Fast Bayesian QU-Fitting with Simulation-Based Inference

Preshanth Jagannathan & Arpan Pal · NRAO

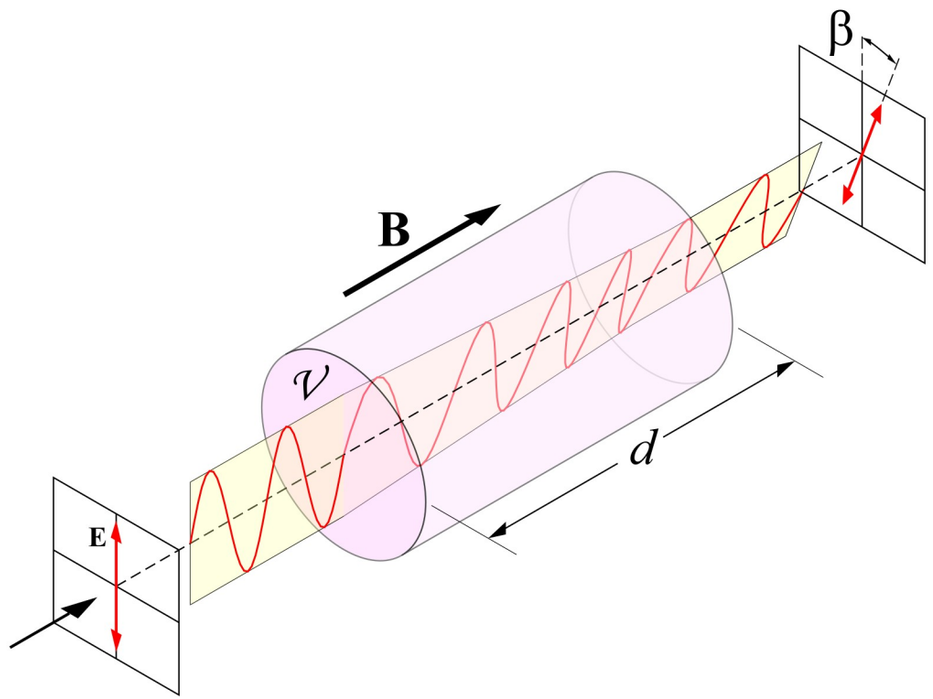
The magnetic universe

Magnetic fields thread galaxies, clusters, and the cosmic web, but they are hard to measure directly.

- Polarized synchrotron light carries the imprint of the magnetized plasma it passed through.
- Traveling through that plasma rotates its polarization angle (Faraday rotation), by an amount set by the line-of-sight field:

$$\phi = 0.81 \int n_e \, B_{\parallel} \, dl \quad [\mathrm{rad\,m^{-2}}]$$

- Measuring that rotation is one of the few direct handles we have on cosmic magnetism.

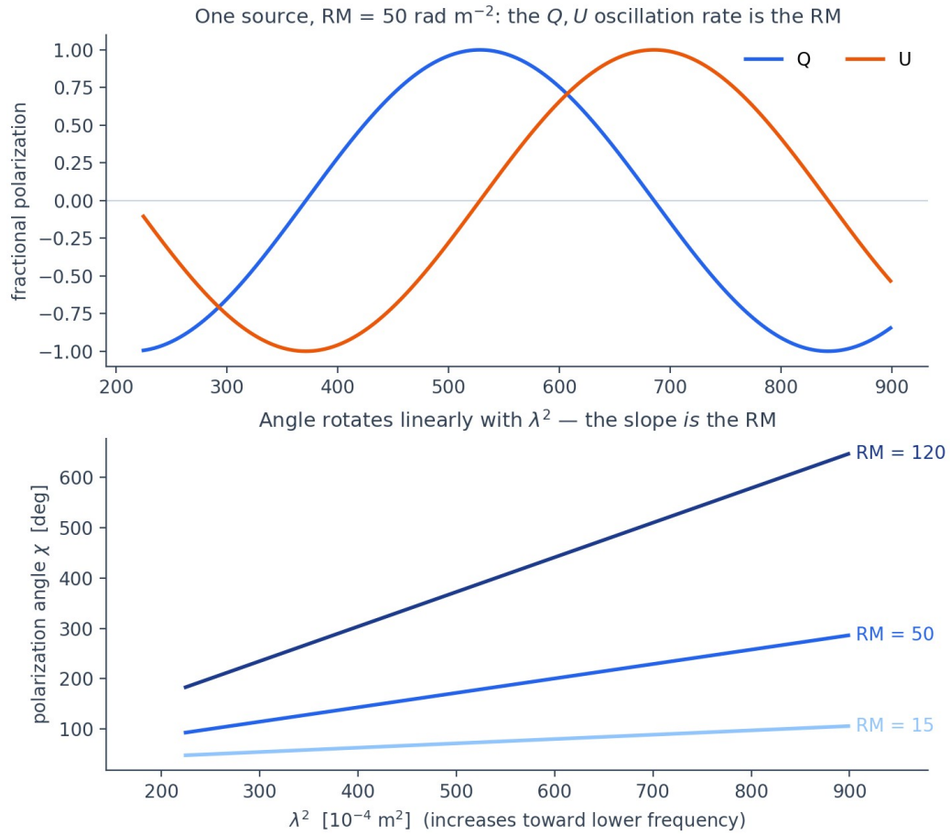


Linear polarization rotating as it traverses a magnetized medium (Wikimedia, Faraday effect)

