



MMORF—FSL’s MultiMOdal Registration Framework

Frederik J. Lange^a, Christoph Arthofer^a, Andreas Bartsch^b, Gwenaëlle Douaud^a, Paul McCarthy^a,
Stephen M. Smith^a, Jesper L. R. Andersson^a

^aWellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

^bDepartment of Neuroradiology, University of Heidelberg, Heidelberg, Germany

Corresponding Author: Frederik J. Lange (frederik.lange@ndcn.ox.ac.uk)

ABSTRACT

We present MMORF—FSL’s MultiMOdal Registration Framework—a newly released nonlinear image registration tool designed primarily for application to magnetic resonance imaging (MRI) images of the brain. MMORF is capable of simultaneously optimising both displacement and rotational transformations within a single registration framework by leveraging rich information from multiple scalar and tensor modalities. The regularisation employed in MMORF promotes local rigidity in the deformation, and we have previously demonstrated how this effectively controls both shape and size distortion, leading to more biologically plausible warps. The performance of MMORF is benchmarked against three established nonlinear registration methods—FNIRT, ANTs, and DR-TAMAS—across four domains: FreeSurfer label overlap, diffusion tensor imaging (DTI) similarity, task-fMRI cluster mass, and distortion. The evaluation is based on 100 unrelated subjects from the Human Connectome Project (HCP) dataset registered to the Oxford-MultiModal-1 (OMM-1) multimodal template via either the T1w contrast alone or in combination with a DTI/DTI-derived contrast. Results show that MMORF is the most consistently high-performing method across all domains—both in terms of accuracy and levels of distortion. MMORF is available as part of FSL, and its inputs and outputs are fully compatible with existing workflows. We believe that MMORF will be a valuable tool for the neuroimaging community, regardless of the domain of any downstream analysis, providing state-of-the-art registration performance that integrates into the rich and widely adopted suite of analysis tools in FSL.

Keywords: registration, nonlinear, multimodal, volumetric, FSL

1. INTRODUCTION

In this paper, we describe and evaluate FSL’s MultiMOdal Registration Framework (MMORF). MMORF is a nonlinear image registration tool, primarily intended for magnetic resonance imaging (MRI) of the brain.

Biomedical image registration is a core component in most neuroimaging processing pipelines. If we assume that all brains, regardless of appearance, are built using the same anatomical components arranged in the same configuration (i.e., they are topologically identical), then we can use image registration to find the set of deforma-

tions that map all brains to a common reference space/template brain in a one-to-one manner. In this case, a template may refer to either a group-average brain or an individual subject brain. Although the assumption of identical topology may not always hold, it is valid enough for registration to be used in both the localisation and quantification of similarities and differences between individuals or population groups—that is, the template allows us to say where the differences occur, and the mappings ensure that what we are comparing is measured at the same place in all subjects. A poor registration will, therefore, impact both the power to detect and

Received: 26 September 2023 Revision: 20 January 2024 Accepted: 5 February 2024 Available Online: 12 February 2024



© 2024 Massachusetts Institute of Technology.
Published under a Creative Commons Attribution 4.0
International (CC BY 4.0) license.

Imaging Neuroscience, Volume 2, 2024
https://doi.org/10.1162/imag_a_00100

the ability to accurately localise any effects of interest across subjects.

However, the reality is that image registration is an ill-posed problem, and therefore the true, one-to-one mapping can generally never be found. By ill-posed, we mean that there is typically not enough information in an image itself to find a mapping that uniquely maximises some measure of similarity between a subject's brain and the template. Therefore, registration methods require regularisation to constrain the solution to be unique. Regularisation achieves this by encoding a model of what types of deformations are considered more likely than others. The challenge in designing a registration tool is then to find the best way to combine image information and regularisation to produce as good an approximation to the true mapping as possible.

MMORF addresses the image information aspect of this challenge by taking a multimodal approach to computing brain similarity. Up until now, we have described registration as if there is only one measurement/image of the brain that we can use to find the correct mapping. However, with MRI we are able to acquire a number of different image modalities—all within the same imaging session, and each with different contrast and information content. In MMORF, we have leveraged this complementary information to reduce the degree to which the registration problem is ill-posed, thereby improving our confidence that we are finding an accurate mapping for each subject. MMORF's registration abilities extend beyond scalar modalities to include diffusion tensor imaging (DTI). When using the full tensor data (rather than scalar, rotationally invariant, derived features such as fractional anisotropy (FA)), it matches the directional information in the diffusion tensor to guide the alignment of, in particular, white matter.

MMORF's regularisation method is one of its most unique attributes. By employing a penalty that aims to preserve the original shape and volume of an object as far as is reasonable, it is able to produce deformations that are far more biologically plausible than those generated by conventional techniques. A detailed description of its implementation in MMORF, and a thorough evaluation of the benefits of this form of regularisation, can be found in [Lange et al. \(2020\)](#). While that work was carried out with an early development version of MMORF that lacked many of the core features found in the released version (most notably, DTI registration, bias field estimation, and masking were not yet implemented), the results were sufficiently encouraging to convince us that this was the regularisation approach we should pursue as we developed the tool further.

Through reducing the degree to which the registration problem is ill-posed (using multimodal data), and by pro-

ducing more realistic deformations in those regions where it remains so (using biomechanically realistic regularisation), MMORF is capable of state-of-the-art registration accuracy.

The cost of combining multiple imaging modalities with a complex regularisation model in this way is that the computational requirements of the method increase. We have addressed this from the outset by designing MMORF to use graphics processing units (GPUs) to parallelise its execution. This allows MMORF to execute 1 mm isotropic registrations with reasonable runtimes of between 5 and 45 minutes (depending on the number of modalities used) when run on a modern GPU (e.g., Nvidia Titan V). Runtimes are approximately doubled on older GPUs (e.g., Nvidia K80).

In the methods section below, we detail the most important design decisions made when developing MMORF, both those aimed at improving registration quality and those aimed at ensuring efficient execution, so that its operation can be clearly understood. We then contextualise these decisions with reference to a subset of comparable current registration tools. Finally, we validate and benchmark MMORF against these tools across structural-, diffusion-, functional-, and distortion-derived metrics, in order to demonstrate its performance and utility. While we draw on approaches and methodologies used in general image registration comparison studies, such as [Klein et al. \(2009\)](#) and [Ou et al. \(2014\)](#), the focus of the comparisons presented here is to contextualise MMORF's performance relative to a few key methods, rather than evaluating the current image registration field as a whole. Furthermore, while those studies are broader in terms of the number of datasets considered, this work aims to be deeper in terms of the types of metrics used, and the modalities from which they are derived.

2. METHODS

As stated in the introduction, the true one-to-one mapping between brain images can never be known, and therefore image registration can only find the “optimal” mapping based on the available information and our prior beliefs about what mappings are more likely than others. Unsurprisingly then, registration is normally formulated as an optimisation problem, requiring a cost function to minimise. MMORF is intended to be used with any number of scalar and tensor images driving the registration, and we therefore choose to minimise a total cost function \mathcal{C}_{tot} that is defined as follows:

$$\mathcal{C}_{tot}(\mathbf{w}) = \sum_{\mathbf{x} \in \Omega} \left(\sum_s \lambda_s \mathcal{C}_s(\mathbf{w}, \mathbf{x}) + \sum_t \lambda_t \mathcal{C}_t(\mathbf{w}, \mathbf{x}) + \lambda_r \mathcal{C}_r(\mathbf{w}, \mathbf{x}) \right) \quad (1)$$

where \mathbf{w} are the warp parameters being optimised, Ω is the domain over which the warp is defined, λ_* are cost function weightings, and subscripts s , t , and r refer to scalar, tensor, and regularisation, respectively.

Each scalar and tensor cost function is based on an image dissimilarity metric between a pair of images—one belonging to the reference subject (often a template), and one to the moving subject. Although Equation 1 shows that these cost functions are separable, and can therefore be evaluated independently, it is critical that within each subject all images are accurately co-registered. If not, then a single warp cannot correctly map all modalities between subjects, and registration accuracy will suffer. We do not attempt to ensure this within MMORF itself. Rather, we rely on running accurate, rigid, between-modality, within-subject registration using FSL FLIRT (Jenkinson & Smith, 2001; Jenkinson et al., 2002), as well as distortion correction using FSL topup (Andersson et al., 2003) and eddy (Andersson & Sotiropoulos, 2016) for relevant modalities such as DTI, before performing nonlinear registration. However, it is not necessary to resample any of the images following rigid alignment before feeding them into MMORF. Instead, MMORF expects a separate affine transformation matrix to be supplied for each image being registered (in both the moving and reference subject datasets). This affine transformation points to a separate, user-specified, “warp-space image,” whose extents define the domain over which MMORF will estimate and output its nonlinear deformation. This eliminates the need for multiple resampling of the data, and requires no matching of the image resolution or dimensions of any of the images being registered.

With this information in hand, the following sections elaborate on some of the key decisions that went into MMORF’s design, and the role they play in its performance.

2.1. Optimisation strategy

Nonlinear image registration tools generally use one of two iterative optimisation approaches—first or second order. First-order methods consider only the gradient of the cost function, with respect to each optimisable parameter, when picking a parameter update step. Second-order methods extend this by also considering how long the derivative is valid for and the interaction between parameters. First-order algorithms tend to have update steps that can be calculated more quickly, but second-order algorithms tend to require fewer update steps to reach convergence. In our experience, the trade-off tends to favour second-order approaches for methods such as MMORF.

We have, therefore, implemented two variants of the (second-order) Gauss-Newton (GN) optimisation strategy—which is itself a variation on Newton’s method.

Newton’s method is an iterative optimisation algorithm that uses a quadratic Taylor approximation of the cost function \mathcal{C} around the current set of parameters \mathbf{w} (Nocedal & Wright, 2006, Ch.10, pp 254). That is:

$$\mathcal{C}(\mathbf{w} + \Delta\mathbf{w}) \approx \mathcal{C}(\mathbf{w}) + \nabla\mathcal{C}(\mathbf{w})^T \Delta\mathbf{w} + \frac{1}{2} \Delta\mathbf{w}^T \mathbf{H}(\mathbf{w}) \Delta\mathbf{w} \quad (2)$$

where $\nabla\mathcal{C}$ and \mathbf{H} are the gradient and the Hessian of the cost function, respectively. The update $\Delta\mathbf{w}$ that minimises this approximation to the cost function (when the cost function is convex) is then:

$$\Delta\mathbf{w} = -\mathbf{H}^{-1}(\mathbf{w}) \nabla\mathcal{C}(\mathbf{w}) \quad (3)$$

During registration, MMORF minimises \mathcal{C} by iteratively solving Equation 3 for $\Delta\mathbf{w}$ and applying the update. This involves both calculating and storing $\nabla\mathcal{C}$ and \mathbf{H} —the latter requiring the majority of both MMORF’s runtime and GPU RAM usage.

The two variants of GN that can be selected in MMORF are known as Levenberg-Marquardt (LM) (Nocedal & Wright, 2006, Ch.10, pp258) and Majorise-Minimisation (MM) (Hunter & Lange, 2004), respectively. A detailed description of these variants is given in Appendix A, but the primary difference between them is in how they approximate \mathbf{H} —with \mathbf{H}_{MM} requiring on the order of 1000 times less memory than \mathbf{H}_{LM} to store. The reason for including both LM and MM is that GPUs often have relatively little RAM and, therefore, storing \mathbf{H}_{LM} becomes impossible for warps beyond a certain resolution. At this point (typically warp resolutions around 2 mm isotropic), MMORF switches from MM to LM optimisation. The price for these vastly reduced memory requirements is that MM optimisation takes more steps to converge—but it requires no other changes to the registration algorithm, which makes it an appealing option.

2.2. Transformation model

MMORF employs a transformation of the small deformation framework variety (Bajcsy & Kovačič, 1989; M. Miller et al., 1997)—that is, it defines the deformation as a displacement field, rather than a velocity field as per the large deformation framework (Bro-Nielsen & Gramkow, 1996; M. Miller et al., 1997). A major difference between these two frameworks is their relationship to diffeomorphism.

Diffeomorphism is a desirable property in image registration. Diffeomorphic transformations are smooth, one-to-one, and onto, and are therefore guaranteed to induce neither folding nor tearing when applied to an image. The large deformation framework, and in particular the large deformation diffeomorphic metric mapping (LDDMM)

(Beg et al., 2005) family of tools, have the advantage that their transformation model can be made inherently diffeomorphic—that is, they are, in principle, diffeomorphic by construction.

However, despite not being inherently diffeomorphic, the small deformation framework can be made diffeomorphic by employing a regularisation penalty that enforces diffeomorphism (Rohlfing et al., 2003). This has the benefit that warp-induced distortions, such as changes in shape and volume, can be calculated (and therefore controlled) directly from the model parameters themselves. In contrast, large deformation models require the vector field that they parametrise to first be integrated over time and converted into a displacement field before being able to calculate such measures. Therefore, explicitly controlling the amount of distortion is harder in the large deformation framework, despite its guarantee of diffeomorphism. This is important in our case, as one advantage of MMORF's regularisation (described in Section 2.4) is that it non-linearly penalises the stretching and compressing effects of the warp directly.

MMORF's transformation is parametrised using cubic B-splines (Unser et al., 1993a, 1993b). This imposes an inherent smoothness to the deformations, as cubic B-splines, and therefore the warps, have C2 continuity. The warp field is then also well defined across all of the image (i.e., not just at voxel centres), eliminating any need to match image subsampling or resolution either before or during registration. Despite their smoothness, B-splines still have compact support (i.e., the effect of a particular spline is exactly zero at a fixed distance from the spline centre), and, therefore, interaction effects between splines are fixed and finite. This has important implications for second-order optimisation (such as the Gauss-Newton method employed by MMORF), since it means that the Hessian matrix \mathbf{H} in Equation 3 (which encodes the interaction between optimisable parameters) is sparse and predictably patterned.

Furthermore, the spatial derivatives of B-splines are smooth and have a closed form solution. Since these derivatives are required to calculate gradients and Hessians for the optimisation of warps in MMORF, this is computationally beneficial.

We now provide a more explicit description of how we have formulated our transformation model, and how this interacts with the calculation of the Hessian during optimisation (the most computationally intensive part of the registration algorithm). A set of B-spline basis functions can be used to transform a set of sample coordinates in a reference image f to their corresponding coordinates in a moving image g as follows:

f = Reference image

g = Moving image

N = Number of sampled voxels in f

$\mathbf{x} / \mathbf{y} / \mathbf{z}$ = x, y and z coordinates of samples in f

$$\mathbf{X} = \underbrace{\begin{bmatrix} \mathbf{x}^T \\ \mathbf{y}^T \\ \mathbf{z}^T \end{bmatrix}}_{3 \times N} = \begin{bmatrix} x_0 & x_1 & \cdots & x_N \\ y_0 & y_1 & \cdots & y_N \\ z_0 & z_1 & \cdots & z_N \end{bmatrix}$$

M = Number of splines per warp direction

$\mathbf{w}_x / \mathbf{w}_y / \mathbf{w}_z$ = x, y and z direction warp parameters

$$\mathbf{W} = \underbrace{\begin{bmatrix} \mathbf{w}_x^T \\ \mathbf{w}_y^T \\ \mathbf{w}_z^T \end{bmatrix}}_{3 \times M} = \begin{bmatrix} w_{x0} & w_{x1} & \cdots & w_{xM} \\ w_{y0} & w_{y1} & \cdots & w_{yM} \\ w_{z0} & w_{z1} & \cdots & w_{zM} \end{bmatrix}$$

\mathbf{b}_m = Vectorised m^{th} B-spline at sample positions \mathbf{X}

$$\mathbf{B} = \underbrace{\begin{bmatrix} \mathbf{b}_0^T \\ \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_M^T \end{bmatrix}}_{M \times N} = \underbrace{\begin{bmatrix} b_{00} & b_{01} & \cdots & b_{0N} \\ b_{10} & & \ddots & \vdots \\ \vdots & & \ddots & \vdots \\ b_{M0} & \cdots & \cdots & b_{MN} \end{bmatrix}}_{\substack{1 \text{ voxel per column}}} \left. \vphantom{\begin{bmatrix} b_{00} \\ b_{10} \\ \vdots \\ b_{M0} \end{bmatrix}} \right\} 1 \text{ spline per row}$$

$$\begin{aligned} \phi(\mathbf{X}, \mathbf{W}) &= \text{Transformed sample coordinates} \\ &= \mathbf{X} + \mathbf{WB} \end{aligned}$$

Note that here we assume that the warp space and the reference image f are the same. We use the same configuration of B-splines to define the warps in all three directions (i.e., the number, order, and spatial extent of splines defining the displacements in each direction (x, y and z) is the same, and they are located identically in space). M is chosen to be the set of splines whose spatial support, wholly or partially, overlaps with the domain over which f is defined. The warp parameters $\mathbf{w}_{x/y/z}$ are then the coefficients of each B-spline, and each parameter only affects the displacement in a single direction.

The compact support of B-splines means that \mathbf{B} is very sparse. Additionally, (disregarding edge cases) each row of \mathbf{B} is simply a shifted version of any other row. We therefore never store \mathbf{B} explicitly, and instead compute \mathbf{WB} using convolution.

