# Cross-Scanner Reliability of Brain MRI Foundation Model Embeddings: A Travelling-Heads Study

Rafael Navarro-González[a,1], Santiago Aja-Fernández[a,b], Álvaro Planchuelo-Gómez[a,b], Rodrigo de Luis-García[a,b]

[a] *Laboratorio de Procesado de Imagen (LPI), Universidad de Valladolid, ETSI Telecomunicación, Campus Miguel Delibes s/n, 47011, Valladolid, Spain*
[b] *LPI-BIVa, Valladolid Health Research Institute (IBioVALL), Edificio Rondilla, C. Rondilla de Sta. Teresa s/n, 47010, Valladolid, Spain*

**Abstract**

Foundation models (FMs) for brain magnetic resonance imaging (MRI) are increasingly adopted as pretrained backbones for clinical tasks such as brain age prediction, disease classification, and anomaly detection. However, if FM embeddings (internal representations) shift systematically across MRI scanners, downstream analyses built on them may reflect acquisition hardware rather than biology. No study has yet quantified this cross-scanner reproducibility. Here, we assess the cross-scanner reliability of brain MRI FM embeddings and investigate which design factors (pretraining strategy, network architecture, embedding dimensionality, and pretraining dataset scale) best explain the observed differences. Using the ON-Harmony travelling-heads dataset (20 participants, eight scanners, three vendors), we evaluate the embeddings of five architecturally diverse FMs and a FreeSurfer morphometric baseline via within- and between-scanner intraclass correlation coefficient (ICC), variance decomposition, and scanner fingerprinting. Reliability spanned the full spectrum: biology-guided models achieved good-to-excellent cross-scanner ICC (AnatCL: 0.97 [95% confidence interval (CI): 0.94, 0.98]; y-Aware: 0.81 [0.63, 0.88]), matching or surpassing FreeSurfer (0.93 [0.83, 0.96]), whereas purely self-supervised models fell below the poor threshold (BrainIAC: 0.45, BrainSegFounder: 0.31, 3D-Neuro-SimCLR: 0.25), with 23–58% of embedding variance attributable to scanner identity. The strongest correlate of cross-scanner reliability among the

[1] Corresponding author: rafael.navarro@uva.es (Rafael Navarro-González)

models evaluated was pretraining strategy: incorporating biological metadata (cortical morphometrics, age) into the contrastive objective produced scanner-robust embeddings, whereas architecture, dimensionality, and dataset scale did not predict reliability.

*Keywords:* foundation models, brain MRI, reliability, travelling heads

## 1. Introduction

Foundation models (FMs) are neural networks pretrained on large, diverse datasets so that their learned representations can be reused across downstream tasks without retraining from scratch. FMs are being increasingly adopted in brain MRI for clinical and research applications, including brain age prediction, disease classification, and anomaly detection (Tak et al., 2026; Kaczmarek et al., 2025; Barbano et al., 2026; Dufumier et al., 2021; Cox et al., 2024). As these models are applied to data acquired across heterogeneous scanners and sites, it becomes essential to assess whether their learned representations are reproducible across acquisition conditions (Grote et al., 2024). A concrete manifestation of this issue is that deep features can carry strong acquisition signatures, including scanner or vendor-related information (Kushol et al., 2023). In structural brain MRI, if scanning the same brain on two different MRI systems yields substantially different FM embeddings (internal representations), then downstream analyses built on those representations may be scanner-dependent rather than biology-driven. Crucially, downstream adaptation may not eliminate this nuisance signal: biases encoded in pretrained features can propagate to downstream predictions (Glocker et al., 2023b), and scanner-sensitive representations may persist unless the adaptation objective explicitly targets scanner invariance (Dinsdale et al., 2021). Despite this growing adoption, the cross-scanner reproducibility of brain MRI FM embeddings has not, to our knowledge, been systematically benchmarked.

We treat this reproducibility question as a reliability problem and quantify it with the intraclass correlation coefficient (ICC). This can be interpreted as the

proportion of total variance attributable to stable between-subject differences rather than measurement error (Shrout and Fleiss, 1979). Across neuroimaging, ICC varies widely: the test-retest reliability of functional connectivity measures has been reported as poor (mean ICC $\approx 0.29$; (Noble et al., 2019)), whereas structural measures such as cortical thickness have shown good-to-excellent reliability depending on the region and protocol (Fortin et al., 2018), following standard interpretive thresholds (Koo and Li, 2016). Deep-learning brain-age estimates have reached the upper end of this range (e.g., ICC = 0.90–0.99 (Cole et al., 2017); ICC $\approx$ 0.94–0.98 for several public packages (Dörfel et al., 2023)), but these are scalar outputs. This motivates a natural question for FMs: do their high-dimensional embedding vectors achieve comparable between-scanner reliability, or do they shift systematically with the scanner?

Recent work in adjacent domains offers partial glimpses. In digital pathology, Thiringer et al. (Thiringer et al., 2026) showed that FM embeddings can carry pronounced scanner-specific signatures across whole-slide scanners, and de Jong et al. (de Jong et al., 2025) found that multiple pathology FM embedding spaces are more strongly organised by medical center than by biological factors. Broader clinical benchmarks of pathology FMs across institutions (Campanella et al., 2025; Neidlinger et al., 2025) have reported substantial cross-cohort performance variability, though the contribution of scanner effects to these differences has not been isolated. Carloni et al. (Carloni et al., 2025) further proposed a scanner-aware contrastive loss (ScanGen) to mitigate this sensitivity. In radiology, Pai et al. (Pai et al., 2024) and Aerts et al. (Aerts et al., 2025) reported high embedding similarity in same-scanner test–retest CT settings, and in structural brain MRI, BrainIAC (Tak et al., 2026) assessed embedding stability under synthetic perturbations (contrast shifts, Gibbs ringing, bias fields). Critically, however, none of these studies employed a travelling-heads design, in which the same individuals are scanned on multiple MRI systems. Such a design is essential for disentangling biological from technical variability and computing per-subject cross-scanner reliability estimates such as ICC, leaving the cross-scanner reliability of brain MRI FM embeddings largely

uncharacterised.

The travelling-heads paradigm, in which the same individuals are scanned on multiple scanners, provides a well-suited design for disentangling biological from technical variability (Voelker et al., 2021; COVID-CNS Consortium, 2022). Here, we leverage the Oxford-Nottingham Harmonisation (ON-Harmony) resource (Warrington et al., 2025, 2023), one of the largest publicly available travelling-heads datasets for 3T brain MRI (20 participants, eight scanners, three vendors), to present a systematic reliability evaluation of FM embeddings for structural brain MRI. We evaluate five architecturally diverse FMs alongside a FreeSurfer morphometric baseline, computing within- and between-scanner ICC, decomposing embedding variance into biological, scanner, and residual components, and testing whether embeddings carry scanner-identifiable signatures. The models span convolutional neural network (CNN) and Transformer architectures with pretraining objectives ranging from pure self-supervision to metadata-guided contrastive learning. Our goal is to determine whether current brain MRI FM embeddings are reproducible across scanners and to identify which design factors best explain the observed reliability differences.

## 2. Methods

### *2.1. Dataset*

We used the Oxford-Nottingham Harmonisation (ON-Harmony) dataset (OpenNeuro accession ds004712, version 2.0.1; (Warrington et al., 2025)), a multi-site, multi-modal travelling-heads resource collected between 2018 and 2024. The dataset comprises multiple MRI modalities. For this study we used the T1-weighted (T1w) structural scans from 20 healthy participants (13 M / 7 F; age 19–50, mean $31.5 \pm 10.5$ years; Table 1), each scanned on six different 3T MRI scanners (from a pool of eight) across five imaging sites in Oxford and Nottingham, UK. All acquisition protocols were aligned to the UK Biobank imaging protocol (Warrington et al., 2023).