What we actually observe

$$P(\lambda^2) = Q + iU = p\,I\,e^{2i(\chi_0 + \phi\lambda^2)}$$

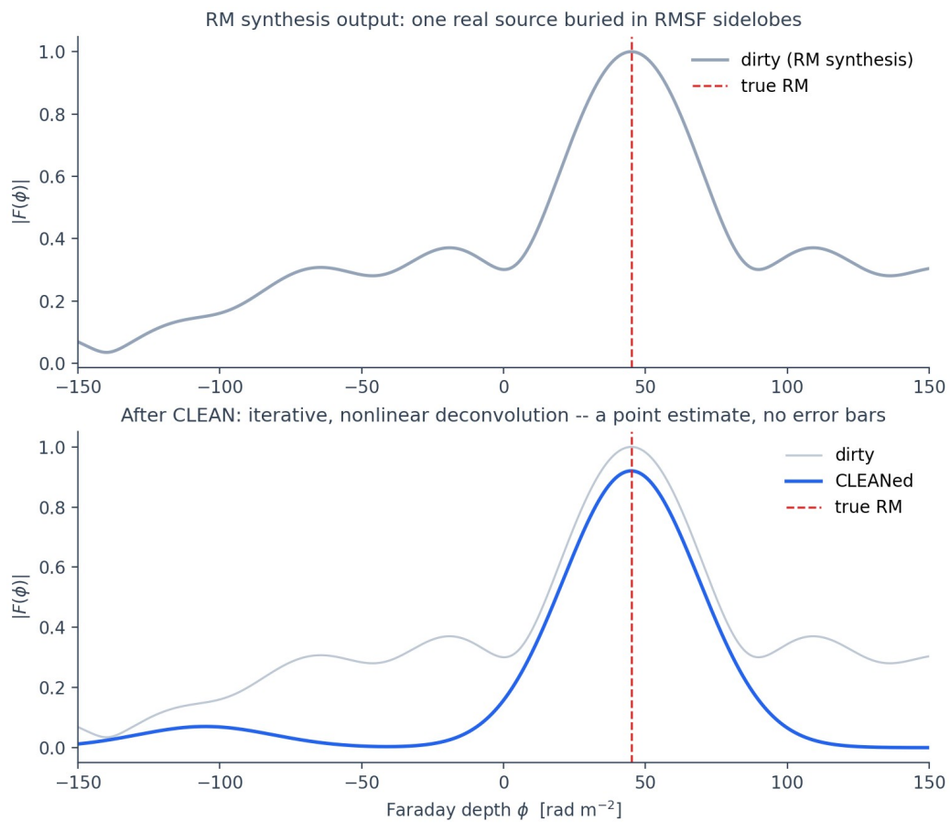
- We sample in **frequency**, not λ^2: the λ^2 sampling this gives us is inherently non-uniform.
- Q, U **oscillate** across the band at a rate set by ϕ: we can't just average channels together to boost SNR without collapsing that oscillation first.
- RM synthesis and QU-fitting are two different ways to do that collapse.



RM Synthesis

RM synthesis Fourier-transforms $P(\lambda^2)$ against trial Faraday depth ϕ, producing a “Faraday spectrum” $F(\phi)$ (Brentjens & de Bruyn 2005; Heald 2009).

- Linear transform (the FT), but nonlinear deconvolution-based recovery (CLEAN).
- No error estimation built in.
- The width of the RMSF main lobe and sidelobes sets how deep you can clean.
- Faraday complexity can also arise from larger uncertainty in the RMSF itself, not just from the source.
- No a priori model needed: this is a non-parametric system, against our Bayesian parametric approach.



QU-Fitting

The alternative is **QU-fitting**: fit a physical depolarization model directly to the observed $Q(\lambda^2), U(\lambda^2)$ spectra in a Bayesian framework, with full posterior uncertainties, and no CLEAN-style deconvolution. The community benchmark rates it more accurate than RM synthesis (Sun et al. 2015), but it costs 1 to 2 minutes per voxel of MCMC-style sampling.

- One SKA-precursor field, thresholded, still costs roughly 1000 core-hours.
- POSSUM, MeerKAT, LOFAR, and ngVLA will detect 10^6 to 10^7 polarized sources.

> QU-fitting's accuracy has been out of reach at survey scale.

The QU-fitting approach

- We have a priori information available during fitting: most of the sky sits within RM $\in [-500, 500]$ rad m^{-2}, and we know the expected distributions of angle and other parameters. This is what we encode in our **prior**.
- The **posterior** is the full uncertainty, with the physical model already in place: every parameter value consistent with the observed spectrum, weighted by consistency.
- To compute that posterior directly, you normally need to write down a likelihood function.
- SBI is a natural fit to this problem.