Another benefit of the sparsity in \mathbf{B} is that it induces sparsity in the Hessian of cost functions based on this transformation. For a mean squared error cost function, each element of the Gauss-Newton approximation of the Hessian can be calculated as:

$$\mathbf{H}_{ij} = \sum_{\substack{\mathbf{x} \in \mathbf{b}_i \cap \mathbf{b}_j \\ \text{spline intersection}}} \mathbf{b}_i(\mathbf{x}) \mathbf{b}_j(\mathbf{x}) \underbrace{\frac{\partial g}{\partial xyz(i)}(\phi(\mathbf{x})) \frac{\partial g}{\partial xyz(j)}(\phi(\mathbf{x}))}_{\text{gradient image}}, \quad (4)$$

where $\frac{\partial g}{\partial xyz(i)}$ means “differentiate with respect to the direction in which B-spline i causes displacement to occur.” As each entry in the Hessian represents the interaction between two of the B-spline basis functions, only those combinations of splines that overlap in their spatial support will ever produce non-zero entries. The number of non-zero entries per row of the Hessian is at most $3 \times 7^3 = 1026$ for 3D images and cubic B-splines. Considering that the number of parameters being optimised over can easily exceed 10^6 , this would lead to a matrix that is at least 99.9% sparse. The redundancy increases as the warp resolution increases (the knot-spacing is reduced), which means that the more parameters one attempts to estimate, the sparser the Hessian becomes. Therefore, the memory requirement for the Hessian scales much better with resolution than might initially be feared.

2.3. Image cost functions

As stated earlier, MMORF optimises the cost function defined in Equation 1, which includes the sum over individual cost functions for each pair of images. In this section, we describe the choice of cost function for scalar and tensor images, as well as how MMORF implements cost function masking/weighting.

2.3.1. Scalar

Scalar image dissimilarity is calculated using the mean squared error (MSE) across the image. That is, for two scalar images f and g :

$$\mathcal{C}_s(\mathbf{w}, f, g) = \frac{1}{\Omega} \sum_{\mathbf{x} \in \Omega} (f(\mathbf{x}) - g(\mathbf{x}, \mathbf{w}))^2 \quad (5)$$

MSE is ideally suited to the GN family of optimisation methods, which have an implicit assumption that the cost function being optimised is some form of squared difference.

One characteristic of MSE is that any differences in intensity between the reference and moving images can be mistaken for differences in alignment (Andersson et al., 2019) leading to unnecessary (and incorrect) image warping during optimisation, which would negatively affect MMORF’s registration accuracy if left unaccounted

for. We therefore use robust mean intensity estimation to scale each image separately in order to correct global, linear, intensity scaling differences between image pairs. However, transmit and receive bias fields in MRI can cause spatially varying inhomogeneities in intensity that cannot be corrected with global intensity scaling (Vovk et al., 2007). To address this, we include the option to explicitly model such inhomogeneities as a smooth, spatially varying multiplicative bias field acting on the reference image intensities. This can be enabled or disabled for each image pair independently. The bias field is parametrised using cubic B-splines in the same way as the warp field, and the resolution (knot-spacing) and smoothness (bending energy; Bookstein, 1997) are set on a per image-pair basis. The optimal bias field is found by minimising the sum of the bending energy of the bias field, and the MSE between the reference image (to which the bias field is applied) and the moving image (to which the warp field is applied). Bias field estimation is interleaved with warp field estimation such that the warp field is kept constant while the bias field is optimised, and the bias field is kept constant while the warp field is optimised.

2.3.2. Tensor

Tensor image dissimilarity is calculated using the mean squared Frobenius norm (MSFN) across the image. This is exactly equivalent to summing the MSE for each of the 9 elements of the diffusion tensor, and therefore fits our GN optimisation strategy. That is, for two tensor images \mathbf{F} and \mathbf{G} :

$$\mathcal{C}_t(\mathbf{w}, \mathbf{F}, \mathbf{G}) = \frac{1}{\Omega} \sum_{\mathbf{x} \in \Omega} \sum_{i=1}^3 \sum_{j=1}^3 (\mathbf{F}_{ij}(\mathbf{x}, \mathbf{w}) - \mathbf{G}_{ij}(\mathbf{x}, \mathbf{w}))^2 \quad (6)$$

Note that, in contrast to the scalar case, the values in the reference image \mathbf{F} are also a function of the warp parameters \mathbf{w} . This is because MMORF applies the rotational effect of the warp only to the reference image, and the displacement effect only on the moving image (as usual).

Rotation of tensors uses the finite strain (FS) method (D. C. Alexander et al., 2001). The Jacobian matrix of the warp at each position $\mathbf{J}(\mathbf{x}, \mathbf{w})$ represents the first-order linear approximation of the deformation at that point, and from it the local rotational effect of the warp $\mathbf{R}(\mathbf{x}, \mathbf{w})$ can be calculated using:

$$\mathbf{R} = (\mathbf{J}\mathbf{J}^T)^{-\frac{1}{2}} \mathbf{J} \quad (7)$$

We include the effects of both displacement and rotation when calculating the gradient and the Hessian

of the tensor cost function, using the derivation by Yeo et al. (2009) for the closed-form of the derivative due to rotation.

Since the diffusion tensor model is quantitative, no intensity rescaling is performed. This implicitly assumes that the tensors are always represented with the same units. Similarly, fitting of the diffusion tensor compensates for any intensity inhomogeneity present in the raw diffusion-weighted images, and therefore no bias field estimation is required.

MMORF assumes that diffusion tensors are stored in FSL dtifit¹ format—that is, a 4D volume where the fourth dimension contains the six upper-diagonal elements vectorised row-wise, and the x-direction is defined in radiological convention (R-L).

2.3.3. Masking

In certain instances, it may be beneficial to focus a registration algorithm on a particular region of an image or, alternatively, for it to ignore a certain region. To this end, we have implemented masking/weighting within MMORF.

Masks can be supplied for all reference and moving images independently, and it is required that these masks be in the same space as their corresponding images. The masks are treated as containing voxelwise multiplicative factors that are applied to the cost function during optimisation. Setting any region of a mask to zero will cause the algorithm to ignore the impact of that region's similarity between reference and moving images, and the deformation in that region will be determined purely by the regularisation. However, masks do not have to be binary, and so a “soft” mask can be used to down-weight certain regions in the image rather than ignoring them completely (e.g., favouring brain alignment over the rest of the head, without disregarding non-brain alignment entirely).

In MMORF, masks can be enabled or disabled for each image at every iteration independently, allowing for maximum flexibility. For example, we have occasionally found it useful to only have a mask be applied during higher resolution iterations of the registration.

2.4. Regularisation

Since image registration is an under-constrained problem, regularisation is essential to ensure that the resulting warp fields are biologically plausible. A plausible warp should, at the very least, be diffeomorphic. Diffeomorphic warps are both one-to-one and onto—that is, each point in the reference space maps to a unique point in the

moving space, and each point in the moving space is reachable from the reference space. They require the mapping to be continuous, smooth (i.e., have a continuous derivative), and invertible (i.e., have a finite, positive Jacobian determinant everywhere). B-spline parametrised warps are, by definition, smooth and continuous. Therefore, provided the Jacobian determinant remains > 0 everywhere, the warp is diffeomorphic.

Diffeomorphism is, however, only one desirable trait in a warp. It guarantees that an image is never torn or folded, but that is all. Typically, penalising the variation of the Jacobian determinant from a value of 1 is used to encourage or ensure (depending on the choice of penalty) that warps remain diffeomorphic in the small deformation framework. However, penalising the Jacobian determinant directly only controls volumetric distortion (i.e., changes in an object's size), and does not in any way control shape distortion (i.e., changes in an object's shape). The singular values of the local Jacobian represent stretches/compressions along three orthogonal directions. Therefore, the difference between them indicates the degree of shape distortion. Because the Jacobian determinant is the product of these singular value, one can control changes in both size and shape by controlling only the singular values. Penalising deviations of the singular values from 1, and ensuring that none become negative, leads to diffeomorphic warps with desirably little distortion in both volume and shape.

In MMORF, the specific penalty used is the mean (across the image) of the sum of the squared logarithms of the singular values at each voxel, as shown in Equation 8 below:

$$C_r(\mathbf{w}) = \frac{1}{\Omega} \sum_{\mathbf{x} \in \Omega} \sum_{i=1}^3 (\log s_i(\mathbf{J}(\mathbf{x})))^2 \quad (8)$$

where s_i is the i^{th} singular value of the local Jacobian matrix \mathbf{J} . This is an adaptation of the penalty first proposed by Ashburner et al. (1999, 2000). Its implementation in MMORF is described (and evaluated) in detail in Lange et al. (2020), where we demonstrate the positive effect that this form of regularisation has on the biological plausibility of the warps MMORF produces.

By taking the squared logarithm, this penalty tends to infinity as any singular value tends to either zero or infinity. Additionally, this does not penalise any transformations that are locally rigid (i.e., regions that are only translated and/or rotated). Therefore, MMORF regularisation enforces diffeomorphism and encourages local rigidity, thereby controlling both volumetric and shape distortions. The highly nonlinear form of the regularisation also allows the weighting to be set low enough such that large deformations are allowed when necessary, while

¹ <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FDT/UserGuide#DTIFIT>

still ensuring diffeomorphism, thereby overcoming one of the perceived limitations of the small deformation framework. In other words, MMORF can generate large magnitude deformations when they are expected (such as in growth, ageing or atrophy), but does so in a way that preserves the original volume and shape of the deformed tissue as far as possible.

2.5. Inverse consistency

A perfectly inverse consistent registration algorithm will produce the inverse of the original warp when the reference and moving images swap roles (Christensen, 1999). Since we often consider the choice of reference and moving image to be arbitrary (e.g., such as when registering two individuals to each other), this is seen as a desirable property.

Some tools, including ANTs (Avants et al., 2008) and DR-TAMAS (Irfanoglu et al., 2016), achieve inverse consistency by design (excluding the affine initialisation) through the use of a transformation model known as Symmetric Normalisation (SyN). SyN proceeds by registering both moving and reference images to a mid-space, and using the inversion and composition of these two warps to define the total forward and inverse transformations. The cost of this is essentially running two registrations per pair of subjects, which takes double the time to compute. In practice, when actually swapping the reference and moving images, this approach leads to very good, but not perfect, inverse consistency—in this work, we found ANTs registration produced mean and max inverse consistency errors (Beg & Khan, 2007) of 0.55 mm and 11.14 mm, respectively, across 10 subjects.

In MMORF we have taken a different approach where, rather than enforcing a symmetric warp, we symmetrise the cost function being minimised. This is achieved by multiplying each cost function (both image similarity and regularisation) by a weighting term of $1+|J|$. Since the cost functions are evaluated on a regularly sampled grid in the space of the reference image, the intuition for this weighting is that it always accounts for the total volume in both images to which each value of the cost function at a location in the grid applies. In the continuous case, this can be shown to exactly symmetrise the cost function (Tagare et al., 2009). However, since we are dealing with discretely sampled data, this correction is only approximate in MMORF. Additionally, the inverse of MMORF's cubic B-spline parametrised transformation cannot be perfectly represented by another cubic B-spline field. In practice, this method produces good levels of inverse consistency, but it does not reach the same levels as SyN—in this work, we found MMORF registration produced mean and max inverse consistency errors of

1.12 mm and 14.95 mm, respectively, across 10 subjects. This approach has the computational benefit that, for a given pair of subjects, only one warp field needs to be calculated during registration.

However, it is not strictly necessary for a registration method to address inverse consistency at all. FNIRT (Andersson et al., 2007), for example, does not explicitly consider inverse consistency, and yet is still able to produce high-quality registration results—in this work, we found FNIRT registration produced mean and max inverse consistency errors of 3.02 mm and 18.7 mm, respectively, across 10 subjects. In some cases, it may even be that there is one correct definition of reference and moving image. For example, the generative model underlying registration to a template is only one-way—that the template transforms into the subject. As such, we do not place too much emphasis on the topic, but have included it in our description of MMORF for the sake of completeness, and because very poor inverse consistency is likely an indicator of other issues.

2.6. Multiresolution pyramid

MMORF employs a coarse-to-fine multi-resolution optimisation strategy (Bajcsy & Kovačič, 1989; Szeliski & Coughlan, 1997). This has been shown to be beneficial in avoiding local minima during optimisation, as well as accelerating convergence, across a wide variety of nonlinear registration tools (Andersson et al., 2007; Ashburner, 2007; Avants et al., 2008; Modat et al., 2010; H. Zhang et al., 2006). In principle, such approaches try to match low-frequency image information first, followed by progressively higher frequency information at each subsequent level. Traditionally, users are required to specify a downsampling factor and amount of smoothing for each level of the pyramid. The lower-resolution warp is then defined on the coarser, downsampled reference image grid. A potential pitfall of this approach is that if insufficient smoothing is applied to the image, then the process of downsampling introduces aliases in the information being aligned due to violation of the Nyquist criterion. MMORF overcomes this by defining the pyramid according to warp resolution (knot spacing) and image smoothing only. An acceptable level of subsampling is then automatically determined by treating the Gaussian smoothing kernel as a low-pass frequency filter. Therefore, the user can iteratively optimise the registration of different frequency content within the image through applying decreasing amounts of smoothing while keeping the warp resolution the same, without any aliasing problems. In practice, we restrict the downsampled resolution to be at least as high as, and at most four times higher than, the warp resolution at each level of the pyramid.

2.7. Reproducibility

MMORF is fundamentally a deterministic registration method, and running a registration with the same inputs and configuration will result in an output warp field that is practically identical. None of the cost functions rely on random undersampling or initialisation, and no stochastic optimisation is performed. The only sources of potential variability are the reduction (summing) operations, where multiple floating point numbers are reduced to a single value. In these cases, the finite precision of floating point numbers introduces a sensitivity to the order in which the values are summed, and this order is subject to a degree of randomness (particularly since these reductions are largely performed on the GPU in MMORF). Additionally, MMORF uses 32 bit (single precision) arithmetic for GPU compatibility and performance reasons, which is more susceptible to these effects than 64 bit (double precision) arithmetic. However, the effect is truly negligible, with maximum differences in displacement across the whole brain typically on the order of 1 μm .

2.8. Other implementation considerations

MMORF is written in C++ and makes extensive use of GPU parallelisation through Nvidia's CUDA framework (NVIDIA, 2023). It uses a mixed computation model, with only the most time-consuming components of the registration algorithm being ported to the GPU. These include:

- Image interpolation
- Cost, gradient and Hessian calculations
- Solving the system of linear equations for update steps

Of these, the Hessian calculation is by far the most computationally complex, with those of the regularisation and the rotational component of the DTI cost functions being particularly burdensome. For these calculations, GPU acceleration is on the order of 20-40x (depending on warp resolution, image dimensions, and image subsampling (Lange et al., 2020; Supplementary Material: GPU Considerations and Code Profiling)). Without the use of GPU parallelisation, these aspects of the registration would be too computationally burdensome to allow MMORF registrations to complete within a reasonable runtime. With GPU acceleration, a typical 1 mm isotropic registration with MMORF takes ≈ 5 minutes for a single scalar, unimodal image pair and ≈ 45 minutes for a single scalar, single tensor, multimodal image pair (when run on an Nvidia Titan V GPU). However, this reliance on CUDA

means that MMORF is only supported on Linux devices with Nvidia GPUs, and there is no CPU implementation currently available. A CPU port is under development, but it will be some time before it is released.

MMORF has been designed from the outset for integration into the FSL suite of neuroimaging tools (Jenkinson et al., 2012). As such, it expects inputs in FSL convention. Specifically, affine matrices between input images and the warp space should be in FLIRT format and DTI images should be in dfit format. MMORF output warp fields follow existing FSL conventions and can therefore serve as drop-in replacements for FNIRT warp fields in FSL commands such as `applywarp`.

As stated previously, a limitation of MMORF is that it requires a CUDA-capable Nvidia GPU to run. MMORF is, however, compiled to be highly backwards compatible with older generation GPUs, and should run (albeit $\approx 2\times$ slower) on devices with microarchitectures as old as 2012's Kepler. As such, even a machine with a basic, consumer-level GPU will be capable of running MMORF. The amount of GPU RAM will dictate at what warp resolution you need to switch from the LM to the MM optimisation strategy. Changing to MM at a 2 mm warp resolution (as we do here) requires ≈ 8 GB of GPU RAM. Changing to MM at 4 mm instead would require only ≈ 1 GB of GPU RAM.

3. THEORETICAL DIFFERENCES BETWEEN METHODS

With an understanding of the design considerations that went into MMORF, we now briefly describe how some of these choices compare to three existing registration tools, namely: FNIRT (Andersson et al., 2007), ANTs (Avants et al., 2008), and DR-TAMAS (Irfanoglu et al., 2016). These three tools are then used to validate and evaluate the performance of MMORF in Section 4.

FNIRT was chosen as it is the predecessor to MMORF, and the current nonlinear registration tool in FSL. Note that, although MMORF draws inspiration from FNIRT in terms of how the image registration problem is formulated, there is almost no code shared between the methods. The largest differences between these methods are MMORF's multimodal capabilities, regularisation, improved inverse-consistency, and GPU parallelisation. In terms of similarities, they share the same transformation and bias field models, and the same optimisation strategy at low resolutions. At higher resolutions, FNIRT switches to a Scaled Conjugate Gradient algorithm (Möller, 1993), in contrast to MMORF's MM approach. Finally, FNIRT performs a simultaneous optimisation of both warp and bias fields, whereas these are optimised in an interleaved manner by MMORF.

(i.e., MMORF assumes the warp field remains constant when estimating the bias field, and vice versa). This is because simultaneous optimisation of the warp and bias fields results in a Hessian without the regular diagonal structure, on which MMORF relies for efficient GPU parallelisation. In practice, despite some similarities in design choices, they perform very differently—even when MMORF is run, like FNIRT, in a unimodal configuration. The resulting warps have a very different character, which we largely attribute to the superior regularisation metric employed in MMORF. MMORF's inputs and output files are fully compatible with those of FNIRT, and it can therefore serve as a drop-in replacement in FSL analysis pipelines.

