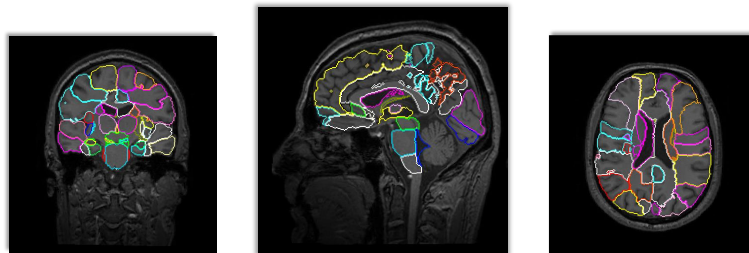




SAHLGRENKA ACADEMY



Evaluation of Synthetic Ground Truth for Semantic Brain Image Segmentation

Developing a Database of Synthetic Images with Ground Truth Segmentations
for Practical Application Assessment

Mardo Mardo

Essay/Thesis:	30 hp
Program and/or course:	Medical physics
Level:	Second Cycle
Term/year:	Spring 2025
Supervisor:	Rolf A. Heckemann
Examiner:	Magnus Båth

Evaluation of Synthetic Ground Truth for Semantic Brain Image Segmentation

Developing a Database of Synthetic Images with Ground-Truth
Segmentations for Practical Application Assessment

Mardo Mardo



UNIVERSITY OF GOTHENBURG

Sahlgrenska Academy

Gothenburg, Sweden 2025

Evaluation of Synthetic Ground Truth for Semantic Brain Image Segmentation

© Mardo Mardo

Supervisor: Rolf A. Heckemann, Department of Medical Radiation Sciences.

Examiner: Magnus Båth, Department of Medical Radiation Sciences.

Master's Thesis 2025
Sahlgrenska Academy
University of Gothenburg

Typeset in L^AT_EX
Gothenburg, Sweden 2025

Abstract

Introduction: Automated brain image segmentation plays an important role in large-scale neuroimaging research and shows growing potential for clinical applications. Traditional segmentation pipelines rely on manually annotated MR-images as atlases, but such data are time-consuming to acquire and may not be readily available in sufficient quantity or variety. This thesis documents a research project where the use of synthetically generated MR-images with known ground truth was explored to support evaluation and benchmarking of automated segmentation methods.

Aim & Purpose: The aim of this thesis is to evaluate the performance and robustness of the multi-atlas segmentation with enhanced registration (MAPER) algorithm using leave-one-out cross-validation, facilitated by the generation of synthetic MR images with controlled structural and intensity alterations. Additionally, this study assesses whether the synthetic images can be used as effective atlases compared to conventional MR-images and find a suitable benchmark for evaluation of different MAPER versions; `onepad (master)` and `no-onepad`.

Method: Five types of synthetic MR-images were generated based on the IXI and Hamers datasets, each with different modes of spatial and intensity modification. Experiments were designed to (1) compare segmentation performance between synthetic atlases and unmodified conventional atlas (2) validate segmentation performance and robustness using leave-one-out cross-validation on synthetic targets, (3) quantify the convergence rate for the different ground truth benchmarks based on a parametric model and perform bootstrapping as a benchmarking measure to find the most sensitive benchmark. Then (4) apply this benchmark to compare two MAPER configurations.

Results and conclusion: Segmentation performance was highest using conventional MR-images atlases, followed by statistical and scrambled synthetic variants. Leave-one-out cross-validation (LOOCV) confirmed similar trends. In benchmarking, the statistical smoothed image (*statsmooth*) type exhibited the fastest convergence rate, making it the most sensitive benchmark. Application of the benchmark revealed that the `no-onepad` version of MAPER outperformed the standard `master` configuration.

Discussion: The results highlight that although synthetic images underperform compared to conventional data as atlases, they provide meaningful insights into algorithm robustness and segmentation behaviour. The ground truth benchmark shows potential as a controlled and reproducible framework for comparing segmentation methods and identifying subtle differences in performance.