The eight scanners span all three major MRI vendors (Table 1): three Siemens systems (Trio, Prisma with 32-channel coil, Prisma with 64-channel coil), two Philips systems (Achieva, Ingenia), and three GE systems (MR750 Discovery, two Premier systems with different coil configurations). The dataset contains 165 T1w scanning sessions. Each of the 20 participants was scanned once on each of six scanners (120 between-scanner sessions). Additionally, nine participants underwent five extra repeat sessions on a single scanner for within-scanner test-retest assessment, yielding 45 additional sessions (Table 1). While the cohort of 20 participants may appear modest in comparison to typical case-control studies, the statistical power of a travelling-heads design derives from its fully crossed structure rather than from sample size alone: 20 subjects $\times$ 6 scanners each yields 120 unique subject–scanner observations, each contributing to the estimation of variance components. This design is consistent with prior travelling-heads studies that have established reliability benchmarks for conventional neuroimaging measures with comparable or smaller samples (Voelker et al., 2021; COVID-CNS Consortium, 2022; Warrington et al., 2023).

### *2.2. Foundation Models*

In neuroimaging, FMs are typically trained with self-supervised learning (SSL), which derives a training signal from the data itself rather than from manual annotations. A common SSL strategy is contrastive learning, in which the model learns to map augmented views of the same image (positive pairs) to nearby points in embedding space while pushing apart representations of different images (negative pairs). SimCLR (Chen et al., 2020) is a widely adopted contrastive framework that applies random augmentations (e.g., cropping, intensity jittering) to generate positive pairs. In Transformer-based architectures, the input volume is divided into patches and processed alongside a special classification (CLS) token. The final representation of this CLS token serves as the global embedding for the entire image. Convolutional architectures instead obtain a global embedding by averaging feature maps across spatial dimensions (global average pooling).

| Vendor | Model | Site | Coil | $N_{\text{sess}}$ | $N_{\text{subj}}$ | $N_{\text{rep}}$ | Within-scanner repeats |
|---|---|---|---|---|---|---|---|
| Siemens | Prisma$^{\dagger}$ | OHBA, Oxford | 32-ch | 30 | 20 | 2 | 2 subj $\times$ 6 sess |
| Siemens | Prisma$^{\dagger}$ | OCMR, Oxford | 64-ch | 30 | 20 | 2 | 2 subj $\times$ 6 sess |
| Siemens | Trio | WIN-FMRIB, Oxford | — | 15 | 10 | 1 | 1 subj $\times$ 6 sess |
| Philips | Achieva | SPMIC, Nottingham | — | 25 | 20 | 1 | 1 subj $\times$ 6 sess |
| Philips | Ingenia | SPMIC, Nottingham | — | 25 | 20 | 1 | 1 subj $\times$ 6 sess |
| GE | MR750 | SPMIC, Nottingham | — | 10 | 10 | 0 | — |
| GE | Premier | SPMIC, Nottingham | 48-ch | 15 | 10 | 1 | 1 subj $\times$ 6 sess |
| GE | Premier | Oxford | 21-ch | 15 | 10 | 1 | 1 subj $\times$ 6 sess |
| **Total** | | | | **165** | **20** | **9** | |

Table 1: ON-Harmony dataset overview. All scanners operate at 3T with UK Biobank-aligned acquisition protocols. $N_{\text{sess}}$ = total sessions; $N_{\text{subj}}$ = unique subjects scanned; $N_{\text{rep}}$ = subjects with 6 within-scanner repeat sessions. $^{\dagger}$Only these two scanners have $\geq$2 repeat subjects, the minimum required for within-scanner ICC(3,1) estimation; the remaining 5 scanners with repeats have only 1 repeat subject each and cannot contribute to ICC. The cohort comprises 20 healthy adults (13 M / 7 F; age 19–50, mean 31.5 $\pm$ 10.5 years) scanned across two phases (A: 10 subjects, 2018–2021; B: expanded to 20, 2023–2024), yielding 165 T1w sessions in total.

The pretraining strategies of current brain MRI FMs range along a supervision spectrum. At one end, purely self-supervised models rely solely on data augmentations to define positive pairs, with no external metadata guiding what the model should learn to be invariant to. At the other end, biology-guided (or metadata-supervised) models incorporate external biological information, such as subject age or cortical morphometric measures, into the contrastive objective, explicitly steering the learned representations toward biologically meaningful variation.

We evaluated five brain MRI FMs spanning diverse architectures and pretraining strategies, plus a conventional morphometric baseline (Table 2). All FM encoders were used in frozen inference mode (no fine-tuning), so that our reliability estimates reflect the properties of the pretrained embedding space itself rather than any task-specific adaptation. The models used are:

1. **BrainIAC (Tak et al., 2026):** is a 3D Vision Transformer (Dosovitskiy

et al., 2021) (ViT-B) encoder pretrained with SimCLR (Chen et al., 2020) on 32,015 brain MRIs (T1w, T2-weighted (T2w), fluid-attenuated inversion recovery (FLAIR), and T1 contrast-enhanced (T1CE)) from 16 datasets spanning 10 neurological conditions (from a curated pool of 48,965 scans across 34 datasets). It produces 768-dimensional embeddings via the CLS token and represents the largest and most modality-diverse publicly available brain MRI FM.

2. **3D-Neuro-SimCLR (Kaczmarek et al., 2025):** is a 3D ResNet-18 (He et al., 2016) encoder pretrained with SimCLR on 44,958 T1w scans from 18,759 patients across 11 publicly available datasets spanning neurological conditions. It produces 512-dimensional embeddings and serves as an independent replication of the SimCLR paradigm with a smaller convolutional architecture.
3. **y-Aware (Dufumier et al., 2021):** is a 3D DenseNet-121 (Huang et al., 2017) encoder pretrained with the y-Aware InfoNCE loss, which uses chronological age as a continuous proxy to modulate positive and negative pair sampling. Pretrained on the BHB-10K aggregation (10,420 T1w sessions from 7,764 subjects across 13 public datasets and 74 acquisition sites), it produces 1,024-dimensional embeddings and represents a metadata-supervised approach where a single biological variable (age) guides contrastive learning. Although originally proposed as a self-supervised pretraining strategy rather than an FM per se, we include it here because its frozen encoder is used in the same feature-extraction paradigm as the other models.
4. **AnatCL (Barbano et al., 2026):** is a 3D ResNet-18 encoder pretrained with a joint loss combining an anatomy-aware contrastive term, which incorporates cortical morphometric metadata (thickness, gray matter volume, surface area from the Desikan-Killiany parcellation) into the sampling strategy, and the y-Aware age-based contrastive term (Dufumier et al., 2021). Pretrained on 3,984 T1w scans from the OpenBHB collection (10 public cohorts of healthy controls), it produces 512-dimensional

embeddings and represents a weakly supervised approach where both anatomical structure and age explicitly guide representation learning. AnatCL is released as five cross-validation fold checkpoints. We evaluate all five independently and report mean ± std across folds to account for checkpoint variability. Per-dimension analyses (e.g., ICC distributions, PCA structure) use fold 0, as embedding dimensions are not aligned across independently trained folds.

5. **BrainSegFounder (Cox et al., 2024):** is a 3D Swin Transformer (Liu et al., 2021) (SwinUNETR) pretrained with a two-stage strategy combining reconstruction, rotation prediction, and contrastive objectives on UK Biobank data. It produces 768-dimensional embeddings and, together with BrainIAC (ViT-B), provides a Transformer comparison against the convolutional architectures. The publicly available Stage 1 (self-supervised) checkpoint accepts single-channel T1w input, so no channel duplication is needed. Embeddings are obtained by global average pooling over the deepest encoder bottleneck features.
6. **FreeSurfer morphometrics (Fischl, 2012):** serve as a conventional baseline. We used 199 pre-extracted morphometric features from the ON-Harmony resource: cortical thickness, surface area, and gray matter volume for 31 regions per hemisphere from the Desikan-Killiany-Tourville (DKT) atlas (186 features), plus 13 subcortical volumes from the automatic segmentation (4 thalamus features excluded due to incomplete data; Supplementary Table S1). These morphometrics serve as both a reliability anchor and a conventional baseline : cortical thickness and subcortical volumes have well-established test-retest ICCs (typically 0.75–0.98) and the ON-Harmony dataset itself has been benchmarked on these features (Warrington et al., 2023, 2025), so reproducing known reliability values validates our analysis framework while placing FM embeddings in the context of the standard currently used in multi-site studies. Additionally, we applied ComBat harmonization (Fortin et al., 2018) to these features (scanner as batch variable, preserving age and

sex) to establish a lower bound for scanner-related variance against which FM embeddings can be compared (Supplementary Table S5).