ANTs was chosen due to its consistently high performance, including in dedicated registration comparisons (Klein et al., 2009). It has become a de facto standard for nonlinear registration in much of medical imaging, and serves as a benchmark against which to compare MMORF's performance. ANTs is a purely scalar registration method, although it can be applied to multiple scalar input modalities simultaneously. It uses a symmetric, greedy approximation of large deformation diffeomorphic metric mapping (LDDMM) known as symmetric normalisation (SyN). At each iteration, an update step is composed with the current warp until convergence is reached. If each update step is diffeomorphic, then the composition of all updates is also diffeomorphic (apart from arithmetic floating point errors). Symmetry is achieved by registering both reference and moving images to a mid-space at each iteration. There are a number of similarity metrics implemented, but we will limit ourselves to locally normalised cross-correlation (LNCC), as this has proven to perform best in previous studies. ANTs regularisation consists of simple Gaussian smoothing, which may be applied independently to both update steps and to the final deformation field.

DR-TAMAS was chosen as it is currently the only other tool, to our knowledge, that can simultaneously register both tensor and scalar modalities in a single framework. It has also proven to match or exceed the best performing diffusion registration tools currently available. As in MMORF, finite-strain reorientation of the tensors is taken into consideration during each update step (i.e., it contributes to the gradient of the cost function, and is not simply applied after each update step). In contrast to MMORF, the DTI dissimilarity is divided into two separate cost functions—the trace similarity (a rotationally invariant scalar), and the deviatoric tensor similarity (sensitive to relative tensor orientation). For scalar inputs, DR-TAMAS uses an LNCC cost function. The transformation model and optimisation strategy are the same as ANTs, and therefore DR-TAMAS can be

considered as a truly multimodal variant of ANTs. In many ways, then, DR-TAMAS is the natural alternative to MMORF.

4. VALIDATION

In order to validate MMORF, we compare its registration performance to the three nonlinear registration tools presented in Section 3, namely, FNIRT, ANTs, and DR-TAMAS. ANTs was run in two configurations—“standard,” using only a T1w image to drive registration (hereafter referred to simply as “ANTs”), and “multimodal,” using both T1w and FA images to drive registration (hereafter referred to as “ANTs-MM”). Additionally, we include results for the linear registration tool FLIRT. The inclusion of FLIRT helps contextualise the results from the nonlinear methods—that is, it acts as a reference value when comparing differences in a metric between nonlinear methods.

Benchmarking a registration tool can be difficult to do well. Methods often appear to perform best when evaluated using metrics based on the same data that drove the registration (Irfanoglu et al., 2016), which risks introducing a degree of circularity in these types of evaluations. We have, therefore, endeavoured to perform a holistic evaluation of registration performance by including a range of structural-, diffusion-, functional-, and morphometry-derived metrics. Structural and diffusion metrics are the most circular, since these are the modalities that drive the registration (either individually or jointly) in all methods—however, they may highlight both the value and the pitfalls of using the data you wish to analyse to drive alignment. Functional metrics are based on a fully held-out modality (that is not seen by any of the registration tools) and therefore serves as the best proxy of true consistency of anatomical alignment across individuals—since overfitting to a driving-modality (e.g., unrealistically deforming the brain to make structural images appear very similar) is likely to have a negative effect on functional alignment (Coalson et al., 2018; Robinson et al., 2018). Morphometry metrics (e.g., measures of distortion that represent how aggressively the images are deformed) are essential to contextualise and interpret the accuracy/similarity metrics—that is, it is important to know how aggressively a tool needs to deform an image to achieve a certain degree of registration accuracy, with less distortion being preferred.

We chose the Human Connectome Project (HCP) young adult 1200 release (100 unrelated subjects subset) dataset as the basis for our testing (Glasser et al., 2013; Van Essen et al., 2012, 2013). The HCP dataset contains high-quality T1w (0.7 mm isotropic), diffusion (1.25 mm isotropic), and task-fMRI (2.0 mm isotropic) scans. This allows both unimodal T1w and multimodal T1w + DTI

registration to be conducted with the same dataset, as well as evaluating registration metrics based on all three modalities. The minimally preprocessed HCP data were used as far as possible (Glasser et al., 2013), which includes motion and distortion correction with FSL topup and eddy (Andersson & Sotiropoulos, 2016; Andersson et al., 2003), and coregistration of the diffusion data to T1w space. The task-fMRI data were, however, reprocessed in subject T1w space with no smoothing (rather than in MNI-152 space with 2.0 mm isotropic smoothing). The diffusion tensor model was fit to the $b = 1000$ s/mm² shell only (since we assume a Gaussian diffusion process without kurtosis) using FSL dtifit.

The Oxford-MultiModal-1 (OMM-1)² template was used as the reference space for all registrations (Arthofer et al., 2023). This template was constructed from 240 UK Biobank (K. L. Miller et al., 2016) subjects, and has both T1w and DTI volumes that are intrinsically co-registered. The OMM-1, therefore, provides a common space in which to compare methods that either use T1w images only, or a combination of T1w and DTI/DTI-derived images to drive the registration.

We calculate two structural accuracy metrics based on pairwise label similarity. The first is a measure of overlap, specifically, the Jaccard Index (JI). The second is a measure of surface distance, specifically, the Modified Hausdorff Distance (MHD) (Dubuisson & Jain, 1994). These are calculated pairwise on automatically generated cortical and subcortical labels for each pair of subjects following resampling to template space (see Section 4.2.3 for details). The JI for two regions A and B is defined as the intersection of those regions over their union. That is:

$$JI = \frac{A \cap B}{A \cup B}, \quad (9)$$

and ranges from 0 for no overlap to 1 for perfect overlap. The MHD is defined as the maximum of the two average directed distances between the surfaces of the regions being compared. That is:

$$MHD = \max \left(\frac{1}{N_a} \sum_{a \in A} \min_{b \in B} \|a - b\|, \frac{1}{N_b} \sum_{b \in B} \min_{a \in A} \|b - a\| \right) \quad (10)$$

The labels are derived from the T1w images, and therefore we expect them to favour scalar registration methods that can match the T1w images without being penalised if that reduces DTI similarity. We have avoided any simple intensity-based or tissue-type overlap metrics, as these are known to correlate poorly with true anatomical consistency (Rohlfing, 2012).

We then calculate three diffusion accuracy metrics that compare the similarity of the template DTI modality with each subject's DTI data after resampling to template space. Each subject is compared voxelwise to the template, and the average across all voxels within a mask is taken as the overall metric for that subject. Overall tensor similarity (OVL, Equation 11) is the first metric. OVL balances directional and magnitude similarity between tensors, and is a good general indicator of the similarity between two tensors. Linear-shape weighted V1 similarity (CLV1, Equation 13) is the second metric. CLV1 is defined as the inner product of the first eigenvector of the diffusion tensor (V1) from the template with V1 from the warped subject, weighted by the coefficient of linear shape (CL) for the template tensor (i.e., how similar V1 is, weighted by how informative V1 is). Planar-shape weighted V3 (CPV3, Equation 15) is the third metric. CPV3 is defined as the inner product of the third eigenvector of the diffusion tensor (V3) from the template with V3 from the warped subject, weighted by the coefficient of planar shape (CP) for the template tensor (i.e., how similar V3 is, weighted by how informative V3 is). As these measures of tensor shape are less commonly encountered in the DTI literature, we provide a visual depiction of CL and CP maps from the OMM-1 template in Figure 1 to aid intuition. In areas where CL is large, the tensor shape is largely prolate (cigar-shaped), and therefore the direction of maximum diffusivity (i.e., V1) is well defined. In areas where CP is large, the tensor shape is largely oblate (plate-shaped), and therefore the direction of minimum diffusivity (i.e., V3) is well defined. CLV1 and CPV3 specifically probe how well orientational information has been aligned, which is most relevant in white matter regions. We expect those tools that utilise the DTI data during registration to perform best in these metrics.

The fMRI accuracy metric used is task-fMRI cluster mass. This measures how consistently the registration tools are able to align those regions of the brain that are significantly activated or deactivated (compared to baseline activity) when performing a task. This assumes a general correspondence between brain structure and function, but this need not be exact. The benefit of this metric is that it is entirely independent of the modalities driving registration (i.e., T1w and DTI) and, therefore, there is little to no circularity in its interpretation, which cannot be said for the previous metrics.

Finally, we calculate both size (log-Jacobian determinant, $\log|J|$) and shape (cube-volume aspect ratio, CVAR—see Section 4.2.6 for definition) distortion metrics to evaluate how much each tool has had to deform the subject's images to match the template. For any given level of accuracy (i.e., the preceding metrics), a smaller amount of distortion is usually preferred, as this indicates

² DOI 10.17605/OSF.IO/S9GE4

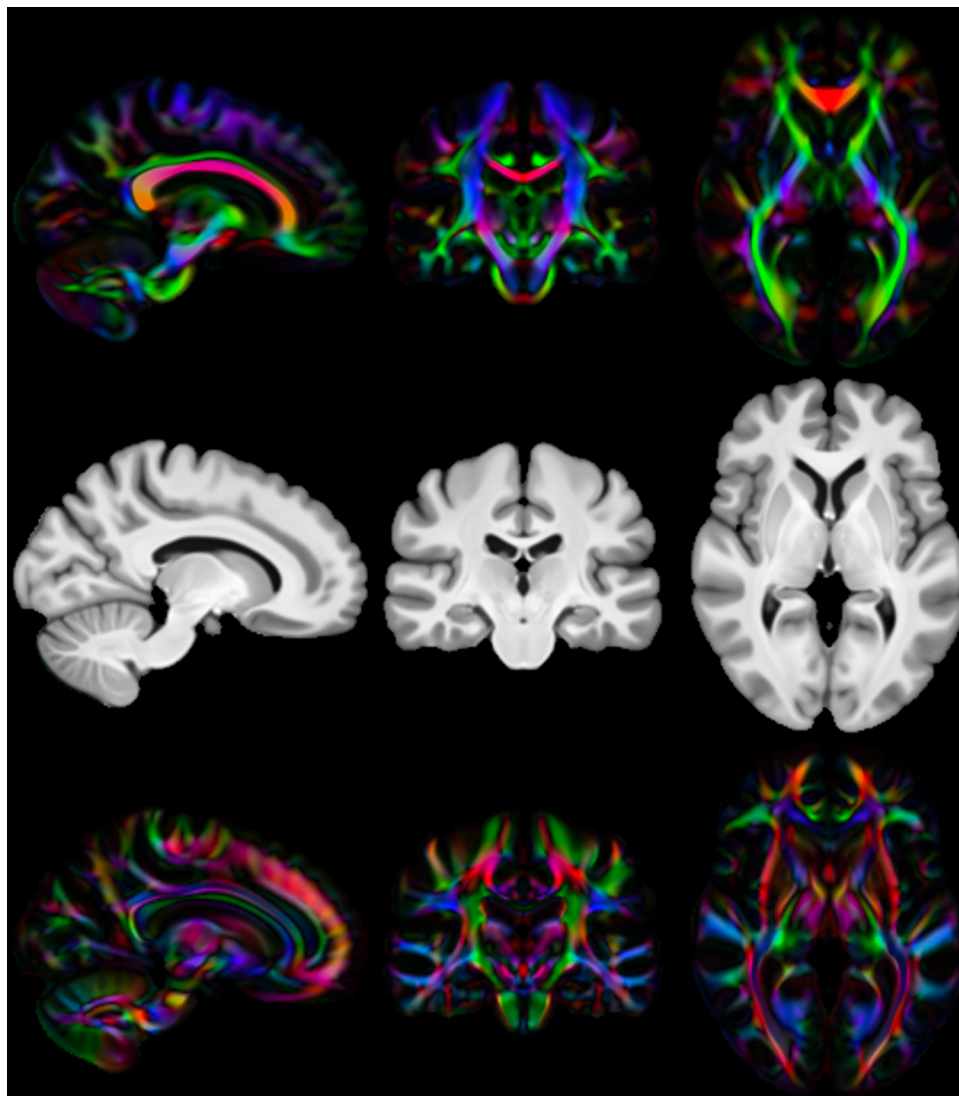


Fig. 1. Visualisation of linear and planar shape coefficients in OMM-1 space: From top to bottom: Linear coefficient (CL) map of the template, T1w template for reference, planar coefficient (CP) map of the template. Images are directionally encoded colour maps of V1 and V3, respectively. Green = Anterior-Posterior, Red = Left-Right, Blue = Inferior-Superior. Both CL and CP values are highest in the white matter, but cover complementary regions therein. CLV1 is most sensitive to how well the primary diffusion direction is matched in voxels where diffusion occurs parallel to that axis only. CPV3 is most sensitive to how well the tertiary diffusion direction is matched in voxels where diffusion occurs within a plane defined by the primary and secondary diffusion directions. Together, these represent the voxels where tensor orientation can be reliably described by a single direction.

that the registration method is changing the original data as little as possible.

4.1. Ethics

All human imaging data used in this work are part of the open access Human Connectome Project Young Adult (HCP-YA) dataset. The data are pseudonymised and identifiable visual features, such as the face and ears, are obscured. Written informed consent to share the data “broadly and openly” was obtained for all participants by the original HCP-YA researchers (Elam et al., 2021).

4.2. Methods

Each of the following steps was performed independently for each registration method.

4.2.1. Registration

All registration was performed to the OMM-1 template. This template contains both T1w and DTI volumes with isotropic resolution of 1 mm. Since both modalities were jointly aligned during template creation, they are intrinsically spatially consistent at the voxel level (Arthofer et al.,

2023). It is therefore a good choice for registration with both unimodal and multimodal tools. Each tool was used to register all of the 100 unrelated HCP subjects to the template. FLIRT, FNIIRT and ANTs used the T1w image only, ANTs-MM used both the T1w and FA images, while MMORF and DR-TAMAS used both the T1w and DTI images. The FLIRT affine matrix was used to initialise both FNIIRT and MMORF, whereas ANTs/ANTs-MM and DR-TAMAS used their own affine registration methods. FNIIRT and MMORF were run with custom configurations identified to empirically produce good results, and bias-field estimation was enabled in both cases. ANTs/ANTs-MM and DR-TAMAS were run with slightly modified configurations that were found to improve registration accuracy over the defaults (where available). The ANTs/ANTs-MM parameters are based on those used in Klein et al. (2009)—although we ran the newer antsRegistration implementation (Avants et al., 2014) and there is, therefore, not an exact one-to-one mapping between options³. DR-TAMAS parameters are based on those provided in the “dtireg_batch_register_to_target_with_structurals” script that is installed with the software. In both cases, the only modification was to make the convergence criteria slightly more strict (allowing for more iterations of the optimisation), which was found to improve visual registration quality in a few cases. All methods were run using a multi-resolution pyramid approach to a final warp resolution of 1 mm isotropic.

4.2.2. Hyperparameter selection

It may be beneficial for readers of this manuscript and users of MMORF to know how we selected the default parameters for MMORF, since this has not been previously documented. Optimisation was carried out over three hyperparameters, namely, regularisation at 1 mm isotropic warp resolution (arbitrary units), regularisation change relative to warp resolution, and smoothing (in mm, FWHM) relative to warp resolution. Only T1w images were used to drive the registration, and their weighting was set to 1 (arbitrary units). Registration quality was evaluated on UKB data (i.e., not the data used for evaluation in the present paper) using tfMRI Cluster Mass from the contrasts generated by the Hariri Faces and Shapes paradigm (Hariri et al., 2002). A coarse search over the three hyperparameters found that the best performance was achieved using a regularisation of 0.18 at 1 mm isotropic warp resolution, dividing regularisation by 0.85 for each doubling of warp

resolution (e.g., 0.21 at 2 mm, 0.25 at 4 mm, etc.), and smoothing by a quarter of warp resolution (e.g., 4 mm FWHM smoothing at 16 mm warp resolution, 2 mm FWHM smoothing at 8 mm warp resolution, etc.). Additional modalities are included with a default weighting of 1 (i.e., the same as T1w).

4.2.3. T1w FreeSurfer label overlap

Automatically segmented subcortical (ASEG atlas; Fischl et al., 2002, 2004) and cortical (Destrieux 2009 atlas; Destrieux et al., 2010) parcellations for each subject were warped into template space. Jaccard indices (measuring label overlap) and modified Hausdorff distances (measuring the average error in label boundaries) of the corresponding warped parcellations were calculated for every possible pair of subjects. A pairwise approach was used because there are no target labels in template space. The average value of JI and MHD for each parcellation was then calculated across pairings for each subject, resulting in 100 values for each parcellation.

Although ASEG segmentations demonstrate good reliability (with Sederevičius et al. (2021) finding scan-rescan Jaccard index values generally >0.82), they are themselves partly registration based and imperfect. Between-subject measures of overlap will therefore be affected both by how accurately the subjects are aligned, and by errors in segmentation. While this prevents us from being able to say anything conclusive about the accuracy of any single registration tool, if we assume that segmentation errors are independent of registration errors for all tools, then we can still compare these overlap measures to gauge the relative performance of each method.

A significant difference in median performance between each method and MMORF was tested using a Wilcoxon signed-rank test. Degrees of freedom of the data were adjusted down to 100 for this testing (a conservative estimate that assumes overlap scores of individual parcellations are not independent within-subject).

4.2.4. DTI similarity

Combined affine and nonlinear (affine only for FLIRT) warp fields were used to resample each subject's DTI volume into template space. Resampling of tensors was performed using the FSL tool vecreg, which includes preservation of principle direction tensor reorientation (D. C. Alexander et al., 2001). This is the most accurate way of accounting for both rotational and shear effects of the warps when resampling DTI data. Tensors were then decomposed into three Eigenvalue-Eigenvector pairs (L1/2/3 and V1/2/3, respectively). OVL, CLV1, and CPV3 were then calculated between each subject's

³ For recommendations on converting between legacy ANTs and modern antsRegistration calls see <https://github.com/ANTsX/ANTs/wiki/ANTS-and-antsRegistration>

warped DTI volume and the template. A pairwise approach was not necessary here since the DTI template volume itself acts as the target. CL and CP weighting coefficients (A. L. Alexander et al., 2000) were generated from the template DTI volume only. V1 and V3 similarity was calculated as the magnitude of the dot product of template and warped subject vectors in each template voxel. Similarity metrics were calculated per voxel according to the following formulae:

$$OVL = \frac{\sum_{i=1}^3 L_i L_m \langle V_i, V_m \rangle^2}{\sum_{i=1}^3 L_i L_m} \quad (11)$$

$$CL = \frac{L1_r - L2_r}{L1_r + L2_r + L3_r} \quad (12)$$

$$CLV1 = CL \langle V1_r, V1_m \rangle \quad (13)$$

$$CP = \frac{2(L2_r - L3_r)}{L1_r + L2_r + L3_r} \quad (14)$$

$$CPV3 = CP \langle V3_r, V3_m \rangle \quad (15)$$

where the subscripts r and m represent the reference and resampled moving image, respectively. All metrics were calculated within the template brain mask only.

