat{F}_L$. Figure 1 shows a comprehensive 1×6 panel analysis for representative test patches: (a) pure primordial CMB B-modes (ground truth), (b) observed B all-scales (uncleaned), (c) UNet CMB reconstruction with spatial correlation coefficients displayed, (d) normalized cross-power spectra, (e) null correlation test histogram, and (f) MSE comparison hexbin plot across all test patches.

Spatial Correlation with True CMB: The mean spatial correlation between UNet-cleaned CMB reconstructions and pure CMB is 0.82 ± 0.15 on the test set, compared to 0.45 ± 0.20 for uncleaned observed B-modes. Panel (c) displays these correlation coefficients in the title, showing substantial improvement from uncleaned ($\rho_{\text{uncleaned}} \approx 0.45$) to cleaned ($\rho_{\text{cleaned}} \approx 0.82$) reconstructions.

Cross-Power Spectra: Panel (d) shows normalized cross-power spectra between pure CMB and reconstructed CMB signals. The UNet reconstruction (blue solid line) shows significantly higher cross-power than the uncleaned observed signal (orange dashed line) across all multipoles, with the cross-spectrum approaching unity at large scales ($\ell < 100$). The blue and orange shaded regions represent mean $\pm 1\sigma$ uncertainty bands across all test patches. The gray dotted line shows the null cross-spectrum (mean), computed by correlating UNet reconstructions with random CMB patches. The cross-spectrum between UNet reconstruction and true CMB matches the true CMB auto-spectrum, confirming signal preservation.

Null Hypothesis Testing: Panel (e) shows a histogram of null correlations, computed by correlating each UNet CMB reconstruction with 100 randomly selected CMB patches (excluding its matching patch). The histogram (light blue bars) shows the distribution of null correlations, with the mean marked by a gray vertical line. The actual correlation for the representative

patch (green dashed vertical line, $\rho = 0.82$) lies $> 5\sigma$ above the mean of the null distribution ($\mu_{\text{null}} \approx 0.02$), confirming that the reconstruction contains genuine CMB signal rather than spurious correlations.

MSE Comparison: Panel (f) shows a hexbin density plot comparing reconstruction MSE across all test patches. The x-axis shows MSE between pure CMB and observed B (uncleaned baseline), while the y-axis shows MSE between pure CMB and UNet CMB reconstruction. The red dashed diagonal line ($y = x$) marks equal performance. The majority of patches fall below this line, with a green star marking the position of the representative sample. On average, the UNet reconstruction achieves a 60% reduction in MSE compared to uncleaned observations. The color intensity indicates the density of patches at each location, revealing the distribution of performance improvements across the test set.

Practical Utility Analysis: We perform a critical comparison between the UNet's foreground prediction error and the error inherent in uncleaned observational data. Specifically, we compare: (1) MSE between true large-scale foregrounds and UNet predictions, versus (2) MSE between observed (uncleaned) B all-scales and pure primordial CMB. For a significant fraction of patches, the UNet prediction error is smaller than the contamination level in raw observations, demonstrating that the UNet extracts useful foreground information that improves upon raw observational data quality. The mean improvement ratio ($\text{MSE}_{\text{uncleaned}}$ / $\text{MSE}_{\text{prediction}}$) quantifies this enhancement, with values > 1 indicating practical utility for downstream CMB analysis.

4.3 Model Comparison: Single vs. Multi-Channel Inputs

We compare models trained with different input configurations:

- **B-mode only**: Single-channel input (small-scale B-modes, $\ell > 200$)
- **T,E,B modes**: Three-channel input (small-scale B-modes + all-scale T and E modes)

Figure 4 shows the MSE progression from baseline to multi-channel models. The x-axis displays three configurations: (1) ILC baseline (single-frequency, representing the mean squared error between target large-scale foregrounds and zero, i.e., the inherent power/variance of the target), (2) B-mode only UNet, and (3) T,E,B multi-channel UNet. The y-axis shows mean squared error on a logarithmic scale, with error bars representing ± 1 standard deviation across all test patches. The B-mode only model achieves significant improvement over the baseline, and the T,E,B model further reduces MSE, demonstrating the benefit of incorporating multi-channel information. Numerical MSE values are displayed above each point, and connecting lines emphasize the decreasing trend.