Simulation-Based Inference

- Simulation-based inference (SBI) trains a neural network to approximate the posterior directly from simulated (θ, x) pairs, without ever writing a likelihood function (Cranmer et al. 2020; Tejero-Cantero et al. 2020).
- Neural posterior estimation (NPE) is the specific SBI variant used here: the network outputs a full posterior density conditioned on a single observation (Papamakarios & Murray 2016; Greenberg et al. 2019).
- Training happens once, up front, on millions of simulated spectra. Inference on a new source is then a single forward pass, amortized to milliseconds.

The network *is* the posterior

- **Simulate, train, infer**: draw θ from the prior, compute the spectrum plus noise, train the network to assign high probability to the θ that actually produced it, then infer the full posterior for a real spectrum in one pass.
- The output is a **normalizing flow**: a fixed base distribution warped by an invertible transform into the posterior's actual shape, differently for every input spectrum (Papamakarios et al. 2019). Invertible means the density is exact: you can both sample from it and evaluate its probability.
- Before the flow sees anything, an **embedding MLP** compresses the ~3000-number spectrum (Q, U, weights, times 1024 channels) into 64 learned summary statistics, the same role as hand-designed statistics (slope, width) astronomers already use, but learned end to end.
- The embedding learns the messy parts too: RFI flags, varying noise, missing sub-bands, since training simulations include all of it.
- Every batch is freshly simulated, so the network never sees the same example twice: overfitting is structurally impossible.

The four physical models

Each is one oscillation, times a depolarization envelope:

Model	$P(\lambda^2) =$	What the parameters mean
Faraday thin	$p_0\, e^{2i(\chi_0+\phi\lambda^2)}$	ϕ: RM (screen); χ_0: intrinsic angle; p_0: polarized fraction
Burn slab	$p_0\, \dfrac{\sin(\Delta\phi\,\lambda^2)}{\Delta\phi\,\lambda^2}\, e^{2i(\chi_0+\phi_c\lambda^2)}$	$\Delta\phi$: Faraday depth through the emitting slab; sinc can change sign
External dispersion	$p_0\, e^{-2\sigma_\phi^2\lambda^4}\, e^{2i(\chi_0+\phi\lambda^2)}$	σ_ϕ: turbulent scatter in a foreground screen, smooth decay
Internal dispersion	$p_0\, \dfrac{1-e^{-S}}{S},\ S = 2\sigma_\phi^2\lambda^4 - 2i\phi\lambda^2$	turbulence mixed into the emitter, decay and rotation entangled

(Up to 5 components summed gives 20 posterior networks, a classifier to pick the family, and a spectral-shape model for the Stokes-I normalization: backup slide.)

What does the network learn

The same physics that helps observers helps, or hurts, the network:

Easy for the network

- Depolarization envelopes (λ^4 decay): a smooth, global amplitude shape, so σ_ϕ is read off the decay rate.
- RM at moderate values: the Q, U oscillation rate *is* ϕ, a strong, distributed signal across the whole band.

Hard for the network

- Pure phase models (Faraday thin): all the information is in a fine oscillation pattern, with nothing smooth to latch onto.
- χ_0, a periodic angle (mod π), on a flow that lives on the real line: the wrap-around is where our under-confidence came from.
- Burn slab's sinc sign flips: a genuine $\chi_0 \pm \pi/2$ degeneracy, so the true posterior is bimodal.

> The calibration matrix mirrors this table closely: the envelope models pass, the phase-dominated models needed the χ_0 fix.

Can we trust a neural posterior?

A flow will always return a confident-looking posterior; it cannot shrug. Are its stated uncertainties honest? Take every source where we report RM $= 30 \pm 5$ (a 1σ interval):

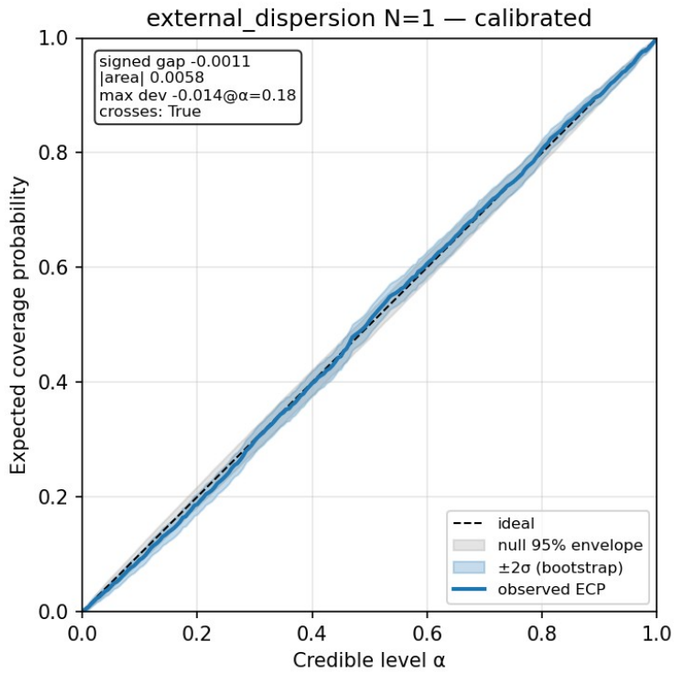
If the network is…	The truth lands inside…	Consequence
Calibrated	1σ of the time	error bars mean what they say
Over-confident	less than 1σ	bars too narrow, false precision, wrong science
Under-confident	more than 1σ	bars too wide, honest but wasteful

Over-confidence is the dangerous failure: it silently corrupts downstream science. Under-confidence only costs efficiency. Our networks err on the under-confident side, and we can show it: coverage testing establishes that where we deviate from calibrated, the deviation is mixed or under-confident, error bars at worst larger than needed, never falsely precise.