A significant difference in median performance between each method and MMORF was tested using a Wilcoxon signed-rank test.

4.2.5. *tfMRI cluster mass*

The HCP task battery (Barch et al., 2013) consists of 7 tasks from which 86 contrasts are derived, and used to generate contrast of parameter estimate (COPE) images in subject T1w space. For each subject, the 86 COPE images were resampled into template space. FSL Randomise (Winkler et al., 2014) was then used to perform a group-level, non-parametric, ordinary-least-squares, random-effects, one-group t -test on the mean COPE image (across subjects) for each contrast. The results of the group-level processing are 86 t -statistic maps/images and 86 family-wise-error (FWE) corrected p -value maps/images (one per contrast). The t -statistic map represents the group activation for a particular contrast, where more accurately aligned activations lead to higher t -statistics. The p -value map represents the statistical significance of the t -statistic at each voxel.

Cluster mass was then calculated as follows. Thresholding was applied to the FWE-corrected p -value map (at $p < 0.05$) for each contrast, to give a binary mask of significantly activated regions. This mask was then applied

voxel-wise to the t -statistic map. The masked t -statistic map was then multiplied voxel-wise with a grey matter mask of the template, and summed across all voxels. The summed value is what we refer to as cluster mass, and is a single number per contrast and registration method. Cluster mass is, therefore, increased by both higher t -statistics and better alignment of grey matter. It can be calculated independently for each registration method, simplifying between-method comparisons.

A significant difference in median percentage difference in cluster mass relative to MMORF was tested using a Wilcoxon signed-rank test. Degrees of freedom of the data were adjusted down to 43 for this testing (assuming positive and negative contrasts are not independent).

4.2.6. *Distortion*

Two measures of distortion are considered, both evaluated within the template brain mask.

The first is the 5th to 95th percentile range of the log-Jacobian determinant ($\log|J|$). This is a measure of volumetric distortion. Since the histogram of $\log|J|$ tends to be centred around zero, the mean is uninformative and, therefore, a range is more useful. This is a fairly robust measure of the extent to which the local effect of the warp field is to expand/contract voxels (and is equally sensitive to both).

The second is the average cube-volume aspect ratio (CVAR; Smith & Wormald, 1998), defined as:

$$CVAR = \sqrt[3]{\frac{s_{i-max}^3}{s_1 \times s_2 \times s_3}} \quad (16)$$

where s_i are the singular values of the local Jacobian matrix. CVAR represents the cube-root of the ratio of the volume of the smallest regular cube (i.e., a cube with sides of equal length) that can fully enclose a cuboid, to that cuboid's own volume. This is a pure shape-distortion measure that is invariant to volumetric changes, and extends the concept of 2D aspect ratio to higher dimensions. For a perfect cube its value is 1, and it is greater than 1 for any shape where one or more sides of the cube are a different length to the others. Since it is always ≥ 1 , the mean across voxels is a meaningful measure of the extent to which the local effect of the warp field is to alter the original shape of the underlying voxels.

Between them, these two measures represent a good summary of the extent to which the warp is distorting the anatomy of the brain to achieve a given degree of registration accuracy. In general, lower distortion for a given level of accuracy is preferred. When both distortion and accuracy increase, then a more nuanced interpretation is required based on the relative changes in each.

Table 1. Summary of results for all domains.

| | FreeSurfer labels | | | | DTI | | | tfMRI | Distortion | |
|----------|-------------------|---------------|---------------|---------------|----------------|----------------|----------------|---------|----------------|----------------|
| | Subcort | | Cort | | OVL | CLV1 | CPV3 | CM | J | CVAR |
| | Jl | MHD | Jl | MHD | | | | | | |
| FLIRT | 0.46* | 0.73* | 0.20* | 2.05* | 0.669* | 0.802* | 0.755* | 53.36* | - | - |
| | (0.24) | (0.81) | (0.09) | (1.03) | (0.030) | (0.022) | (0.021) | (58.36) | (-) | (-) |
| FNIRT | 0.61* | 0.37* | 0.41* | 1.12* | 0.776* | 0.873* | 0.854* | 11.94* | 1.86* | 1.50* |
| | (0.27) | (0.60) | (0.15) | (0.80) | (0.017) | (0.012) | (0.010) | (14.22) | (0.140) | (0.029) |
| ANTs | 0.65 | 0.35 | 0.40 | 1.10 | 0.802* | 0.886* | 0.867* | 5.50* | 1.20 | 1.37* |
| | (0.24) | (0.60) | (0.14) | (0.79) | (0.018) | (0.011) | (0.010) | (8.58) | (0.071) | (0.021) |
| ANTS-MM | 0.64* | 0.35* | 0.39* | 1.10 | 0.815* | 0.894* | 0.872 | 2.01* | 1.35* | 1.41* |
| | (0.23) | (0.54) | (0.13) | (0.79) | (0.014) | (0.008) | (0.007) | (3.40) | (0.069) | (0.018) |
| DR-TAMAS | 0.64* | 0.37* | 0.37* | 1.19* | 0.817* | 0.896* | 0.870* | 1.31* | 1.39* | 1.37* |
| | (0.26) | (0.61) | (0.14) | (0.820) | (0.012) | (0.008) | (0.008) | (4.83) | (0.089) | (0.022) |
| MMORF | 0.66 | 0.31 | 0.40 | 1.10 | 0.825 | 0.900 | 0.872 | - | 1.20 | 1.35 |
| | (0.27) | (0.60) | (0.14) | (0.76) | (0.015) | (0.008) | (0.007) | (-) | (0.062) | (0.014) |

Structural (FreeSurfer labels): median Jaccard Index (Jl) and Modified Hausdorff Distance (MHD) across subjects for both subcortical and cortical labels. Diffusion (DTI): median overall tensor similarity (OVL), linear coefficient-weighted V1 similarity (CLV1) and planar coefficient-weighted V3 (CPV3) similarity. Functional (tfMRI): Median percentage difference in cluster mass (CM) for all contrasts compared to MMORF. Distortion: median 5th to 95th log-Jacobian determinant range (|J|) and cube-volume aspect ratio (CVAR) across subjects. The best performing method in each metric is shown in bold. Methods that perform significantly differently to MMORF (Wilcoxon Signed-Rank test, $p < 0.05$) for a given measure are indicated with an asterisk*. Interquartile ranges are shown in brackets below the median values. In all cases, apart from cortical Jl, MMORF is the best or joint-best performing method (although not always significantly so).

A significant difference in median performance between each method and MMORF was tested using a Wilcoxon signed-rank test.

4.3. Results

The results across all domains and metrics are summarised in Table 1. The median of each measure is quoted, with the interquartile range given in brackets below. The best performing method for each task is highlighted in bold. Results that differ significantly from MMORF (Wilcoxon signed-rank test, $p < 0.05$) are indicated with an asterisk*. More detailed descriptions with accompanying figures are presented in the subsections that follow. Note that we do not discuss the affine (FLIRT) performance in detail, as it is expected that all methods will outperform it. However, we include the results as a reference against which to compare differences between nonlinear methods.

Many figures use raincloud plots (Allen et al., 2021) that simultaneously show the raw data, a box and whiskers plot, and a density estimate. In all cases, the interpretation of the box and whisker part of the plots is the same—the box shows the quartiles (25th, 50th, and 75th percentiles), the whiskers extend to the final non-outlier data point, and the diamonds show outliers (any point more than $1.5 \times$ the interquartile range from the edges of the box).

MMORF registration took 30 to 45 minutes per subject, and was run on a mixture of Nvidia A100 and V100 (Titan V equivalent) GPUs.

4.3.1. T1w FreeSurfer label overlap

The Jaccard index and modified Hausdorff distance results are presented in Figures 2 and 3, respectively. In all cases, nonlinear registration leads to a major improvement in both metrics. No nonlinear method performs so well as to indicate that there is an interaction that would cause inaccuracies in FreeSurfer's automated segmentation to be systematically better aligned. We therefore assume that any segmentation inaccuracies affect all methods equally.

MMORF and ANTs produce the best Jaccard index results in the subcortex, with FNIRT narrowly outperforming them in cortical regions. Modified Hausdorff distance performance is slightly better in both subcortical and cortical regions for MMORF, although not significantly different from that of ANTs/ANTS-MM. It is worth noting that subcortical contrast is relatively poor in T1w images from the HCP dataset, and certain structures (the left pallidum in particular) are poorly segmented in a number of subjects. This is the cause of the relatively heavy tails towards low Jaccard indices in Figure 2a. Since subcortical regions are far easier to define and align than cortical regions, such low Jaccard indices would be very unexpected for correctly segmented subcortical structures.

4.3.2. DTI similarity

OVL, CLV1, and CPV3 similarity results are presented in Figures 4 to 6, respectively. For OVL and CLV1, the trend

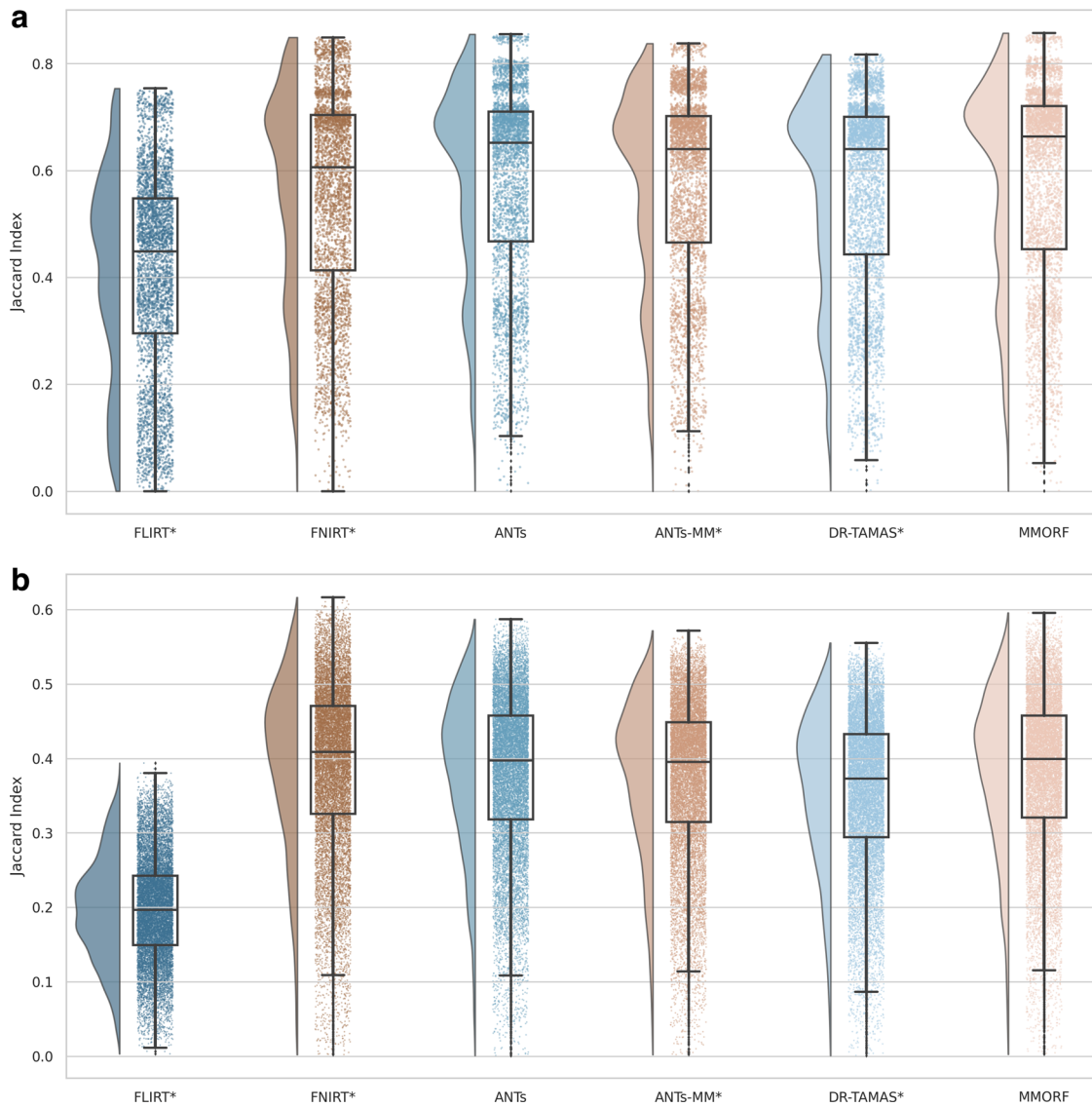


Fig. 2. Subcortical (a) and cortical (b) Jaccard indices for FreeSurfer segmentation overlaps (larger is better): All nonlinear methods improve over affine only, with the greatest improvement being in the cortex. Across all labels, MMORF and ANTs perform similarly (and best), with only FNIRT slightly outperforming them in the cortex. Methods that perform significantly differently to MMORF are shown with an asterisk* next to their name.

is the same: affine registration with FLIRT performs worst followed by FNIRT, ANTs, ANTs-MM, DR-TAMAS, and MMORF (in that order). For CPV3 the ordering is the same, except that ANTs-MM is able to match the performance of MMORF (this is the only comparison where the difference is not statistically significant).

4.3.3. *tfMRI* cluster mass

Cluster mass results are presented in the form of percentage difference plots, and warrant some guidance on their interpretation. Each point on the plot represents one contrast. When comparing method *A* to method *B*, the x-axis represents the cluster mass of method *A*. The x-axis is log-transformed to account for the large range in cluster

mass across contrasts. The y-axis represents the percentage improvement in cluster mass by method *A* over method *B*. Therefore, a point at location $(x,y)=(10,10)$ represents a contrast with a cluster mass of $\exp(10)$ for method *A*, and a 10 % improvement in cluster mass when using method *A* over method *B*. Similarly, a point at location $(x,y)=(15,-5)$ represents a contrast with a cluster mass of $\exp(15)$ for method *A*, and a 5 % reduction in cluster mass when using method *A* over method *B*. By choosing percentage difference for the y-axis, we remove the bias towards large cluster mass contrasts, which would otherwise dominate if *y* were instead the simple difference in cluster mass between methods.

Cluster mass comparisons for MMORF vs. FLIRT, FNIRT, ANTs, ANTs-MM. and DR-TAMAS are presented

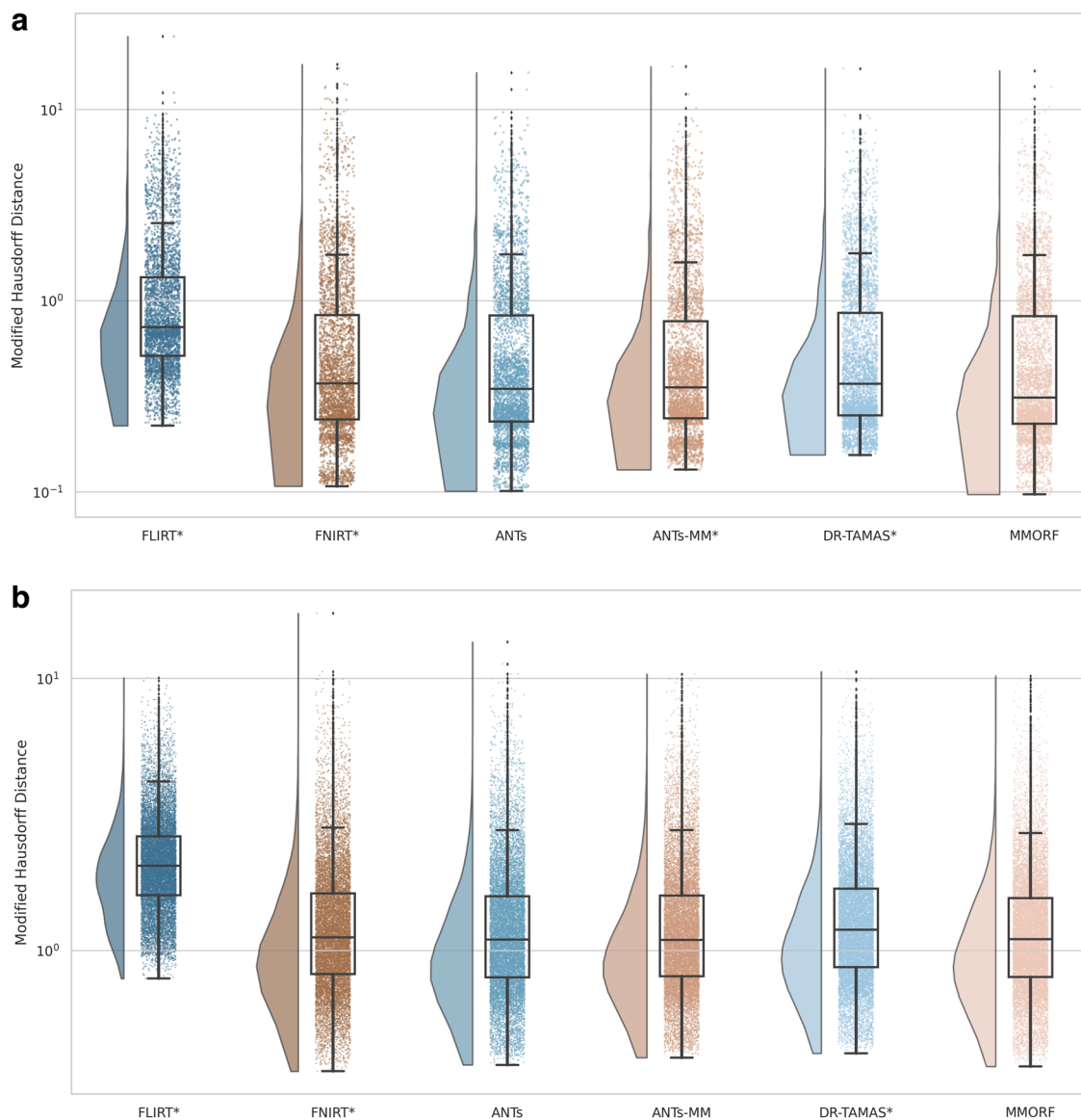


Fig. 3. Subcortical (a) and cortical (b) modified Hausdorff distances for FreeSurfer segmentation overlaps (smaller is better): All nonlinear methods improve over affine only, with larger improvements evident in the cortex. Performance is very similar across all methods, with ANTs and MMORF slightly outperforming the others. Methods that perform significantly differently to MMORF are shown with an asterisk* next to their name.

in [Figures 7 to 11](#), respectively. MMORF produces a large improvement across all contrasts compared to FLIRT, confirming that the CM metric is sensitive to improved registration accuracy. MMORF improves CM across most contrasts compared to FNIRT and, to a slightly lesser extent, ANTs. The distributions of the differences in CM for ANTs-MM are similar in shape to unimodal ANTs, but those differences are reduced across all CM sizes. MMORF and DR-TAMAS produce very similar results for contrasts with high CM (towards the right of the x-axis), but MMORF performs consistently better for small and medium CM contrasts. In all cases, the improvements in CM when using MMORF are statistically significant.