Sammanfattning

Introduktion: Automatiserad segmentering av hjärnavbildningar spelar en viktig roll inom storskalig neuroimaging-forskning och visar en växande potential för kliniska tillämpningar. Traditionella segmenteringsmetoder är beroende av manuellt annoterade MR-bilder som atlas, vilket är tidskrävande att samla in och sådana data finns ofta inte tillgängliga i tillräcklig mängd eller variation. Denna avhandling dokumenterar ett forskningsprojekt där användningen av syntetiskt genererade MR-bilder med känd grundsanning ”ground truth” har undersökts för att stödja utvärdering och benchmarking av automatiserade segmenteringsmetoder.

Syfte och mål: Syftet med detta arbete var att generera syntetiska MR-bilder med kontrollerade strukturella och intensitetsmässiga förändringar för att utvärdera segmenteringsalgoritmen multi-atlas segmentation with enhanced registration (MAPER) avseende både prestanda och robusthet med leave-one-out cross-validation. Vidare utvärderas om de syntetiska bilderna kan användas som effektiva atlaser i jämförelse med konventionella MR-bilder samt om lämpliga riktmärke-konfigurationer kan identifieras för att utvärdera olika versioner av MAPER: onepad (master) och no-onepad.

Metod: Fem typer av syntetiska MR-bilder genererades utifrån IXI- och Hammersdatamängderna, med varierande grad av spatiala och intensitetsrelaterade modifieringar. Följande experiment genomfördes: (1) jämförelse av segmenteringsprestanda mellan syntetiska atlaser och omodifierade konventionella atlaser, (2) utvärdering av prestanda och robusthet med leave-one-out cross-validation på syntetiska och konventionellt mål (3) kvantifiera konvergenshastigheten för de olika grundsanningsriktmärkena utifrån en parametrisk modell och tillämpa bootstrapping som ett riktmärkesmått för att identifiera det mest känsliga riktmärket, samt (4) tillämpa detta riktmärke för att jämföra två olika MAPER-konfigurationer.

Resultat och slutsats: Segmenteringsprestandan var högst med konventionella MR-bilder atlaser, följt av statistiska och scrambled syntetiska varianter. Leave-one-out cross-validation (LOOCV) bekräftade liknande mönster. Vid riktmärkning visade den statistiskt utjämnade syntetiska bilderna (*statsmooth*) medföra högst konvergenshastighet, vilket gjorde den till det mest känsliga riktmärke alternativet. Vid tillämpning av riktmärket framkom att no-onepad-versionen av MAPER presterade bättre än standardversionen master.

Diskussion: Resultaten visar att även om syntetiska bilder presterar sämre än konventionella data som atlaser, ger de värdefull insikt i algoritmens robusthet och segmenteringsbeteende. Grundsanningriktvärde visar lovande potential som en kontrollerad och reproducerbar metod för att jämföra segmenteringsmetoder och identifiera subtila prestandaskillnader.

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Professor Rolf A. Heckemann, for his invaluable support, guidance, and expertise throughout the course of this thesis. His insightful feedback, structured mentorship, and deep understanding of the subject matter have been essential to the successful completion of this project. I am especially thankful for his patience, clarity, and dedication during our discussions which significantly enriched the quality and depth of the work.

I also wish to thank the Department of Medical Radiation Sciences at the Sahlgrenska Academy, University of Gothenburg, for providing the necessary resources and a stimulating academic environment in which to carry out this research.