These five models were selected to span two key design axes: architecture (CNNs: AnatCL and 3D-Neuro-SimCLR on ResNet-18, y-Aware on DenseNet-121; versus Transformers: BrainIAC on ViT-B, BrainSegFounder on Swin-T) and pretraining supervision (pure self-supervised learning in BrainIAC, 3D-Neuro-SimCLR, and BrainSegFounder versus biology-guided objectives in AnatCL and y-Aware). The set also covers a wide range of pretraining data scales, from 3,984 scans (AnatCL) to ∼32,000 (BrainIAC). FreeSurfer morphometrics serve as a conventional baseline with well-established test-retest reliability, anchoring the FM results against a known reference.

| Model | Arch. | Pretraining | Train mod. | Dim | $N_{\mathrm{pt}}$ |
|---|---|---|---|---|---|
| BrainIAC | ViT-B | SimCLR | T1w, T2w, FLAIR, T1CE | 768 | 32,015 |
| 3D-Neuro-SimCLR | ResNet-18 | SimCLR | T1w | 512 | 44,958 |
| AnatCL | ResNet-18 | Anat.-aware CL | T1w | 512 | 3,984 |
| y-Aware | DenseNet-121 | y-Aware InfoNCE | T1w | 1,024 | 10,420 |
| BrainSegFounder* | Swin-T | Recon+rot+CL | T1w, T2-FLAIR | 768 | UKB |
| FreeSurfer (DKT) | — | — | T1w | 199 | — |

Table 2: Foundation models and baseline evaluated in this study. All FM encoders are used frozen (no fine-tuning). Dim = embedding dimensionality; $N_{\mathrm{pt}}$ = number of pretraining scans (UKB = UK Biobank subset); Train mod. = MRI modalities used during pretraining. *BrainSegFounder Stage 1 was pretrained on paired T1w + T2-FLAIR from UK Biobank; the publicly released Stage 1 checkpoint accepts single-channel T1w input.

### *2.3. Preprocessing*

Because each FM was developed with a specific preprocessing protocol, we used model-specific pipelines to ensure faithful reproduction of each model's intended input distribution.

- **Pipeline A, CAT12 voxel-based morphometry (AnatCL, y-Aware):** Bias correction, nonlinear registration to the MNI152 template (CAT12 Geodesic Shooting), skull stripping, and gray matter segmentation

via SPM12/CAT12. AnatCL input: $121 \times 145 \times 121$ modulated normalized gray matter maps (native CAT12 MNI resolution). y-Aware input: $121 \times 145 \times 121$, z-score normalized, saved as NumPy arrays.

- **Pipeline B, TurboPrep (3D-Neuro-SimCLR):** N4 bias field correction (ANTs), skull stripping (SynthStrip (Hoopes et al., 2022)), rigid registration to the skull-stripped MNI152 ICBM 2009c nonlinear symmetric template (Fonov et al., 2011) (ANTs (Avants et al., 2009) `antsRegistrationSyN.sh` with rigid transform type), and WhiteStripe (Shinohara et al., 2014) intensity normalization. At inference time, volumes are transposed and center-cropped to $150 \times 192 \times 192$, then per-sample masked z-score normalized, following the original training data loader (Kaczmarek et al., 2025). Negative voxel intensities present in 15 sessions from Oxford Siemens scanners were clamped to zero prior to registration.

- **Pipeline C, Custom (BrainIAC, BrainSegFounder):** BrainIAC: N4 bias correction, rigid registration to the MNI-space template shipped with the BrainIAC repository (`temp_head.nii.gz`, resampled to 1 mm isotropic), HD-BET (Isensee et al., 2019) skull stripping, resampled to $96 \times 96 \times 96$ voxels, z-score normalized. BrainSegFounder: skull stripping (FSL BET), affine and nonlinear registration to the MNI152 T1 1 mm template (FSL FLIRT + FNIRT (Jenkinson et al., 2012)), MNI brain mask applied to the FNIRT output, right-anterior-superior (RAS) reorientation, percentile-based intensity scaling to $[0, 1]$ (0.5th–99.5th percentile), cropped/padded to $96 \times 96 \times 96$ voxels. The original BrainSegFounder code defaults to a fixed intensity range of $[-1000, 1000]$ inherited from the SwinUNETR codebase, which was designed for CT Hounsfield units. We replaced this with percentile-based scaling, since UK Biobank volumes (on which BrainSegFounder was pretrained) are already intensity-normalized, but externally preprocessed MRI data can exhibit scanner-dependent intensity ranges spanning orders of magnitude.

For the morphometric baseline, we used FreeSurfer imaging-derived phenotypes from the ON-Harmony resource (Warrington et al., 2025): cortical thickness, surface area, and gray matter volume for 31 DKT atlas regions per hemisphere (186 features), plus 17 subcortical volumes from the automatic segmentation (aseg; full feature list in Supplementary Table S1). Four thalamus-related aseg features were excluded due to incomplete data across processing phases (∼50% missing; see Supplementary Table S1), yielding 199 features for all analyses. These features were z-score normalized across the dataset.

### *2.4. Reliability Metrics*

Reliability was quantified with the ICC (Shrout and Fleiss, 1979), using the interpretive thresholds of Koo and Li (2016): poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), and excellent (>0.90). We report (i) between-scanner reliability and (ii) within-scanner test–retest reproducibility.

**Between-scanner reliability (ICC(2,1)):** Between-scanner reliability was assessed with ICC(2,1) (two-way random effects, absolute agreement), treating both subjects and scanners as random effects (McGraw and Wong, 1996; Koo and Li, 2016). For each subject–scanner pair, if multiple sessions were available on the same scanner, embedding vectors were first averaged across sessions to yield a single value per subject–scanner observation before computing ICC. ICC(2,1) was then computed independently for each embedding dimension from the resulting subject $\times$ scanner matrix.

**Within-scanner test–retest reproducibility (ICC(3,1)):** Within-scanner reproducibility was assessed with ICC(3,1) (two-way mixed effects, consistency) (Shrout and Fleiss, 1979; Koo and Li, 2016). ICC(3,1) was computed separately for each scanner with $\geq 2$ repeat subjects, using the subject $\times$ session matrix formed by the six repeat scans (i.e., without averaging across repeats). Scanner-specific ICC estimates were pooled as a weighted average across scanners (weights proportional to the number of repeat subjects per scanner).

ICC was computed independently for each embedding dimension. We report

the median ICC across dimensions and the proportion of dimensions above the good and excellent thresholds. Dimensions with degenerate variance (e.g., near-zero activation across all sessions) can yield undefined ICC values. These were excluded from summary statistics.

Uncertainty was quantified with bootstrap 95% confidence intervals. For between-scanner ICC, we performed 1,000 subject-level resamples (resampling subjects with replacement and recomputing per-dimension ICCs); the reported interval bounds are the medians of the per-dimension 2.5th and 97.5th percentiles across bootstrap samples. For within-scanner ICC, resampling was restricted to the repeat-subject set, and confidence intervals are therefore wider due to the limited number of repeat subjects. Between-model differences in median ICC were assessed by bootstrapping per-dimension ICC values (5,000 resamples) and computing the proportion of resamples in which the difference in medians crossed zero (two-sided).

### *2.5. Variance Decomposition*

To partition embedding variance into biological and technical components, we estimated variance components for each embedding dimension independently via ANOVA-based method-of-moments decomposition, assuming the following two-way random-effects model:

$$y_{ijk} = \mu + u_i + v_j + \epsilon_{ijk} \tag{1}$$

where $y_{ijk}$ is the value of a given embedding dimension for subject $i$, scanner $j$, and session $k$; $\mu$ is the grand mean; $u_i \sim \mathcal{N}(0, \sigma^2_{\text{subject}})$ is the random effect of subject; $v_j \sim \mathcal{N}(0, \sigma^2_{\text{scanner}})$ is the random effect of scanner; and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_{\text{residual}})$ is the residual term capturing within-scanner variability and other unexplained sources. From the estimated variance components, we computed the proportion of total variance attributable to each source:

$$p_{\text{source}} = \frac{\sigma^2_{\text{source}}}{\sigma^2_{\text{subject}} + \sigma^2_{\text{scanner}} + \sigma^2_{\text{residual}}}, \quad \text{source} \in \{\text{subject, scanner, residual}\} \tag{2}$$

We report the mean proportion across all embedding dimensions for each FM. A high $p_{\text{subject}}$ indicates that the embedding captures predominantly biological variability (desirable); a high $p_{\text{scanner}}$ indicates scanner contamination.

