HIDDEN MISALIGNMENT IN LOW-FIELD MRI BENCHMARK REFERENCES: THE LEADERBOARD DEPENDS ON HOW THE REFERENCE IS BUILT

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep-learning enhancement is the most promising route to clinical-quality low-field MRI. Its progress is measured on M4Raw, the only public low-field (0.3 T) raw k-space benchmark. There, every method is scored against the average of repeated acquisitions. We show that this reference is systematically misaligned, and that the leaderboard built on it depends on how the reference is constructed. An audit of all 240 subject-contrast groups finds sub-pixel displacements between repetitions. Registering them lowers the reference's own noise level by 6–14% (validation) and 15–25% (test). The displacement lies along the readout axis, has the same sign in every validation group and 99% of test groups, and grows linearly with acquisition order (about 2.3 ppm per repetition). This is the signature of an uncompensated frequency drift of the permanent magnet, with head motion on top in the labelled motion subset. We build a corrected reference by exact k-space phase-ramp registration and phase-aligned complex averaging, with a per-volume error budget. It is validated by synthetic-truth recovery and leave-one-repetition-out prediction. We also measure the reference's own resolution. Two disjoint halves of the published reference agree only to 28.8–31.7 dB, with a per-subject spread of 1.0–1.7 dB. Registration raises the agreement by 1.1–1.8 dB and shrinks the spread two- to three-fold. We re-score the publicly released baselines under a decision rule frozen before unblinding. For every network-vs-BM3D pair the reference$\times$method interaction is Holm-significant and same-signed under each of the three reference constructions we compare. Registering the reference under the published protocol, and changing nothing else, removes about half of NAFNet's published lead over BM3D ($+0.94/+0.87/+0.37 \rightarrow +0.50/+0.49/+0.20$ dB). Whether the ordering then inverts depends on the construction (five, four or one pair) and on the metric and region. The published reference cannot order these methods. We release the corrected reference (CC BY), the error budgets and the code.

1 INTRODUCTION

Low-field MRI (below 0.5 T) is where learned image enhancement has the most room to help and the least clean supervision to learn from. Signal-to-noise ratio scales with field strength (Marques et al., 2019; Wald et al., 2020), so a 0.3 T scan of a few minutes is dominated by noise. There is no high-field copy of the same patient to serve as a label. Progress is therefore measured on one benchmark. M4Raw (Lyu et al., 2023a) is the only public low-field (0.3 T, 4-channel, 2-D multi-slice) raw multi-coil k-space dataset with repeated acquisitions of every slice. It provides T1w, T2w and FLAIR with 3/3/2 repetitions per contrast in the train and validation splits, and 6/6/4 in the V1.6 test split (Lyu et al., 2023b). Every published enhancement result on it is scored against an unregistered average of those repetitions; the official protocol averages per-slice min-max normalised magnitudes. No published training or evaluation pipeline registers the repetitions before averaging them.

The gaps being adjudicated on this reference are small. Adjacent entries of the maintainers' own MIDL 2024 leaderboard differ by 0.05–0.06 dB (Li et al., 2024). The self-supervised state of the art leads the supervised baselines it retrained by 0.1–0.5 dB (Tu et al., 2025). This reference is the one variable nobody has corrected or budgeted. Its defect is hidden not from the dataset pa-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

per, which measured it, but from every label, score and leaderboard built on it. The dataset paper measured a systematic inter-repetition translation along the frequency-encoding axis (-0.39 ± 0.26, -0.31 ± 0.18 and -0.31 ± 0.17 px for T1w/T2w/FLAIR, Table 2 of Lyu et al. (2023a)). It attributed the translation to permanent-magnet field drift. It states that the motion was "only estimated without performing actual correction" (Lyu et al., 2023a). The maintainers' 2026 complex-averaging demo (M4Raw maintainers, 2026) aligns phase but not position. Their 2026 data card (Lyu et al., 2026) disclaims a noise-free ground truth. The MIDL short paper motivates the 6/6/4-repetition test reference purely by SNR (Li et al., 2024). Nobody has registered the repetitions or budgeted the resulting reference error. Nobody has tested whether the published rankings survive the correction.

Prior work showed that hidden properties of a public reference bias learned-method rankings (Shimron et al., 2022; Wang et al., 2024; 2026). This paper is their misalignment analogue for the low-field benchmark, analysed as Metrics Reloaded and challengeR recommend (Maier-Hein et al., 2024; Wiesenfarth et al., 2021; Maier-Hein et al., 2018). Our contributions are:

1. **An audit** of every subject×contrast group of the validation split, the labelled motion subset and the V1.6 test split (90 + 75 + 75 groups). Sub-pixel inter-repetition displacement inflates the reference's own noise level σ_1. After registration it falls by 6–14 % (validation) and 15–25 % (test), with null controls at 0.00 % (Sec. 2).
2. **A mechanism.** The displacement lies along the readout axis and has the same sign in every validation group. It grows linearly with acquisition order (2.33 ppm of B_0 per repetition) and follows one monotone drift across the whole session ($R^2 = 0.989$). This is the signature of an uncompensated frequency drift of the permanent magnet. Head motion is added on top in the labelled motion subset (Sec. 2).
3. **A corrected reference** built by exact k-space phase-ramp registration and phase-aligned complex averaging, with a per-volume error budget. It is validated by synthetic-truth recovery and leave-one-repetition-out prediction (Sec. 3).
4. **The reference's own resolution.** Two disjoint halves of the published reference agree with each other only to 28.8–31.7 dB in the official convention. The draw-to-draw spread is 1.0–1.7 dB per subject. Registration raises the agreement by 1.1–1.8 dB and shrinks the spread to 0.4–0.7 dB, a two- to three-fold reduction (Sec. 4).
5. **A pre-registered re-scoring** of the three publicly released baselines. It shows that the published leaderboard is not robust to how its reference is built. For every network-vs-BM3D pair the reference×method interaction is Holm-significant and has the same sign under each of the three reference constructions of Sec. 5. Registering the reference under the published protocol, and changing nothing else, removes about half of NAFNet's published lead over BM3D ($+0.94/+0.87/+0.37 \rightarrow +0.50/+0.49/+0.20$ dB in T1w/T2w/FLAIR). Whether the order then inverts depends on the construction: the frozen comparison (published label vs the registered, averaged, then normalised reference) inverts five pairs, the same registration under the average-then-normalise convention four, and registration alone under the published protocol one. Gaps of the size published are decided by reference-construction choices, not by the methods (Sec. 5, Table 1, and Table 4 in App. G.1).
6. **A release**: the corrected reference (CC BY 4.0, a derivative of M4Raw), the per-volume error budgets and displacement tables, and the audit and re-scoring code (MIT) (Sec. 6).

2 THE M4RAW REFERENCE AND ITS AUDIT

Data. Each M4Raw volume is a $(18, 4, 256, 256)$ complex k-space array with axes (slice, coil, frequency encoding, phase encoding). 195 (T1w/T2w) or 198 (FLAIR) phase-encode lines are acquired and symmetrically zero-filled to 256. The scanner is a 0.3 T permanent-magnet system with a 4-channel head coil, 240 mm field of view and 5 mm slices. The readout bandwidth is 122 Hz/px, so 1 ppm of B_0 (12.8 Hz at 12.77 MHz) corresponds to 0.105 px along readout. For the spin-echo splits, rows of the stored image are the readout axis. The file headers label them the other way round, a header error confirmed by the maintainers' 2026 data card. One caveat governs every ppm figure: 122 Hz/px reads the header's 31.25 kHz as the total bandwidth over the 256 samples (no readout oversampling, per the maintainers); if it is the ± half-bandwidth instead, all ppm values double and the pixel shifts are unaffected. Every ppm figure uses the least-squares test drift, -0.244 px per

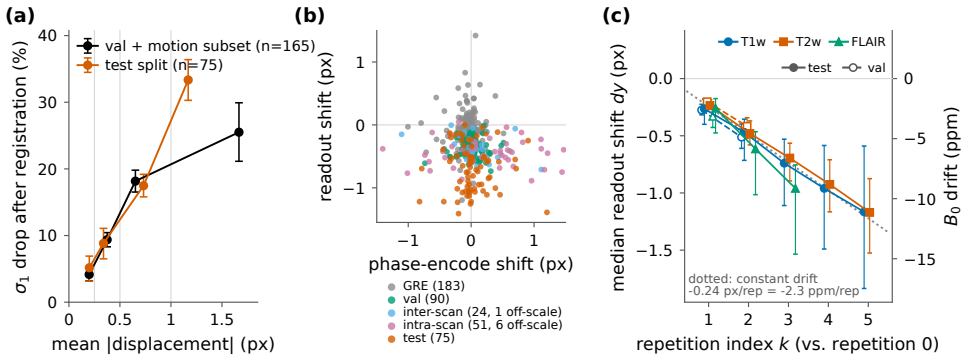

Figure 1: **Hidden misalignment in the M4Raw reference is B_0 drift plus head motion. (a)** Dose-response. Relative drop of σ_1 after per-slice rigid registration of the repetitions vs the mean absolute inter-repetition displacement of the groups in each bin (< 0.25, 0.25–0.5, 0.5–1, ≥ 1 px). Black: val pooled with the labelled motion subset ($n = 165$ groups). Vermilion: V1.6 test ($n = 75$). Error bars: 95 % percentile bootstrap CI of the bin mean (5000 resamples, seed 0). **(b)** Signed median (over slices) shift of the repetitions relative to repetition 0, phase-encode vs readout component, one point per group. The components are dx and dy for the spin-echo splits and are swapped for GRE, whose readout is the stored column axis. Groups: val ($n = 90$), inter-scan (24) and intra-scan (51) motion cohorts, V1.6 test (75) and the V1.5 GRE arm (183). $dy < 0$ in 100 % of val and 99 % of test groups. The GRE readout shift is -0.065 px in median, negative in 72 % of pairs and larger than the phase-encode shift in 66 %. **(c)** Median readout shift of repetition k relative to repetition 0 per contrast (T1w circles, T2w squares, FLAIR triangles), test (filled, solid) and val (open, dashed); error bars: 95 % bootstrap CI of the median over subjects. Dotted: constant drift $dy = -0.244\,\text{px} \times k$ (-2.33 ppm per repetition) fitted to the test medians. Right axis: the same shift as B_0 drift in ppm (1 ppm = 0.105 px at 122 Hz/px).

repetition $= -2.33$ ppm. The audit covers the validation split (30 subjects $\times$ 3 contrasts $= 90$ groups; 3/3/2 repetitions), the labelled motion subset (8 inter-scan and 17 intra-scan participants; 24 and 51 groups) and the V1.6 test split (25 subjects, 6/6/4 repetitions, 75 groups). The V1.5 GRE data (183 subjects $\times$ 2 repetitions) are a replication arm. Their readout lies along the stored column axis, orthogonal to that of the spin-echo splits, so GRE shifts are reported by readout and phase-encode component.

Endpoint and controls. The audit needs no ground truth. For a group of n repetitions $x_1, \ldots, x_n$ (root-sum-of-squares magnitudes in the units of the released images) the noise level is the pairwise-difference estimate $\sigma_1 = \text{median}_{i<j}\ \text{std}_{\text{mask}}(x_i - x_j)/\sqrt{2}$. It is unbiased if the repetitions differ only by noise, and any displacement between them inflates it. The registration sensitivity is the relative drop of σ_1 after per-slice rigid (translation) registration of the repetitions. The translation is estimated by upsampled phase correlation (Guizar-Sicairos et al., 2008, upsample factor 100, unnormalised cross-power spectrum) and applied as an exact k-space phase ramp, so no interpolation kernel enters. Four controls bound the estimator. C1 passes the identity through the operator; C1b measures its round-trip cost at the measured displacement (≈ 0.005 relative RMS). C2 and C2b are null runs on repetitions that differ only by synthetic noise: displacement floor ≈ 0.01 px, σ_1 drop 0.00 %. Two independent implementations of the audit agree bit-for-bit on all 165 groups. Three estimators, phase correlation with an unnormalised and with a normalised cross-power spectrum and a normalised cross-correlation rigid search, differ by 0.016 px in the median on shifts of 0.31–0.39 px. They agree on the sign of the readout shift in every case, and all three reproduce the σ_1 drop.

Results (Fig. 1). Registering the repetitions lowers σ_1 by 13.8/6.3/10.6 % (T1w/T2w/FLAIR) on the validation split, where the inter-repetition displacement is 0.36–0.45 px. The drop is 23.8/10.8/17.1 % on the labelled motion subset (0.6–1.0 px). It is 25.4/15.5/18.9 % on the V1.6 test split, where the drift accumulates over 6/6/4 repetitions. It is 9.0 % on the GRE arm at 0.24 px.

The effect is dose-dependent (Fig. 1a). Binned by mean absolute displacement (< 0.25, 0.25–0.5, 0.5–1, ≥ 1 px), the drop is 4.2/9.4/18.2/25.5 % over the 165 validation-plus-motion groups (26/64/56/19 per bin). The test split (Fig. 1a) and the GRE arm (App. H) show the same shape.