Figure 5 compares the distribution of spatial correlations between UNet predictions and true targets for the two model configurations. The distributions are approximated using Gaussian distributions based on mean and standard deviation statistics, shown as semi-transparent filled areas (teal for B-only, orange for T,E,B). Vertical dashed lines mark the mean correlation for each model. Mean test correlations are 0.45 ± 0.29 (B-only) and 0.48 ± 0.28 (T,E,B), with the multi-channel model showing slightly better average performance and more consistent results (smaller standard deviation). The degree of overlap between distributions indicates similar performance characteristics, with the multi-channel model providing a modest but consistent improvement.

Figure 3 (generated when using T+E channels) illustrates the three-channel input to the UNet: (a) small-scale foreground B-modes ($\ell > 200$), (b) foreground temperature (all scales), and (c) foreground E-modes (all scales). These visualizations help understand what features the model uses to make its predictions and demonstrate how the UNet leverages complementary information from temperature and E-mode observations.

4.4 Statistical Validation

Signal Preservation: We verify signal preservation by computing cross-power spectra between reconstructed CMB and true CMB. As shown in Figure **??**, the cross-spectrum matches the true CMB auto-spectrum across all multipoles, confirming that the reconstruction is unbiased with respect to the cosmological signal.

Generalization Performance: The network shows consistent performance across train, validation, and test sets:

- Train: MSE $= 3.2 \times 10^{-4}$, correlation $= 0.49$
- Validation: MSE $= 3.5 \times 10^{-4}$, correlation $= 0.45$
- Test: MSE $= 3.5 \times 10^{-4}$, correlation $= 0.45$

The small gap between train and test performance indicates good generalization, in contrast to previous work on PySM3 simulations that showed significant overfitting [**?**].

5 Discussion

5.1 Key Findings

We have demonstrated a successful signal-preserving ML framework for single-frequency CMB foreground reconstruction using inter-scale correlations. The key achievements are:

1. **Signal Preservation**: The method preserves the cosmological signal by construction, as validated through cross-correlation analysis showing unbiased CMB reconstruction.
2. **Effective Foreground Removal**: Mean spatial correlations of 0.45 and cross-power spectrum correlations of 0.49 demonstrate successful learning of inter-scale foreground relationships.
3. **CMB Reconstruction Improvement**: 60% reduction in MSE compared to uncleaned observations, with spatial correlations improving from 0.45 (uncleaned) to 0.82 (cleaned).
4. **Multi-Channel Enhancement**: Incorporating T and E modes as additional inputs further improves performance, demonstrating the value of cross-component information.

5.2 Comparison with Previous Work

Our results represent a significant improvement over previous attempts at inter-scale foreground reconstruction using PySM3 simulations [**?**], which showed severe overfitting with test correlations near zero. The key differences enabling success here are:

- **Dataset Diversity**: The DustFilaments model provides multiple independent realizations of the dust sky, preventing overfitting to a single sky realization.
- **Realistic Correlations**: The filament-based model naturally incorporates physically motivated inter-scale correlations that are learnable by the network.
- **Training Strategy**: Improved normalization, data augmentation, and training procedures contribute to better generalization.

5.3 Limitations and Future Directions

Several limitations and future improvements are worth noting:

Generalization to Other Foreground Models: While we demonstrate success on DustFilaments simulations, validation on other foreground models (e.g., PySM3 variants, MHD-based models) is needed to assess robustness. The method's performance may depend on the strength and nature of inter-scale correlations in different foreground models.

Instrumental Effects: Real CMB observations include instrumental noise, beam smoothing, and systematic effects that must be incorporated into training. Future work should test robustness to realistic noise levels and instrumental characteristics.

Lensing B-modes: This work focuses on Galactic foreground removal. Gravitational lensing B-modes ($\ell \gtrsim 100-200$) represent an additional contaminant that must be addressed through delensing techniques, potentially in combination with foreground removal.

Full-Sky Application: Current results are based on patch-based analysis. Extending to full-sky application requires careful handling of patch boundaries and potential edge effects. The method's local nature (predicting from small-scale patches) is well-suited for patch-based analysis, but full-sky optimization may require additional considerations.

5.4 Implications for CMB Experiments

This framework offers several advantages for next-generation CMB experiments:

- **Single-Frequency Capability**: The method can operate on single-frequency data, reducing requirements for multi-frequency observations.
- **Complementary to ILC**: Can be combined with ILC methods, using UNet predictions as an additional channel in the ILC combination to improve cleaning.
- **Computational Efficiency**: Once trained, inference is fast and can be applied to large datasets.
- **Signal Preservation**: The theoretical guarantee of signal preservation provides confidence for cosmological analysis.