4.3.4. Distortion

Comparisons of the 5th to 95th percentile Jacobian determinant range (volumetric distortion) and the average CVAR (shape distortion), which are both evaluated only within the template brain mask, are presented in [Figures 12 and 13](#), respectively.

FNIRT shows significantly more volumetric and shape distortion than the other methods. ANTs and MMORF produce the lowest amounts of volumetric distortion. ANTs-MM produces more volumetric distortion than its unimodal counterpart, but slightly less than DR-TAMAS. MMORF produces less shape distortion than all the other

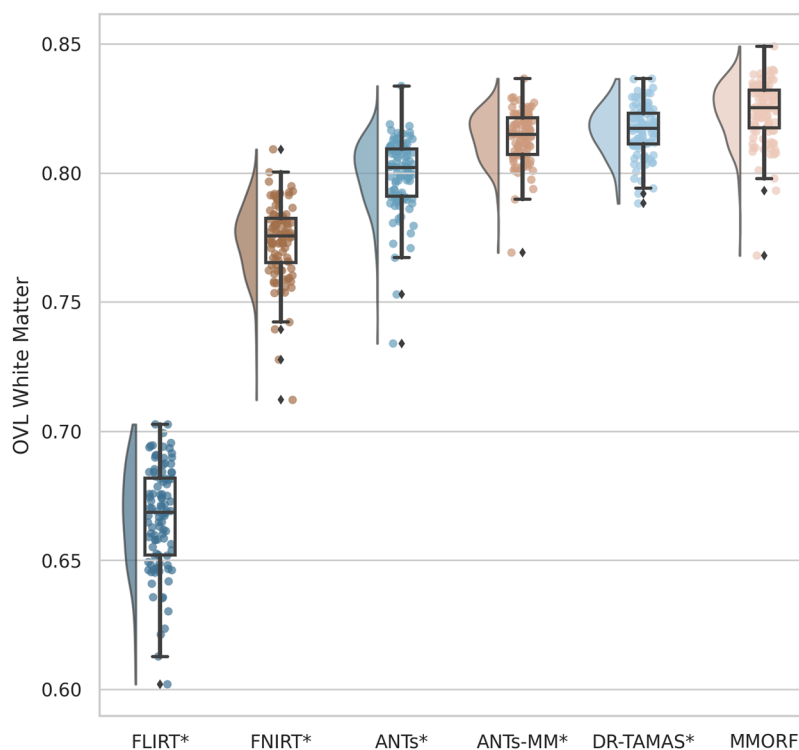


Fig. 4. Overall tensor similarity (OVL): Calculated within a mask of the template white matter, defined as $FA > 0.2$. All nonlinear methods improve over affine only, with those methods that include the full DTI-tensor data in the registration (MMORF and DR-TAMAS) outperforming the method that includes DTI-derived FA (ANTs-MM), which outperforms the T1w-only methods (FNIRT and ANTs). MMORF performs best overall. Methods that perform significantly differently to MMORF are shown with an asterisk* next to their name.

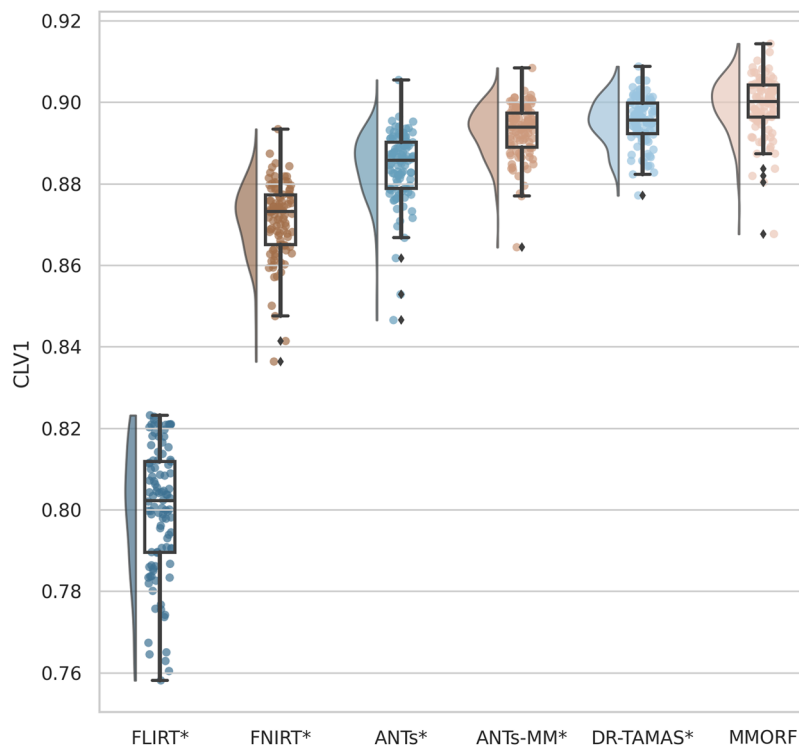


Fig. 5. Linear shape weighted V1 similarity (CLV1): All nonlinear methods improve over affine only, with those methods that include the full DTI-tensor data in the registration (MMORF and DR-TAMAS) outperforming the method that includes DTI-derived FA (ANTs-MM), which outperforms the T1w-only methods (FNIRT and ANTs). MMORF performs best overall. Methods that perform significantly differently to MMORF are shown with an asterisk* next to their name.

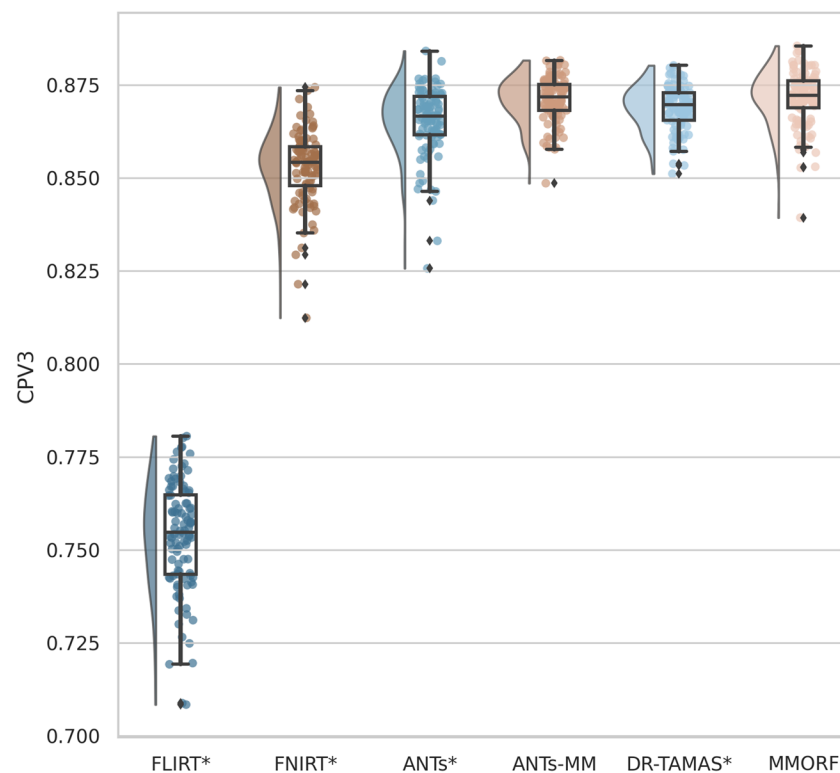


Fig. 6. Planar shape weighted V3 similarity (CPV3): All nonlinear methods improve over affine only, with those methods that include any DTI data in the registration (MMORF, DR-TAMAS and ANTs-MM) outperforming the T1w-only methods (FNIRT and ANTs). MMORF and ANTs-MM perform best overall. Methods that perform significantly differently to MMORF are shown with an asterisk* next to their name.

methods, followed by ANTs and DR-TAMAS. ANTs-MM produces the most shape distortion out of the three SyN-based methods.

The differences between ANTs, on the one hand, and ANTs-MM and DR-TAMAS, on the other, are likely attributable to larger deformations within the white matter, due to diffusion information (FA and DTI, respectively) driving the registration harder in that region.

4.4. Summary

Of the methods tested, MMORF is the most consistently high-performing across the full range of evaluation metrics.

It is expected that the T1w-only-driven registration methods would perform well when evaluated with a T1w-derived similarity metric (particularly in the cortex) and, indeed, both FNIRT and ANTs perform very well in the label overlap metric in this region. It is also possible that the inclusion of DTI information during registration might negatively affect such a metric and, again, we do indeed see that DR-TAMAS performs relatively poorly in the cortex, despite good subcortical performance. By contrast, ANTs-MM performs nearly as well as its unimodal counterpart in these evaluations, which may be

due to the FA modality having low signal in cortical areas, and therefore largely aligning those areas based on the same T1w information as unimodal ANTs. MMORF, does not seem to suffer cortically in the same way as DR-TAMAS either—performing on par with ANTs both cortically and subcortically. We may, therefore, conclude that MMORF is a good choice of method, on par with ANTs, when a structurally-derived segmentation comparison (e.g., VBM (Ashburner & Friston, 2000) or FSL-VBM (Douaud et al., 2007)) is the type of study for which registration is being employed.

When evaluating alignment in the white matter, the value of including diffusion information in the registration is clear, if unsurprising. MMORF, DR-TAMAS, and ANTs-MM noticeably outperform both FNIRT and ANTs across all DTI similarity metrics. Of the DTI-tensor registration methods, MMORF has the edge over DR-TAMAS across all metrics. Interestingly, the inclusion of FA has brought the performance of ANTs-MM close to that of DR-TAMAS for OVL and CLV1, and even exceeded it for CPV3 (matching MMORF in this metric), although this does come at the expense of an increase in shape-distortion. These results demonstrate that there are two factors that improve DTI similarity. The first is improved correspondence of the location of white-matter macrostructure (information that

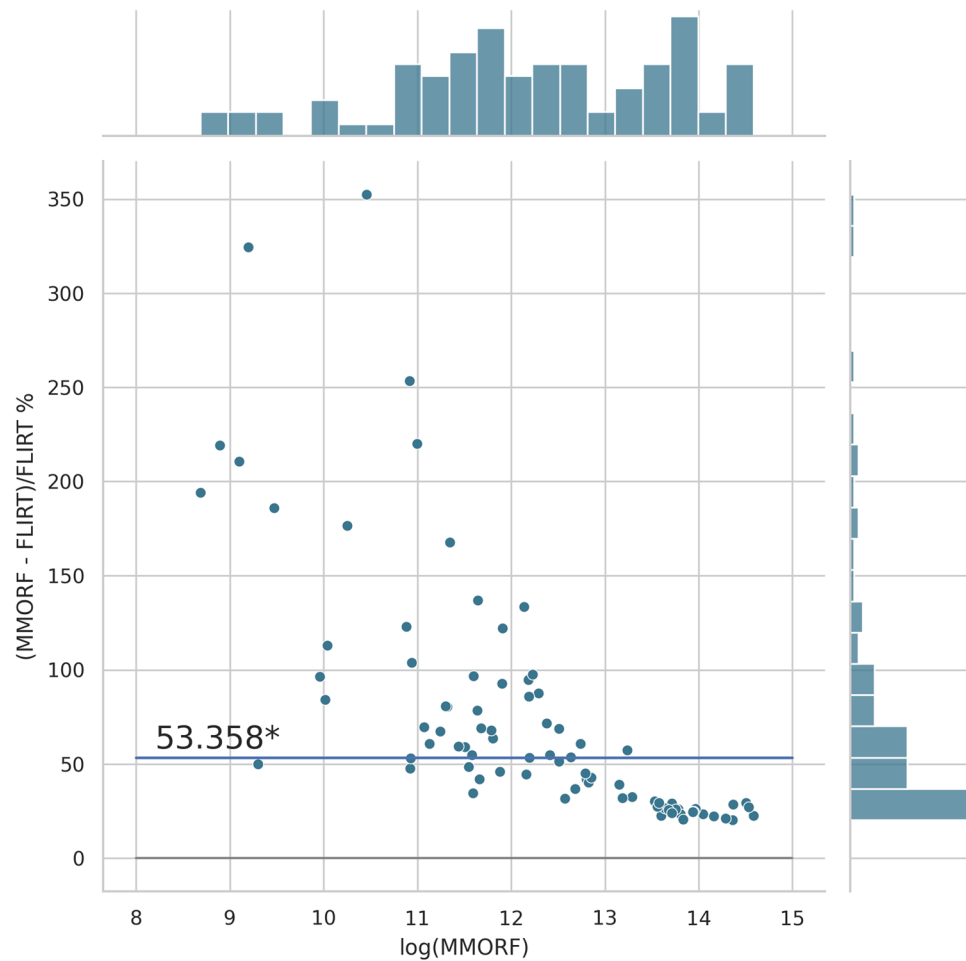


Fig. 7. MMORF vs. FLIRT cluster mass: Across clusters of all sizes, MMORF (nonlinear) outperforms FLIRT (linear), with median improvement in cluster mass of $\approx 53\%$ (the asterisk* indicating that the improvement is significant). This is not surprising, but serves to demonstrate how cluster mass is sensitive to registration accuracy.

is present in the FA modality). The second is improved correspondence of the directionality of white-matter microstructure (information that is not present in the FA modality). The addition of directional information (contained in the DTI modality) would appear to be most beneficial in matching the combined shape and direction of the tensors (measured by OVL), and the primary diffusion direction (measured by CLV1), but including FA seems sufficient to ensure the orientation of oblate tensors is well matched (measured by CPV3). Overall, including the full DTI modality, rather than the simpler tensor-derived FA, in MMORF appears beneficial to registration performance. Note, however, that this is the most circular of our evaluations, and while it is informative, care should be taken to interpret these results using the rest of our evaluations to add context (e.g., looking at the relative amount of distortion required to produce an improvement in DTI similarity).

We believe the tfMRI results to be the most unbiased assessment of registration accuracy presented here, since they are evaluated using a modality that was never seen by any of the registration methods being tested.

Since MMORF outperforms all other methods under test, this is a compelling indication that it is really achieving high registration accuracy (i.e., correct anatomical correspondence and not just image similarity). It is notable that FNIRT, the best performing method in terms of cortical label overlaps, does not perform as well in this (also cortical) metric. Similarly, DR-TAMAS is the closest performing method to MMORF in the tfMRI evaluation, but is the poorest performing method in cortical label overlap. This is a clear example of the value in not only considering the more circular evaluation metrics (i.e., those derived from the same modalities that are driving the registration) when comparing registration tools. Furthermore, the substantial increase in performance of the multimodal methods over the unimodal methods indicates that improved alignment within white matter leads to improvements in alignment of functional areas in the cortex.