The displacement has a definite geometry (Fig. 1b). It lies along the readout axis (dy) with $dx \approx 0$. $dy < 0$ holds in 100 % of validation groups and 99 % of test groups (medians -0.34 and -0.67 px). The across-slice interquartile range is 0.04–0.09 px, so the whole volume moves as one. The median in-plane rotation per contrast is 0.04–0.11° (validation) to 0.07–0.20° (test), reported but not corrected. The two-repetition GRE arm reads out along the orthogonal stored axis. Its readout shift is -0.065 px in median, negative in 131 of 181 pairs ($z = 6.0$) and larger than the phase-encode shift in 66 % of subjects. This is about half the amplitude the measured drift rate predicts for its 53 s scan. With $n = 2$ no growth can be tested. The shift grows with acquisition order (Fig. 1c). $|dy(\text{rep}\,2)| > |dy(\text{rep}\,1)|$ in 96 % of the 110 validation-plus-motion groups, with a median ratio of 1.99 where a constant drift rate predicts 2.0. On the test split the T1w medians for repetitions 1–5 relative to repetition 0 are $-0.27/-0.47/-0.74/-0.96/-1.17$ px (T2w $-0.24/-0.48/-0.70/-0.93/-1.17$; FLAIR $-0.26/-0.62/-0.96$ for repetitions 1–3). This is a least-squares drift of -0.244 px per repetition. The drift is a property of the session, not of the sequence. We registered all eight volumes of each validation subject across contrasts on gradient-magnitude images (App. H) and fitted one monotone drift to every candidate acquisition order. The order "file name within contrast, blocks T1w → FLAIR → T2w" wins in 30/30 subjects (median R^2 0.989) at -0.28 px (≈ 2.7 ppm) per volume position. The header series counter fits worse (0.775). The shift does not reset at sequence boundaries, so no per-sequence re-calibration is visible. The maintainers confirmed (M. Lyu, personal communication, 4 September 2026) that the file suffix is the creation-time order, that the series counter is an export artefact and that the centre frequency was calibrated manually once per session.

Head motion sits on top of the drift rather than replacing it. The inter-scan cohort adds a phase-encode component of $dx = +0.03$ to $+0.09$ px. The intra-scan cohort reaches dx up to $+0.5$ px with 2–5× the across-slice spread. All seven groups exceeding 1.5 px belong to the labelled cohorts. The picture is a disclosed but uncorrected misalignment: an uncompensated B_0 frequency drift of the permanent magnet, already recorded in the dataset paper's Table 2 ($-0.39/-0.31/-0.31$ px). NdFeB remanence falls by about 0.1 % per kelvin (Cooley et al., 2021; O'Reilly et al., 2021), so 2.3 ppm per repetition is a few millikelvin. Head motion in a minority of participants adds to it. We describe the drift by its signature: readout-only, sign-consistent, monotone in acquisition time, continuous across sequence boundaries. No direct field measurement is available. The GRE arm reproduces the σ_1 inflation and a same-signed, readout-dominant shift at about half the predicted amplitude; its two repetitions cannot show growth.

3 A CORRECTED REFERENCE WITH AN ERROR BUDGET

Construction. The corrected reference registers the repetitions, aligns their phase and averages them in the complex domain, and it declares what it cannot correct. Registration is a per-slice translation estimated by upsampled phase correlation (Guizar-Sicairos et al., 2008, upsample 100, unnormalised). It is refined by a joint least-squares fit over all $n(n-1)/2$ repetition pairs with repetition 0 fixed. The pair residual is a consistency diagnostic. A split-coil estimate (coils $\{0,1\}$ vs $\{2,3\}$) gives an empirical residual bound. The shifts are applied as exact per-coil k-space phase ramps. No interpolation kernel is chosen and the 195/198-line acquired support is preserved. A phase ramp on a zero-filled column is still zero. In-plane rotation is estimated by log-polar phase correlation (three-shear Fourier rotation (Unser et al., 1995; Larkin et al., 1997), 0.05° precision) and reported, not applied. Its per-contrast median is 0.04–0.11° on validation and 0.07–0.20° on test, which at the cortex (90 px) is 0.1–0.35 px, larger than the translation residual below. Phase alignment removes, per slice and coil, a global phase $\angle \sum_{\text{mask}} X_i \overline{X_0}$, in the spirit of real-valued averaging (Eichner et al., 2015). It then removes a low-pass ($\sigma = 8$ px) residual phase-difference map. The coherence map $|\sum_i X_i| / \sum_i |X_i|$ is compared with its noise-null model as a quality check. The coil-wise complex mean is then combined by root-sum-of-squares.

Every variant is built in the frame of repetition 0 and re-expressed exactly in the frame of the repetition a method receives by the phase ramp of $-\text{shift}_k$. Four variants form a 2×2 factorial. R0 is the published construction (no registration, magnitude average). R1 registers and averages mag-

nitudes, isolating misalignment. R2 registers in k-space and averages phase-aligned complex data. It is the corrected reference and also removes the noncentral-χ floor. R3 averages complex data without registration and isolates the averaging domain. Sec. 5 adds a fifth construction, R1_official: the published normalise-then-average label with registration as the only change.

Error budget. For every volume the reference carries a budget of what it can still be wrong by (App. C). The first term is the translation residual. It is the largest of the estimator's null floor at the group's SNR, the split-coil half-difference RMS and the joint-fit residual, converted to intensity by the RMS gradient. The second is the unexplained residual $\sqrt{\sigma^2_{\text{after}} - \sigma^2_{\text{noise}}}$ between registered, phase-aligned repetitions beyond thermal noise (through-plane motion on 5 mm slices, residual in-plane error, physiology). It is reported, not corrected. The drift's own through-plane part, 0.08–0.15 mm per repetition (App. C), falls into this term. The third is the noise the average still carries, $\sigma_1/\sqrt{n}$. The fourth is the noncentral-χ bias in closed form for four coils (Koay & Basser, 2006). The fifth is the operator's round-trip cost. The components are summed in quadrature and expressed as the PSNR change an evaluation at the reference PSNR could see from reference error alone. A variant excludes the noise term. On the test split the translation residual is 0.014–0.017 px. The budget is 3.5/4.4/4.7 dB (T1w/T2w/FLAIR; 2.6/3.6/3.3 dB excluding the noise term) at a reference PSNR of 33 dB. On validation it is 3.6/4.4/4.9 dB. On validation the air background drops from 7.5 image units under R0 to 4.5 ($n = 3$) and 5.5 ($n = 2$) under R2, within 3–8 % of the noncentral-χ prediction. The budget is dominated by the reference's own noise and by the unexplained residual. The second pre-registered criterion thus uses the registration residual and the unexplained residual only (Sec. 5).

Validation. Four checks say the correction is real and introduces no misalignment of its own. First, on the validation split (90 groups) we inject known shifts of up to 1.5 px and matched noise into registered coil stacks. The shift-recovery RMSE is 0.016/0.018/0.021 px (T1w/T2w/FLAIR), with a bias below 0.03 px in every displacement bin. The measured σ_1 drop equals the true drop within 0.05 %. The RMSE of R2 to the known truth is 0.33/0.41/0.48× that of R0. Second, in a leave-one-repetition-out test (validation T1w/T2w; 60 groups × 3 held-out repetitions) the registered average of $m - 1$ repetitions predicts the held-out repetition better than the unregistered average. In the official convention the gain is $+0.91$ dB (T1w; CI 0.74–1.08; 97.8 % of repetitions) and $+0.53$ dB (T2w; 0.44–0.63; 100 %). FLAIR has two repetitions and admits no such test; the test split, with 6/6/4 repetitions, does (App. I). Third, complex averaging requires the registration. R3, our phase-aligned but unregistered complex average, is the analogue of the maintainers' phase-corrected virtual-coil combination (M4Raw maintainers, 2026), which likewise has no spatial registration step. Without registration the drift-induced k-space phase ramps destroy coherence and R3 collapses to 28–30 dB. Fourth, the re-scored methods' residual maps say the same (App. G.3). Against the published label, every method's residual carries the same apparent readout shift (-0.13 px) as the input itself. Against R1 it carries none (0.00 px on both axes).

4 THE RESOLUTION OF THE REFERENCE

A reference built from n noisy, drifting repetitions has a finite resolution. We measure it as the reference's agreement with an independent copy of itself. On the V1.6 test split (25 subjects per contrast) two disjoint m-subsets of the n repetitions are averaged separately and compared, using all disjoint pairs: 45 at $m = 2$ and 10 at $m = 3$ for T1w/T2w ($n = 6$), 3 at $m = 2$ for FLAIR ($n = 4$). From the two averages we take $\sigma_m = \text{std}(\text{avg}_A - \text{avg}_B)/\sqrt{2}$ (the difference method of Dietrich et al., 2007), the gain $g(m) = \sigma_1/\sigma_m$. We also take the test-retest PSNR and SSIM of the two halves with the official-convention scorer of Sec. 5. The unregistered (published) and registered magnitude constructions use the same draws, so their comparison is paired. The subject bootstrap gives the intervals.

Two things follow (Fig. 3, App. F). The gain $g(m)$ lies within 4 % of $\sqrt{m}$ for both constructions, slightly above it unregistered and at or slightly below it registered. This is not evidence that the repetitions are still. The disjoint-split estimator returns $\sqrt{m}$ for any exchangeable, uncorrelated perturbation (App. D). The level of σ_m, not its scaling, carries the misalignment. Registration lowers σ_m by 18–29 % (T2w least, T1w most). It raises the reference's test-retest agreement with

Table 1: **Outcome of the pre-registered decision rule.** M4Raw test split, $n = 25$ subjects per contrast. Unit: subject×contrast (mean of per-slice PSNR over the slices of one input repetition), official convention. Family: U-Net, NAFNet, BM3D. For each pair (k, l), $D_s = \mathrm{PSNR}_k - \mathrm{PSNR}_l$ per subject and $E_s = D_{\mathrm{new},s} - D_{\mathrm{old},s}$. Old = R0_official (the published label: each repetition per-slice normalised, then averaged); New = R1 (registered, averaged, then normalised). Overbars: means over subjects; CI: 95% percentile subject bootstrap (2000 resamples, seed 0). Holm p: two-sided Wilcoxon on E_s, Holm-corrected within contrast ($\alpha = 0.05$). ***flip***: the sign of $\bar{D}$ changes and E is Holm-significant (criterion 1); †: also under the stricter variant; ***sig., no flip***: Holm-significant E without a sign change; ***n.s.***: not significant. Criterion (2), $|\bar{E}| > \max(\text{budget}, |\bar{D}_{\mathrm{old}}|)$, held for no pair (budgets 3.8–4.8 dB). Frozen before unblinding. Settings as pre-registered: yes. Criterion (1) held for five pairs. Bold rows: the network-vs-BM3D pairs on which the main claim rests; their interaction keeps its sign under each of the three constructions of Sec. 5 (Table 4, Fig. 4).

Contrast	Pair (k, l)	$\bar{D}_{\mathrm{old}}$	$\bar{D}_{\mathrm{new}}$	$\bar{E}$ [95% CI]	Holm p	Verdict
T1w	U-Net vs NAFNet	-1.12	-0.43	$+0.68$ [$+0.54$, $+0.81$]	3.6×10^{-7}	sig., no flip
	U-Net vs BM3D	-0.18	-0.84	$\mathbf{-0.66}$ **[$\mathbf{-0.88}$, $\mathbf{-0.46}$]**	6.6×10^{-6}	**sig., no flip**
	NAFNet vs BM3D	$+0.94$	-0.41	$\mathbf{-1.34}$ **[$\mathbf{-1.49}$, $\mathbf{-1.21}$]**	1.8×10^{-7}	**flip**†
T2w	U-Net vs NAFNet	-0.82	$+0.08$	$+0.90$ [$+0.78$, $+1.03$]	1.8×10^{-7}	flip†
	U-Net vs BM3D	$+0.05$	-0.14	$\mathbf{-0.20}$ **[$\mathbf{-0.39}$, $\mathbf{0.00}$]**	1.6×10^{-2}	**flip**
	NAFNet vs BM3D	$+0.87$	-0.22	$\mathbf{-1.10}$ **[$\mathbf{-1.22}$, $\mathbf{-0.99}$]**	1.8×10^{-7}	**flip**†
FLAIR	U-Net vs NAFNet	-0.75	-0.49	$+0.26$ [$+0.12$, $+0.39$]	2.5×10^{-3}	sig., no flip
	U-Net vs BM3D	-0.38	-0.79	$\mathbf{-0.41}$ **[$\mathbf{-0.58}$, $\mathbf{-0.26}$]**	2.4×10^{-4}	**sig., no flip**
	NAFNet vs BM3D	$+0.37$	-0.30	$\mathbf{-0.67}$ **[$\mathbf{-0.72}$, $\mathbf{-0.61}$]**	1.8×10^{-7}	**flip**†

itself by 1.1–1.8 dB at the same m (T1w $m = 3$: 31.7 → 33.5 dB; T2w $m = 3$: 30.6 → 31.8; FLAIR $m = 2$: 28.8 → 30.2).

Two caveats apply. A method is scored against the full n-average, whose noise is lower than that of an $m = n/2$ half. The family scores 33.4–36.2 dB against the published label (Table 2), above the half-average agreement. The test-retest value is therefore a resolution statement about the reference, not an upper bound on method PSNR. The draw-to-draw standard deviation of the reference's own agreement is 1.0–1.7 dB per subject unregistered and 0.4–0.7 dB registered, a two- to three-fold reduction. A paired comparison of two methods removes the part of this uncertainty that both share, and the pre-registered test resolves pairwise differences of 0.06–0.26 dB (Sec. 5). What pairing cannot remove is a reference error that interacts with the method, and Sec. 5 tests for one with three reference constructions. We do not extrapolate to $m = n$.

5 Re-scoring the published methods under a pre-registered decision rule

Reproduce first. Following Reinke et al. (2023), we first ran the authors' released NAFNet weights through their official evaluation code. The protocol min-max normalises each repetition's root-sum-of-squares image per slice to $[0, 1]$, averages the normalised repetitions into the label, feeds the fixed middle repetition as input, clips the outputs to $[0, 1]$ and computes PSNR and SSIM per slice with data range 1. The scores are 36.67/35.63/34.27 dB on validation (T1w/T2w/FLAIR; noisy input 33.77/32.66/32.65) and 36.15/35.15/34.12 dB on test (32.59/31.17/31.37). The same NAFNet outputs score $+2.68/+2.87/+1.91$ dB (T1w/T2w/FLAIR, test) higher against the published label than against the average-then-normalise R0 of the same unregistered repetitions. The published label normalises each repetition before averaging. We call it R0_official and keep the two apart.