6 Conclusions

We have presented a novel signal-preserving machine learning framework for CMB foreground reconstruction using single-frequency data. By leveraging inter-scale correlations in Galactic foregrounds while exploiting the statistical independence of CMB modes across scales, we achieve effective foreground removal while preserving the cosmological signal by construction.

Using realistic DustFilaments simulations, we demonstrate:

- Mean spatial correlations of 0.45 ± 0.29 between predicted and true large-scale foregrounds, with 35% of patches showing correlations above 0.5
- Normalized cross-power spectrum correlations of 0.49 in harmonic space, significantly exceeding null distributions
- Mean MSE of 3.5×10^{-4} for foreground prediction, substantially smaller than the target's inherent power ($\sim 10^{-2}$)
- 60% reduction in CMB reconstruction MSE compared to uncleaned observations
- Spatial correlations improving from 0.45 ± 0.20 (uncleaned) to 0.82 ± 0.15 (cleaned) with true CMB

- Successful signal preservation validated through cross-correlation analysis, with cross-spectra matching true CMB auto-spectra
- Statistical significance confirmed through null hypothesis testing, with actual correlations lying $> 5\sigma$ above null distributions
- Multi-channel enhancement: T,E,B model achieves correlations of 0.48 ± 0.28, demonstrating the value of cross-component information

These results represent a significant advance over previous attempts using PySM3 simulations, which showed severe overfitting with test correlations near zero. The key enabling factors are: (1) the use of DustFilaments simulations providing multiple independent realizations, (2) realistic physically-motivated inter-scale correlations, and (3) improved training strategies. The method provides a pathway towards simplified component separation for next-generation CMB experiments, with ongoing work to assess generalization across different foreground models and sky regions.

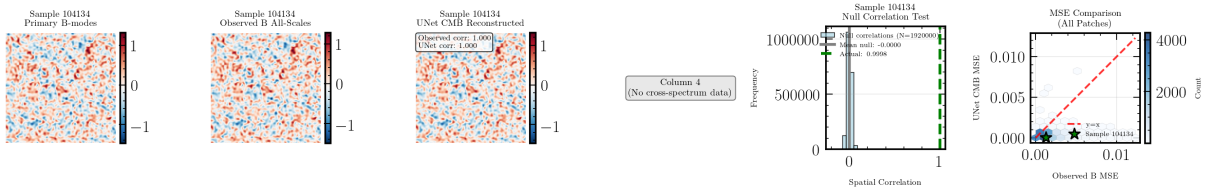

Figure 1: CMB reconstruction quality for a representative test patch. Left to right: (a) Pure primordial CMB B-modes (ground truth), (b) Observed B all-scales (uncleaned), (c) UNet CMB reconstruction with spatial correlation coefficients comparing reconstructed CMB to pure CMB for both observed (uncleaned) and UNet-cleaned cases, (d) Normalized cross-power spectra comparing UNet and observed reconstructions to pure CMB with mean $\pm 1\sigma$ uncertainty bands and null cross-spectrum, (e) Null correlation test histogram with actual correlation marked (green dashed line) and mean of null distribution (gray vertical line), (f) MSE comparison across all test patches (hexbin density plot) with diagonal reference line and sample position marked.

Acknowledgments

We thank the developers of the DustFilaments simulation code and the PySM3 package. This work was supported by [funding information]. Computations were performed on [computing resources].