The computations for this thesis work were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Contents

1	Introduction	1
2	Preliminaries	5
2.1	Fundamental Definitions	5
2.2	Multi-Atlas Segmentation	6
2.2.1	Multi-Atlas Propagation with Enhanced Registration	7
2.3	Evaluation of Segmentation Algorithm	7
2.3.1	Benchmark performance & Synthesizing MR-images	9
3	Method	11
3.1	Participants and MRI dataset	11
3.2	Generation of Synthetic Ground Truth Images	11
3.3	Experimental Setup	14
3.3.1	Synthetic Images as Atlases	14
3.3.2	Leave-One-Out Cross-Validation on Synthetic Images	15
3.3.3	Benchmarking the Ground Truth Benchmark	16
3.3.4	Example Application of Ground Truth Benchmark	18
4	Results	19
4.1	Synthetic Images as Atlases	19
4.2	Leave-One-Out Cross-Validation on Synthetic Images	20
4.3	Benchmarking the Ground Truth Benchmark	21
4.4	Example Application of Ground Truth Benchmark	23
5	Discussion	24
5.1	Synthetic Images as Atlases	24
5.2	Leave-One-Out Cross-Validation on Synthetic Images	25
5.3	Benchmarking the Ground Truth Benchmark	26
5.4	Example Application of Ground Truth Benchmark	27
6	Conclusion	28
	References	29

1

Introduction

Spatial anatomical representation of human anatomy in *atlases* has undergone various improvements throughout history, with origins tracing back to ancient Egyptian times (1700s BC) [1]. These early attempts to document, communicate and understand the human body laid the groundwork for what would eventually become sophisticated tools for exploring the brain's complex structures and functions. The creation of visual anatomical atlases began to flourish in the 16th century and advanced significantly in the 19th century with the legalization of cadaver dissection in Europe [2]. These developments not only deepened the understanding of human anatomy but also established a foundation for modern medicine and psychology by linking anatomical structures to physiological and cognitive functions.

A major milestone in brain mapping was reached in the early 20th century, with Korbinian Brodmann [3], who divided the cerebral cortex into 52 distinct regions based on cytoarchitectonic differences. Building on Brodmann's work, the **Talairach system** (Talairach atlas), developed by Talairach and Tournoux in 1988, introduced a three-dimensional coordinate system that became a standard for brain mapping [4]. This system allowed researchers to localize brain structures independently of individual differences in brain size and shape. However, the Talairach atlas was based on a single left hemisphere of a post-mortem brain, making it inherently limited in representing population-level anatomical variability. This individual-specific approach also made it difficult to generalize findings across studies or accurately transform data between individuals [5].

To address the limitations of the Talairach system, Montreal Neurological Institute created the MNI coordinate system (*MNI space*) in the 1990s. By co-registering 305 MRI scans and creating an average brain template, the MNI space provided a more generalized and population-based coordinate system [6]. This allowed for more robust qualitative and quantitative analyses of brain structures, facilitating comparisons across studies and improving the reproducibility of neuroimaging research [6][7].

Automated segmentation within *magnetic resonance imaging* (MRI) has since become an essential tool for performing quantitative image analysis and providing contextual information, such as detecting lesions or delineating specific brain structures. While manual segmentation by an expert radiologist is considered the "gold standard" for distinguishing between different brain tissues, it is not without its limitations [8]. Manual segmentation is an inherently time-consuming and complex task, making it impractical for analyzing the large volumes of MRI data commonly encountered in research practice [8].

Moreover, although manual approaches are often regarded as accurate, they are prone to intra- and inter-rater variability, where the same radiologist may produce different results on separate occasions and different radiologists may disagree on the segmentation [8].

These inconsistencies make it challenging to establish a definitive standard of correctness, as there is no absolute ground truth against which manual segmentations can be validated [8].

These challenges inherent in manual segmentation have driven the development of various automated segmentation models, each offering different levels of performance and complexity. Automated methods help standardize the segmentation process, reduce variability, and handle the growing demand for high-throughput analysis of large-scale MRI datasets in clinical and research settings [8]. Among the various automated segmentation approaches, *single atlas segmentation* (SAS) represents the simplest form of atlas-based segmentation, where a single reference atlas, consisting of an intensity image and its corresponding labeled image, is used to segment a new, unlabelled image (Figure 1.1) [9]. This process typically involves nonrigid image registration to align the atlas with the target image, followed by label propagation to assign tissue labels to the corresponding regions [10]. More specifically, a “labeled image” is a medical image in which each voxel or pixel is assigned a category representing a specific tissue type, anatomical region, or other relevant object class (such as background), whereas “modality image” refers to the original grayscale image containing voxel values that reflect modality-dependent signal characteristics, such as signal intensity in MRI, radiodensity in CT or echogenicity in ultrasound [10].