Family, protocol and decision rule (frozen). The confirmatory family is the publicly available M4Raw-trained weights. These are the official U-Net (Ronneberger et al., 2015; Lyu et al., 2023a, release v20230206, 1.99 GB, 256 channels) and the official NAFNet (Chen et al., 2022, 117 MB), both checksummed. The family also contains BM3D (Dabov et al., 2007) with a frozen σ rule: a per-slice wavelet-based noise estimate on the raw-unit input, with no per-split or per-contrast tuning. C2S (Tu et al., 2025), the MIDL 2024 baselines (Li et al., 2024; Zhang et al., 2023; Zamir et al.,

Table 2: **Mean PSNR of the pre-registered family under three reference constructions** (V1.6 test, $n = 25$ subjects per contrast, official convention). R0_official (R0_off.): the published label. R1: registered, averaged, then normalised. R1_official (R1_off.): the published label with registration as the only change. Bold: best method per column. Pairwise statistics: Tables 1 and 4.

	T1w			T2w			FLAIR		
Method	R0_off.	R1	R1_off.	R0_off.	R1	R1_off.	R0_off.	R1	R1_off.
U-Net (official)	35.04	33.78	35.95	34.33	32.43	34.59	33.37	32.49	33.98
NAFNet (official)	**36.15**	34.21	**37.17**	**35.15**	32.35	**35.37**	**34.12**	32.98	**34.88**
BM3D	35.22	**34.62**	36.67	34.28	**32.58**	34.88	33.75	**33.28**	34.69
noisy input	32.59	32.40	33.49	31.17	30.39	31.59	31.37	31.25	32.02

2022) and ImT-MRD (Zhu et al., 2024) had no public weights at the freeze; author-provided and retrained versions are exploratory and cannot satisfy the rule (App. G.4). The unit of analysis is the subject×contrast volume, the mean of per-slice PSNR over the 18 slices of the official middle input repetition. The primary split is the V1.6 test split ($n = 25$ per contrast); the validation split ($n = 30$) is descriptive. The reference of the rule is R1, registered magnitudes re-expressed exactly in the input repetition's frame. R0 is never shifted, because the published protocol never moves it. For every ordered pair (k, l) within a contrast we form, per subject, $D_{\mathrm{old},s}$ and $D_{\mathrm{new},s}$. These are the PSNR differences between the two methods under the old and the new reference. The interaction is $E_s = D_{\mathrm{new},s} - D_{\mathrm{old},s}$. The rule is met under either criterion. Criterion (1): the sign of $\bar{D}$ changes and a two-sided Wilcoxon test on E_s is Holm-significant over the three pairs within the contrast ($\alpha = 0.05$; Holm, 1979). Criterion (2): $|\bar{E}| > \max(\mathrm{budget}, |\bar{D}_{\mathrm{old}}|)$ with the same significance. A stricter variant, reported alongside, additionally requires the ordering under R0 itself to be Holm-significant. The budget is the registration residual plus the unexplained residual in normalised units at the method's own R0 PSNR. The text was frozen on 2026-09-01, before unblinding (full text in App. A). A predicted-displacement table was fixed before the first test scoring. A pilot on 10 training subjects, never on test, gave $\mathrm{SD}_E = 0.09$–0.40 dB (minimal detectable difference 0.06–0.26 dB at $n = 25$).

Outcome (Tables 1 and 2, Fig. 2). The pre-registered decision rule was satisfied, with five Holm-significant ranking flips. NAFNet-vs-BM3D inverts in all three contrasts ($\bar{E} = -1.34/-1.10/-0.67$ dB for T1w/T2w/FLAIR; raw Wilcoxon $p = 6 \times 10^{-8}$, Holm $p = 1.8 \times 10^{-7}$, all 25 subjects moving the same way). In T2w, U-Net-vs-NAFNet and U-Net-vs-BM3D invert as well, so the T2w point-estimate order reverses (Kendall $\tau_b = -1.00$; $+0.33$ for T1w and FLAIR). Under R1 BM3D has the highest point estimate in every contrast. Two of the three T2w new-order intervals include zero (inversion probabilities 0.87 and 0.96 under R1, Fig. 2b). The interaction holds for all five pairs, the new ordering only where the $\bar{D}_{\mathrm{new}}$ interval excludes zero. The rank-neutral shift is $-1.27/-2.13/-0.83$ dB. Criterion (2) was never met (budgets 3.8–4.8 dB, conservative by construction). Pooled Holm over all nine pairs, a declared sensitivity check, leaves every interaction significant (largest raw $p = 0.016$).

The "old" reference of this test is R0_official, the published label: each repetition normalised, then averaged. The "new" one is R1: registered, averaged, then normalised. Registration and the normalisation order therefore changed together, and the rank-neutral shift contains the normalisation-order component. The like-for-like average-then-normalise R0→R1 shift is *positive* for every family method ($+0.74/+0.07/+0.77$ dB for NAFNet on test). The controls below separate the two.

Control: registration as the only change (Table 4 and Fig. 4, App. G.1). R1_official keeps the published construction and changes only the geometry. Each repetition is per-slice normalised exactly as in R0_official, translated into the input repetition's frame with the audit shifts, then averaged, so every scale factor is unchanged. It is aligned: the input's apparent readout residual shift is 0.000 px against R1_official and -0.11 to -0.16 px against R0_official. All methods score higher against it (uniform shift $+1.13/+0.36/+0.77$ dB). Under this control NAFNet-vs-BM3D does *not* invert: $+0.94 \rightarrow +0.50$ / $+0.87 \rightarrow +0.49$ / $+0.37 \rightarrow +0.20$ dB (T1w/T2w/FLAIR), $\bar{E} = -0.44/-0.39/-0.17$ dB, all Holm $p \leq 8 \times 10^{-7}$. Registration alone removes about half of NAFNet's published lead over BM3D but not the lead itself. U-Net-vs-BM3D inverts in T2w

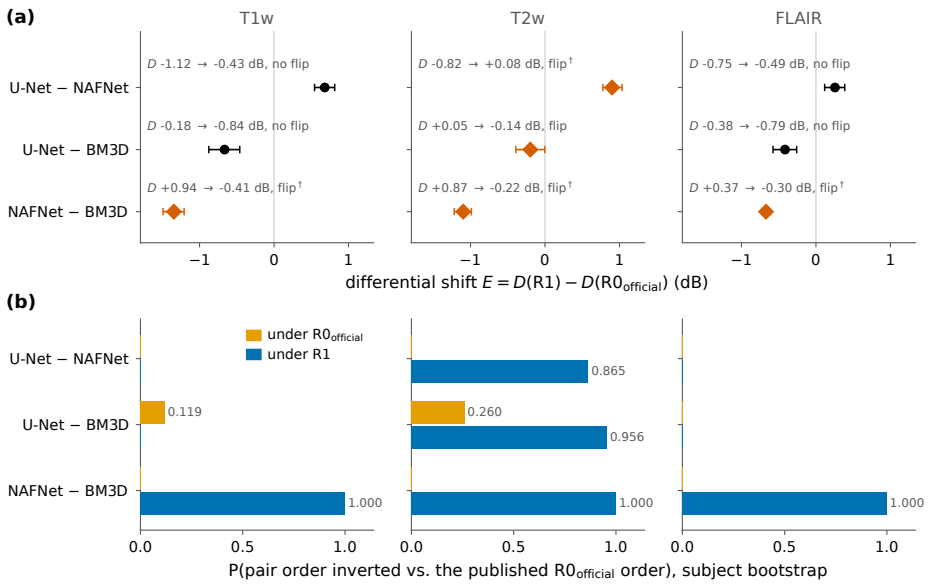

Figure 2: **The pre-registered decision rule (V1.6 test split).** Columns T1w, T2w, FLAIR; $n = 25$ subjects per contrast; unit and family as in Table 1; mean PSNRs in Table 2. **(a)** Pairwise interaction $E = D(R1) - D(R0_official)$ for the three family pairs; error bars: 95 % percentile subject-bootstrap CI (2000 resamples, seed 0). Vermilion diamonds: pre-registered flips (sign of $\bar{D}$ changes and Wilcoxon on E Holm-significant within the contrast). Dagger: also a flip under the stricter variant. Black: Holm-significant E without a sign change. Open grey: not significant. **(b)** Probability, over the same resamples, that the pair's order is inverted relative to the full-data order under R0_official, when the resample is scored under R0_official (orange) or R1 (blue). Fig. 4 (App. G.1) separates registration from the normalisation order.

($+0.05 \rightarrow -0.29$, Holm $p = 1.8 \times 10^{-7}$). Under the like-for-like average-then-normalise R0→R1 pairing, registration alone inverts four pairs with Holm significance ($\bar{E} = -0.50/-0.27/-0.31$ for NAFNet-vs-BM3D). Along that path the frozen T1w NAFNet-vs-BM3D interaction of -1.34 dB decomposes into -0.50 dB from registration and about -0.85 dB from the normalisation order; along the R1_official path the registration part is -0.44 dB. Across the three constructions the sign of the interaction never changes for any network-vs-BM3D pair. The inversion count is five, four and one for the frozen comparison, the like-for-like pairing and registration alone, or four, zero and zero under the stricter variant (Fig. 4, App. G.1). The same holds across metrics and regions (App. G.2). Inside a brain mask the interaction persists ($\bar{E} = -1.02/-0.97/-0.50$ for NAFNet-vs-BM3D, R0_official→R1) but NAFNet stays ahead ($+0.15/+0.17/+0.03$ dB). Under SSIM BM3D overtakes both networks in T1w only. The validation split ($n = 30$; smaller misalignment, 0.36–0.45 px) reproduces the T1w and T2w NAFNet-vs-BM3D inversions but not FLAIR ($+0.92 \rightarrow +0.64$).

What the interaction is. The defensible reading is not that BM3D is the better denoiser. It is that the published reference cannot order these methods. The two M4Raw-trained networks lose 0.1–1.3 dB relative to BM3D whenever the reference is aligned, under each of the three constructions. Whether that loss inverts the order depends on the normalisation order, the metric region and the metric. Relative to the registered reference the two networks' outputs are spectrally shrunk *anisotropically*, more along the readout axis than along phase encode. In the 0.15–0.35 cycles/px band the ratio a_{ro}/a_{pe} is 0.93/0.93/0.96 for NAFNet and 0.96/0.96/0.97 for U-Net (T1w/T2w/FLAIR), below 1 in every subject for NAFNet and in 74 of 75 volumes for U-Net. BM3D (1.02–1.03) and the noisy input (1.07–1.09) are isotropic or readout-heavy (App. G.3, Fig. 5). A blur confined to the readout axis is the drift signature of the labels (Sec. 2). This is consistent with the networks having absorbed part of the labels' drift blur. The two external methods not trained on M4Raw's averaged labels behave accordingly: registration displaces ImT-MRD least of the network rows and displaces the self-supervised C2S upwards (App. G.4). This is consistency, not proof. The maintainers expected

registration to change absolute values rather than conclusions (M. Lyu, personal communication, 4 September 2026); E measures exactly that.

Attribution, prediction and disclosure. The 2×2 factorial of Sec. 3 separates misalignment from the averaging domain. R2 (registered complex) sits 2.2–2.8 dB below R1 for every family method, because it removes the noncentral-χ floor that magnitude-trained networks learned. R3 (unregistered complex) collapses to 28–30 dB. The linear spectral-shrinkage model fixed before unblinding predicted interactions of -0.30 to $+0.10$ dB. The registration-only component measured on test ($-0.50/-0.27/-0.31$ dB for NAFNet-vs-BM3D) is within about twice that range. The excess in the frozen test is the normalisation-order component the model did not include (two of nine predicted signs were wrong). Disclosure: before the freeze, an exploratory single-method NAFNet re-scoring (2026-08-24) exposed its uniform displacements ($+0.74/+0.07/+0.77$ dB, R0→R1 on test); no pairwise quantity and no second method were computed. The validation split, SSIM, brain-masked PSNR, R2 and R3, the bootstrap, the stricter sensitivity variant, the controls above and all non-family methods are descriptive (App. G). The raw-unit family convention, the other input positions and the leave-input-out references named in the frozen text were not run for this version. Re-training three architectures on either label (App. G.4) keeps the sign of every network-vs-BM3D interaction and, on the registered label, removes the readout-axis shrinkage.

6 Related work, limitations and release

Related work. Audits of MRI benchmark flaws have so far concerned processing hidden in public data (Shimron et al., 2022), noise hidden in root-sum-of-squares references (Wang et al., 2024), metric insensitivity to lost resolution (Wang & Haldar, 2026) and reference-based evaluation of motion correction (Wang et al., 2026). The challenge literature has shown rankings to be unstable (Maier-Hein et al., 2018; Reinke et al., 2023) and has prescribed case-level aggregation, bootstrap stability and pre-specified analyses (Maier-Hein et al., 2024; Wiesenfarth et al., 2021). The metrics themselves are known to reward blur (Dohmen et al., 2025). A precedent about noise rather than geometry, at 6.5 mT, reversed an IFFT-vs-unrolled ranking by scoring against a 16-average reference that shared the input's k-space instead of an independent 256-average one (Shimron et al., 2024). Photographic denoising benchmarks align their multi-shot references before averaging (Plötz & Roth, 2017; Abdelhamed et al., 2018). Misalignment of the MRI reference is the geometric member of this family of hidden reference flaws. Unlike hidden noise, it cannot be repaired by a noise-aware metric. Among users of M4Raw, C2S (Tu et al., 2025), ImT-MRD (Zhu et al., 2024), NexOP (Oved & Shimron, 2026), the hybrid, unified and diffusion pipelines (Pei et al., 2026; Shao et al., 2026; Demir et al., 2025) and the maintainers' own benchmark and complex-averaging demo (Li et al., 2024; M4Raw maintainers, 2026) all score against unregistered averages. None budgets the reference or re-ranks. The ingredients of the correction are classical: real-valued averaging after phase removal (Eichner et al., 2015), the noncentral-χ bias (Koay & Basser, 2006; Tisdall et al., 2011) and sub-pixel registration by upsampled cross-correlation (Guizar-Sicairos et al., 2008).