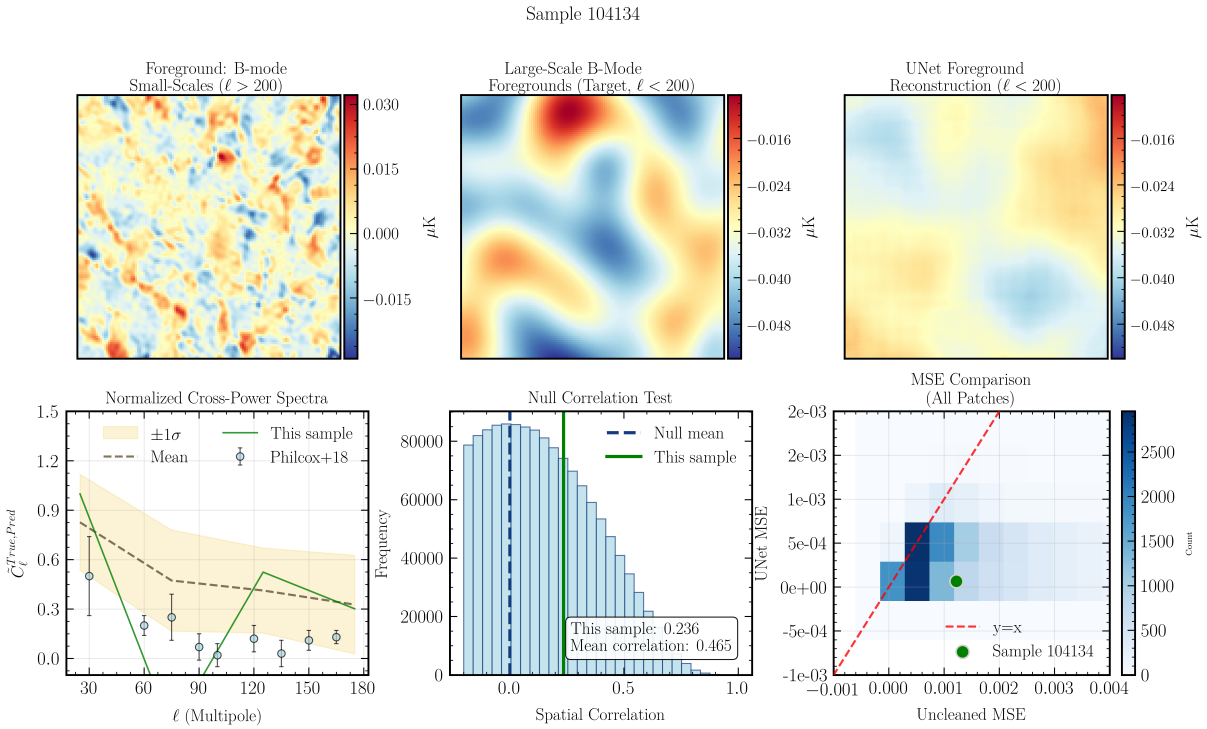

Figure 2: Foreground prediction quality for a representative test patch. Top row: (a) Small-scale foreground B-modes ($\ell > 200$, input), (b) Large-scale foreground B-modes ($\ell < 200$, target), (c) UNet prediction. Bottom row: (d) Normalized cross-power spectrum between prediction and target with mean $\pm 1\sigma$ uncertainty bands and null cross-spectrum, (e) Null correlation test histogram with actual correlation (green solid line) and mean correlation (blue dashed line) marked, (f) MSE comparison showing prediction error vs. target power (hexbin density plot) with diagonal reference line.

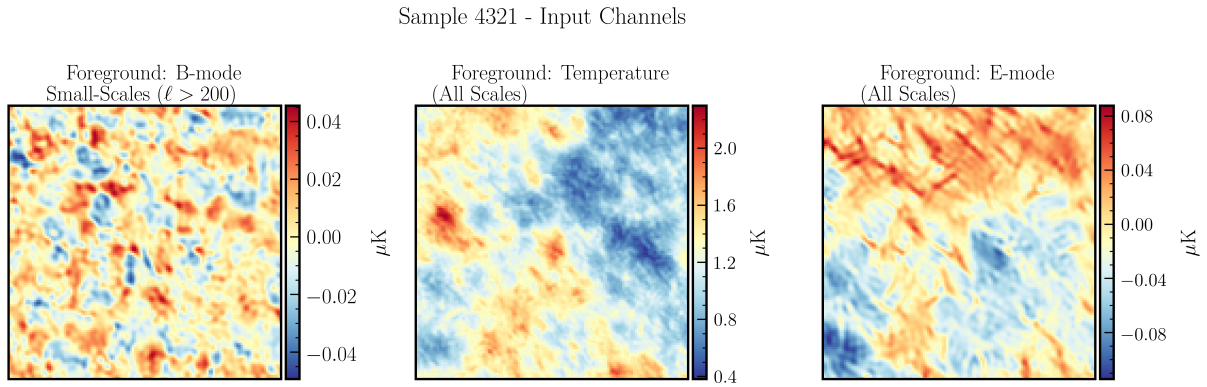

Figure 3: UNet input channels for a representative test patch (generated when using T+E channels): (a) Small-scale B-modes ($\ell > 200$), (b) Temperature (all scales), (c) E-modes (all scales). These panels illustrate the three-channel input to the UNet, showing the spatial structure of the information available to the model for predicting large-scale B-mode foregrounds.