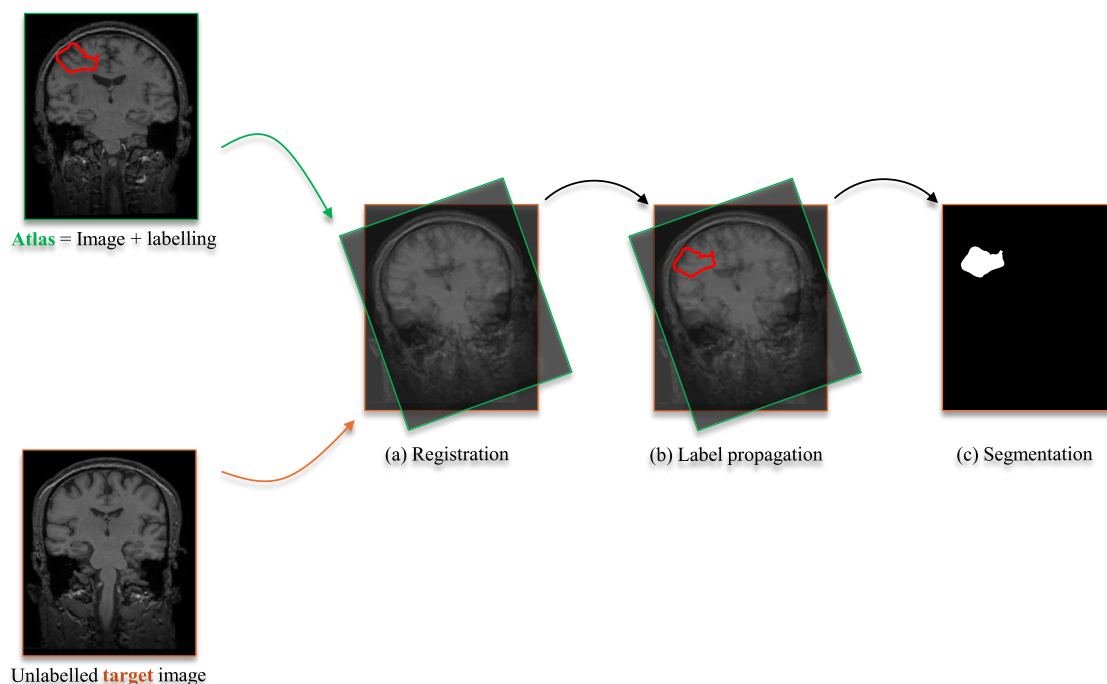


Figure 1.1: Steps in the single-atlas segmentation process with brain MR-images. The atlas image is registered (a) to the target image, and its labels (within the red contour) are propagated (b) to produce the final segmentation (c).

Although SAS was one of the earliest automated segmentation models, it comes with significant limitations [8]. Since anatomical structures exhibit natural inter-individual variability, a single atlas cannot capture the full spectrum of anatomical differences present across a population [8]. Consequently, segmentation errors can arise when the test image deviates significantly from the atlas due to factors such as individual variation, age,

pathology, or imaging artifacts. Furthermore, errors introduced during the registration process can propagate through the segmentation pipeline, leading to suboptimal results [8][11].

To address these shortcomings, more advanced approaches, such as *multi-atlas segmentation* (MAS), have been developed. MAS improves upon SAS by utilizing multiple atlases instead of a single reference image [11]. By incorporating multiple atlases, MAS can better account for anatomical variability, increase robustness and improve segmentation performance through statistical label fusion techniques [11][12].