Limitations. The correction is a 2-D in-plane translation; rotation is reported, not applied, and the through-plane part of the drift (0.08–0.15 mm per repetition on 5 mm slices) enters the budget only through the unexplained residual. The data come from one site, one scanner, one coil and healthy volunteers, though unregistered averaging is not specific to this dataset (Tibrewala et al., 2024). The GRE arm has two repetitions and cannot test growth. The confirmatory family is limited to the three methods with public M4Raw weights. The inversions, though not the sign of the network-vs-BM3D interaction, depend on the reference construction, the normalisation convention and the metric region (Sec. 5). No downstream, perceptual or reader-study endpoint is presented. The reference still contains noise and the input repetition; the latter is in both R0 and R1 and largely cancels in E. The mechanism is inferred from its signature, not from a field measurement.

Release. We release the corrected reference for the validation and V1.6 test splits under CC BY 4.0 as a derivative of M4Raw (Lyu et al., 2023a;b): one HDF5 file per subject×contrast (R1, R2, shifts, mask, coherence, provenance) with manifest, checksums, validator and per-volume tables. The code (MIT) includes the frozen pre-registration.

AI USE STATEMENT

The AI use statement required by the conference policy is given in full in App. K.

ETHICS STATEMENT

This work uses only M4Raw (Lyu et al., 2023a;b). M4Raw is a public, de-identified dataset of healthy volunteers, collected by its authors under their institutional approval and released under CC BY 4.0. We use the publicly released or author-provided weights of the methods we re-score. The corrected reference we release is a derivative of M4Raw under the same licence and adds no personal information. The paper questions published rankings. To avoid selective reporting, the confirmatory family was restricted to methods with publicly released M4Raw weights. The analysis was frozen before unblinding (App. A). Every deviation and every post-hoc analysis is disclosed. The maintainers of the dataset and the authors of the re-scored methods were contacted before submission. No competing interests are declared. Funding sources are named in the de-anonymised version.

REPRODUCIBILITY STATEMENT

Two artefacts accompany this paper. The first is the anonymised supplementary archive: the code (MIT) for the audit, the corrected reference, the error budget, the resolution measurement, the re-scoring, the pre-registered test and the post-decision controls, together with every result table the paper reads and a checksum list over them. Every number in the paper is read from one of those files, and an index in the archive maps each paragraph, table and caption to its source files. The figures and tables are regenerated from the same files by deterministic scripts. The second artefact is the corrected reference itself (CC BY 4.0, a derivative of M4Raw), released separately with a manifest, per-file checksums and a validator that re-derives it from the raw M4Raw files (Sec. 6). The pre-registration was frozen before unblinding and is reproduced in App. A; the record of the pre-registered test carries checksums of the code and its input scores. The evaluation protocol is the vendored official M4Raw demo, pinned by tests. The checkpoints are identified by checksums (App. B). The Python environment that produced the scores is frozen in an environment specification.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1692–1700, 2018. doi: 10.1109/CVPR.2018.00182.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision (ECCV)*, volume 13667 of *Lecture Notes in Computer Science*, pp. 17–33, 2022. doi: 10.1007/978-3-031-20071-7_2.

Clarissa Zimmerman Cooley, Patrick C. McDaniel, Jason P. Stockmann, Sai Abitha Srinivas, Stephen F. Cauley, Monika Śliwiak, Charlotte R. Sappo, Christopher F. Vaughn, Bastien Guerin, Matthew S. Rosen, Michael H. Lev, and Lawrence L. Wald. A portable scanner for magnetic resonance imaging of the brain. *Nature Biomedical Engineering*, 5(3):229–239, 2021. doi: 10.1038/s41551-020-00641-5. States a temperature coefficient of NdFeB remanence of about $-0.1\,\%/^{\circ}\mathrm{C}$ and of coercivity of about $-0.5\,\%/^{\circ}\mathrm{C}$ (Discussion).

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16 (8):2080–2095, 2007. doi: 10.1109/TIP.2007.901238.

Başar Demir, Yikang Liu, Xiao Chen, Eric Z. Chen, Lin Zhao, Boris Mailhe, Terrence Chen, and Shanhui Sun. DiffDenoise: Self-supervised medical image denoising with conditional diffusion models. arXiv:2504.00264, 2025.

Olaf Dietrich, José G. Raya, Scott B. Reeder, Maximilian F. Reiser, and Stefan O. Schoenberg. Measurement of signal-to-noise ratios in MR images: Influence of multichannel coils, parallel

imaging, and reconstruction filters. *Journal of Magnetic Resonance Imaging*, 26(2):375–385, 2007. doi: 10.1002/jmri.20969.

Melanie Dohmen, Mark A. Klemens, Ivo M. Baltruschat, Tuan Truong, and Matthias Lenga. Similarity and quality metrics for MR image-to-image translation. *Scientific Reports*, 15:3853, 2025. doi: 10.1038/s41598-025-87358-0.

Cornelius Eichner, Stephen F. Cauley, Julien Cohen-Adad, Harald E. Möller, Robert Turner, Kawin Setsompop, and Lawrence L. Wald. Real diffusion-weighted MRI enabling true signal averaging and increased diffusion contrast. *NeuroImage*, 122:373–384, 2015. doi: 10.1016/j.neuroimage.2015.07.074.

Manuel Guizar-Sicairos, Samuel T. Thurman, and James R. Fienup. Efficient subpixel image registration algorithms. *Optics Letters*, 33(2):156–158, 2008. doi: 10.1364/OL.33.000156.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Cheng Guan Koay and Peter J. Basser. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. *Journal of Magnetic Resonance*, 179(2):317–322, 2006. doi: 10.1016/j.jmr.2006.01.016.

Kieran G. Larkin, Michael A. Oldfield, and Hanno Klemm. Fast Fourier method for the accurate rotation of sampled images. *Optics Communications*, 139(1–3):99–106, 1997. doi: 10.1016/S0030-4018(97)00097-7.

Yi Li, Mengye Lyu, Guanxiong Luo, Jingwei Guan, and Jingyu Li. Recent updates of the M4Raw dataset and applications in evaluating MRI denoising methods. In *Medical Imaging with Deep Learning (MIDL)*, 2024. URL `https://openreview.net/forum?id=ldvtPOCS0i`. Short paper, 4 pages.

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X. Wu. M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. *Scientific Data*, 10:264, 2023a. doi: 10.1038/s41597-023-02181-4.

Mengye Lyu et al. M4Raw V1.6: multi-coil test subset (25 subjects, 6/6/4 repetitions) and V1.5 GRE data. Zenodo, 2023b. doi: 10.5281/zenodo.8056074; CC BY 4.0. The V1.1 train, validation and motion splits are doi: 10.5281/zenodo.7523691.

Mengye Lyu et al. M4Raw_brain v1.6.0 dataset card. Hugging Face, `https://huggingface.co/datasets/mylyu/M4Raw_brain`, 2026. Card dated 2026-08-20; states that repeat averages are higher-SNR references rather than independent anatomical ground truth, and that the header axis labels are reversed for T1w/T2w/FLAIR.

M4Raw maintainers. `data_correction_demo.ipynb`: phase-corrected complex-domain repetition averaging (BART `cc`), pull request #5 of `mylyu/M4Raw`. GitHub, `https://github.com/mylyu/M4Raw`, 2026. Opened 2026-04-28, merged 2026-06-03; no spatial registration step.

Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9:5217, 2018. doi: 10.1038/s41467-018-07619-7.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z.

José P. Marques, Frank F. J. Simonis, and Andrew G. Webb. Low-field MRI: An MR physics perspective. *Journal of Magnetic Resonance Imaging*, 49(6):1528–1542, 2019. doi: 10.1002/jmri.26637.

Thomas O'Reilly, Wouter M. Teeuwisse, Danny de Gans, Kirsten Koolstra, and Andrew G. Webb. In vivo 3D brain and extremity MRI at 50 mT using a permanent magnet Halbach array. *Magnetic Resonance in Medicine*, 85(1):495–505, 2021. doi: 10.1002/mrm.28396. Thermal sensitivity of the 50 mT Halbach magnet about 2.5 kHz per degree Celsius (about 1200 ppm/°C at 2.15 MHz, our conversion); in vivo drift about 800 Hz per hour.

Tal Oved and Efrat Shimron. NexOP: Joint optimization of NEX-aware k-space sampling and image reconstruction for low-field MRI. arXiv:2605.11583, 2026.

Haoyang Pei, Nikola Janjušević, Renqing Luo, Ding Xia, Xiang Xu, William Moore, Yao Wang, Hersh Chandarana, and Li Feng. Hybrid learning: a combination of self-supervised and supervised learning for joint MRI reconstruction and denoising in low-field MRI. *Physics in Medicine & Biology*, 71(12):125029, 2026. doi: 10.1088/1361-6560/ae792b. arXiv:2505.05703.

Tobias Plötz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2750–2759, 2017. doi: 10.1109/CVPR.2017.294.

Annika Reinke, Georg Grab, and Lena Maier-Hein. Challenge results are not reproducible. In *Bildverarbeitung für die Medizin (BVM)*, Informatik aktuell, pp. 198–203, 2023. doi: 10.1007/978-3-658-41657-7_43.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Yuchen Shao, Hongyan Huang, Lingyan Zhang, Dongsheng Li, Zhiguang Ding, Fan Wang, Shengli Chen, Shiwei Lin, Yuning Gu, Mu Du, et al. Real-world unified denoising for multi-organ fast MRI: a large-scale prospective validation. *npj Digital Medicine*, 9:366, 2026. doi: 10.1038/s41746-026-02548-y.

Efrat Shimron, Jonathan I. Tamir, Ke Wang, and Michael Lustig. Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences*, 119(13):e2117203119, 2022. doi: 10.1073/pnas.2117203119. arXiv:2109.08237.

Efrat Shimron, Shanshan Shan, James Grover, Neha Koonjoo, Sheng Shen, Thomas Boele, Annabel J. Sorby-Adams, John E. Kirsch, Matthew S. Rosen, and David E. J. Waddington. Accelerating low-field MRI: From compressed sensing to deep learning reconstruction with CNNs and transformers. arXiv:2411.06704, 2024. Sec. 3.3 and Fig. 5 of v2 score undersampled 6.5 mT reconstructions against a NEX = 16 reference built from the same k-space and against an independent NEX = 256 reference; the IFFT-versus-unrolled ranking reverses between the two.

Radhika Tibrewala, Tarun Dutt, Angela Tong, Luke Ginocchio, Riccardo Lattanzi, Mahesh B. Keerthivasan, Steven H. Baete, Sumit Chopra, Yvonne W. Lui, Daniel K. Sodickson, Hersh Chandarana, and Patricia M. Johnson. FastMRI Prostate: A public, biparametric MRI dataset to advance machine learning for prostate cancer imaging. *Scientific Data*, 11:404, 2024. doi: 10.1038/s41597-024-03252-w.

M. Dylan Tisdall, Richard A. Lockhart, and M. Stella Atkins. The bias/variance trade-off when estimating the MR signal magnitude from the complex average of repeated measurements. *Magnetic Resonance in Medicine*, 66(5):1456–1467, 2011. doi: 10.1002/mrm.22910.

Jiachen Tu, Yaokun Shi, and Fan Lam. Score-based self-supervised MRI denoising. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=uNd289HjLi`. arXiv:2505.05631; public code and M4Raw checkpoints (MIT).

Michael Unser, Philippe Thévenaz, and Leonid Yaroslavsky. Convolution-based interpolation for fast, high-quality rotation of images. *IEEE Transactions on Image Processing*, 4(10):1371–1381, 1995. doi: 10.1109/83.465102.

Lawrence L. Wald, Patrick C. McDaniel, Thomas Witzel, Jason P. Stockmann, and Clarissa Zimmerman Cooley. Low-cost and portable MRI. *Journal of Magnetic Resonance Imaging*, 52(3):686–696, 2020. doi: 10.1002/jmri.26942.

Jiayang Wang and Justin P. Haldar. Well-designed k-space coverage is important for good MRI denoising. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 2026. doi: 10.1007/s10334-026-01339-6. arXiv:2511.05735.

Jiayang Wang, Di An, and Justin P. Haldar. The "hidden noise" problem in MR image reconstruction. *Magnetic Resonance in Medicine*, 92(3):982–996, 2024. doi: 10.1002/mrm.30100.

Kun Wang, Tobit Klug, Stefan Ruschke, Jan S. Kirschke, and Reinhard Heckel. Reliable evaluation of MRI motion correction: Dataset and insights. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=5PY8HR2Zz6. arXiv:2506.05975.

Manuel Wiesenfarth, Annika Reinke, Bennett A. Landman, Matthias Eisenmann, Laura Aguilera Saiz, M. Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific Reports*, 11: 2369, 2021. doi: 10.1038/s41598-021-82017-6. arXiv:1910.05121; https://github.com/wiesenfa/challengeR.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5718–5729, 2022. doi: 10.1109/CVPR52688.2022.00564.

Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via Swin-Conv-UNet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023. doi: 10.1007/s11633-023-1466-0.

Zheren Zhu, Azaan Rehman, Xiaozhi Cao, Congyu Liao, Yoo Jin Lee, Michael Ohliger, Hui Xue, and Yang Yang. Advancing low-field MRI with a universal denoising imaging transformer: towards fast and high-quality imaging. arXiv:2404.19167, 2024. Evaluates on the M4Raw V1.6 test subset against an unregistered complex-average reference; weights by request only.