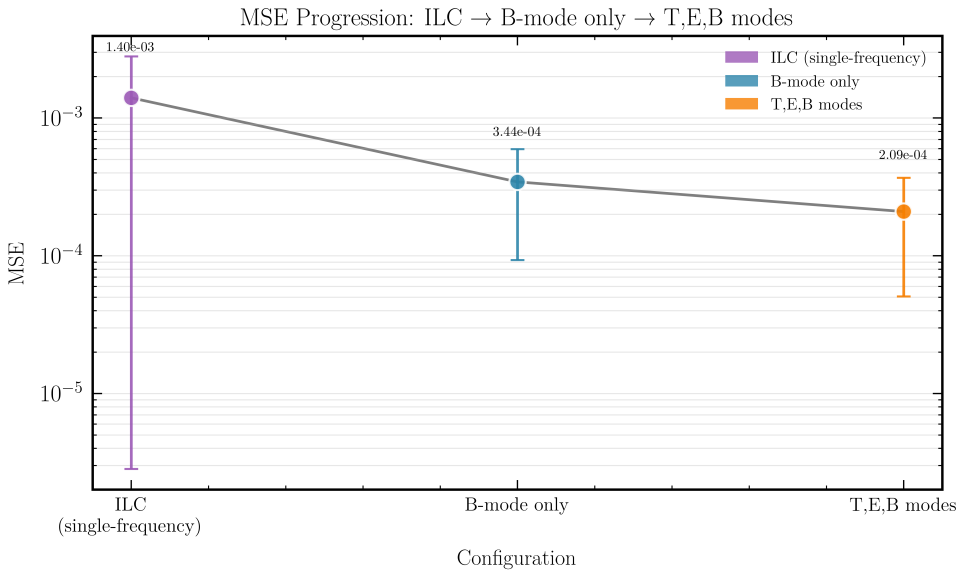

Figure 4: MSE progression showing systematic improvement from ILC baseline (single-frequency, no prediction) to B-mode only UNet to T,E,B multi-channel UNet. Error bars represent ±1 standard deviation across all test patches. Y-axis uses logarithmic scaling. Connecting lines emphasize the decreasing trend. Numerical MSE values are displayed above each point for precise quantification.

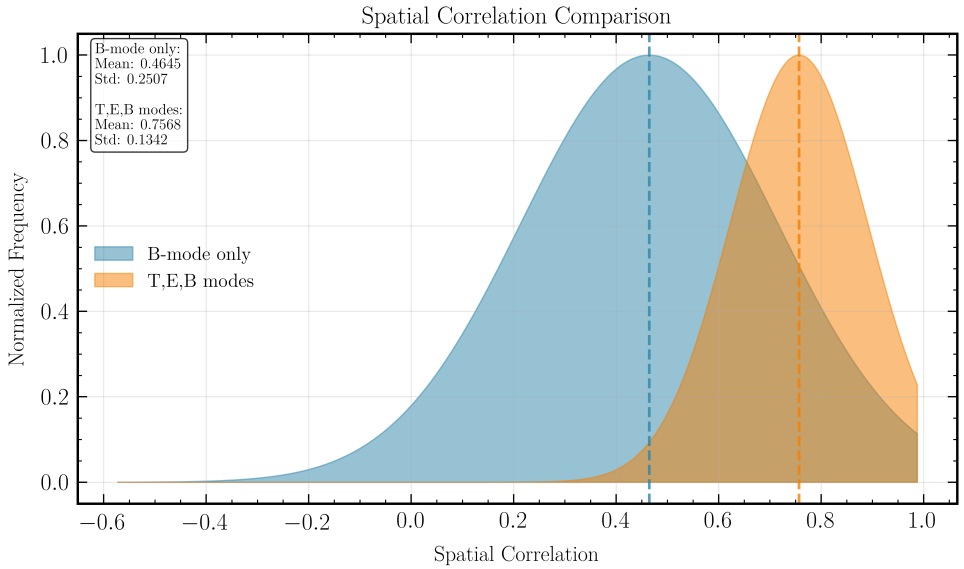

Figure 5: Spatial correlation distributions for B-mode only (teal) and T,E,B multi-channel (orange) UNet models. Distributions are approximated from mean and standard deviation statistics using Gaussian distributions, shown as semi-transparent filled areas. Vertical dashed lines mark the mean correlation for each model. Higher mean correlation with smaller standard deviation indicates better and more consistent performance.