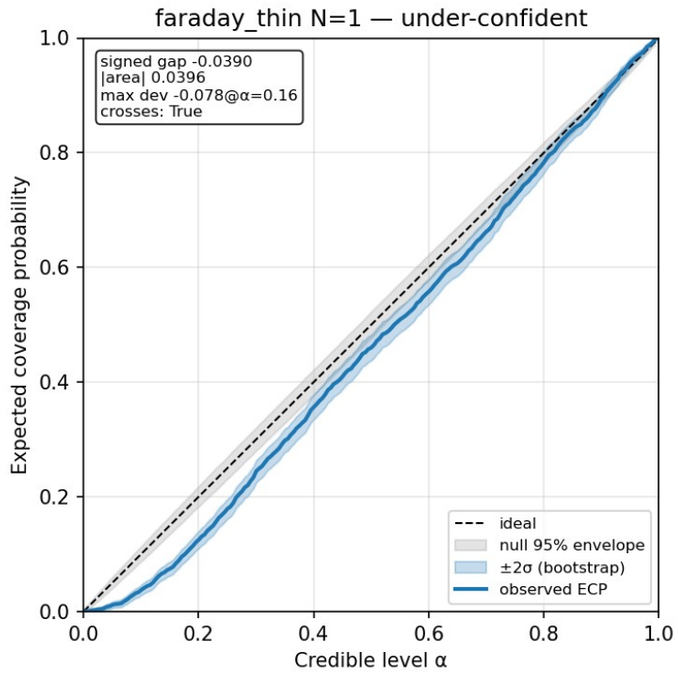
TARP: coverage testing at scale

- On thousands of simulations where we know the truth: of all the times we claimed an α% credible region, did the truth land inside it α% of the time, for every α at once?
- This is the TARP diagnostic (Lemos et al. 2023), a joint-coverage complement to simulation-based calibration (Talts et al. 2018).
- Calibrated means the curve lies on the diagonal; above is under-confident, below is over-confident.

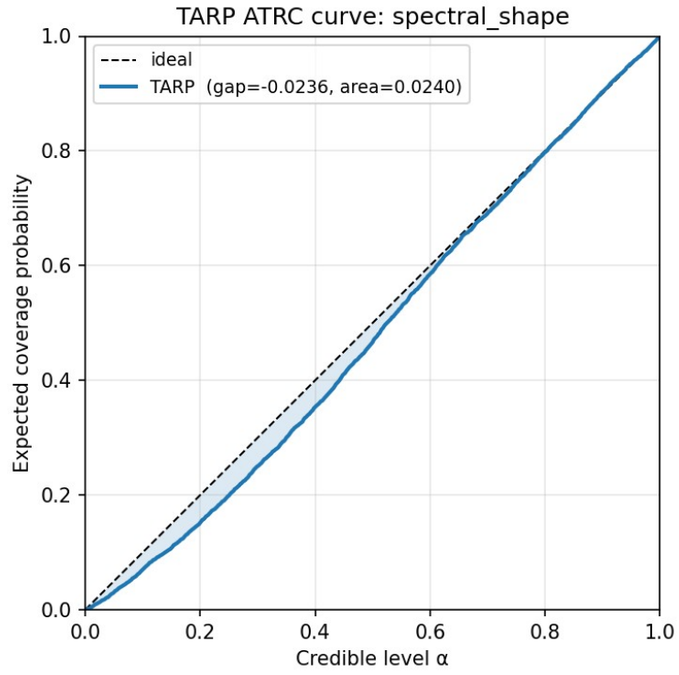
Calibrated, external dispersion.



Under-confident, Faraday thin.