### *2.6. Scanner Fingerprinting and Subject Identification*

To assess whether FM embeddings carry identifiable scanner signatures, inspired by the connectome fingerprinting paradigm (Finn et al., 2015), we trained a multiclass linear support vector machine (SVM) to predict scanner identity (eight classes) from embedding vectors using leave-one-session-out cross-validation.

Conversely, to assess whether subject identity is preserved across scanner changes, we performed a subject identification analysis: for each between-scanner session, we identified the nearest neighbor (by cosine similarity) among all sessions from other scanners and recorded whether the correct subject was retrieved. High subject identification accuracy indicates that biological signal dominates over scanner noise in the embedding space, even when the scanner changes.

We additionally performed two supplementary analyses. First, cross-scanner sex classification using leave-one-scanner-out logistic regression on frozen embeddings tested whether embeddings retain biologically relevant information across scanners (note that the same subjects appear in train and test folds, so this measures cross-scanner stability of sex-related signal rather than generalization to unseen individuals). Second, principal component analysis (PCA) of each model's embedding space characterised intrinsic dimensionality and whether leading components are dominated by subject or scanner identity (assessed via $\eta^2$). Full details and results are reported in the Supplementary Material.

All analyses were implemented in Python (NumPy, SciPy, scikit-learn). Code and analysis scripts are publicly available (see Code availability).

## 3. Results

Table 3 summarises the reliability metrics for all five FMs and the FreeSurfer baseline. The models span the full spectrum from excellent to near-zero cross-scanner reliability.

| | FreeSurfer | AnatCL | y-Aware | BrainIAC | BrainSeg.* | 3D-Sim.† |
|---|---|---|---|---|---|---|
| **Intraclass correlation** | | | | | | |
| ICC(2,1) between | **0.93** | **0.97** | 0.81 | 0.45 | 0.31 | 0.25 |
| 95% CI | [0.83, 0.96] | [0.94, 0.98] | [0.63, 0.88] | [0.22, 0.63] | [0.16, 0.42] | [0.09, 0.39] |
| % good (≥0.75) | 86.4 | 100.0 | 70.2 | 0.4 | 0.0 | 0.4 |
| % poor (<0.50) | 2.0 | 0.0 | 3.5 | 64.1 | 93.5 | 91.0 |
| ICC(3,1) within | 0.90 | 0.99 | 0.71 | 0.81 | 0.53 | 0.52 |
| 95% CI | [−0.20, 0.99] | [−0.20, 1.00] | [−0.20, 1.00] | [−0.20, 1.00] | [−0.20, 1.00] | [−0.20, 1.00] |
| **Variance decomposition (%)** | | | | | | |
| Subject | **84.3** | **88.1** | 77.1 | 51.1 | 35.2 | 27.0 |
| Scanner | 11.8 | 11.8 | 7.9 | 23.1 | **57.9** | 32.0 |
| Residual | 3.9 | 0.1 | 15.0 | 25.8 | 6.8 | 40.9 |
| **Fingerprinting & identification** | | | | | | |
| Scanner fingerprint | 80.0% | 45.5% | 57.6% | 33.3% | **90.9%** | 89.1% |
| Subject ID | **100%** | **100%** | **100%** | 49.7% | 86.1% | 62.4% |

Table 3: Reliability and reproducibility metrics across models. FreeSurfer morphometrics serve as the reference baseline; FMs are ordered by decreasing between-scanner ICC. ICC values are medians across embedding dimensions; % good and % poor refer to the proportion of dimensions reaching the indicated threshold. 95% CIs are bootstrap estimates (1,000 subject-level resamples). Variance proportions are means across dimensions. Scanner fingerprint = leave-one-out SVM accuracy; Subject ID = cross-scanner nearest-neighbour (cosine similarity) accuracy. Within-scanner CIs are wide because only two scanners have ≥2 repeat subjects, limiting the precision of ICC(3,1) estimates. AnatCL values are means across 5 cross-validation folds. †3D-Neuro-SimCLR; *BrainSegFounder.

### *3.1. Between-scanner reliability*

Between-scanner ICC(2,1) revealed a clear three-tier structure (Figure 1a). AnatCL achieved the highest reliability among all models, including the FreeSurfer baseline, with a median ICC(2,1) of 0.97 (95% CI [0.94, 0.98]) and 100% of dimensions reaching good reliability (≥0.75). FreeSurfer morphometrics

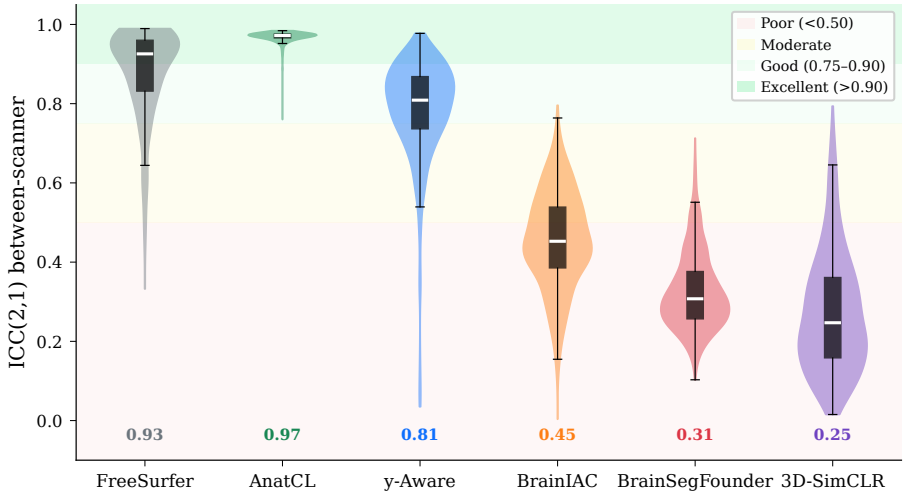


Figure 1: Between-scanner reliability of FM embeddings. Distribution of per-dimension ICC(2,1) values for each model (violin plots with inner box plots). Color bands mark standard ICC thresholds. Median ICC values are annotated below each violin. AnatCL shows tight clustering above 0.90, while 3D-Neuro-SimCLR and BrainSegFounder cluster in the poor range.

confirmed known reliability benchmarks with ICC(2,1) = 0.93 [0.83, 0.96], with 86.4% of features achieving good reliability. y-Aware achieved good reliability (ICC 0.81 [0.63, 0.88]) with 70.2% of dimensions above the good threshold.

Bootstrap paired difference tests confirmed that all adjacent model rankings are statistically significant ($p < 0.001$ for all pairwise comparisons; Supplementary Table S2), including the AnatCL–FreeSurfer difference ($\Delta_{\text{ICC}} = +0.05$ [0.03, 0.07]).

In contrast, three models exhibited poor between-scanner reliability (Figure 1b; per-dimension distributions in Supplementary Figure S1). BrainIAC (ICC 0.45 [0.22, 0.63]) showed predominantly poor reliability, with 64.1% of dimensions below 0.50 and virtually none (0.4%) reaching good levels. BrainSegFounder (ICC 0.31 [0.16, 0.42]) showed poor reliability, with 93.5% of dimensions rated as poor. 3D-Neuro-SimCLR was the least reliable (ICC 0.25 [0.09, 0.39]), with 91.0% of dimensions rated as poor.