A Pre-registration and disclosures

The pre-registered decision rule was written in two steps, both before any pairwise quantity existed. The first text named a method family that was not executable, since no public C2S or MIDL-baseline weights exist, and its metric-displacement criterion could be met trivially by a rank-neutral shift. The amendment below made three changes. The family became the three publicly released weights. The criterion became the differential displacement E_s with Holm correction within contrast. The reference of the rule became R1. The amendment is reproduced verbatim from the frozen document. Its operationalisation fixes the unit of analysis, input repetition, BM3D σ rule, budget, declared exploratory analyses and "what we will not do". The archived pre-registration document holds it, with the full list of pre-freeze exposures and post-decision descriptive analyses. The rule is also pinned in code and tested. The freeze came first, then the rule's code, then the predicted-displacement table, then the first test scoring, the rule's evaluation and the recorded outcome, in that order. Timestamps are in the archive.

```
Published-method family (only weights that exist publicly today, sha256 recorded):
  official U-Net (M4Raw release v20230206), official NAFNet (same release), BM3D
  (deterministic, sigma rule fixed in docs/preregistration.md). C2S (no public code
  or weights), SCUNet / Restormer / MPRNet (no public weights) are NOT in the family;
  retrained or zero-shot versions may be reported only as exploratory, never for the gate.

Unit of analysis: subject x contrast volume (mean over slices, per-slice official
metric convention: per-slice min-max normalisation, data_range = 1). Primary split:
M4Raw V1.6 test (n = 25 subjects); secondary: val (n = 30). Input repetition: the
official fixed middle repetition; all-positions analysis is exploratory.

Corrected reference for the gate: R1 = registered magnitude average in the input
repetition's frame (isolates misalignment; same averaging domain as R0). R2
(registered complex) and R3 are attribution/ablation rows, not gate inputs.

For every ordered method pair (k, l) within a contrast:
```

```
  D_old_s = M_k(R0)_s - M_l(R0)_s,  D_new_s = M_k(R1)_s - M_l(R1)_s,
  E_s = D_new_s - D_old_s  (the reference x method interaction, per subject).

GO if at least one of:
  (1) Ranking flip: sign(mean D_old) != sign(mean D_new) AND the paired test on E_s
      is significant after Holm correction over all pairs within the contrast
      (alpha = 0.05); OR
  (2) Differential displacement: |mean E| > max(budget_dB, |mean D_old|) AND
      Holm-significant as in (1), where budget_dB is computed from the A1 error
      budget using ONLY the registration residual and the through-plane lower
      bound (the residual-noise term sigma_1^2/m is reported, not budgeted).
The uniform (rank-neutral) shift, bootstrap flip probabilities, Kendall tau_b,
challengeR plots, SSIM, val, leave-input-out references and all non-family methods
are descriptive. A per-pair predicted-displacement table (from the spectral
shrinkage of each method's outputs under R0/R1) is committed before unblinding.
NO-GO consequences: unchanged from v1.
```

Of the descriptive analyses named in the frozen text, the raw-unit family convention, the other input positions and the leave-input-out references were not run for this version. The pooled Holm correction over all nine pairs (pre-registration 2.8) leaves every interaction significant, the largest raw p being 0.016.

B Official protocol, checkpoints and the BM3D rule

- The maintainers' official evaluation code (CC BY 4.0) defines the protocol. Each repetition's root-sum-of-squares image is min-max normalised per slice. The label is the mean of the normalised repetitions, input included. The test input is the repetition at position count/2, rounded down, in sorted order. On the test split this is 0-based index 3 of 6 for T1w/T2w and 2 of 4 for FLAIR. Outputs are clipped to $[0, 1]$. PSNR and SSIM are scikit-image, per slice, data range $= 1$. Our implementation reproduces it to within 10^{-6} and is pinned by tests.
- Checkpoints: the released U-Net (1,985,501,085 B) and NAFNet (116,866,513 B) weights, release v20230206, with checksums recorded and verified on load.
- BM3D runs per 2-D slice of the raw-unit input repetition with package defaults. σ comes from the scikit-image wavelet-MAD estimator (db2). A slice with $\sigma = 0$ passes through. No per-split or per-contrast tuning.
- One protocol correction preceded test scoring. The first BM3D run used repetition 0 as input instead of the official middle repetition. It was fixed and recorded per pre-registration 2.9.

C Error budget definitions

The budget is computed per subject$\times$contrast volume and stored with each released reference and in a per-volume table. Intensities are in the units of the released magnitude images. The mean signal is the mean of R2 inside the brain mask shared with the audit. The components are:

1. *Registration residual* (px): the $\max$ of three independent bounds on the error of the applied translation. The first is the estimator's null floor at the group's SNR (control C2). The second is half the RMS difference between the shifts from coils $\{0, 1\}$ and $\{2, 3\}$. The third is the residual of the joint pairwise least-squares fit. In intensity it is the residual times the RMS gradient of R2 over the mask.
2. *Unexplained residual*: $\sqrt{\max(\sigma^2_{\text{after}} - \sigma^2_{\text{noise}}, 0)}$. Here σ_{after} is the pairwise noise level of the registered, phase-aligned repetitions in the complex domain. σ_{noise} is the per-coil background noise: the standard deviation of real and imaginary parts over the lowest-10 % air region, eroded 3 px, excluding the zero-fill ringing rows. It bounds through-plane motion, residual in-plane error and physiological change. It is reported, not corrected. The drift itself has a through-plane part: at an assumed 1–2 kHz slice-selection bandwidth, 30 Hz per repetition moves a 5 mm slice by 0.08–0.15 mm per repetition, 0.4–0.75 mm over six repetitions. It cannot be corrected in plane and enters this component.

3. *Noise after averaging*: $\sigma_1/\sqrt{n}$, with σ_1 the pre-registration magnitude noise level (the larger, more conservative choice).
4. *Noncentral-χ bias*: the mask mean of the closed-form $\mathbb{E}[\mathrm{RSS}] - |\mathrm{signal}|$ for four coils (Koay & Basser, 2006). Where the signal is zero, the same effect is the air background of R0 minus that of R2.
5. *Operator cost*: the relative RMS change of a round trip through the shift operator at the observed displacement. On the k-space path it is $\sim 10^{-15}$ (not summed). On the magnitude path it is what R1 pays (control C1b).

Components 1–4 are summed in quadrature relative to the mean signal, and $\mathrm{budget}_{\mathrm{dB}} = 10\log_{10}\big(1 + (\mathrm{total}\cdot\overline{\mathrm{signal}})^2/\mathrm{mse}_{\mathrm{ref}}\big)$ with $\mathrm{mse}_{\mathrm{ref}}$ the MSE at the evaluation PSNR. This is the PSNR change an evaluation could see from reference error alone. A second variant omits component 3. Components 3 and 4 are common to R0 and R2 to first order. Including them makes the budget larger and criterion (2) harder to meet, never easier. This was deliberate and declared before the rule was evaluated. The rule uses only components 1 and 2 in normalised units (pre-registration 2.6). The frozen phrase "through-plane lower bound" refers to component 2, which is an upper bound on the through-plane part. Per-contrast test means: translation residual (the frozen text's registration residual) 0.014–0.017 px. Unexplained residual 2.4–3.1, noise after averaging 1.8–2.5 and noncentral-χ bias 0.8–1.7 image units. The budget is 3.5/4.4/4.7 dB (T1w/T2w/FLAIR).

D WHY $g(m) = \sqrt{m}$ IS AN IDENTITY

Write the repetitions as $x_i = s + \delta_i$, $i = 1, \ldots, n$, where s is the common image. The perturbations δ_i (noise, and any displacement of the anatomy) are exchangeable with pairwise-difference variance $\sigma_1^2 = \frac{1}{2}\mathrm{Var}(\delta_i - \delta_j)$. For two disjoint index sets A and B of size m, $\mathrm{avg}_A - \mathrm{avg}_B = \frac{1}{m}\sum_{i\in A}\delta_i - \frac{1}{m}\sum_{j\in B}\delta_j$. By exchangeability its variance is $\frac{2}{m}\cdot\frac{1}{2}\mathrm{Var}(\delta_i - \delta_j) = 2\sigma_1^2/m$ when the perturbations are uncorrelated across repetitions. Hence $\sigma_m^2 = \mathrm{Var}(\mathrm{avg}_A - \mathrm{avg}_B)/2 = \sigma_1^2/m$ and $g(m) = \sigma_1/\sigma_m = \sqrt{m}$, whatever the δ_i are. The misalignment enters the *level* of σ_1 and σ_m, not the $\sqrt{m}$ scaling. So Fig. 3a cannot detect it and Fig. 3b can. A deterministic drift breaks exchangeability, because the difference between two halves then depends on which repetitions they contain. This is the small excess above $\sqrt{m}$ of the unregistered T1w and FLAIR points. Registration removes it.

E PER-SUBJECT TABLE

Table 3 covers every subject and contrast of the test split. It lists the PSNR of each family method under the published label, the registered reference and the registration-only control. It also lists the three pairwise interactions E_s that Table 1 aggregates. It is generated deterministically from the released score tables.

Table 3: Per-subject PSNR (dB; mean over the 18 slices of the official middle input repetition) of the family under the published label R0_official, the registered reference R1 and the registration-only control R1_official, and the interaction $E_s = D_{\mathrm{new}} - D_{\mathrm{old}}$ (R0_official → R1) for the three ordered pairs. U = official U-Net, N = official NAFNet, B = BM3D.

contrast	subject	R0_official			R1			R1_official			E_s (R0_official→R1)		
		U	N	B	U	N	B	U	N	B	U−N	U−B	N−B
T1w	2022101101	36.80	39.55	38.97	34.21	36.30	37.37	37.24	39.96	39.77	+0.66	-1.00	-1.65
	2022101102	35.14	38.02	36.75	33.76	36.01	36.87	36.53	39.30	38.80	+0.63	-1.50	-2.13
	2022101103	35.95	38.56	38.08	34.24	36.11	36.82	36.26	39.07	38.61	+0.74	-0.45	-1.18
	2022101201	34.74	35.80	34.90	33.09	33.09	33.21	35.01	36.03	35.47	+1.06	+0.04	-1.02
	2022101204	34.22	35.44	34.24	34.32	34.74	35.37	36.47	38.11	37.74	+0.80	-1.02	-1.83
	2022101301	34.18	35.34	34.42	33.25	33.80	33.92	35.20	36.57	35.91	+0.61	-0.44	-1.05
	2022101302	34.53	35.25	34.39	32.71	32.67	32.73	34.65	35.37	34.64	+0.76	-0.16	-0.92
	2022101303	36.53	37.65	37.07	34.64	34.32	34.82	36.74	37.96	37.44	+1.44	+0.36	-1.08
	2022112702	35.91	36.79	35.51	34.57	34.84	35.29	37.18	38.16	37.55	+0.62	-1.11	-1.73
	2022112704	35.87	36.95	36.36	33.81	33.94	34.57	36.09	37.08	36.79	+0.95	-0.28	-1.23
	2022120401	36.24	36.63	36.22	33.73	33.48	34.13	36.33	36.72	36.51	+0.63	-0.42	-1.06
	2022120402	35.28	35.93	34.76	34.58	34.62	35.00	36.58	37.81	37.20	+0.61	-0.94	-1.55
	2022120403	35.80	36.42	35.04	34.12	34.11	34.64	37.02	37.61	37.24	+0.62	-1.28	-1.91
	2022120404	33.70	34.62	33.37	32.45	32.65	32.76	34.47	35.32	34.86	+0.73	-0.64	-1.37
	2022120406	35.25	36.12	35.05	34.09	34.30	34.68	35.96	36.85	36.41	+0.66	-0.79	-1.45
	2022120407	36.86	37.45	36.43	34.92	34.47	35.06	37.12	37.60	37.28	+1.04	-0.57	-1.61