Finally, in terms of distortion, MMORF demonstrates levels that are the same as or lower than those of ANTs, despite matching it in label overlap and bettering it in both the DTI and tfMRI metrics. DR-TAMAS and ANTs-MM

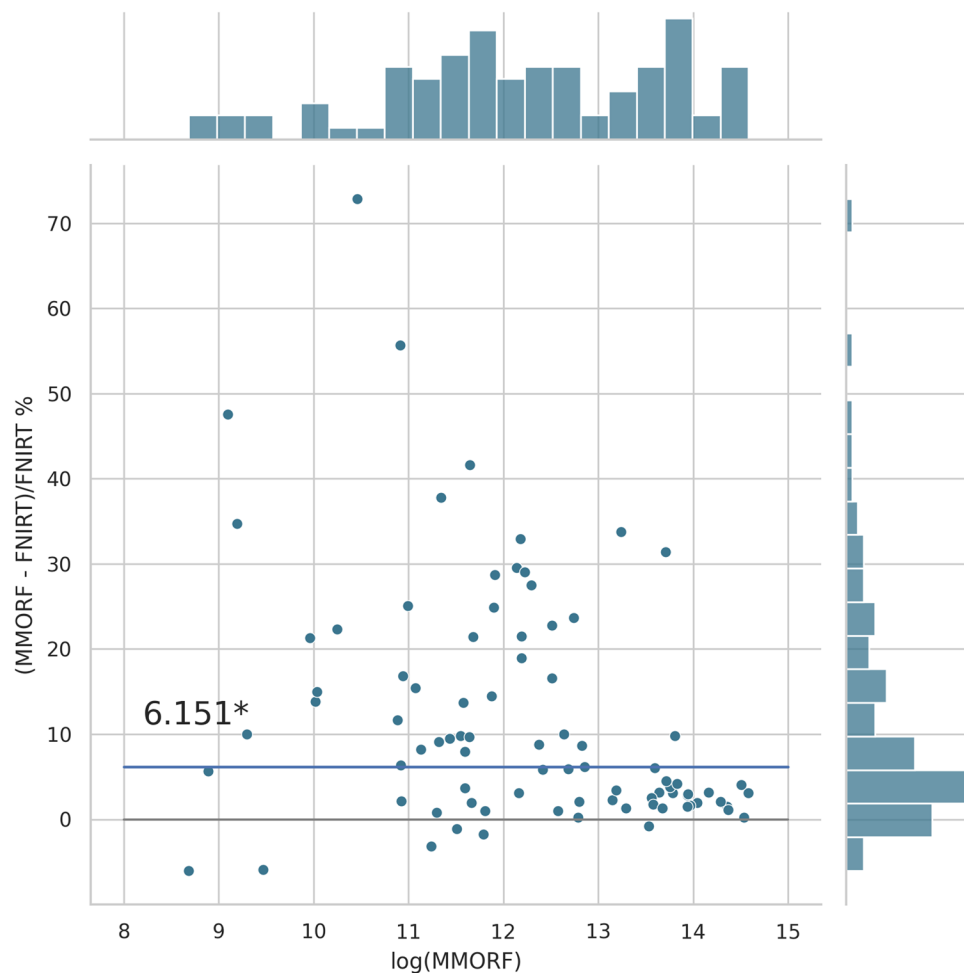


Fig. 8. MMORF vs. FNIRT cluster mass: MMORF outperforms FNIRT with median improvement in cluster mass of $\approx 6.2\%$ (the asterisk* indicating that the improvement is significant). Improvements are seen across all CM sizes, but the largest improvements are in the mid-sized clusters.

produce more distortion than their sibling method, ANTs, which is not unexpected given that they are also being driven by DTI information in the white matter. The substantial increase in shape distortion of ANTs-MM over DR-TAMAS would reasonably suggest that penalising misalignment of the direction of diffusion has a regularising effect on the warps when compared to the simpler task of matching FA. FNIRT produces larger distortions than the other nonlinear methods. Without MMORF for comparison, this might easily be interpreted as an inevitable consequence of driving a small deformation framework method very hard to maximise accuracy metrics. However, given that MMORF and FNIRT have the same transformation model, this should instead highlight the value of using a more biologically plausible regularisation model within the small-deformation framework.

5. DISCUSSION

The primary objective in developing MMORF was to exploit the rich information available in multimodal datasets in

order to align brain images with maximal accuracy. A truly multimodal tool must be able to leverage information about both magnitude and directionality in each voxel, since we are able to capture and represent both of these with MRI. To this end, MMORF is able to explicitly, and simultaneously, optimise both the displacement and rotational effects of a warp field. Importantly, not optimising for rotation (as is the case for any scalar registration method) does not mean that there is no rotation, only that it cannot be controlled to either improve directional alignment or prevent the introduction of misalignment. Simultaneously optimising over multiple modalities through a single warp both allows the unique information from each modality to influence the resulting transformation (as when each modality is aligned separately), and ensures that all modalities remain co-registered following warping (as when using only a single modality to drive alignment).

But even when combining information across modalities, image registration is still an underdetermined problem and, therefore, alignment accuracy (as measured by how well anatomy are co-localised) is highly dependent

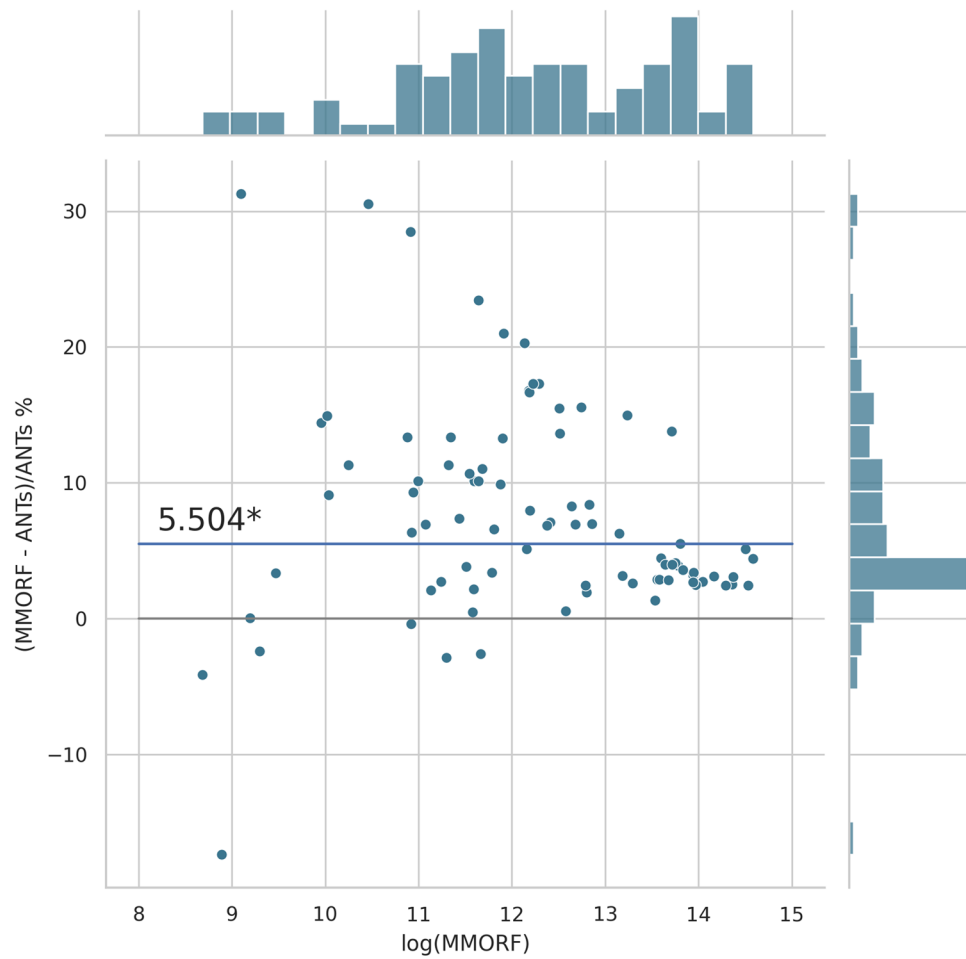


Fig. 9. MMORF vs. ANTs cluster mass: MMORF outperforms ANTs with median improvement in cluster mass of $\approx 5.5\%$ (the asterisk* indicating that the improvement is significant). Improvements are seen across all CM sizes, but the largest improvements are in the mid-sized clusters.

on the choice of warp regularisation. We therefore included in MMORF a regularisation model that promotes highly biologically plausible deformations, thereby effectively controlling excessive levels of distortion in both shape and size.

The combination of our multimodal approach and our regularisation metric is computationally intensive, particularly when it comes to calculating the Hessian matrix required by MMORF's Gauss-Newton optimisation strategy. We address this through GPU parallelisation of the most computationally intensive parts of the algorithm, allowing MMORF to run to 1 mm warp resolutions within 5 to 45 minutes (depending on image resolution and number of modalities).

We have evaluated MMORF across four domains—T1w-label overlap, DTI similarity, tfMRI cluster mass, and image distortion. Performance was benchmarked against several established registration tools—FLIRT, FNIRT, ANTs, ANTs-MM, and DR-TAMAS. These tools represent an affine reference, MMORF's predecessor, a state-of-the-art scalar method (run in both unimodal and multimodal

configurations) and the most similar multimodal alternative, respectively. A common theme was that methods performed well when tested on metrics derived from modalities used in their respective cost functions—for example, FNIRT performed well on T1w-label overlap, and DR-TAMAS performed well on DTI similarity. MMORF performed best or near-best in both scalar and tensor evaluations, best in the held-out tfMRI evaluation, and best in terms of distortion.

FNIRT was able to slightly outperform MMORF in cortical T1w-label overlap, but was outperformed by MMORF across all other metrics. FNIRT induces the most distortion out of the methods tested. This likely contributed to the relatively poor DTI similarity results, since excessive deformations can cause incorrect rotation of the tensors. Interestingly, having the best cortical label overlap has not translated to the best tfMRI performance, with FNIRT actually showing the poorest performance, despite both domains being evaluated in the cortex. The strong cortical T1w-label performance is, therefore, likely due to overfitting of the T1w image similarity metric, and highlights the

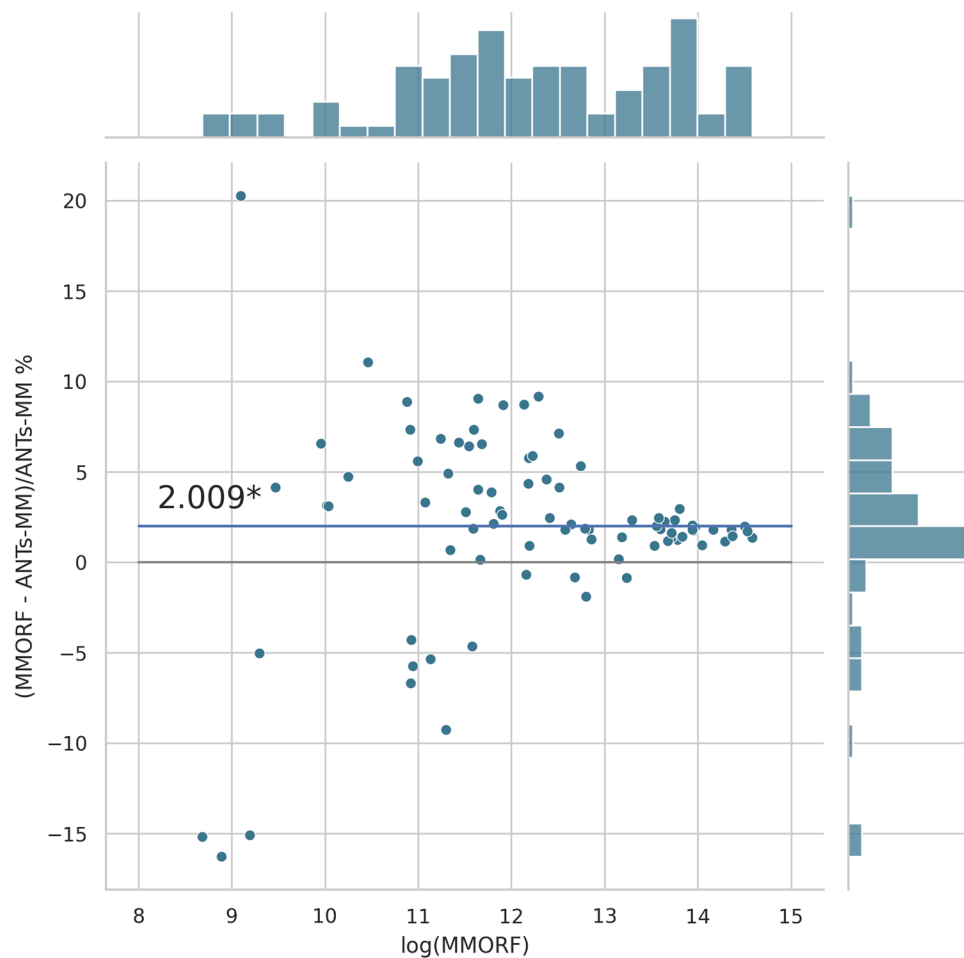


Fig. 10. MMORF vs. ANTs-MM cluster mass: MMORF outperforms ANTs-MM with median improvement in cluster mass of $\approx 2.0\%$ (the asterisk* indicating that the improvement is significant). The distribution of the differences is similar in shape to that of ANTs in [Figure 9](#), but the magnitude is reduced across clusters of all sizes.

importance of holistically evaluating registration performance, including the use of a held-out modality. These results demonstrate the significant improvement in performance of MMORF over its predecessor. Since both the inputs and outputs to MMORF are fully compatible with FNIRT and FSL, the benefits of this improvement can be realised by simply substituting in the new method.

ANTs performed near-identically to MMORF in both cortical and subcortical T1w-label overlap measures, making them the best performing methods in this domain. The inclusion of DTI data in MMORF registration has, therefore, not compromised the alignment of the T1w channel, which is not necessarily a given (see the comparison to DR-TAMAS below). The benefits of including the DTI data are evident in the DTI similarity results, where MMORF clearly has the advantage over ANTs. MMORF consistently outperforms ANTs in the tfMRI evaluation, indicating better anatomical consistency in grey matter, despite similar label overlap performance. Levels of distortion are very comparable between these two methods,

despite there being more information to drive white matter deformations harder in MMORF.

ANTs-MM performed slightly worse in the T1w-label evaluation than the best performing methods, but the inclusion of FA did not have any major detrimental effects. As might be expected, the inclusion of a scalar DTI-derived modality improved its performance over unimodal ANTs in the DTI similarity evaluation, but with overall performance still below that of DR-TAMAS and MMORF. The inclusion of FA was sufficient to bring tfMRI performance closer to that of DR-TAMAS and MMORF, indicating an overall improvement in true registration accuracy compared to unimodal ANTs. Finally, comparing levels of distortion between ANTs-MM and DR-TAMAS shows an interesting difference. While ANTs-MM produces slightly less volumetric distortion than DR-TAMAS, it produces quite a bit more shape distortion. This suggests that using a cost function that is sensitive to differences in orientation has a regularising effect on how the shape of the brain is distorted, even though

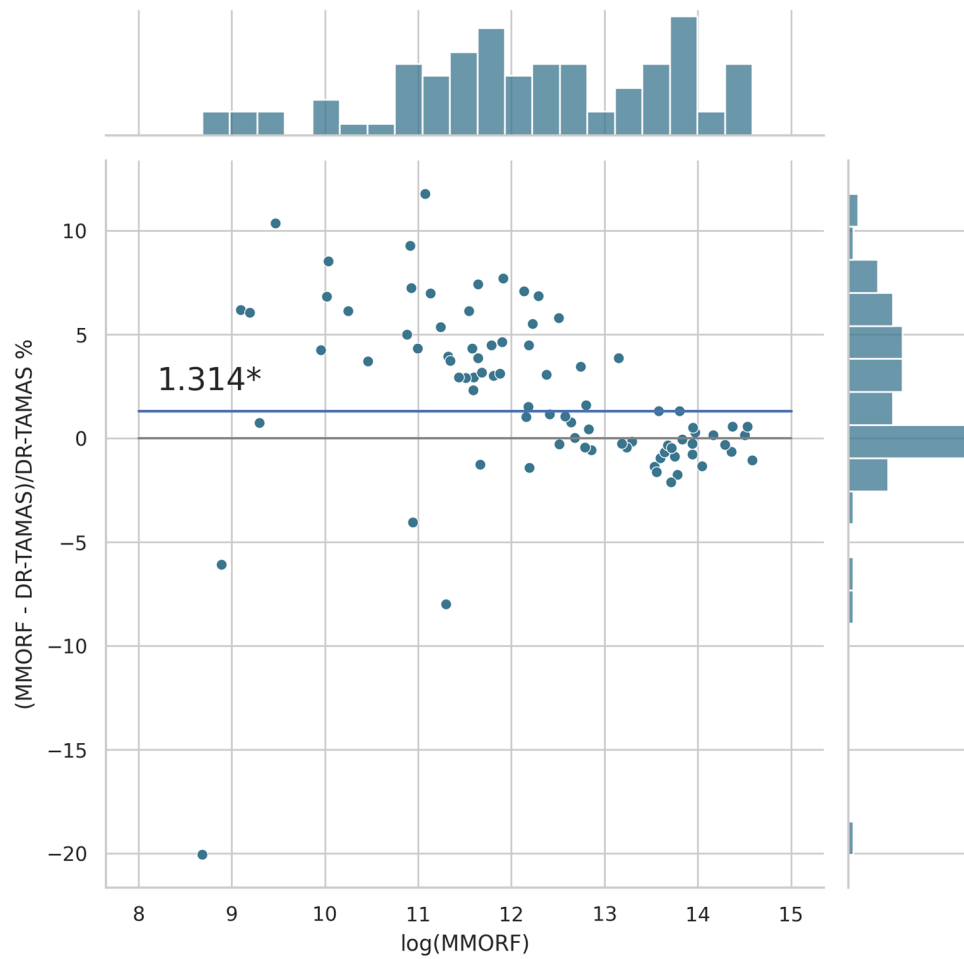


Fig. 11. MMORF vs. DR-TAMAS cluster mass: MMORF outperforms DR-TAMAS with median improvement in cluster mass of $\approx 1.3\%$ (the asterisk* indicating that the improvement is significant). They perform similarly for high CM contrasts, with improvements by MMORF in the low to medium CM contrasts.

there is potentially more information to drive the registration harder.

DR-TAMAS was the poorest performing nonlinear method in terms of cortical T1w-label overlap, and second-poorest subcortically. Since it largely shares its SyN scalar registration algorithm and implementation with ANTs, this may be partly due to the inclusion of DTI data in driving the registration, but different parameter choices likely also play a role. As has already been noted, this was not the case for MMORF. This could be attributed to the differences in DTI cost functions employed by DR-TAMAS and MMORF—where MMORF uses the whole tensor, while DR-TAMAS uses a combination of the mean diffusivity (a scalar) and the deviatoric tensor (shape and direction information only). Similarly, including FA led to only slightly lower overlap performance for ANTs-MM compared to ANTs. DR-TAMAS came a close second to MMORF across all DTI similarity metrics, with performance clearly better than the T1w-only methods and better than ANTs-MM for OVL and CLV1 (but slightly worse for CPV3). DR-TAMAS was also the joint closest

performing method to MMORF in the tfMRI evaluation, demonstrating again that cortical segmentation performance does not necessarily translate to well matched cortical activations. In terms of the level of deformation, DR-TAMAS induced more than MMORF, despite having been outperformed across all accuracy measures.