### 3.2. Within-scanner reproducibility and variance decomposition

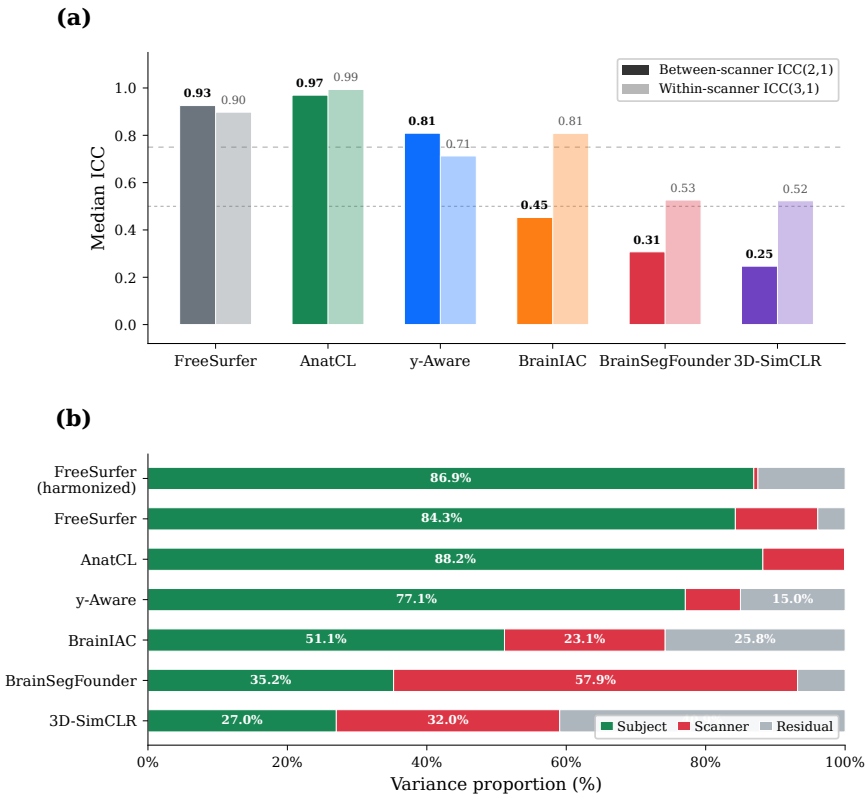


Figure 2: Within- vs. between-scanner reliability and variance decomposition. **(a)** Grouped bar chart comparing within-scanner ICC(3,1) (lighter bars) and between-scanner ICC(2,1) (solid) per model. BrainIAC exhibits the largest discrepancy (0.81 within vs. 0.45 between), indicating stable embeddings that shift systematically across scanners. **(b)** Variance decomposition into subject (biological), scanner (technical), and residual components, averaged across all embedding dimensions. ComBat-harmonized FreeSurfer (top bar) provides a lower-bound reference for scanner variance (0.6%); AnatCL and raw FreeSurfer allocate >84% of variance to subject differences but retain an order of magnitude more scanner variance than the harmonized reference, whereas BrainSegFounder and 3D-Neuro-SimCLR are dominated by scanner and residual variance, respectively.

Within-scanner ICC(3,1) showed a different pattern (Figure 2a). AnatCL again led with near-perfect within-scanner reproducibility (median ICC 0.99). BrainIAC achieved good within-scanner reliability (ICC 0.81), in contrast

with its poor between-scanner performance (0.45). FreeSurfer maintained good within-scanner reliability (0.90), while y-Aware showed moderate within-scanner performance (0.71), lower than its between-scanner ICC (0.81). BrainSegFounder achieved moderate within-scanner reliability (0.53) but poor between-scanner reliability (0.31), and 3D-Neuro-SimCLR showed similar within-scanner reliability (0.52) with the poorest between-scanner reliability (0.25).

Variance decomposition quantified the source of variability in each model's embedding space (Table 3, Figure 2b). AnatCL (88.1% subject, 11.8% scanner, 0.1% residual) and FreeSurfer (84.3%, 11.8%, 3.9%) allocated the vast majority of variance to between-subject differences. y-Aware showed a similar profile with more residual noise (77.1% subject, 7.9% scanner, 15.0% residual). BrainIAC showed roughly balanced variance (51.1% subject, 23.1% scanner, 25.8% residual). BrainSegFounder allocated the majority of variance to scanner identity (57.9%), with subject variance (35.2%) and residual noise (6.8%). 3D-Neuro-SimCLR allocated the largest share to residual noise (40.9%), with scanner variance (32.0%) exceeding subject variance (27.0%).

### *3.3. Scanner fingerprinting and subject identification*

Scanner fingerprinting accuracy varied widely across models (Figure 3; per-scanner distance heatmaps in Supplementary Figure S2): BrainSegFounder achieved the highest scanner classification (90.9%), followed by 3D-Neuro-SimCLR (89.1%), FreeSurfer (80.0%), y-Aware (57.6%), AnatCL (45.5%), and BrainIAC (33.3%).

Subject identification showed a partially converse pattern. AnatCL, y-Aware, and FreeSurfer all achieved perfect subject identification (100%): the nearest cross-scanner neighbour always matched the correct subject. In contrast, BrainIAC (49.7%) and 3D-Neuro-SimCLR (62.4%) achieved subject identification well below perfect. BrainSegFounder reached 86.1%, higher than the other two pure SSL models but still short of perfect. Notably, FreeSurfer combined high scanner fingerprinting (80.0%) with perfect subject

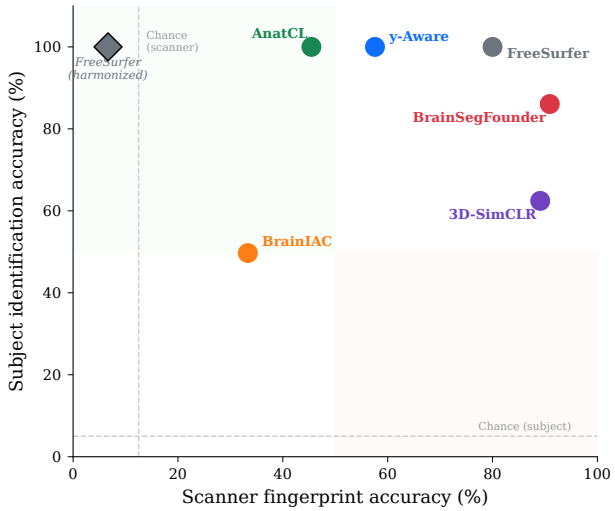


Figure 3: Scanner fingerprint accuracy vs. subject identification accuracy for each model. Dashed lines indicate random chance levels ($1/K$: 12.5% for scanner fingerprinting with $K = 8$ scanners; 5% for subject identification with $K = 20$ subjects). Note that class imbalance across scanners (10 to 30 sessions per scanner) means that a majority-class classifier would already achieve 18.2%, so fingerprinting accuracies close to chance should be interpreted cautiously. Models in the upper-left quadrant (low scanner fingerprinting, high subject ID) have biologically dominated embeddings. Models in the lower-right (high scanner fingerprinting, low subject ID) are scanner-dominated.

identification.

Figure 4: Two-dimensional t-distributed stochastic neighbor embedding (t-SNE) projections of embedding spaces for all six models, ordered by decreasing between-scanner ICC. Top row: points colored by subject identity (20 distinct colors); bottom row: same projections colored by scanner vendor (Siemens, Philips, GE). Reliable models (AnatCL, FreeSurfer, y-Aware) show tight per-subject clusters with vendors intermixed. Unreliable models (BrainIAC, BrainSegFounder, 3D-Neuro-SimCLR) show increasing vendor separation with subjects scattered across vendor groups.

t-SNE projections (Figure 4) illustrate these patterns: AnatCL embeddings formed tight per-subject clusters with scanner vendors intermixed, BrainIAC showed diffuse structure with partial vendor separation, and BrainSegFounder exhibited clear vendor-level clustering with subject identity scattered within

vendor groups.

### *3.4. Supplementary analyses*

Three additional analyses are reported in the Supplementary Material. In cross-scanner sex classification (leave-one-scanner-out logistic regression; Supplementary Figure S3), FreeSurfer, AnatCL, y-Aware, and BrainSegFounder achieved near-perfect balanced accuracy ($\geq$98.7%), while BrainIAC (83.0%) and 3D-Neuro-SimCLR (89.5%) showed lower performance. PCA of embedding spaces (Supplementary Table S3, Figure S4) showed that BrainSegFounder concentrates 62.4% of variance in a single scanner-dominated principal component, whereas reliable models concentrate variance in subject-dominated components. A dimensionality-normalised robustness check (Supplementary Table S4), in which all models were reduced to the same number of principal components before recomputing ICC, confirmed that the top three models (AnatCL, FreeSurfer, y-Aware) retained good-to-excellent reliability. Among the bottom three, BrainSegFounder improved substantially from 0.31 to 0.76 at 10 PCs, surpassing BrainIAC (0.38) and 3D-Neuro-SimCLR (0.29), though all three remained below the top tier.