Table 3 continued

		U	N	B	U	N	B	U	N	B	U−N	U−B	N−B
	2022120408	33.90	34.42	33.92	33.58	33.63	33.80	34.73	35.47	35.02	+0.47	-0.20	-0.67
	2022120409	36.23	37.42	36.61	34.06	34.44	34.69	36.48	37.84	37.06	+0.82	-0.25	-1.07
	2022121701	34.82	35.69	34.86	34.29	34.96	35.32	36.50	37.42	36.76	+0.19	-0.99	-1.18
	2022121702	35.09	35.98	34.66	33.83	34.51	35.08	36.19	36.92	36.57	+0.21	-1.69	-1.90
	2022121703	31.78	32.09	31.03	33.19	33.61	33.50	35.62	36.16	35.18	-0.11	-1.05	-0.95
	2022121705	33.29	34.12	32.74	32.88	33.70	33.76	34.93	36.42	35.51	+0.01	-1.43	-1.44
	2022121706	33.50	34.57	33.31	33.05	33.57	33.77	34.70	35.96	35.41	+0.55	-0.91	-1.46
	2022121707	35.34	36.95	36.30	33.61	33.91	34.48	35.74	37.27	37.06	+1.31	+0.09	-1.22
	2022121708	35.02	36.10	35.41	33.51	33.59	33.86	35.10	36.18	35.84	+1.00	+0.04	-0.96
T2w	2022101101	33.78	35.26	34.50	31.85	32.01	32.25	33.80	35.16	34.81	+1.33	+0.31	-1.01
	2022101102	31.09	32.02	30.97	30.53	31.09	30.72	31.96	32.85	31.93	+0.37	-0.31	-0.68
	2022101103	34.16	35.11	34.58	32.50	31.51	31.53	34.13	35.07	34.60	+1.94	+1.38	-0.56
	2022101201	35.26	36.39	35.57	33.03	33.40	33.59	35.12	36.17	35.80	+0.76	-0.25	-1.01
	2022101204	34.60	35.64	34.23	32.92	32.96	33.31	35.36	36.36	35.80	+1.01	-0.75	-1.75
	2022101301	34.38	35.00	34.11	32.99	32.57	32.64	35.25	35.95	35.32	+1.04	+0.08	-0.96
	2022101302	32.95	33.62	32.50	31.97	32.32	32.16	34.11	34.84	34.00	+0.32	-0.64	-0.96
	2022101303	34.50	35.04	34.44	33.07	32.82	33.03	35.56	36.33	35.74	+0.79	-0.02	-0.81
	2022112702	35.32	35.89	34.72	33.15	32.69	32.87	35.48	36.00	35.44	+1.03	-0.32	-1.34
	2022112704	34.31	34.96	34.27	32.40	31.94	32.26	34.40	35.02	34.59	+1.11	+0.09	-1.01
	2022120401	35.19	35.98	35.18	32.79	32.76	33.04	35.04	35.74	35.31	+0.81	-0.26	-1.07
	2022120402	34.35	34.99	33.97	32.35	32.37	32.68	34.51	35.08	34.59	+0.63	-0.70	-1.32
	2022120403	34.99	35.59	34.09	32.88	32.79	33.05	35.40	35.88	35.25	+0.69	-1.07	-1.76
	2022120404	34.96	35.42	34.58	32.63	32.17	32.48	34.94	35.29	34.97	+0.92	-0.23	-1.15
	2022120406	35.02	35.33	34.79	32.73	32.30	32.70	35.05	35.32	35.14	+0.75	-0.20	-0.95
	2022120407	33.50	35.25	34.65	31.53	32.20	32.75	33.63	35.31	35.02	+1.09	-0.06	-1.15
	2022120408	34.02	35.16	34.70	32.40	32.13	32.36	34.06	35.21	34.79	+1.41	+0.72	-0.70
	2022120409	34.48	35.23	34.69	32.51	32.38	32.70	34.51	35.27	34.76	+0.87	+0.01	-0.86
	2022121701	34.35	35.31	34.66	32.31	32.42	32.67	34.22	35.16	34.81	+0.86	-0.05	-0.91
	2022121702	34.38	35.12	33.97	32.39	32.35	32.70	34.62	35.20	34.77	+0.78	-0.72	-1.50
	2022121703	34.07	35.01	33.77	32.18	32.20	32.24	34.27	35.13	34.62	+0.93	-0.35	-1.28
	2022121705	34.67	35.32	34.13	32.53	32.75	33.10	35.11	35.67	35.10	+0.42	-1.11	-1.53
	2022121706	34.64	35.32	34.35	32.62	32.25	32.65	34.66	35.24	34.73	+1.04	-0.32	-1.36
	2022121707	35.65	36.48	35.92	32.53	32.62	32.97	35.61	36.33	35.99	+0.75	-0.16	-0.90
	2022121708	33.67	34.36	33.59	32.07	31.87	31.99	34.04	34.67	34.16	+0.89	+0.00	-0.88
FLAIR	2022101101	35.61	36.71	36.63	34.63	35.50	36.22	36.15	37.36	37.46	+0.24	-0.57	-0.80
	2022101102	32.24	33.96	33.89	31.83	33.44	34.03	33.15	35.03	35.21	+0.11	-0.55	-0.66
	2022101103	35.71	36.33	36.25	33.94	34.17	34.79	35.82	36.48	36.43	+0.39	-0.31	-0.69
	2022101201	33.44	34.54	33.99	32.69	33.04	33.14	33.64	34.81	34.39	+0.74	+0.09	-0.66
	2022101204	33.25	33.64	33.15	33.61	33.93	34.18	35.04	35.75	35.59	+0.06	-0.67	-0.73
	2022101301	32.90	33.78	33.30	31.65	32.60	32.74	33.31	34.32	33.99	-0.07	-0.68	-0.62
	2022101302	30.62	31.13	30.84	30.06	30.75	30.73	31.58	32.43	32.13	-0.18	-0.45	-0.26
	2022101303	33.50	35.01	34.61	32.15	32.72	32.95	33.70	35.24	34.94	+0.94	+0.31	-0.63
	2022112702	33.01	33.47	32.99	32.71	32.70	32.83	34.12	34.66	34.38	+0.47	-0.15	-0.62
	2022112704	34.88	34.79	34.43	33.16	33.02	33.39	35.06	35.00	34.69	+0.06	-0.67	-0.73
	2022120401	33.69	34.37	33.89	32.22	32.74	33.12	33.96	34.69	34.40	+0.17	-0.69	-0.86
	2022120402	32.58	33.90	33.36	32.17	33.12	33.42	33.31	35.21	34.88	+0.37	-0.47	-0.84
	2022120403	32.51	32.76	32.21	32.44	32.94	33.20	34.18	34.78	34.49	-0.24	-1.06	-0.81
	2022120404	32.51	33.03	32.53	31.60	32.04	32.26	33.23	33.84	33.63	+0.08	-0.65	-0.73
	2022120406	32.63	33.20	32.70	31.96	32.55	32.67	33.47	34.20	33.96	-0.01	-0.63	-0.62
	2022120407	35.14	34.93	34.75	34.01	33.20	33.57	35.38	35.19	35.15	+0.60	+0.05	-0.55
	2022120408	33.82	34.80	34.64	33.03	33.06	33.36	34.05	35.13	34.94	+0.94	+0.48	-0.46
	2022120409	33.20	34.98	34.80	32.01	33.13	33.43	33.26	35.14	34.95	+0.66	+0.18	-0.48
	2022121701	34.50	34.85	34.55	33.40	33.21	33.61	34.77	35.15	34.94	+0.54	-0.16	-0.70
	2022121702	32.53	33.41	32.89	31.96	32.69	32.96	33.58	34.63	34.48	+0.16	-0.64	-0.80
	2022121703	33.05	33.19	32.75	32.83	33.23	33.62	34.39	34.70	34.63	-0.25	-1.09	-0.84
	2022121705	32.77	33.64	33.03	32.20	32.71	32.89	33.45	34.70	34.38	+0.37	-0.43	-0.79
	2022121706	33.54	34.06	33.64	31.93	32.81	33.12	33.99	34.53	34.39	-0.36	-1.09	-0.72
	2022121707	33.18	34.12	33.79	31.77	32.57	32.77	33.49	34.54	34.30	+0.14	-0.39	-0.53
	2022121708	33.50	34.37	34.13	32.26	32.67	33.05	33.55	34.52	34.42	+0.47	-0.15	-0.62

F THE REFERENCE'S AGREEMENT WITH ITSELF (FIGURE)

G ADDITIONAL METRICS AND EXPLORATORY ANALYSES

G.1 CONTROLS: REGISTRATION AS THE ONLY CHANGE

Table 4 gives the pairwise statistics of the R1_official control discussed in Sec. 5. Fig. 4 plots the nine interactions under the three reference constructions side by side. A shift-first variant, in which each repetition is translated before its per-slice normalisation, gives $\bar{E} = +0.04/-0.13/+0.21$ dB for NAFNet-vs-BM3D (T1w/T2w/FLAIR; Holm $p = 0.6$, 0.06 and 3×10^{-6}) and the same T2w

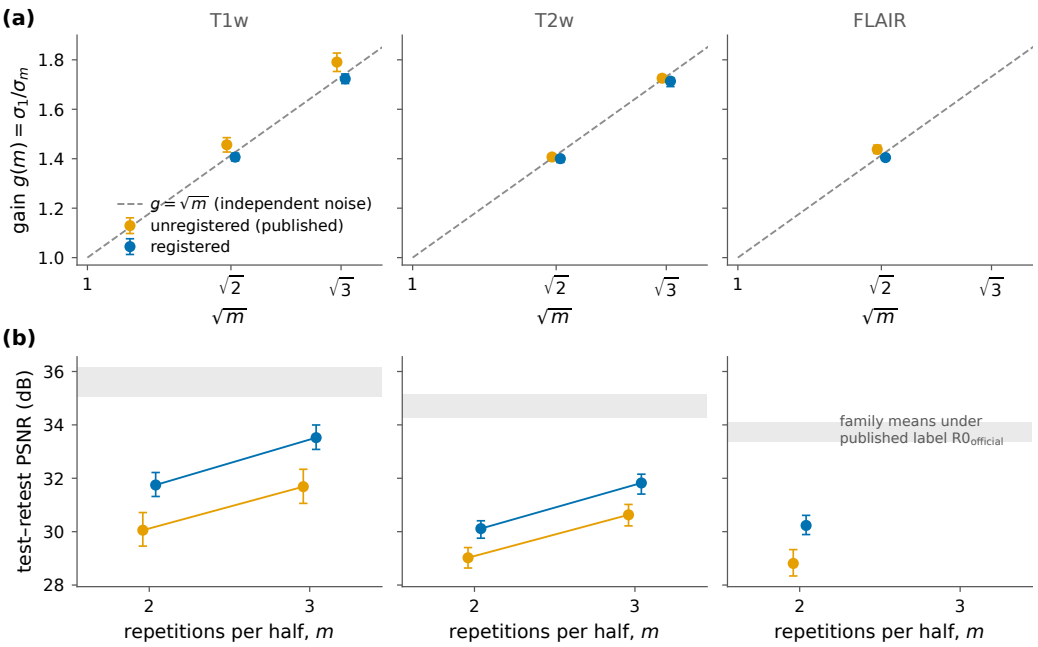

Figure 3: **The reference's agreement with itself (V1.6 test split; Sec. 4).** Columns T1w, T2w, FLAIR ($n = 25$ subject volumes each). Two disjoint m-subsets of the n repetitions are averaged and compared ($n = 6$ for T1w/T2w, $m \leq 3$; $n = 4$ for FLAIR, $m = 2$; all disjoint pairs). Orange: unregistered magnitude average as published. Blue: registered magnitude average. **(a)** Noise gain $g(m) = \sigma_1/\sigma_m$ against $\sqrt{m}$. The dashed line $g = \sqrt{m}$ is the independent-noise prediction. It holds for both constructions (an exchangeability identity, App. D). **(b)** Test-retest PSNR between the two disjoint m-averages in the official M4Raw metric convention. Both halves are per-slice min-max normalised with data range $= 1$, scored by the same pinned scorer as Sec. 5. This is how well the reference agrees with an independent copy of itself (28.8–31.7 dB unregistered, 30.1–33.5 dB registered). Grey band per column: the range of the family's mean PSNR against the published label in this paper's own scoring (T1w 35.0–36.2, T2w 34.3–35.1, FLAIR 33.4–34.1 dB). A method is compared with the full n-average, so the half-average agreement is a resolution statement about the reference, not an upper bound on method PSNR. Registration raises the agreement by 1.1–1.8 dB at the same m. Error bars: 95 % percentile bootstrap CI over the 25 subject volumes.

U-Net-vs-BM3D inversion. It changes the scale factors as well as the geometry, so it is not a pure-geometry control and is not one of the three constructions counted in Sec. 5; under it the FLAIR NAFNet-vs-BM3D interaction is positive.

Prediction and power. The predicted-displacement table was frozen before unblinding. It lists the spectral shrinkage $a_M(k)$ of each method's outputs per frequency band, the blur term $b(k)$ of the reference and the predicted pairwise E. It predicted $E \in [-0.30, +0.10]$ dB and got 7 of 9 signs right. The pilot on 10 training subjects gave $\mathrm{SD}_E = 0.09$–0.40 dB, hence a minimal detectable difference of 0.06–0.26 dB at $n = 25$. The pilot was run before the test scoring with the rep-0 BM3D input ($\mathrm{SD}_E = 0.19$–1.12 dB). The run with the corrected input, which gives the values quoted, was committed four minutes after the decision and did not inform it. Both runs are kept.

G.2 Secondary metrics

Tables 5 and 6 run the pairwise machinery of pre-registration 2.8 on SSIM and on the brain-masked PSNR of the error-map analysis (R0_official → R1, descriptive). The interaction keeps its sign for every network-vs-BM3D pair in both metrics. The inversions do not all follow. Under SSIM, BM3D moves ahead of both networks in T1w only. Inside the brain mask, NAFNet stays ahead of BM3D

Table 4: **Control: registration as the only change (descriptive, post-decision).** Same split, unit, metric, family (U-Net, NAFNet, BM3D), pairs, bootstrap (95% CI, 2000 resamples), Holm-corrected Wilcoxon on E_s and verdict symbols as Table 1. $n = 25$ subjects per contrast. Old = R0_official (the published label: each repetition per-slice normalised, then averaged); New = R1_official (the same normalised repetitions translated into the input repetition's frame with the audit shifts, then averaged). The normalisation order and every scale factor are unchanged. Only the geometry differs. This run was made after the pre-registered decision rule had been evaluated and is not an input to it. Settings as pre-registered: no. Criterion (1) held for one pair. Bold rows: the network-vs-BM3D pairs on which the main claim rests, as in Table 1.