There are a number of other nonlinear registration tools that are not included in this comparison, but that are worth noting here in order to form a more complete picture of the current state of nonlinear registration. *mrregister* (Raffelt et al., 2011), part of the MRtrix3 software suite, is another variant of the ANTs/SyN method, that uses fibre orientation distributions (FODs) to drive registration. FODs provide a continuous model of the directionality of the underlying white matter fibres in each voxel, and are better suited to modelling the microstructure of crossing fibre regions of the brain than the diffusion tensor. While *mrregister* performs reorientation of the diffusion data (FOD) after each update step, this is not included in the gradient calculation as in MMORF and DR-TAMAS. *NiftyReg* (Modat et al., 2010) is another registration tool that, like

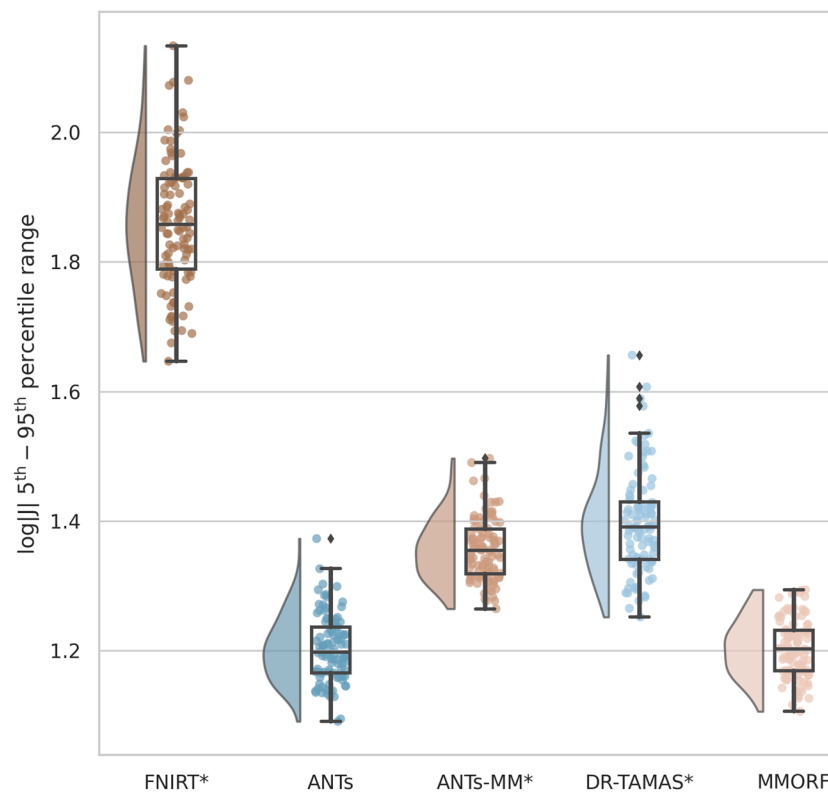


Fig. 12. 5th to 95th percentile Jacobian determinant range: FNIRT shows the largest amount of distortion. MMORF and ANTs have the lowest (and very similar) levels of volumetric distortion. This is in spite of the fact that the MMORF warps are also being driven to align the DTI information in the white matter, which would be expected to increase the amount of distortion (as can be seen has happened for ANTs-MM and DR-TAMAS). Methods that perform significantly differently to MMORF are shown with an asterisk* next to their name.

MMORF, uses GPU parallelisation to accelerate optimisation of a B-spline parametrised warp field—although it uses a simpler gradient-based optimisation and bending energy regularisation. The deformation model and regularisation penalty mean that NiftyReg can be expected to produce qualitatively similar results to FNIRT. Lastly, there are a number of deep-learning based methods becoming available that aim to improve both the speed and accuracy of image registration. Probably the best known of these is VoxelMorph (Balakrishnan et al., 2018), which also forms the basis of a number of extensions of the method. Since its initial release, VoxelMorph has been adapted to tackle some of the challenges associated with deep-learning based methods, including amortised hyperparameter tuning with HyperMorph (Hoopes et al., 2022) and insensitivity to differences in image contrast with SynthMorph (Hoffmann et al., 2022). It has even been extended to include novel diffusion registration approaches—DDMReg (F. Zhang et al., 2022) first models the diffusion data as a set of 40 tract orientation maps (Wasserthal et al., 2019), and then estimates an individual deformation for each map using multiple VoxelMorph networks, before fusing them into a single deformation. While these other methods mentioned all have at least one

unique contribution to the field of image registration, and knowledge of them helps contextualise our work, we believe that the subset of methods we have chosen to evaluate above allow us to effectively benchmark the performance of MMORF in a way that remains both informative and interpretable.

Note that, although our primary objective was human brain alignment, MMORF does not rely on any human brain priors (e.g., tissue maps or assumed brain size), and may be applied to any domain of medical imaging. MMORF has, for example, been successfully used in the generation of multimodal, non-human primate brain templates of several phylogenetically distant species for the purpose of performing comparative anatomy of white matter tracts (Roumazeilles et al., 2022). Similarly, while the HCP enables us to compare the relative performance of each registration tool across multiple imaging modalities using a single dataset, our evaluation here is largely confined to the cerebrum as there are little to no labelled data in the cerebellum, brain stem, and spinal cord. MMORF may, in fact, be particularly beneficial for studying these non-cerebral regions (note their increased contrast in the diffusion data relative to T1w in Fig. 1), and we hope to explore this in future.

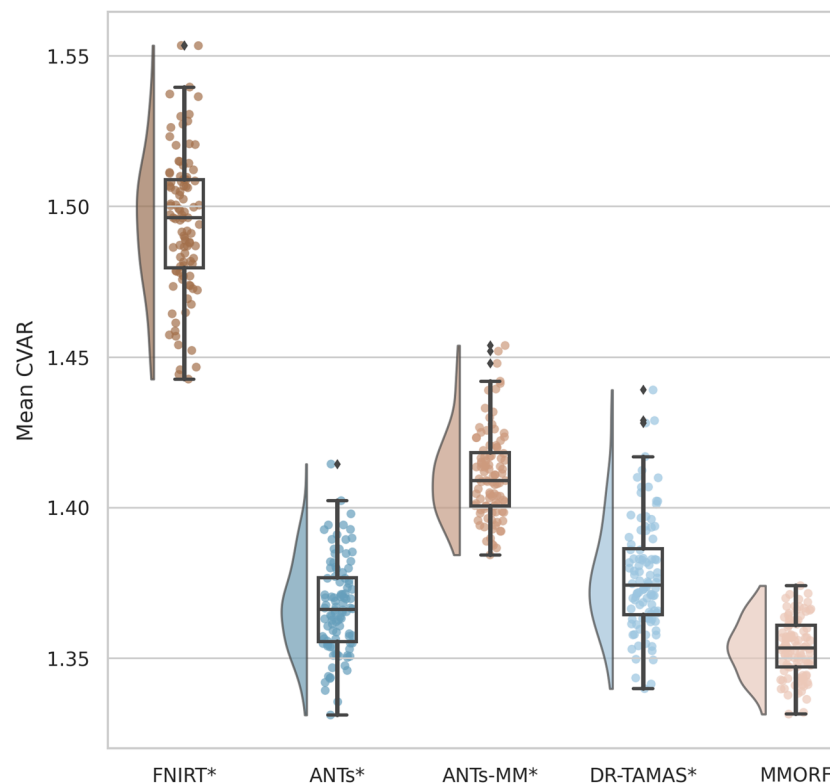


Fig. 13. Mean cube-volume aspect ratio (CVAR) shape distortion: MMORF produces the lowest levels of shape distortion, followed by ANTs, DR-TAMAS, ANTs-MM, and FNIRT (in that order). As with the volumetric distortion, this is in spite of the fact that the MMORF warps are also being driven to align the DTI information in the white matter, which would be expected to increase the amount of distortion (as can be seen has happened for ANTs-MM and, to a slightly lesser extent, DR-TAMAS). This also demonstrates that MMORF's low levels of volumetric distortion (see Fig. 12) do not come at the cost of increased shape distortion, which has been shown to occur when using a regularisation that penalises only the Jacobian determinant (Lange et al., 2020). Methods that perform significantly differently to MMORF are shown with an asterisk* next to their name.

6. LIMITATIONS

One limitation of most evaluations of registration methods is that developers are likely to be able to better tune their own method than those against which they are compared. However, starting from the default settings, we have here made an effort to optimise the performance of each, and kept any adjustments that proved beneficial. As such, we believe that our results are representative of what can be expected in real-world use. While there was always at least one tool with similar performance to MMORF in each test, none were consistently as high-performing across the board.

Similarly, it is possible that using the OMM-1 template, which was constructed using MMORF, might slightly bias the registration accuracy in its favour. We believe that this is mitigated, to an extent, by performing our evaluation using data from a completely separate study and age-group (the HCP) than the one from which the OMM-1 template was derived (the UKB). However, it remains true that templates produced using ANTs scripts will have contrast that is better suited to its preferred LNCC cost

function (Avants et al., 2010; Joshi et al., 2004). A final point regarding templates is that we intend to use the OMM-1 as the default template for FSL analyses in the future, and so this is a good representation of our planned use-case for MMORF going forward.

We have made a point of noting that many evaluation metrics have an inherent circularity to them, and this is true of both the FreeSurfer label and DTI similarity metrics presented here. Potential ways of unbiasing these evaluations (at least to an extent) would be to use overlap of labels derived from the T2w images from the HCP (to break the link with the T1w images used during registration), and tract segmentations derived from the diffusion weighted images (to break the link with the tensor images used during registration). This is only a partial solution to circularity though, since the T1w and T2w channels share much of the same information, and automated methods for doing tractography tend to require registration to a standard space for initialisation (an aspect we are interested in exploring with MMORF in the future). However, we believe that while the metrics used

in this paper are not perfect, they are still useful—particularly when considering multiple metrics holistically, as we do here. Provided the results are not over-interpreted, they aid in creating an overall picture of how these methods compare with each other.

Finally, in this work, we have only considered MMORF's performance in comparison to other volumetric registration tools. There is evidence that applying surface-based registration has a beneficial effect on cortical alignment accuracy (Coalson et al., 2018), and we are not claiming that it is better to use MMORF instead of those methods when performing purely cortical or surface-based cross-subject fMRI analyses, for example. However, volumetric registration is, as yet, still the only way to align non-cortical brain regions, and is likely to remain just as valuable as surface registration in neuroimaging research for the foreseeable future, especially when data are registered intra-individually and when surface reconstructions fail, for example, due to the presence of intra-axial lesions (such as brain tumours).

7. FUTURE WORK

As mentioned in the introduction, there are many other nonlinear registration tools against which we could have compared, and therefore we cannot unequivocally state that MMORF is the best option for every registration scenario. Going forward, we intend to continue benchmarking MMORF against other methods, and hope that we (and other users) might find ways in which we can improve our tool. We would be particularly interested in comparing the performance of MMORF to that of mrregister in the future, in order to gauge the impact of using more complex models of diffusion to drive the registration. However, this would require a change of methodology, as the OMM-1 does not currently include an FOD modality.

Although we cover a broad range of evaluation metrics across a large number of subjects, they are restricted to healthy, young volunteers. This approach has allowed us to balance the requirements of both sufficiently describing MMORF's implementation and demonstrating its utility within a single software toolbox paper. We have also used a template based on middle-aged healthy volunteers from an entirely different dataset, which has several morphological differences to HCP (including ventricle sizes that are $\approx 2\times$ larger, on average). While we believe that the HCP dataset is well suited to this task (as it includes a large range of data types that allow a holistic set of evaluations to be performed), it would certainly be of future interest to benchmark MMORF's performance in other datasets with unique characteristics—for example, assessing the impact of using DTI data and our regulari-

sation on the alignment of pathological brain images, such as the tumour example in Ou et al. (2014).

Finally, we are continuing to develop and improve MMORF, including the implementation of a CPU version. We plan on incorporating it into existing FSL tools such as FEAT and TBSS in the near future. We are also working on ways to combine cortical surface alignment and volumetric alignment within the MMORF framework, with the hope of leveraging the unique advantages of each of these approaches to further improve registration accuracy.

8. CONCLUSION

Given MMORF's strong performance, we believe it to be an excellent choice of tool for volumetric registration—regardless of the domain of any follow-on analyses. As such, it is easy to recommend MMORF to all users working with neuroimaging data. Due to its compatibility with FSL, existing FSL users will further benefit from state-of-the-art registration accuracy with very little modification to any existing pipelines. MMORF is available as part of FSL, versions 6.0.7 and up.

DATA AND CODE AVAILABILITY

MMORF binaries and the associated source code are available for download via the standard FSL installer⁴. A stand-alone Singularity image, instructions for running MMORF, and example configuration files are available on the FMRIB GitLab server⁵. All analysis scripts, figures, figure source code, and latex source code that were used to generate this manuscript are available on the FMRIB GitLab server⁶.

AUTHOR CONTRIBUTIONS

F.J.L.: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing (original draft, review and editing), and Visualisation. C.A.: Methodology, Software, Validation, and Writing (review and editing). A.B. and G.D.: Validation, Writing (review and editing). P.M.: Software, Data Curation. S.M.S.: Conceptualisation, Methodology, Validation, Resources, Writing (review and editing), Supervision, Project administration, and Funding acquisition. J.L.R.A.: Conceptualisation, Methodology, Software, Validation, Formal analysis, Writing (review and editing), Supervision, and Project administration.

⁴ <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FslInstallation>

⁵ https://git.fmrib.ox.ac.uk/flange/mmorf_beta

⁶ https://git.fmrib.ox.ac.uk/flange/mmorf_toolbox

DECLARATION OF COMPETING INTERESTS

We have no competing interests to declare.

ACKNOWLEDGEMENTS

This research was funded by a Wellcome Trust Collaborative Award (215573/Z/19/Z). The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

REFERENCES

- Alexander, A. L., Hasan, K., Kindlmann, G., Parker, D. L., & Tsuruda, J. S. (2000). A geometric analysis of diffusion tensor measurements of the human brain. *Magnetic Resonance in Medicine*, 44, 283–291. [https://doi.org/10.1002/1522-2594\(200008\)44:2<283::aid-mrm16>3.0.co;2-v](https://doi.org/10.1002/1522-2594(200008)44:2<283::aid-mrm16>3.0.co;2-v)
- Alexander, D. C., Pierpaoli, C., Basser, P. J., & Gee, J. C. (2001). Spatial transformations of diffusion tensor magnetic resonance images. *IEEE Transactions on Medical Imaging*, 20, 1131–1139. <https://doi.org/10.1109/42.963816>
- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., Van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, 4, 63. <https://doi.org/10.12688/wellcomeopenres.15191.2>
- Andersson, J. L. R., Jenkinson, M., & Smith, S. M. (2007). *Non-linear registration aka spatial normalisation*. Technical Report June. <https://www.fmrib.ox.ac.uk/datasets/techrep/tr07ja2/tr07ja2.pdf>
- Andersson, J. L. R., Jenkinson, M., & Smith, S. M. (2019). High resolution nonlinear registration with simultaneous modelling of intensities. *bioRxiv*, 646802. <https://doi.org/10.1101/646802>
- Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage*, 20, 870–888. [https://doi.org/10.1016/s1053-8119\(03\)00336-7](https://doi.org/10.1016/s1053-8119(03)00336-7)
- Andersson, J. L. R., & Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, 125, 1063–1078. <https://doi.org/10.1016/j.neuroimage.2015.10.019>
- Arthofer, C., Smith, S. M., Douaud, G., Bartsch, A., Alfaro-Almagro, F., Andersson, J., & Lange, F. J. (2023). Internally-consistent and fully-unbiased multimodal MRI brain template construction from UK Biobank: Oxford-MM. *bioRxiv*. <https://doi.org/10.1101/2023.11.30.569378>
- Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38, 95–113. <https://doi.org/10.1016/j.neuroimage.2007.07.007>
- Ashburner, J., Andersson, J. L. R., & Friston, K. J. (1999). High-dimensional image registration using symmetric priors. *NeuroImage*, 9, 619–628. <https://doi.org/10.1006/nimg.1999.0437>
- Ashburner, J., Andersson, J. L. R., & Friston, K. J. (2000). Image registration using a symmetric prior—In three dimensions. *Human Brain Mapping*, 9, 212–225. <https://onlinelibrary.wiley.com/doi/epdf/10.1002/%28SICI%291097-0193%28200004%299%3A4%3C212%3A%3AAID-HBM3%3E3.0.CO%3B2-%23>
- Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—The methods. *NeuroImage*, 11, 805–821. <https://doi.org/10.1006/nimg.2000.0582>
- Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12, 26–41. <https://doi.org/10.1016/j.media.2007.06.004>
- Avants, B. B., Tustison, N. J., Stauffer, M., Song, G., Wu, B., & Gee, J. C. (2014). The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, 8, 1–13. <https://doi.org/10.3389/fninf.2014.00044>
- Avants, B. B., Yushkevich, P. A., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., & Gee, J. C. (2010). The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*, 49, 2457–2466. <https://doi.org/10.1016/j.neuroimage.2009.09.062>
- Bajcsy, R., & Kovačič, S. (1989). Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46, 1–21. [https://doi.org/10.1016/s0734-189x\(89\)80014-3](https://doi.org/10.1016/s0734-189x(89)80014-3)
- Balakrishnan, G., Zhao, A., Sabuncu, M. R., Dalca, A. V., & Guttag, J. (2018). An unsupervised learning model for deformable medical image registration. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9252–9260). IEEE. <https://doi.org/10.1109/cvpr.2018.00964>
- Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., & Van Essen, D. C. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 80, 169–189. <https://doi.org/10.1016/j.neuroimage.2013.05.033>
- Beg, M. F., & Khan, A. (2007). Symmetric data attachment terms for large deformation image registration. *IEEE Transactions on Medical Imaging*, 26, 1179–1189. <https://doi.org/10.1109/tmi.2007.898813>
- Beg, M. F., Miller, M., Trounev, A., & Younes, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61, 139–157. <https://doi.org/10.1023/b:visi.0000043755.93987.a>
- Bookstein, F. L. (1997). Quadratic variation of deformations. In: *Lecture Notes in Computer Science (Including*

- Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (vol. 1230, pp. 15–28). https://doi.org/10.1007/3-540-63046-5_2
- Bro-Nielsen, M., & Gramkow, C. (1996). Fast fluid registration of medical images. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (vol. 1131, pp. 267–276). <https://doi.org/10.1007/BFb0046964>
- Christensen, G. E. (1999). Consistent linear-elastic transformations for image matching. In: A. Kuba, M. Šámal, & A. Todd-Pokropek (Eds.), *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 224–237). Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/3-540-48714-X_17
- Chun, I. Y., & Fessler, J. A. (2018). Convolutional dictionary learning: Acceleration and convergence. *IEEE Transactions on Image Processing*, 27, 1697–1712. <https://doi.org/10.1109/tip.2017.2761545>
- Coalson, T. S., Van Essen, D. C., & Glasser, M. F. (2018). The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proceedings of the National Academy of Sciences*, 115, E6356–E6365. <https://doi.org/10.1073/pnas.1801582115>
- Destrieux, C., Fischl, B., Dale, A., & Hagren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53, 1–15. <https://doi.org/10.1016/j.neuroimage.2010.06.010>
- Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., James, S., Voets, N., Watkins, K., Matthews, P. M., & James, A. (2007). Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*, 130, 2375–2386. <https://doi.org/10.1093/brain/awm184>
- Dubuisson, M. P., & Jain, A. (1994). A modified Hausdorff distance for object matching. In: *Proceedings of 12th International Conference on Pattern Recognition* (pp. 566–568). IEEE. <https://doi.org/10.1109/icpr.1994.576361>
- Elam, J. S., Glasser, M. F., Harms, M. P., Sotiropoulos, S. N., Andersson, J. L. R., Burgess, G. C., Curtiss, S. W., Oostenveld, R., Larson-Prior, L. J., Schoffelen, J. M., Hodge, M. R., Cler, E. A., Marcus, D. M., Barch, D. M., Yacoub, E., Smith, S. M., Ugurbil, K., & Van Essen, D. C. (2021). The Human Connectome Project: A retrospective. *NeuroImage*, 244, 118543. <https://doi.org/10.1016/j.neuroimage.2021.118543>
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33, 341–355. [https://doi.org/10.1016/s0896-6273\(02\)00569-x](https://doi.org/10.1016/s0896-6273(02)00569-x)
- Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23, S69–S84. <https://doi.org/10.1016/j.neuroimage.2004.07.016>
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L. R., Xu, J., Jbabdi, S., Webster, M. A., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80, 105–124. <https://doi.org/10.1016/j.neuroimage.2013.04.127>
- Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The amygdala response to emotional stimuli: A comparison of faces and scenes. *NeuroImage*, 17, 317–323. <https://doi.org/10.1006/nimg.2002.1179>
- Hoffmann, M., Billot, B., Greve, D. N., Iglesias, J. E., Fischl, B., & Dalca, A. V. (2022). SynthMorph: Learning contrast-invariant registration without acquired images. *IEEE Transactions on Medical Imaging*, 41, 543–558. <https://doi.org/10.1109/tmi.2021.3116879>
- Hoopes, A., Hoffmann, M., Greve, D. N., Fischl, B., Guttag, J., & Dalca, A. V. (2022). Learning the effect of registration hyperparameters with hyper-morph. *Machine Learning for Biomedical Imaging*, 1, 1–30. <https://doi.org/10.59275/j.melba.2022-74f1>
- Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *American Statistician*, 58, 30–37. <https://doi.org/10.1198/0003130042836>
- Irfanoglu, M. O., Nayak, A., Jenkins, J., Hutchinson, E. B., Sadeghi, N., Thomas, C. P., & Pierpaoli, C. (2016). DR-TAMAS: Diffeomorphic registration for tensor accurate alignment of anatomical structures. *NeuroImage*, 132, 439–454. <https://doi.org/10.1016/j.neuroimage.2016.02.066>
- Jenkinson, M., Bannister, P. R., Brady, J. M., & Smith, S. M. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17, 825–841. <https://doi.org/10.1006/nimg.2002.1132>
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, 62, 782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>
- Jenkinson, M., & Smith, S. M. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5, 143–156. [https://doi.org/10.1016/s1361-8415\(01\)00036-6](https://doi.org/10.1016/s1361-8415(01)00036-6)
- Joshi, S., Davis, B., Jomier, M., & Gerig, G. (2004). Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23, S151–S160. <https://doi.org/10.1016/j.neuroimage.2004.07.068>
- Klein, A., Andersson, J. L. R., Ardekani, B. A., Ashburner, J., Avants, B. B., Chiang, M. C., Christensen, G. E., Collins, D. L., Gee, J. C., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P. M., Vercauteren, T., Woods, R. P., Mann, J. J., & Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46, 786–802. <https://doi.org/10.1016/j.neuroimage.2008.12.037>
- Lange, F. J., Ashburner, J., Smith, S. M., & Andersson, J. L. R. (2020). A symmetric prior for the regularisation of elastic deformations: Improved anatomical plausibility in nonlinear image registration. *NeuroImage*, 219, 116962. <https://doi.org/10.1016/j.neuroimage.2020.116962>
- Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., ..., Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19, 1523–1536. <https://doi.org/10.1038/nn.4393>
- Miller, M., Banerjee, A., Christensen, G., Joshi, S., Khaneja, N., Grenander, U., & Matejic, L. (1997). Statistical methods in computational anatomy. *Statistical Methods in Medical Research*, 6, 267–299. <https://doi.org/10.1177/096228029700600305>
- Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., & Ourselin, S. (2010). Fast free-form deformation using graphics

- processing units. *Computer Methods and Programs in Biomedicine*, 98, 278–284. <https://doi.org/10.1016/j.cmpb.2009.09.002>
- Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6, 525–533. [https://doi.org/10.1016/s0893-6080\(05\)80056-5](https://doi.org/10.1016/s0893-6080(05)80056-5)
- Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*. Springer Science+Business Media, USA, New York. <https://doi.org/10.1007/b98874>
- NVIDIA. (2023). CUDA C programming guide. <https://docs.nvidia.com/cuda/archive/12.3.1/>
- Ou, Y., Akbari, H., Bilello, M., Da, X., & Davatzikos, C. (2014). Comparative evaluation of registration algorithms in different brain databases with varying difficulty: Results and insights. *IEEE Transactions on Medical Imaging*, 33, 2039–2065. <https://doi.org/10.1109/tmi.2014.2330355>
- Raffelt, D., Tournier, J. D., Frupp, J., Crozier, S., Connelly, A., & Salvado, O. (2011). Symmetric diffeomorphic registration of fibre orientation distributions. *NeuroImage*, 56, 1171–1180. <https://doi.org/10.1016/j.neuroimage.2011.02.014>
- Robinson, E. C., Garcia, K., Glasser, M. F., Chen, Z., Coalson, T. S., Makropoulos, A., Bozek, J., Wright, R., Schuh, A., Webster, M., Hutter, J., Price, A., Cordero Grande, L., Hughes, E., Tumor, N., Bayly, P. V., Van Essen, D. C., Smith, S. M., Edwards, A. D., ..., Rueckert, D. (2018). Multimodal surface matching with higher-order smoothness constraints. *NeuroImage*, 167, 453–465. <https://doi.org/10.1016/j.neuroimage.2017.10.037>
- Rohlfing, T. (2012). Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *IEEE Transactions on Medical Imaging*, 31, 153–163. <https://doi.org/10.1109/tmi.2011.2163944>
- Rohlfing, T., Maurer, C. R., Bluemke, D. A., & Jacobs, M. A. (2003). Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint. *IEEE Transactions on Medical Imaging*, 22, 730–741. <https://doi.org/10.1109/tmi.2003.814791>
- Roumazeilles, L., Lange, F. J., Benn, R. A., Andersson, J. L. R., Bertelsen, M. F., Manger, P. R., Flach, E., Khrapitchev, A. A., Bryant, K. L., Sallet, J., & Mars, R. B. (2022). Cortical morphology and white matter tractography of three phylogenetically distant primates: Evidence for a simian elaboration. *Cerebral Cortex*, 32, 1608–1624. <https://doi.org/10.1093/cercor/bhab285>
- Sederevičius, D., Vidal-Piñeiro, D., Sørensen, Ø., Van Leemput, K., Iglesias, J. E., Dalca, A. V., Greve, D. N., Fischl, B., Bjørnerud, A., Walhovd, K. B., & Fjell, A. M. (2021). Reliability and sensitivity of two whole-brain segmentation approaches included in FreeSurfer—ASEG and SAMSEG. *NeuroImage*, 237, 118113. <https://doi.org/10.1016/j.neuroimage.2021.118113>
- Smith, W., & Wormald, N. (1998). Geometric separator theorems and applications. In: *Proceedings 39th Annual Symposium on Foundations of Computer Science* (Cat. No.98CB36280, pp. 232–243). IEEE. <https://doi.org/10.1109/sfcs.1998.743449>
- Szeliski, R., & Coughlan, J. (1997). Spline-based image registration. *International Journal of Computer Vision*, 22, 199–218. <https://doi.org/10.1023/a:1007996332012>
- Tagare, H. D., Groisser, D., & Skrinjar, O. (2009). Symmetric non-rigid registration: A geometric theory and some numerical techniques. *Journal of Mathematical Imaging and Vision*, 34, 61–88. <https://doi.org/10.1007/s10851-008-0129-7>
- Unser, M., Aldroubi, A., & Eden, M., 1993a. B-spline signal processing. I. Theory. *IEEE Transactions on Signal Processing*, 41, 821–833. <https://doi.org/10.1109/78.193220>
- Unser, M., Aldroubi, A., & Eden, M., 1993b. B-spline signal processing. II. Efficiency design and applications. *IEEE Transactions on Signal Processing*, 41, 834–848. <https://doi.org/10.1109/78.193221>
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, 80, 62–79. <https://doi.org/10.1016/j.neuroimage.2013.05.041>
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., ..., Yacoub, E. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62, 2222–2231. <https://doi.org/10.1016/j.neuroimage.2012.02.018>
- Vovk, U., Pernuš, F., & Likar, B. (2007). A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 26, 405–421. <https://doi.org/10.1109/tmi.2006.891486>
- Wasserthal, J., Neher, P. F., Hirjak, D., & Maier-Hein, K. H. (2019). Combined tract segmentation and orientation mapping for bundle-specific tractography. *Medical Image Analysis*, 58, 101559. <https://doi.org/10.1016/j.media.2019.101559>
- Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92, 381–397. <https://doi.org/10.1016/j.neuroimage.2014.01.060>
- Yeo, B. T., Vercauteren, T., Fillard, P., Peyrat, J. M., Pennec, X., Golland, P., Ayache, N., & Clatz, O. (2009). DT-REFinD: Diffusion tensor registration with exact finite-strain differential. *IEEE Transactions on Medical Imaging*, 28, 1914–1928. <https://doi.org/10.1109/tmi.2009.2025654>
- Zhang, F., Wells, W. M., & O'Donnell, L. J. (2022). Deep diffusion MRI registration (DDMReg): A deep learning method for diffusion MRI registration. *IEEE Transactions on Medical Imaging*, 41, 1454–1467. <https://doi.org/10.1109/tmi.2021.3139507>
- Zhang, H., Yushkevich, P. A., Alexander, D. C., & Gee, J. C. (2006). Deformable registration of diffusion tensor MR images with explicit orientation optimization. *Medical Image Analysis*, 10, 764–785. <https://doi.org/10.1016/j.media.2006.06.004>

APPENDIX A. LEVENBERG-MARQUARDT AND MAJORISE-MINIMISATION OPTIMISATION

Appendix A.1. Levenberg-Marquardt

Given a quadratic Taylor approximation of a cost function \mathcal{C} around the point \mathbf{w} in parameter space,

$$\mathcal{C}(\mathbf{w} + \Delta\mathbf{w}) \approx \mathcal{C}(\mathbf{w}) + \nabla\mathcal{C}(\mathbf{w})^T \Delta\mathbf{w} + \frac{1}{2} \Delta\mathbf{w}^T \mathbf{H}(\mathbf{w}) \Delta\mathbf{w} \quad (\text{A.1})$$

where $\nabla\mathcal{C}$ and \mathbf{H} are the gradient and the Hessian of the cost function, respectively. The update $\Delta\mathbf{w}$ that minimises this approximation to the cost function (when the cost function is convex) is then:

$$\Delta\mathbf{w} = -\mathbf{H}^{-1}(\mathbf{w}) \nabla\mathcal{C}(\mathbf{w}) \quad (\text{A.2})$$

For cost functions that can be written in the form:

$$\mathcal{C}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N c_i^2(\mathbf{w}) \quad (\text{A.3})$$

The gradient is then:

$$\nabla\mathcal{C}(\mathbf{w}) = \sum_{i=1}^N c_i(\mathbf{w}) \frac{\partial c_i}{\partial \mathbf{w}} \quad (\text{A.4})$$

And the Hessian is:

$$\mathbf{H}(\mathbf{w}) = \sum_{i=1}^N \frac{\partial c_i}{\partial \mathbf{w}}^T \frac{\partial c_i}{\partial \mathbf{w}} + c_i(\mathbf{w}) \frac{\partial^2 c_i}{\partial \mathbf{w}^2} \quad (\text{A.5})$$

Gauss-Newton approximates \mathbf{H} by dropping the second, mixed partial derivative, term in Equation A.5. Close to the optimum set of \mathbf{w} , the Newton and Gauss-Newton Hessians tend towards equivalence. Further from the optimum, the Gauss-Newton Hessian has the benefit of being positive definite, and therefore the approximation of the cost function is always convex. This ensures that each step is always in a direction that would improve the original cost function.

However, the Gauss-Newton step can still lead to an increase in the cost function if it oversteps. This can be addressed by the Levenberg-Marquardt (LM) extension (Nocedal & Wright, 2006, Ch.10, pp258), which replaces the Gauss-Newton Hessian (\mathbf{H}_{GN}) with:

$$\mathbf{H}_{LM} = \mathbf{H}_{GN} + \lambda \mathbf{I}. \quad (\text{A.6})$$

For small values of λ , the update $\Delta\mathbf{w}$ behaves much like a Gauss-Newton update, and for large values of λ it

behaves more like gradient descent. After choosing a sensible starting value for λ (e.g., 0.001), a typical Levenberg-Marquardt iteration then proceeds as follows:

- Calculate a candidate parameter update $\Delta\mathbf{w}$ using Equation A.2.
- If $\mathcal{C}(\mathbf{w} + \Delta\mathbf{w}) < \mathcal{C}(\mathbf{w})$ accept the update and set $\lambda = \lambda + 10$.
- If $\mathcal{C}(\mathbf{w} + \Delta\mathbf{w}) > \mathcal{C}(\mathbf{w})$ set $\lambda = \lambda \times 10$ and recompute $\Delta\mathbf{w}$ and $\mathcal{C}(\mathbf{w} + \Delta\mathbf{w})$, repeating until $\mathcal{C}(\mathbf{w} + \Delta\mathbf{w}) < \mathcal{C}(\mathbf{w})$ and then accept the update.

Appendix A.2. Majorise-minimisation

The Majorise-Minimisation (MM) algorithm (Hunter & Lange, 2004) is a method for cost function minimisation that can be used when storing the full Gauss-Newton Hessian is infeasible (e.g., due to memory constraints, which prove to be important in this application). In essence, what this algorithm states is that if $z(\mathbf{w}) \geq y(\mathbf{w}), \forall \mathbf{w}$, with both y and z convex and touching, then $\mathbf{w}_{k+1} = \arg \min_{\mathbf{w}} z(\mathbf{w})$ will either reduce y or leave it unchanged.

Since the quadratic Taylor approximation of $\mathcal{C}(\mathbf{w})$ using \mathbf{H}_{GN} in Equation A.1 is convex, it may serve as y . Chun & Fessler (2018) showed that because the diagonal elements of \mathbf{H}_{GN} are all positive:

$$\mathbf{H}_{MM} = \text{diag}(|\mathbf{H}_{GN}| \mathbf{1}) \succeq \mathbf{H}_{GN} \quad (\text{A.7})$$

where $|\mathbf{H}_{GN}|$ is the matrix of the absolute values of \mathbf{H}_{GN} and $\mathbf{1}$ is a column vector of ones. In simpler terms, \mathbf{H}_{MM} is a matrix that is zero everywhere except along its main diagonal, and the value on the diagonal for any row is equal to the sum of the absolute values of the corresponding row in \mathbf{H}_{GN} . The \succeq symbol means that \mathbf{H}_{MM} is at least as positive definite as \mathbf{H}_{GN} . Therefore, substituting \mathbf{H}_{MM} for \mathbf{H}_{GN} in the quadratic Taylor approximation of $\mathcal{C}(\mathbf{w})$ majorises y , and may serve as z . In other words simply, replacing \mathbf{H}_{GN} with \mathbf{H}_{MM} in the LM algorithm and solving as usual satisfies the MM algorithm requirements.

The major advantage of using \mathbf{H}_{MM} is that it is only non-zero along its main diagonal, and therefore requires far less memory to store than \mathbf{H}_{GN} . This is particularly important when the number of parameters in \mathbf{w} is large, as it may not be possible to store \mathbf{H}_{GN} .