## 4. Discussion

In this study, we assessed the cross-scanner reliability of brain MRI FM embeddings using a travelling-heads design and investigated which model design factors best explained the observed differences. The clearest pattern was that pretraining strategy, rather than architecture or dataset scale, was the strongest correlate of cross-scanner reliability among the models evaluated. Biology-guided contrastive models (AnatCL, y-Aware) achieved good-to-excellent cross-scanner reliability, while all three pure self-supervised models (BrainIAC, 3D-Neuro-SimCLR, BrainSegFounder) fell below the poor threshold.

To our knowledge, this is the first travelling-heads reliability evaluation of FM embedding vectors in neuroimaging. AnatCL (ICC 0.97, 88.1% subject

variance) exceeded even FreeSurfer morphometrics (ICC 0.93), placing it alongside cortical thickness (typically 0.75–0.95; (Fortin et al., 2018)) and comparable to the best scalar brain age predictions (ICC 0.90–0.98; (Cole et al., 2017; Dörfel et al., 2023)), but now for 512-dimensional embedding vectors rather than single scalars. y-Aware (ICC 0.81) also exceeded the good-reliability threshold (ICC $\geq 0.75$; (Koo and Li, 2016)). At the other extreme, 3D-Neuro-SimCLR (ICC 0.25) fell below measurements from functional connectivity (mean ICC 0.29; (Noble et al., 2019)), a modality known for poor reliability (ICC $< 0.50$; (Koo and Li, 2016)), and BrainSegFounder (ICC 0.31) and BrainIAC (ICC 0.45) remained in the poor range, below fMRI activation maps (mean 0.44; (Bennett and Miller, 2010)). These findings are consistent with emerging evidence from pathology (Thiringer et al., 2026; de Jong et al., 2025) and radiology (Aerts et al., 2025) FMs, where scanner-specific signatures have been documented. Notably, Carloni et al. (Carloni et al., 2025) showed that this scanner sensitivity persists even after standard task-specific fine-tuning, requiring a scanner-aware contrastive loss (i.e., one that explicitly treats scanner identity as a nuisance factor) to mitigate.

BrainIAC showed a particularly informative pattern: good within-scanner ICC (0.81) combined with poor between-scanner ICC (0.45), suggesting it has learned intensity or contrast characteristics that are reproducible within a scanner but shift across hardware. This contrasts with its published "stability analysis" (Tak et al., 2026), which evaluated downstream task performance under synthetically injected contrast shifts, Gibbs ringing, and bias field artifacts rather than real cross-scanner variation. Our results suggest that such synthetic perturbations do not capture the systematic embedding shifts introduced by different scanner hardware. AnatCL showed the smallest gap between within- and between-scanner ICC (0.02), consistent with scanner-invariant representations.

*Pretraining strategy as the strongest correlate.* AnatCL's training objective combines an anatomy-aware contrastive term, which uses cortical thickness,

volume, and surface area to define positive and negative pairs, with the y-Aware age-based contrastive term (Dufumier et al., 2021). This joint loss likely encourages the model to treat scanner-related appearance differences as nuisance variation. Notably, since AnatCL subsumes y-Aware's age signal, the reliability gap between the two models (ICC 0.97 vs. 0.81) can be attributed to the additional anatomical grounding: anchoring representations to morphometric features that are themselves scanner-robust appears to confer stronger invariance than age-based supervision alone. At the other end of the spectrum, BrainSegFounder's particularly poor reliability may reflect a mismatch between its segmentation-oriented training and our use case of global embedding extraction. BrainSegFounder is based on a SwinUNETR-style encoder–decoder designed for dense prediction, and its pretraining combines masked volume inpainting, 3D rotation, and contrastive coding (Cox et al., 2024). Such objectives can emphasise spatially detailed features that are useful for segmentation but may be less stable as subject-level signature. Moreover, in our pipeline we summarize the model's bottleneck feature map into a single vector via global average pooling. This aggregation can discard spatially distributed subject-specific information, potentially leaving a pooled embedding that is dominated by scanner-dependent statistics. The PCA structure of BrainSegFounder's embeddings supports this interpretation: 62.4% of total variance concentrates in a single scanner-dominated principal component ($\eta^2_{\text{scanner}} = 0.86$), while subject-driven variation emerges only in later PCs. Once this dominant scanner axis is factored out by projection onto lower-rank PCs, median ICC rises from 0.31 to 0.76 (Supplementary Table S4), confirming that subject-relevant information is present but overwhelmed by scanner variance in the native embedding space.

Neither architecture, dimensionality, nor dataset scale explain the observed differences. The most and least reliable FMs both use ResNet-18 (AnatCL: 0.97 vs. 3D-Neuro-SimCLR: 0.25). Models sharing the same dimensionality exhibit opposite reliability profiles (AnatCL and 3D-Neuro-SimCLR, both 512-dim), and reducing all models to the same number of principal components preserves

the top-three ranking (Supplementary Table S4). Pretraining scale does not guarantee reliability either: AnatCL was pretrained on only 3,984 scans, an order of magnitude fewer than BrainIAC (32,015) or 3D-Neuro-SimCLR (44,958), yet achieved the highest reliability. Training-data composition also fails to explain the pattern: biology-guided models were pretrained exclusively on healthy controls, whereas the pure SSL models included mixed pathological data, yet BrainSegFounder, pretrained on the largely healthy UK Biobank, achieved the worst reliability. With only five FMs evaluated, each differing in multiple design choices, we cannot isolate the causal contribution of any single factor. Nonetheless, the consistent pattern across models points to pretraining strategy as the most explanatory variable.

Based on these results, AnatCL and y-Aware exhibited substantially higher cross-scanner robustness than the remaining models. However, scanner fingerprinting remained above chance for all models, raising the question of how far even the best FMs are from truly scanner-invariant representations. ComBat harmonization of the FreeSurfer features provides a useful lower-bound reference (Supplementary Table S5; Figure 3): after explicit removal of scanner effects, scanner variance dropped from 11.8% to 0.6% and scanner fingerprinting fell from 80.0% to 6.7%, effectively below chance for eight scanners (12.5%). By comparison, AnatCL (11.8% scanner variance, 45.5% fingerprinting) and y-Aware (7.9%, 57.6%) remain well above this floor, indicating that even the most reliable FMs still carry substantial scanner-related signal and that further validation on downstream multi-site tasks would be needed before deployment without any form of domain adaptation. The poor within-scanner ICC of 3D-Neuro-SimCLR (0.52) and BrainSegFounder (0.53) further suggests that these models would also be unreliable in single-scanner longitudinal designs. For feature extraction workflows, post-hoc statistical harmonization of the extracted embeddings (e.g., ComBat; (Fortin et al., 2018; Jia et al., 2025)) may mitigate scanner effects. For end-to-end fine-tuning, scanner-aware contrastive objectives offer a more principled alternative (Carloni et al., 2025; Huang et al., 2025). Given the magnitude of the differences observed here (ICC 0.25 to

0.97), model selection itself may be the single most consequential design choice for multi-site reliability. Harmonization cannot substitute for a representation that already conflates biology with acquisition. Whether either approach can recover reliable representations from the less stable models, and whether poor embedding reliability degrades performance on downstream multi-site tasks such as classification or regression, are important open questions that fall outside the scope of the present work.

Pursuing scanner-invariant representations might not be without costs: Wang et al. (Wang et al., 2022) showed that enforcing domain invariance can reduce predictive accuracy, because clinically relevant variation may be partly entangled with scanner- and site-specific effects. At the same time, the mere presence of scanner information in an embedding does not automatically imply poor reliability: what matters for reproducibility is whether between-subject differences dominate scanner-related variation. FreeSurfer illustrates this point, combining high scanner fingerprinting (80.0%) with perfect subject identification. Of course, any residual scanner signal can still become a confound in downstream prediction if it correlates with target labels. AnatCL achieved excellent cross-scanner reliability without explicit invariance constraints, suggesting that biologically grounded supervision during pretraining can steer models away from scanner shortcuts. Whether the next generation of FMs can learn such robustness implicitly from scale and diversity alone remains an open question.