Contrast	Pair (k, l)	$\bar{D}_{\text{old}}$	$\bar{D}_{\text{new}}$	$\bar{E}$ [95% CI]	Holm p	Verdict
T1w	U-Net vs NAFNet	-1.12	-1.21	-0.10 [-0.18, -0.03]	7.5×10^{-2}	n.s.
	U-Net vs BM3D	-0.18	-0.71	**-0.53 [-0.67, -0.41]**	1.8×10^{-7}	**sig., no flip**
	NAFNet vs BM3D	$+0.94$	$+0.50$	**-0.44 [-0.56, -0.32]**	1.8×10^{-7}	**sig., no flip**
T2w	U-Net vs NAFNet	-0.82	-0.78	$+0.04$ [$+0.01$, $+0.07$]	3.4×10^{-3}	sig., no flip
	U-Net vs BM3D	$+0.05$	-0.29	**-0.34 [-0.43, -0.27]**	1.8×10^{-7}	**flip**
	NAFNet vs BM3D	$+0.87$	$+0.49$	**-0.39 [-0.48, -0.29]**	1.8×10^{-7}	**sig., no flip**
FLAIR	U-Net vs NAFNet	-0.75	-0.90	-0.15 [-0.20, -0.10]	1.8×10^{-7}	sig., no flip
	U-Net vs BM3D	-0.38	-0.70	**-0.32 [-0.40, -0.25]**	1.8×10^{-7}	**sig., no flip**
	NAFNet vs BM3D	$+0.37$	$+0.20$	**-0.17 [-0.22, -0.13]**	8.3×10^{-7}	**sig., no flip**

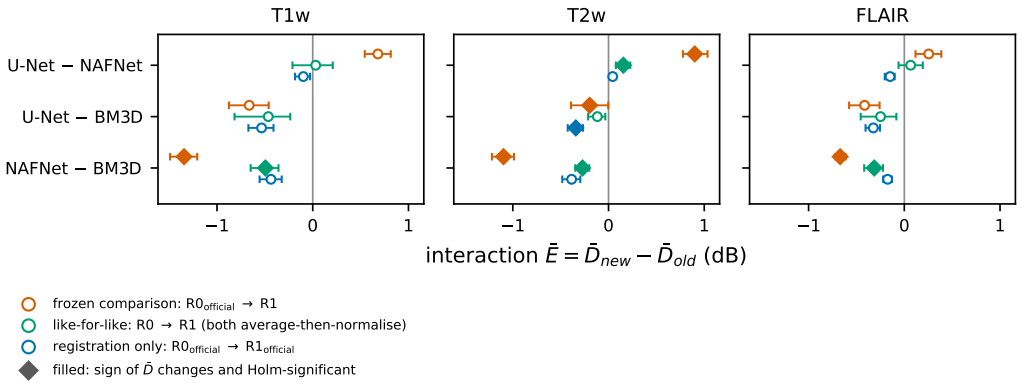

Figure 4: **The same pairwise interaction under three reference constructions** (test, $n = 25$ per contrast). Markers: $\bar{E}$ with 95 % subject-bootstrap CI. Filled diamond: sign change and Holm-significant. Vermilion: the frozen pre-registered comparison (published label vs registered, averaged, then normalised). Green: like-for-like, both average-then-normalise. Blue: registration as the only change under the published normalisation (Table 4). The interaction has the same sign for every network-vs-BM3D pair in each of the three constructions. Its size and the inversion count (five, four and one for the frozen, like-for-like and registration-only comparisons; four, zero and zero under the stricter variant) depend on the construction.

in all three contrasts ($+0.15/+0.17/+0.03$ dB), while U-Net-vs-BM3D inverts in T2w. This is why the main text claims an interaction and a loss of resolving power, not a new winner.

G.3 Spectral shrinkage and error maps

Spectral shrinkage and its anisotropy. For each output volume the slice-summed power spectrum is compared with that of the reference in radial bins, $a(f) = \sqrt{P_{\text{out}}(f)/P_{\text{ref}}(f)}$. The same ratio along the two k-space axes ($|k_\perp| < 0.05$ cycles/px bands) gives the anisotropy a_{ro}/a_{pe} (Fig. 5). Relative to R1 in the 0.15–0.35 cycles/px band, the noisy input has $a > 1$ (noise) and anisotropy 1.07–1.09. BM3D has $a \approx 0.97$–1.00 and anisotropy 1.02–1.03. NAFNet and U-Net have $a = 0.87$–0.92 and anisotropy 0.93–0.97. It is below 1 in every subject for NAFNet and in

Table 5: **Exploratory: the reference×method interaction in SSIM.** Same unit, pairs, paired Wilcoxon on E_s, Holm within contrast and subject bootstrap CI as Table 1; old = R0_official, new = R1. Descriptive, declared in pre-registration 2.8; not an input to the decision rule. *flip* = sign of $\bar{D}$ changes and Holm-significant.

Contrast	Pair (k, l)	$\bar{D}_{\text{old}}$	$\bar{D}_{\text{new}}$	$\bar{E}$ [95% CI]	Holm p	Verdict
T1w	U-Net vs NAFNet	-0.004	-0.000	$+0.004$ [$+0.002$, $+0.005$]	8.1×10^{-4}	*sig., no flip*
	U-Net vs BM3D	$+0.004$	-0.002	-0.005 [-0.008, -0.003]	1.5×10^{-4}	*flip*
	NAFNet vs BM3D	$+0.008$	-0.001	-0.009 [-0.010, -0.008]	1.8×10^{-7}	*flip*
T2w	U-Net vs NAFNet	-0.003	$+0.002$	$+0.005$ [$+0.004$, $+0.006$]	1.8×10^{-7}	*flip*
	U-Net vs BM3D	$+0.005$	$+0.004$	-0.001 [-0.002, $+0.000$]	0.027	*sig., no flip*
	NAFNet vs BM3D	$+0.008$	$+0.001$	-0.006 [-0.007, -0.006]	1.8×10^{-7}	*sig., no flip*
FLAIR	U-Net vs NAFNet	$+0.003$	$+0.005$	$+0.002$ [$+0.001$, $+0.003$]	0.006	*sig., no flip*
	U-Net vs BM3D	$+0.011$	$+0.011$	-0.001 [-0.002, $+0.001$]	0.491	*n.s.*
	NAFNet vs BM3D	$+0.009$	$+0.006$	-0.002 [-0.003, -0.002]	1.1×10^{-4}	*sig., no flip*

Table 6: **Exploratory: the reference×method interaction in brain-masked PSNR (MSE inside a brain mask, from the error-map analysis).** Same unit, pairs, paired Wilcoxon on E_s, Holm within contrast and subject bootstrap CI as Table 1; old = R0_official, new = R1. Post-decision descriptive analysis (disclosure item 4 of the pre-registration record); not an input to the decision rule. *flip* = sign of $\bar{D}$ changes and Holm-significant.

Contrast	Pair (k, l)	$\bar{D}_{\text{old}}$	$\bar{D}_{\text{new}}$	$\bar{E}$ [95% CI]	Holm p	Verdict
T1w	U-Net vs NAFNet	-1.35	-1.20	$+0.15$ [-0.01, $+0.29$] dB	0.063	*n.s.*
	U-Net vs BM3D	-0.18	-1.05	-0.87 [-1.08, -0.67] dB	1.8×10^{-7}	*sig., no flip*
	NAFNet vs BM3D	$+1.17$	$+0.15$	-1.02 [-1.17, -0.88] dB	1.8×10^{-7}	*sig., no flip*
T2w	U-Net vs NAFNet	-1.10	-0.57	$+0.52$ [$+0.40$, $+0.66$] dB	1.2×10^{-6}	*sig., no flip*
	U-Net vs BM3D	$+0.04$	-0.40	-0.44 [-0.65, -0.23] dB	5.6×10^{-4}	*flip*
	NAFNet vs BM3D	$+1.14$	$+0.17$	-0.97 [-1.09, -0.84] dB	1.8×10^{-7}	*sig., no flip*
FLAIR	U-Net vs NAFNet	-1.10	-1.37	-0.27 [-0.44, -0.11] dB	0.007	*sig., no flip*
	U-Net vs BM3D	-0.57	-1.34	-0.77 [-0.95, -0.59] dB	2.4×10^{-7}	*sig., no flip*
	NAFNet vs BM3D	$+0.53$	$+0.03$	-0.50 [-0.56, -0.46] dB	1.8×10^{-7}	*sig., no flip*

74 of 75 volumes for U-Net (paired Wilcoxon against BM3D $p = 6 \times 10^{-8}$). The extra shrinkage of the networks is therefore concentrated along the readout axis, the axis of the label drift. This is the spectral fingerprint of a blur learnt from the misaligned labels rather than of stronger isotropic denoising. The label itself predicts the size of the effect. On the training split the transfer function of the published label, the plain average of the displaced repetitions, has a band-averaged readout-to-phase-encode anisotropy of 0.969/0.966/0.978 (T1w/T2w/FLAIR; mean over the 128 groups per contrast, medians 0.969/0.967/0.975), computed from the audited shifts alone. U-Net matches it (0.96/0.96/0.97), NAFNet shrinks more (0.93/0.93/0.96) and the registered-label twins of App. G.4 lie above it (0.98–1.00). The input's ratio above 1 reflects the anisotropy of the reference's in-band signal-to-noise ratio, not a property of the input. Relative to R0_official every row shows anisotropy > 1, because the label is itself readout-blurred. The networks show the least (1.07–1.15 vs 1.13–1.21 for BM3D and 1.20–1.26 for the input). The external rows add a caveat. ImT-MRD, supervised on other data, is isotropic like BM3D (1.02/1.02/1.01). The self-supervised C2S is readout-heavy in T2w and FLAIR but not T1w (1.04/0.95/0.87). Readout-preferential shrinkage alone is therefore not a unique fingerprint of label blur. What distinguishes the M4Raw-trained family is that the pattern is uniform across contrasts and holds in every subject.

Error maps: the registered reference removes edge mismatch and adds none. A natural objection is that registration introduces its own degrees of freedom (estimator, interpolation, choice of fixed image) and could itself create edge mismatch. Our operator has no interpolation kernel, since shifts are exact Fourier translations of each coil's k-space. The fixed image is the repetition the method receives. Three estimators agree to 0.016 px (Sec. 2). Figure 6 checks the residuals directly.

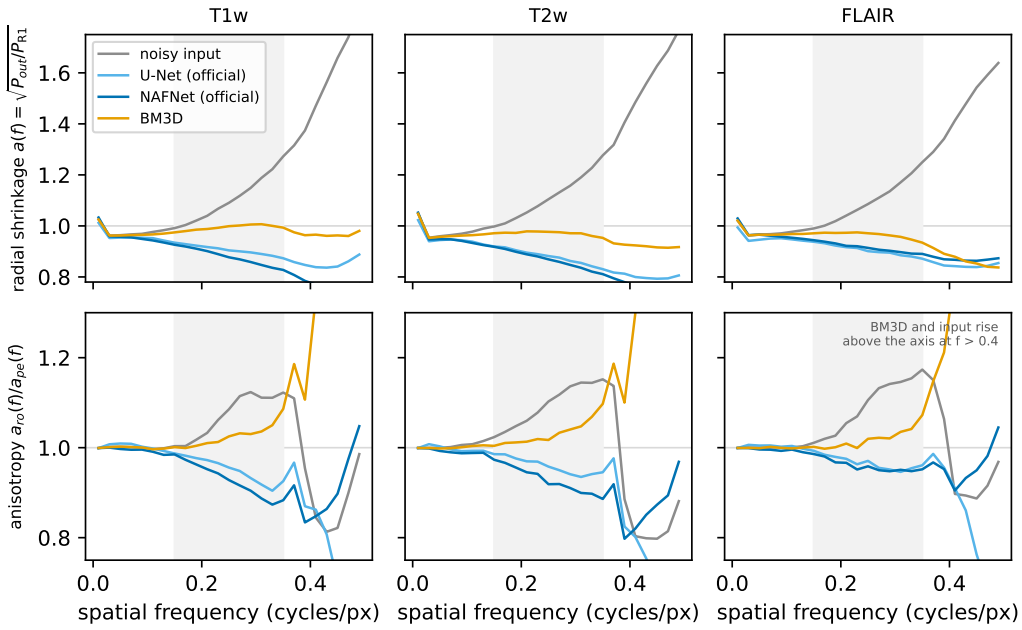

Figure 5: Spectral shrinkage of the family outputs relative to the registered reference R1 (test, median over 25 subjects per contrast). Top: radial $a(f)$. Bottom: readout/phase-encode anisotropy $a_{ro}(f)/a_{pe}(f)$. Grey band: the 0.15–0.35 cycles/px band summarised in the text. The networks (blues) shrink readout frequencies more than phase-encode frequencies. BM3D (orange) and the input (grey) do not.

For every method and for the noisy input alike, the residual against the published label carries the same apparent readout shift. It is $-0.11/-0.12/-0.16$ px per contrast, -0.13 px on average over the 75 test volumes. This is what the drift rates of Sec. 2 predict for a label that sits half a repetition's drift from the middle repetition ($-0.12/-0.12/-0.16$ px). The apparent shift is the least-squares coefficient of the residual on the readout derivative of the reference, inside a brain mask. The residual energy is concentrated on edges (edge-to-background MSE ratio 2–3). Against R1 the apparent shift is 0.00 px along both axes and the input's edge ratio falls to 1.3. What remains under R1 is an edge-concentrated residual for the two networks (1.7–2.0 vs 1.3 for the input and 1.6 for BM3D). This is the blur they learnt from the misaligned labels. Residuals in the orbits persist under both references. Eye motion between repetitions is not rigid and is part of the unexplained term of the budget.

G.4 External and retrained checkpoints

External checkpoints (exploratory; outside the pre-registered family). After the pre-registered decision, the authors of ImT-MRD (Zhu et al., 2024) provided its inference weights under a non-commercial licence. They are not redistributed. ImT-MRD was *not* trained on M4Raw labels. Their code ran unmodified under their own protocol. It takes repetition 01 as input, applies RSS separately to real and imaginary channels and scores against the unregistered mean over repetitions with stack-level metrics. It reproduces their published M4Raw figure (≈32 dB) to within 0.5 dB. Table 7 re-scores the same outputs under the official per-slice convention against R0 and R1 (Sec. 5), for the authors' input position and for the official middle repetition. Two observations follow, both descriptive. First, under the authors' protocol the unregistered reference is offset from repetition 01 by about half a pixel along the readout axis on the test split. Repetition 01 sits at one end of the drift. This offset alone pushes every score, including the noisy input, 1–3 dB below the registered value. Second, with the comparable middle-repetition input, ImT-MRD is displaced least of all network rows (every row but BM3D and the noisy input) by the change of reference ($\Delta = -0.21/-1.29/-0.15$ dB). Among the family rows it is last of the networks in T1w and T2w and second behind NAFNet in

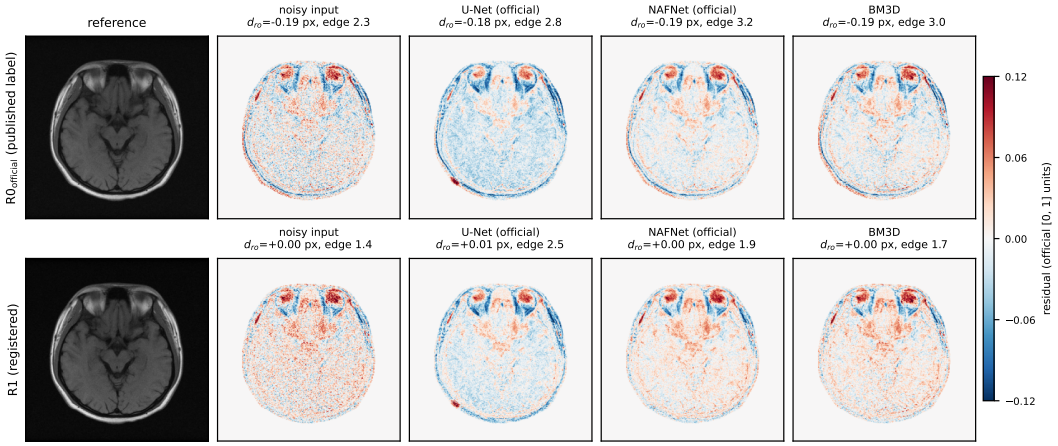

Figure 6: Residual maps of one test slice (T1w, subject 2022120408, middle-repetition input) against the published label R0_official (top) and the registered reference R1 (bottom). Same colour scale (± 0.12 in the official $[0, 1]$ units). d_{ro} is the apparent residual shift along the readout axis. "Edge" is the edge-to-background MSE ratio. The bipolar rim signature present for every column under R0 disappears under R1.