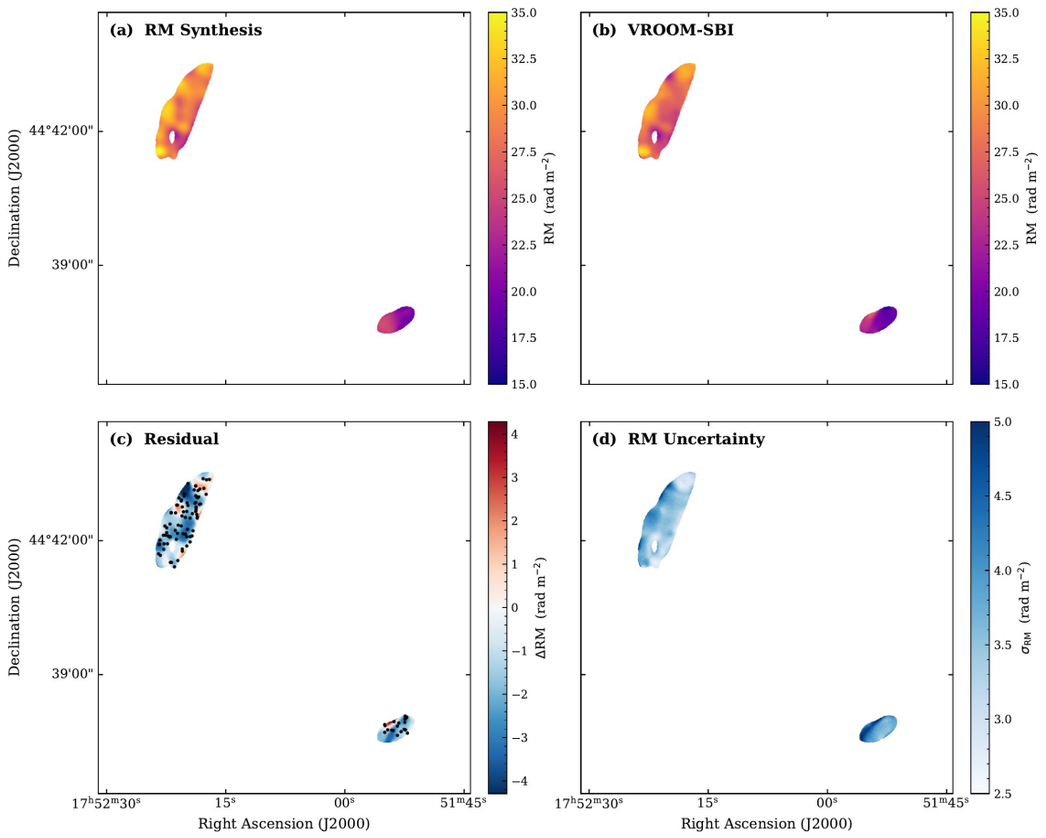


Under-confident, spectral model; its mean still normalizes Stokes-I reliably.

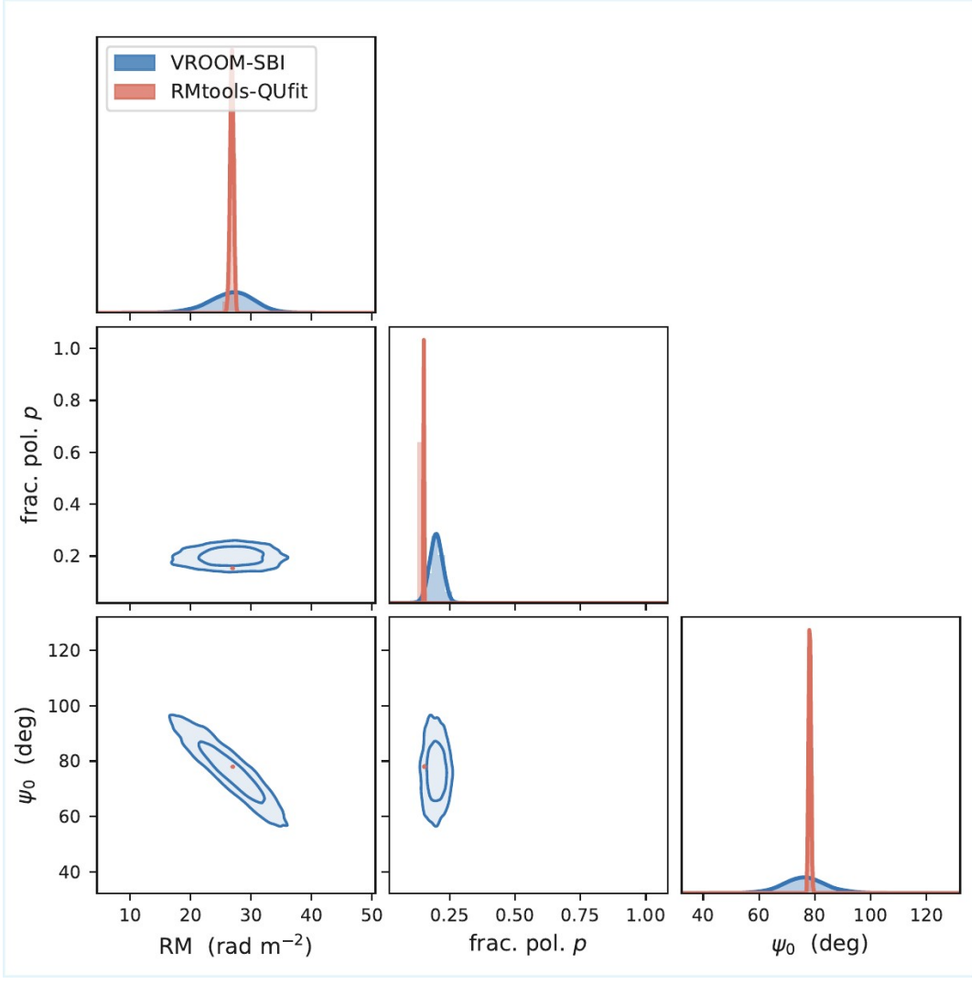


> TARP identifies which side of calibrated each model sits on, and for ours the answer is the safe side.