In MAS, atlases function as independent raters, each consisting of a manually segmented image and its corresponding labels, which are separately registered (*pairwise registration*) to the unlabelled target image. This registration is performed using a rigid or affine transformation for global alignment, to align the atlas and target image to the same coordinate system, followed by a nonlinear transformation to capture local anatomical variations. Furthermore, to create the final segmentation, the labels transferred from each atlas are fused using a consensus-based approach in a process known as *label fusion*. The most common types of label fusion methods are majority voting, STAPLE (Simultaneous Truth and Performance Level Estimation) and Bayesian fusion models [11][12]. In *majority voting*, each voxel is assigned the most frequently occurring label among the registered atlases (*i.e.* voxelwise label likelihoods), assuming all atlas labels are equally weighted. This would yield a final segmentation output, but some fusion schemes utilize *voting maps*, which map the occurrence of each label per voxel for further processing and refinement using machine learning (*i.e.* *voxel classification*) or statistical modeling. For a more thorough summary of the multi-atlas pipeline see Figure 1.2.

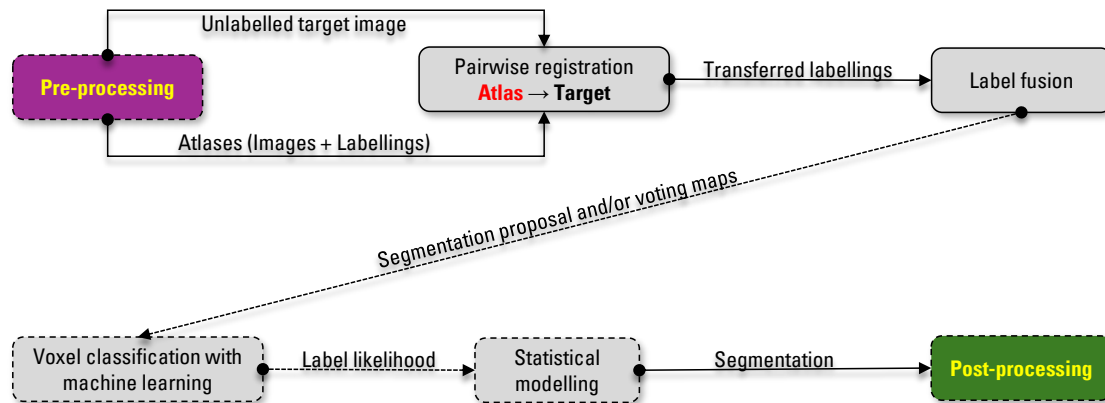


Figure 1.2: A schematic overview of a multi-atlas process, where optional sub-steps are indicated with dashed blocks and arrows. Solid lines represent mandatory sub-steps which include pairwise registration of the atlas images and label propagation onto the unlabelled target, followed by label fusion. The label fusion step integrates these labellings to generate the final segmentation. Alternatively, a voxel classification approach (dashed boxes) may be employed, where label likelihoods are estimated and subsequently refined using statistical modeling prior to segmentation.

While MAS has proven to be a robust segmentation method, it still faces challenges when the target anatomy deviates significantly from the atlas population [12][13]. Large anatomical variations, such as ventricular enlargement in neurodegenerative diseases, can lead to misalignment during registration, reducing segmentation performance [12]. This led to the development of *MAPER* (Multi-Atlas Propagation with Enhanced Registration) [12].

MAPER improves the registration process by incorporating tissue classification alongside traditional intensity-based registration[8][12]. By aligning tissue probability maps in the early registration stages, MAPER increases robustness against anatomical variability and improves segmentation performance, particularly in cases with structural deformations [12]. This enhancement allows for more reliable segmentation across both healthy and pathological datasets, extending the applicability of MAS in clinical and research settings [12].

Study Aim and Purpose

Problem: The absence of genuine ground-truth segmentations for real MR images prevents rigorous benchmarking and validation of atlas-based anatomical segmentation methods.