FLAIR under the published label, and ahead of every family row in every contrast under R1. This is consistent with the interpretation that the M4Raw-trained networks have learnt the drift blur of their labels. No test is claimed for these rows. The pre-registered analysis (Table 1) is unaffected. The Corruption2Self authors (Tu et al., 2025) released their per-contrast M4Raw checkpoints under an MIT licence. Their own evaluation on our test copy reproduces their reported figures to the last digit ($+2.18/+2.89/+2.58$ dB over the repetition-02 input, i.e. 32.61/32.30/33.02 dB). Their protocol feeds repetition 02, which sits off the drift centre, and scores against the published label. This costs the method 1.6/1.4/-0.1 dB relative to the middle repetition and explains most of the gap to the family's numbers. With the middle repetition, C2S is the only row displaced *upwards* by the registered average-then-normalise reference ($+1.16/+0.10/+0.54$ dB; first of all rows under R1 in T1w and T2w, and within 0.01 dB of ImT-MRD in FLAIR). Yet under R1_official it scores 35.16/34.47/33.72 dB, below every family row (ImT-MRD was not scored under R1_official). A self-supervised model trained on single normalised repetitions follows a denoised repetition's intensity scale, not the normalise-then-average label's. Its rank therefore depends on the normalisation convention even more than the family's. The two external rows are the methods in Table 7 not trained on M4Raw's averaged labels. Both are displaced least or in the opposite direction by registration.

Table 7: Exploratory: authors' ImT-MRD weights next to the pre-registered family. PSNR in dB on the test split, $n = 25$ subjects per contrast, official per-slice convention. R1 = registered magnitude reference in the input repetition's frame; Δ = R1 − R0_official. The external rows are outside the pre-registered family; symbols are explained below the table.

	T1w			T2w			FLAIR		
method	R0_off.	R1	Δ	R0_off.	R1	Δ	R0_off.	R1	Δ
U-Net (official)	35.04	33.78	-1.26	34.33	32.43	-1.90	33.37	32.49	-0.88
NAFNet (official)	36.15	34.21	-1.94	35.15	32.35	-2.80	34.12	32.98	-1.14
BM3D	35.22	34.62	-0.60	34.28	32.58	-1.70	33.75	33.28	-0.47
noisy input (middle rep.)	32.59	32.40	-0.19	31.17	30.39	-0.78	31.37	31.25	-0.12
exploratory (outside the pre-registered family; not an input to the pre-registered decision rule)									
C2S†	34.39	35.55	+1.16	33.98	34.08	+0.10	33.15	33.68	+0.54
ImT-MRD†	34.94	34.73	-0.21	34.26	32.97	-1.29	33.85	33.69	-0.15
C2S†, rep. 02	32.81	35.26	+2.45	32.62	33.95	+1.33	33.21	33.69	+0.48
ImT-MRD†, rep. 01	30.45	33.45	+3.01	31.14	32.54	+1.40	31.26	33.10	+1.83

R0_off. = R0_official, the published label. †authors' weights. Δ for the two authors'-protocol rows is in that repetition's frame.

Retrained checkpoints (exploratory): the same recipe on the published and on the registered label. We re-trained NAFNet (width 32), SCUNet and a 64-channel U-Net with the maintainers' released recipe (51 epochs, joint contrasts, one seed, best epoch selected on the validation split), twice each: once on the published label and once on a registered label that changes only the geometry. The registered label normalises each repetition per slice as the published label does, translates the other repetitions into the frame of the repetition that serves as input, and averages. Both runs of an architecture share the batch size. The test split was scored once per checkpoint, after the validation-derived predicted-displacement tables had been committed. Table 8 gives the numbers under the published label and under R1; the R1_official column and the scores of the epoch-51 checkpoints are in the released tables. Four observations follow, all descriptive. First, the re-run reproduces the released checkpoints' level under the published label (NAFNet-w32 36.20 against the official NAFNet's 36.15 dB in T1w) but sits 0.42–0.54 dB above them under R1. The released NAFNet is more sensitive to registration than a one-seed re-run of its own architecture, and its inversions against BM3D are not reproduced by the re-run in any contrast; the re-run stays 0.13–0.19 dB ahead of BM3D under R1. Second, every network-vs-BM3D interaction keeps the sign of Sec. 5 and is Holm-significant over the 36 pairs of the enlarged set: $-0.85/-0.72/-0.86$ dB for the published-label NAFNet-w32 / SCUNet / U-Net-64 in T1w, $-0.75/-0.71/-0.62$ in T2w and $-0.41/-0.35/-0.54$ in FLAIR. Third, the registered-label models reduce that interaction by a quarter to a half, ratios 0.54–0.77 ($-0.51/-0.44/-0.49$, $-0.50/-0.55/-0.41$, $-0.22/-0.23/-0.31$) and are nearly immune when only the geometry of the reference changes ($-0.06/0.00/-0.12$ dB in T1w under R1_official, two of the three not significant). They pay 0.20–0.36 dB under the published label and are within 0.11 dB of their twins under either registered reference. Each network is best on the label it was trained toward. Fourth, the anisotropic shrinkage of Sec. 5 is a property of the label. In the 0.15–0.35 cycles/px band the ratio a_{ro}/a_{pe} of the published-label re-runs is 0.93–0.97, below 1 in every test subject, as for the released networks. For the registered-label models it is 0.98–1.00. The paired difference is positive in all nine architecture-contrast cases (Wilcoxon $p = 6 \times 10^{-8}$). Training on the registered label removes the readout-axis blur that the networks learn from the drifted labels, while the PSNR under the registered references barely moves. The validation-derived prediction had anticipated the reduction of the BM3D interaction (predicted ratio 0.4–0.5, observed 0.54–0.77) but overstated the twins' gain under R1 (predicted $+0.06$ to $+0.15$ dB, observed -0.11 to $+0.09$). Restormer, re-trained on the published label only (batch 2 on a 40 GB card, four chained segments), is the strongest row under the published label in T1w and T2w (36.28/35.41/34.29 dB) and stays ahead of BM3D under R1 (34.85/32.81/33.30 dB); its interaction against BM3D is $-0.83/-0.91/-0.53$ dB, Holm-significant, and its anisotropy ratio 0.96/0.96/0.98 is the least anisotropic of the published-label rows, still below 1 in 25/25, 24/25 and 22/25 subjects. No test is claimed for any of these rows.

H SESSION TIMELINE AND REPLICATION ARMS

The per-subject timeline fits register all eight volumes of a validation subject across contrasts on gradient-magnitude images. On the 210 same-contrast pairs the gradient-magnitude shifts agree with the magnitude-image shifts of Sec. 2 within 0.039 px in the median. One monotone drift is fitted to every candidate acquisition order. The file-name order within contrast with blocks T1w $\rightarrow$ FLAIR $\rightarrow$ T2w wins in 30/30 subjects (median $R^2 = 0.989$). The header series counter reaches 0.775 and the five other block permutations 0.18–0.67. The fitted drift is -0.28 px per volume position. The GRE arm's dose-response is 4.5/16.0/22.8/31.8 % for the four displacement bins (125/37/19/2 groups). Its subject identifiers place 179 of 183 GRE acquisitions in the same acquisition campaign (June to September 2022) as the training and validation sessions. Read along its own readout axis, the stored column axis, the GRE pair shows a same-signed readout-dominant shift (-0.065 px median, negative in 131/181 pairs) at about half the predicted amplitude. The shortfall is not explained by a different campaign and remains open. The labelled motion cohorts add a phase-encode component and 2–5$\times$ the across-slice spread to the same drift.

I VALIDATION OF THE CORRECTION ON THE TEST SPLIT

The two validations of Sec. 3 were repeated on the V1.6 test split (75 groups, 6/6/4 repetitions, the split the published numbers use), with the shifts the released reference carries. Synthetic-truth

Table 8: Exploratory: architectures re-trained by us with the maintainers' recipe, once on the published label and once on the registered label, next to the pre-registered family. PSNR in dB on the test split, $n = 25$ subjects per contrast, official per-slice convention, best epoch on validation. R1 = registered magnitude reference in the input repetition's frame; Δ = R1 − R0_official. The retrained rows are outside the pre-registered family; symbols, batch sizes, best epochs and wall-clock times are given below the table.

	T1w			T2w			FLAIR		
method	R0_off.	R1	Δ	R0_off.	R1	Δ	R0_off.	R1	Δ
U-Net (official)	35.04	33.78	-1.26	34.33	32.43	-1.90	33.37	32.49	-0.88
NAFNet (official)	36.15	34.21	-1.94	35.15	32.35	-2.80	34.12	32.98	-1.14
BM3D	35.22	34.62	-0.60	34.28	32.58	-1.70	33.75	33.28	-0.47
noisy input (middle rep.)	32.59	32.40	-0.19	31.17	30.39	-0.78	31.37	31.25	-0.12
exploratory (outside the pre-registered family; not an input to the pre-registered decision rule)									
NAFNet-w32‡	36.20	34.75	-1.45	35.22	32.77	-2.45	34.34	33.46	-0.87
NAFNet-w32§	35.83	34.73	-1.10	34.93	32.73	-2.20	34.04	33.35	-0.69
SCUNet‡	36.24	34.93	-1.31	35.24	32.83	-2.41	34.25	33.43	-0.82
SCUNet§	35.93	34.89	-1.04	35.03	32.78	-2.25	34.04	33.35	-0.70
U-Net-64‡	35.81	34.36	-1.46	34.93	32.61	-2.32	34.13	33.13	-1.00
U-Net-64§	35.53	34.45	-1.08	34.66	32.55	-2.11	33.86	33.08	-0.78
Restormer‡	36.28	34.85	-1.43	35.41	32.81	-2.61	34.29	33.30	-1.00

R0_off. = R0_official, the published label. ‡retrained by us on the published label. §retrained by us on the registered label. Batch size / best epoch / wall-clock on one A100-40G. Published label: NAFNet-w32 24 / 27 / 2.5 h; SCUNet 16 / 35 / 3.6 h; U-Net-64 24 / 41 / 1.1 h; Restormer 2 / 32 / 37.8 h in four chained segments. Registered-label twins: NAFNet-w32 24 / 30 / 2.6 h; SCUNet 16 / 43 / 3.7 h; U-Net-64 24 / 45 / 1.2 h.

recovery: shift RMSE 0.011/0.012/0.015 px (T1w/T2w/FLAIR), the measured σ_1 drop equal to the true drop within 0.02 %, and the RMSE of R2 to the known truth 0.20/0.28/0.29× that of R0. Leave-one-repetition-out prediction, now possible for all three contrasts (400 held-out repetitions, 25 subjects): the registered average of the other repetitions predicts the held-out one better than the unregistered average by $+1.98$ dB (T1w; CI 1.56–2.44; 100 % of repetitions), $+1.09$ dB (T2w; 0.90–1.30; 99.3 %) and $+1.52$ dB (FLAIR; 1.20–1.88; 100 %). The gains are larger than on the validation split because the drift accumulates over six repetitions and the reference is better averaged.

J Compute

Everything except the exploratory re-training runs on a workstation. The audit and corrected references take CPU minutes (validation split 3 min, test split 5 min on 24 cores; 42–47 s per split on a 32-core cluster node). Resolution measurement over all disjoint pairs takes about 8 min. The synthetic-truth, leave-one-out and estimator-robustness validations take CPU minutes on 16 cores (test split 6 min) and a few hours on one. Re-scoring the family takes about 5–10 hours on one RTX 4080 workstation, most of it BM3D on the CPU: official U-Net and NAFNet inference plus BM3D on 75 test and 90 validation volumes. The pre-registered test, its controls, the error maps and the spectral analysis take CPU minutes. External checkpoints (ImT-MRD, C2S) take a few GPU minutes. The pre-registered analysis trained no model. The exploratory re-training of App. G.4 used about 47 GPU-hours on A100 cards for the four architectures on the published label and about 7.5 GPU-hours for the three on the registered label.

K AI use statement

In this work we used a generative AI coding assistant (Claude Code, Anthropic). It wrote and refactored analysis code, unit tests and the figure and table scripts. It polished the pre-registration text and this manuscript from the authors' criteria, result tables and notes. It performed literature triage, locating and summarising candidate related work, which the authors then read. After the pre-registered decision it proposed and, under the authors' direction, executed the descriptive control analyses (the registration-only reference, the like-for-like pairing, the error maps and the spectral analysis) and an internal readiness review. It also proved the derivation of the $\sqrt{m}$ identity in App. D, which the authors checked. It generated no data. The method family, the frozen decision rule and every claim and experiment that entered the paper were decided by the authors. The remaining disclosure categories of the conference policy are not applicable. We have reviewed all AI-assisted

work. Every quantitative statement in the paper is traced to a file in the checksummed release. The manuscript text was read and edited by the authors. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.