VROOM-SBI vs. other methods

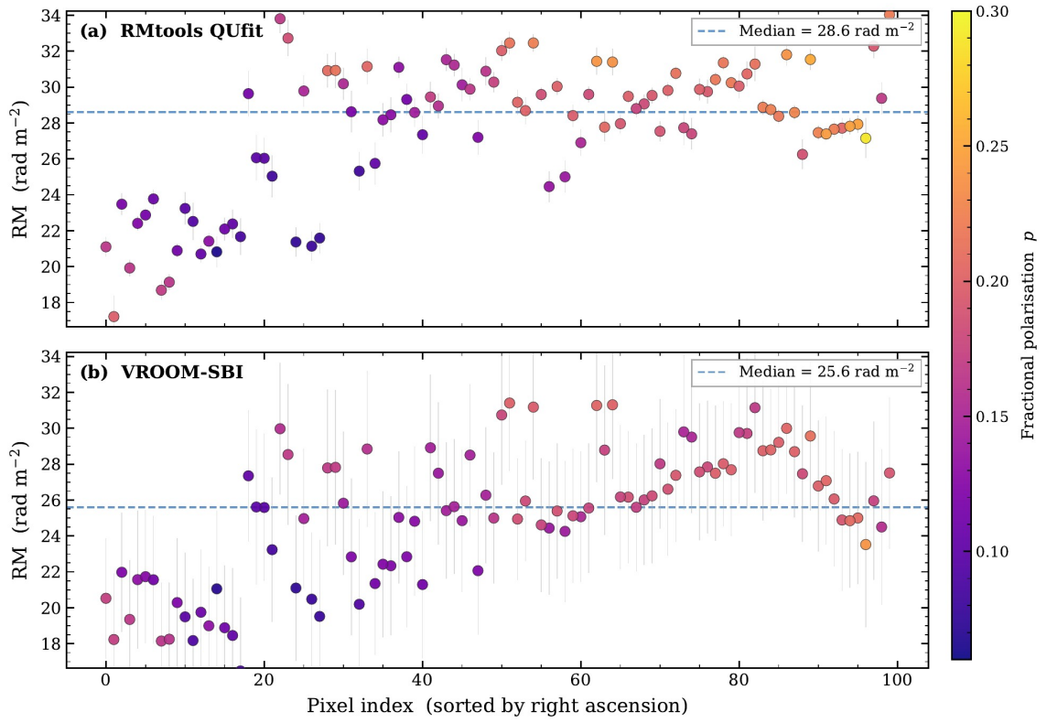


RM synthesis vs. VROOM-SBI, residual and uncertainty

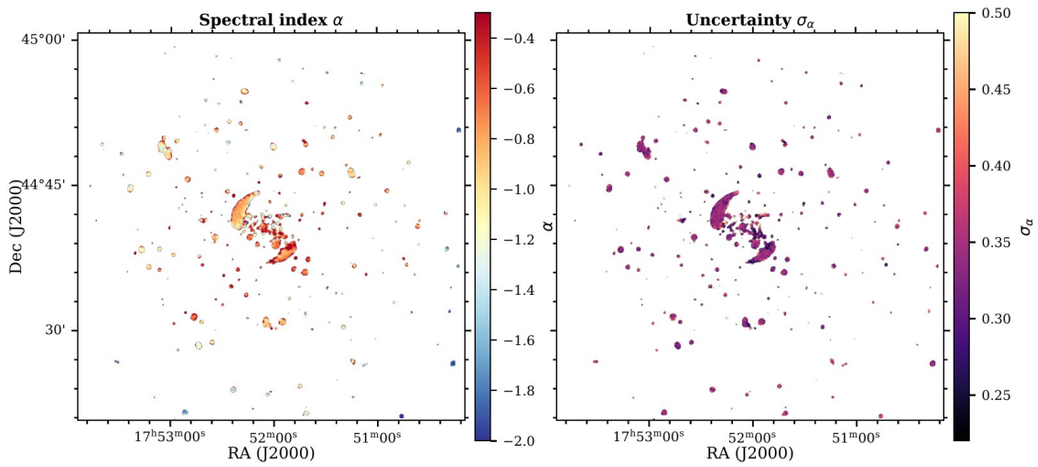


Per-pixel posterior: VROOM-SBI vs. RMtools-QUfit

More real-data results



100-pixel RM comparison: RMtools-QUfit vs. VROOM-SBI



Spectral-index map of the same relic, from the VROOM-SBI spectral-shape posterior

The payoff: about 500x, and honest

Full posterior inference over 35,700 polarized voxels:

Method	Per voxel	Full field
VROOM (RM synth, GPU)	~25 μs	~min (no posteriors)
VROOM-SBI (GPU)	**~0.2 s**	**~125 min**
QU-fitting (dynesty)	1.5 to 2 min	~900 to 1200 CPU-hr

> A million-source catalog with full Bayesian posteriors in about a day on one GPU.

- QU-fitting's accuracy at RM-synthesis-like speed, via SBI. Posteriors, not point estimates.
- Calibration validated and reported (TARP/SBC): error bars err wide, never falsely precise.
- Next: out-of-distribution flagging, RMSF-aware 3D Faraday synthesis.