*Relevance of frozen embedding reliability.* It is fair to ask whether frozen-embedding reliability matters, given that FMs are often fine-tuned for specific tasks. In practice it matters in at least two common settings. First, FMs are frequently used as off-the-shelf feature extractors. In this regime, any scanner-dependent drift in frozen embeddings is inherited directly by downstream models (Kora et al., 2022; Glocker et al., 2023a,b). Second, even with end-to-end adaptation, pretraining sets the optimisation starting point: if the representation already mixes biology with acquisition artefacts, downstream

training must "unlearn" these shortcuts, which may not fully succeed without explicit objectives designed to suppress scanner signal (Dinsdale et al., 2021; Carloni et al., 2025). Frozen reliability provides a simple pretraining audit: embeddings that make scanner identity easily predictable are a warning sign that the invariances required for multi-site deployment have not been learned (Kushol et al., 2023; Dinsdale et al., 2021). Moreover, embeddings are increasingly treated as biomarkers in their own right (for retrieval, clustering, or anomaly detection) where reproducibility without task-specific adaptation is the entire point (Tripathi et al., 2025; Denner et al., 2025; Schulthess and Konukoglu, 2025).

*Limitations.* A few caveats are worth keeping in mind when interpreting these reliability estimates. First, ON-Harmony includes 20 healthy adults (aged 19–50), so uncertainty is unavoidably larger than in a bigger cohort, and the homogeneous sample can yield conservative ICCs because ICC depends on between-subject heterogeneity as well as measurement error (Shrout and Fleiss, 1979; Koo and Li, 2016). Similarly, subject identification accuracy may be somewhat inflated by the small pool of 20 candidates, as distinguishing among fewer individuals is inherently easier. Nonetheless, the stark contrast between models achieving perfect identification (AnatCL, y-Aware, FreeSurfer) and those falling well below (BrainIAC: 49.7%) indicates that the relative ranking is robust. The within-scanner analysis is even more constrained: although nine participants completed six repeat sessions on a single scanner, repeats are spread across scanners, leaving only two scanners with $\geq 2$ repeat participants and therefore limited precision for ICC(3,1).

Second, model comparisons are partly entangled with preprocessing. To follow published protocols, each FM was run with its recommended preprocessing pipeline, but the two most reliable models (AnatCL, y-Aware) both used aggressive nonlinear processing (CAT12 registration and tissue segmentation; Pipeline A), which may attenuate some scanner differences before the encoder sees the data. In contrast, less reliable models used simpler

spatial normalization: rigid registration for BrainIAC and 3D-Neuro-SimCLR, and affine plus nonlinear (FLIRT + FNIRT) for BrainSegFounder. As a result, we cannot fully attribute reliability differences to pretraining objectives alone. Notably, FreeSurfer, which applies its own nonlinear processing, achieved similarly high ICC while still exhibiting substantial scanner fingerprinting (80.0%), suggesting that nonlinear preprocessing by itself does not guarantee scanner invariance.

Finally, our conclusions are scoped to 3T T1-weighted MRI and to frozen embeddings: fine-tuning may alter the reliability profile, other modalities may behave differently, and for BrainSegFounder we evaluated only the publicly available Stage 1 (self-supervised) checkpoint rather than task-specific Stage 2.

## 5. Conclusion

Among the models evaluated, incorporating biological metadata into pretraining was the strongest correlate of whether brain MRI FM embeddings were reliable across scanners. Models trained with anatomy- or age-guided contrastive losses matched or surpassed the reliability of conventional morphometrics, whereas pure self-supervised models fell below the poor reliability threshold, with two of three (BrainSegFounder, 3D-Neuro-SimCLR) capturing more scanner-related than subject-related variance in their embeddings. AnatCL and y-Aware showed the highest cross-scanner robustness, but whether their embeddings are sufficient for multi-site deployment without harmonization remains to be validated on downstream tasks. Results from downstream assessments based on multi-site data and FM-derived features should be interpreted with caution, particularly when using models whose pretraining did not incorporate biological supervision. As FMs enter clinical neuroimaging pipelines, routine cross-scanner reliability checks, even on small travelling-heads cohorts, can flag representations that primarily reflect acquisition hardware before they propagate into clinical conclusions.

## Acknowledgements

We thank the ON-Harmony consortium for making the travelling-heads dataset publicly available. Graphical abstract icons by Hasymi (warning sign), Octopocto (tick sign), kornkun (people), pojok d (neural network), Dave Gandy (embedding), and Kalashnyk (brain), from Flaticon (www.flaticon.com). Some illustrations adapted from Servier Medical Art (smart.servier.com), licensed under CC BY 4.0.

## Funding

This work was supported by MCIN/AEI/10.13039/501100011033 through grants PID2021-124407NB-I00, PID2024-162453NB-I00, and PID2024-158963NB-I00, co-funded by the European Union, European Regional Development Fund (ERDF, 2021–2027); and by the Junta de Castilla y León (grant VA156P24), co-financed by ERDF.

## CRediT authorship contribution statement

**Rafael Navarro-González:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Santiago Aja-Fernández:** Supervision, Funding acquisition, Writing – review & editing. **Álvaro Planchuelo-Gómez:** Funding acquisition, Writing – review & editing. **Rodrigo de Luis-García:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

## Declaration of competing interest

The authors declare no competing interests.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used Claude (Anthropic) for assistance with code debugging, LaTeX formatting, and editorial suggestions. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of the publication.

## Ethics approval

This study used publicly available, de-identified data from OpenNeuro (ds004712). No additional ethics approval was required.

## Data availability

The ON-Harmony dataset is publicly available at `https://openneuro.org/datasets/ds004712/versions/2.0.1`.

## Code availability

Analysis code is available at `https://github.com/rafaloz/Reliability_FMs`.

## References


Aerts, H.J.W.L., Pai, S., Hadzic, I., Fedorov, A., Mak, R., 2025. Foundation model embeddings for quantitative tumor imaging biomarkers. Research Square preprint, DOI: 10.21203/rs.3.rs-6630446/v1.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data 8, 53. doi:10.1186/s40537-021-00444-8.

Avants, B.B., Tustison, N., Song, G., 2009. Advanced normalization tools (ANTS). Insight Journal 2, 1–35.

Barbano, C.A., Brunello, M., Dufumier, B., Grangetto, M., 2026. Anatomical foundation models for brain MRIs. Pattern Recognit. Lett. 199, 178–184. doi:10.1016/j.patrec.2025.11.028.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155. doi:10.1111/j.1749-6632.2010.05446.x.

Campanella, G., et al., 2025. A clinical benchmark of public self-supervised pathology foundation models. Nat. Commun. doi:10.1038/s41467-025-58796-1.

Carloni, G., Brattoli, B., Keum, S., Park, J., Lee, T., Ahn, C.H., Pereira, S., 2025. Pathology foundation models are scanner sensitive: Benchmark and mitigation with contrastive ScanGen loss, in: MICCAI Workshop on Foundation Models for General Medical AI. ArXiv:2507.22092.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning (ICML), pp. 1597–1607. doi:10.48550/arXiv.2002.05709.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6, 284–290. doi:10.1037/1040-3590.6.4.284.

Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G., 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124. doi:10.1016/j.neuroimage.2017.07.059.

COVID-CNS Consortium, 2022. Reliability of multi-site UK Biobank MRI brain phenotypes for the assessment of neuropsychiatric complications of SARS-CoV-2 infection: The COVID-CNS travelling heads study. PLOS ONE 17, e0273704. doi:10.1371/journal.pone.0273704.

Cox, J., Liu, P., Stolte, S.E., Yang, Y., Liu, K., See, K.B., Ju, H., Fang, R., 2024. BrainSegFounder: Towards 3D foundation models for neuroimage segmentation. Med. Image Anal. 97, 103301. doi:10.1016/j.media.2024.103301.

Denner, S., Zimmerer, D., Bounias, D., Bujotzek, M., Xiao, S., Stock, R., Kausch, L., Schader, P., Penzkofer, T., Jäger, P.F., Maier-Hein, K.H., 2025. Leveraging foundation models for content-based image retrieval in radiology. Comput. Biol. Med. 196, 110640. doi:10.1016/j.compbiomed.2025.110640.

Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., 2021. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. NeuroImage 228, 117689. doi:10.1016/j.neuroimage.2020.117689.

Dörfel, R.P., Arenas-Gomez, J.M., Fisher, P.M., Ganz, M., Knudsen, G.M., Svensson, J.E., Plavén-Sigray, P., 2023. Prediction of brain age using structural magnetic resonance imaging: A comparison of accuracy and test-retest reliability of publicly available software packages. Hum. Brain Mapp. 44, 6139–6148. doi:10.1002/hbm.26502.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR). doi:10.48550/arXiv.2010.11929.

Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., McDonald, C., Piguet, C.M., Duchesnay, E., 2021. Contrastive learning with continuous proxy meta-data for 3D MRI classification, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pp. 58–68. doi:10.1007/978-3-030-87196-3_6.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671. doi:10.1038/nn.4135.

Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54, 313–327. doi:10.1016/j.neuroimage.2010.07.033.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage 167, 104–120. doi:10.1016/j.neuroimage.2017.11.024.

Glocker, B., Jones, C., Bernhardt, M., Winzeck, S., 2023a. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. eBioMedicine 89, 104467. doi:10.1016/j.ebiom.2023.104467.

Glocker, B., Jones, C., Roschewitz, M., Winzeck, S., 2023b. Risk of bias in chest radiography deep learning foundation models. Radiology Artif. Intell. 5, e230060. doi:10.1148/ryai.230060.

Grote, T., Freiesleben, T., Berens, P., 2024. Foundation models in healthcare require rethinking reliability. Nat. Mach. Intell. 6, 1421–1423. doi:10.1038/s42256-024-00924-5.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M., 2022. SynthStrip: skull-stripping for any brain image. NeuroImage 260, 119474. doi:10.1016/j.neuroimage.2022.119474.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. arXiv preprint doi:10.1109/CVPR.2017.243. arXiv:1608.06993.

Huang, Y., et al., 2025. Knowledge-guided adaptation of pathology foundation models effectively improves cross-domain generalization and demographic fairness. Nat. Commun. doi:10.1038/s41467-025-66300-y.

Isensee, F., Schell, M., Pflueger, I., Bruber, G., Spinoit, D., Full, Y., Maier-Hein, K.H., 2019. Automated brain extraction of multisequence MRI using artificial neural networks. Hum. Brain Mapp. 40, 4952–4964. doi:10.1002/hbm.24750.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62, 782–790. doi:10.1016/j.neuroimage.2011.09.015.

Jia, W., Li, H., Ali, R., Shanbhogue, K.P., Masch, W.R., Aslam, A., Harris, D.T., Reeder, S.B., Dillman, J.R., He, L., 2025. Investigation of ComBat harmonization on radiomic and deep features from multi-center abdominal

MRI data. J. Imaging Inform. Med. 38, 1016–1027. doi:10.1007/s10278-024-01253-0.

de Jong, E.D., Marcus, E., Teuwen, J., 2025. Current pathology foundation models are unrobust to medical center differences. ArXiv:2501.18055.

Kaczmarek, E., Szeto, J., Nichyporuk, B., Arbel, T., 2025. Building a general SimCLR self-supervised foundation model across neurological diseases to advance 3D brain MRI diagnoses, in: ICCV 2025 Workshop on Computer Vision for Automated Medical Diagnosis (CVAMD). ArXiv:2509.10620.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15, 155–163. doi:10.1016/j.jcm.2016.02.012.

Kora, P., Ooi, C.P., Faust, O., Raghavendra, U., Mohanasen, A., Barua, P.D., Fujita, H., Acharya, U.R., 2022. Transfer learning techniques for medical image analysis: A review. Biocybern. Biomed. Eng. 42, 79–107. doi:10.1016/j.bbe.2021.11.004.

Kushol, R., Parnianpour, P., Wilman, A.H., Kalra, S., Yang, Y.H., 2023. Effects of MRI scanner manufacturers in classification tasks with deep learning models. Sci. Rep. 13, 16791. doi:10.1038/s41598-023-43715-5.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022. doi:10.1109/ICCV48922.2021.00986.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46. doi:10.1037/1082-989X.1.1.30.

Neidlinger, P., et al., 2025. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nat. Biomed. Eng. doi:10.1038/s41551-025-01516-3.

Noble, S., Scheinost, D., Constable, R.T., 2019. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. NeuroImage 203, 116157. doi:10.1016/j.neuroimage.2019.116157.

Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J., Aerts, H.J.W.L., 2024. Foundation model for cancer imaging biomarkers. Nat. Mach. Intell. 6, 354–367. doi:10.1038/s42256-024-00807-9.

Schulthess, N., Konukoglu, E., 2025. Anomaly detection by clustering DINO embeddings using a Dirichlet process mixture, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. ArXiv:2509.19997.

Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Reich, D.S., Crainiceanu, C.M., 2014. Statistical normalization techniques for magnetic resonance imaging. NeuroImage Clin. 6, 9–19. doi:10.1016/j.nicl.2014.08.008.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. 86, 420–428. doi:10.1037/0033-2909.86.2.420.

Tak, D., Garomsa, B.A., Chaunzwa, T.L., Zapaishchykova, A., Climent Pardo, J.C., Ye, Z., Zielke, J., Ravipati, Y., Vajapeyam, S., Mahootiha, M., Smith, C., Familiar, A.M., Liu, K.X., Prabhu, S., Bandopadhayay, P., Nabavizadeh, A., Mueller, S., Aerts, H.J.W.L., Huang, R.Y., Poussaint, T.Y., Kann, B.H., 2026. A generalizable foundation model for analysis of human brain MRI. Nat. Neurosci. doi:10.1038/s41593-026-02202-6.

Thiringer, E., Gustafsson, F.K., Ledesma Eriksson, K., Rantalainen, M., 2026. Scanner-induced domain shifts undermine the robustness of pathology foundation models. ArXiv:2601.04163.

Tripathi, A., Waqas, A., Schabath, M.B., Yilmaz, Y., Rasool, G., 2025. HoneyBee: A scalable modular framework for creating multimodal oncology

datasets with foundational embedding models. npj Digit. Med. 8, 622. doi:10.1038/s41746-025-02003-4.

Voelker, M.N., Kraff, O., Goerke, S., Laun, F.B., Hanspach, J., Pine, K., Ehses, P., Zaiss, M., Liebert, A., Straub, S., Eckstein, K., Robinson, S., Nagel, A.M., Stefanescu, M.R., Wollrab, A., Klix, S., Felder, J., Hock, M., Bosch, D., Weiskopf, N., Speck, O., Ladd, M.E., Quick, H., 2021. The traveling heads 2.0: Multicenter reproducibility of quantitative imaging methods at 7 Tesla. NeuroImage 232, 117910. doi:10.1016/j.neuroimage.2021.117910.

Wang, R., Chaudhari, P., Davatzikos, C., 2022. Embracing the disharmony in medical imaging: A simple and effective framework for domain adaptation. Med. Image Anal. 76, 102309. doi:10.1016/j.media.2021.102309.

Warrington, S., Ntata, A., Mougin, O., Campbell, J., Torchi, A., Craig, M., Alfaro-Almagro, F., Miller, K.L., Morgan, P.S., Jenkinson, M., Sotiropoulos, S.N., 2023. A resource for development and comparison of multimodal brain 3T MRI harmonisation approaches. Imaging Neurosci. 1, 1–22. doi:10.1162/imag_a_00042.

Warrington, S., Torchi, A., Mougin, O., Campbell, J., Ntata, A., Craig, M., Assimopoulos, S., Alfaro-Almagro, F., Miller, K.L., Jenkinson, M., Morgan, P.S., Sotiropoulos, S.N., 2025. A multi-site, multi-modal travelling-heads resource for brain MRI harmonisation. Sci. Data 12, 609. doi:10.1038/s41597-025-04822-2.

You, K., Kou, Z., Long, M., Wang, J., 2020. Co-tuning for transfer learning, in: Advances in Neural Information Processing Systems (NeurIPS).