Everything is open:

- Models: huggingface.co/skunkworks-ra/vroomsbi
- Code: github.com/skunkworks-ra/vroom-sbi

Training at scale

Every posterior trains on simulations generated on the fly, directly on GPU, never touching disk: prior draw, forward model, RFI flagging and weight augmentation, noise, all on-device. About 33 million simulations per epoch, so the network never sees the same example twice.

We scaled and validated this pipeline across GPU generations before committing to production runs: RTX 3080 for fast local iteration, L40 for mid-scale validation sweeps, and A100 for the full 20-posterior production matrix.

Computations were supported by:

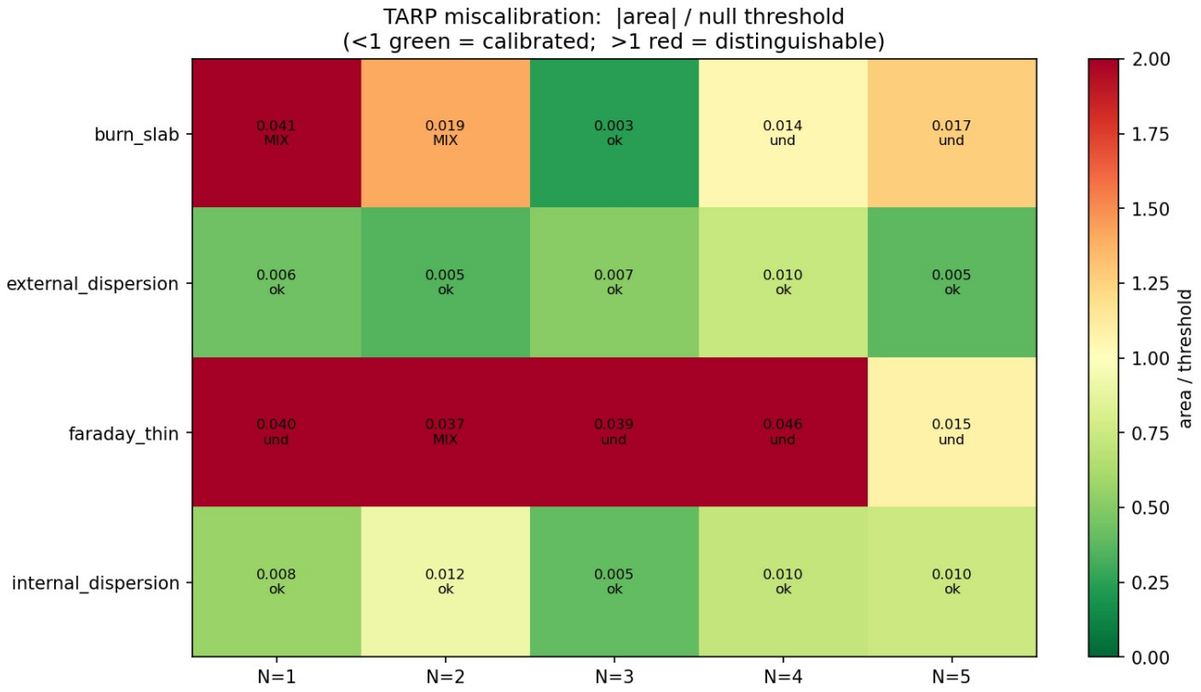
- **NAIRR Pilot**, allocation **NAIRR250508**.
- **ACCESS** program, allocation **PHY260060**, providing close to **200,000 service units**.

Each posterior trains in roughly 8 hours on a single A100.

Thank you

The full calibration matrix

Every model times component count, tested the same way.



> Calibration is a first-class deliverable here, not an afterthought: it shows where the posteriors are tight versus deliberately broad, rather than hiding it.

Anatomy of the network

Two networks in series, trained end to end, plus a referee:

$$\underbrace{[Q, U, w]}_{3\times 1024=3072 \text{ numbers}} \xrightarrow{\text{embedding MLP}} \underbrace{s(x)}_{64 \text{ summaries}} \xrightarrow{\text{conditions}} \underbrace{q_w(\theta \mid s(x))}_{\text{Neural Spline Flow}}$$

- Embedding MLP ($3072 \to 256 \to 128 \to 64$, ~0.8M weights): the learned summary statistics.
- Neural Spline Flow (15 spline transforms, ~5M weights) (Durkan et al. 2019): this is the NPE, the learned probability distribution over θ, reshaped per spectrum by the 64 summaries.
- One such pair per (model family times component count), giving 20 posterior networks.
- A separate CNN classifier arbitrates which model family a spectrum belongs to (model selection).

About 5.9M parameters per posterior, small by ML standards. It trains in roughly 8 hours on one GPU, generating simulations on the fly, never touching disk.

How we train

Simulation-on-the-fly, entirely on the GPU: no training set ever touches disk.

- Every step draws a fresh batch of 65,536 simulated spectra: prior draw, forward model, RFI flagging and weight augmentation, noise, all on-device.
- About 33 million simulations per epoch; the network never sees the same example twice, so overfitting is structurally impossible.
- Adam (lr 2×10^{-4}), plateau-triggered LR decay, early stopping on a fixed 50k validation set, TF32 matmuls.
- Each (model times component) posterior trains in roughly 8 hours on a single A100. Development and scaling tests ran across RTX 3080, L40, and A100 hardware, on time from the NAIRR Pilot (NAIRR250508) and ACCESS (PHY260060) allocations.

Augmentations: per-channel noise scaling 2 to 300x, scattered/gap/block flagging; this is why it survives real RFI.

Stokes I: the spectral-shape model

The depolarization models describe the fractional polarization, but we observe Q, U in flux units. Stokes I sets the normalization, and it has its own spectrum:

$$F(\nu) = \exp\left(\alpha x + \beta x^2 + \gamma x^3\right), \qquad x = \log(\nu/\nu_0), \quad F(\nu_0) \equiv 1$$

- A log-log polynomial: α is the spectral index, β, γ capture curvature, enough for realistic synchrotron SEDs across the band.
- Modeled separately, with its own NPE: same machinery, simulator, flow, TARP-tested.
- Inference composes: the spectral-shape posterior normalizes $I(\nu)$, the depolarization posterior handles $q = Q/I$, $u = U/I$.

> I is explicitly modeled, just factorized so each network solves one clean problem.

Why under-confident? Diagnosing the flow

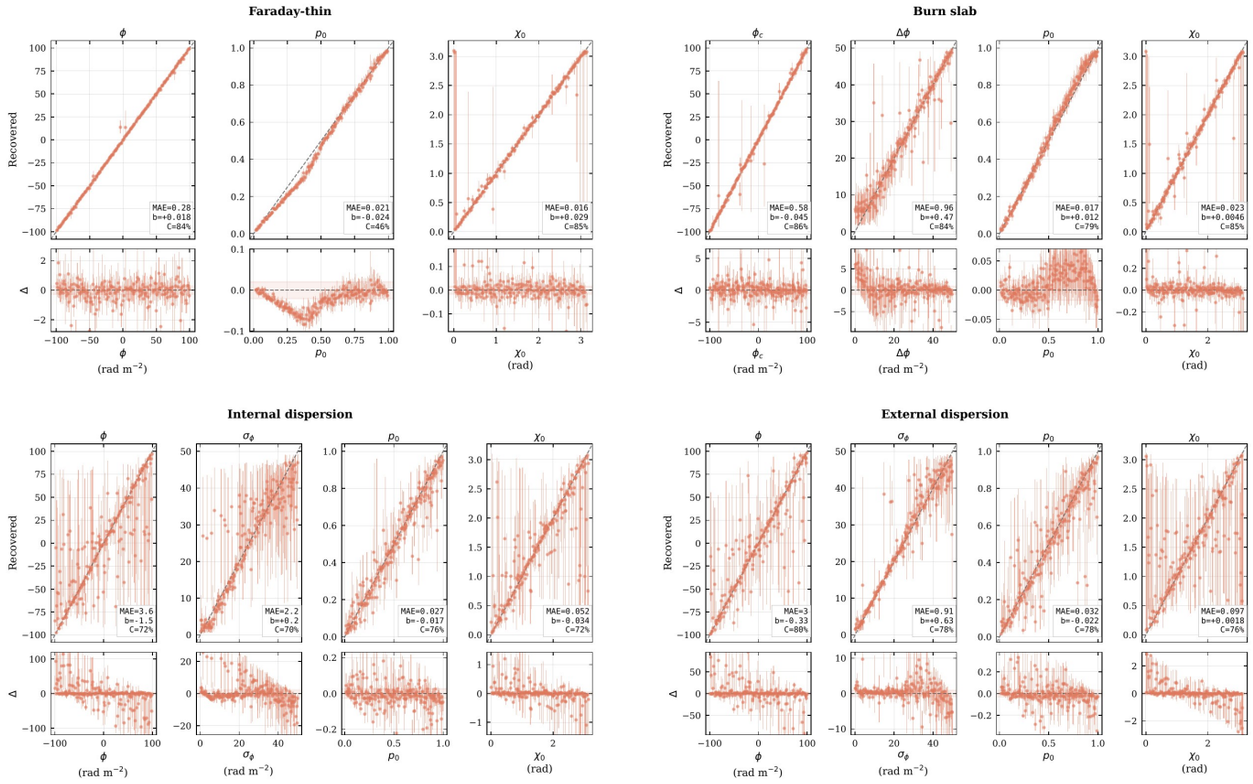
The calibration matrix didn't just grade the models, it localized the bug:

- The two model families that fail TARP are exactly the ones whose identifiability leans on the phase structure, that is, on the polarization angle χ_0.
- χ_0 is periodic (defined mod π), but the flow trains it on an unbounded line with a hard $[0, \pi]$ box.
- Truths near the wrap ($\chi_0 \approx 0$ or π) force the flow to put near-zero density on one side, producing rare, catastrophic validation losses.
- Early stopping keys on the mean loss, so it selects deliberately blurry checkpoints. Under-confidence, mechanically explained.

The fix is classical: train the flow on $(\sin 2\chi_0, \cos 2\chi_0)$, continuous across the wrap, and map samples back with $\frac{1}{2}\arctan 2$. Full-matrix retrain in progress.

SBI vs. truth

On held-out simulations with known ground truth: posterior median vs. injected value, all four models.



- Rotation measure and polarization angle recovered tightly across all models.
- Error bars are the posterior 16 to 84%, and they cover the truth.