Solution: This thesis generates synthetic MR-like images using pre-existing anatomical label sets. These label sets provide known ground-truth segmentations corresponding directly to the synthetic MR images. Several distinct image synthesis methods are developed and compared.

Validation: Different synthetic benchmark configurations derived from the developed synthesis methods are systematically tested. The sensitivity and reliability of each configuration are evaluated by varying experimental conditions, such as the size of the training atlas sets, and performing leave-one-out cross-validation across three image datasets—two synthetically altered and one conventional, unmodified.

Demonstration: The practical utility of the validated benchmark framework is demonstrated by comparing two variants of the automated segmentation algorithm MAPER, onepad versus no-onepad. This serves as an applied example of the benchmark framework’s capability, rather than constituting an independent research aim.

2

Preliminaries

2.1 Fundamental Definitions

- **Atlas:** An *atlas* generally consists of a manual segmentation paired with its corresponding intensity image.
- **Classes:** The term *classes* refers to the predetermined object categories, such as anatomical structures or tissue types (i.e white and gray matter or cerebrospinal fluid), that each voxel in a medical image may be labeled as during the segmentation process.
- **Classification:** Image *classification* assigns a single label to an entire image (i.e car, bicycle or truck *etc.*), while voxelwise classification assigns a label to each individual voxel, enabling detailed spatial segmentation.
- **Ground truth/Gold standard:** A *ground truth* (or ground truth labelling) refers to a reference image that has been manually segmented by a qualified medical expert, such as a radiologist, and is considered the standard for evaluating automated methods.
- **Image:** An *image* can have different definitions depending on the field, but in the context of this thesis, it refers to a three-dimensional matrix composed of elements (voxels), each representing a gray-scale intensity value corresponding to the underlying anatomical structures measured with MR-scanner.
- **Label:** Voxels can be assigned a specific *label*, represented as an integer value that refers to an object class. For instance, a label value of 0 might represent *background class*, while 1 could represent *bone tissue class* in a medical imaging dataset.
- **Labelling:** Image *labelling* refers to an integer matrix matching the dimensions of its reference image, in which each voxel is assigned a specific label corresponding to an object class, typically done automatic or manually.
- **Probability map:** A *Probability map*, is a spatial map where each voxel holds a value indicating the likelihood of belonging to a specific class or structure.
- **Segmentation:** Synonymous with labelling, but *segmentation* typically refers to a labelling process that is performed automatically or semi-automatically.
- **Target:** The *target* image is an unlabelled image selected for segmentation.
- **Voting map:** A *voting map* that aggregates label predictions from multiple atlases, where each voxel's final label is determined based on a majority vote or weighting scheme.
- **Voxel:** A 3D pixel representing the smallest unit of volume in a 3D image.

2.2 Multi-Atlas Segmentation

Multi-atlas segmentation is a more sophisticated and advanced approach compared to single-atlas segmentation. Here, the atlas refers to the intensity image and its associated labelling. In single-atlas segmentation, a modality image (“atlas image” or “source image”) is registered to the unlabelled target image, followed by label propagation in which the labels are transferred based on the estimated transformation (Figure 1.1). This means that the target image occupies the same anatomical space as the source image after registration, making segmentation inherently dependent on registration accuracy; hence, the method is termed registration-based segmentation. In contrast, multi-atlas segmentation involves the use of multiple atlas images, each registered individually to the unlabelled target image (Figure 2.1). This approach captures anatomical variability across different atlases and enhances the robustness of the segmentation process. After registration, the corresponding labels from all atlases are propagated to the target space (according to pairwise atlas-target registration), followed by a *label fusion* step that integrates the multiple segmentations into a final consensus segmentation.

Various label fusion approaches are available. One of the simplest is the *vote-rule decision fusion*, where each voxel in the target image is assigned the label that occurs most frequently (a plurality) across the registered atlas segmentations.

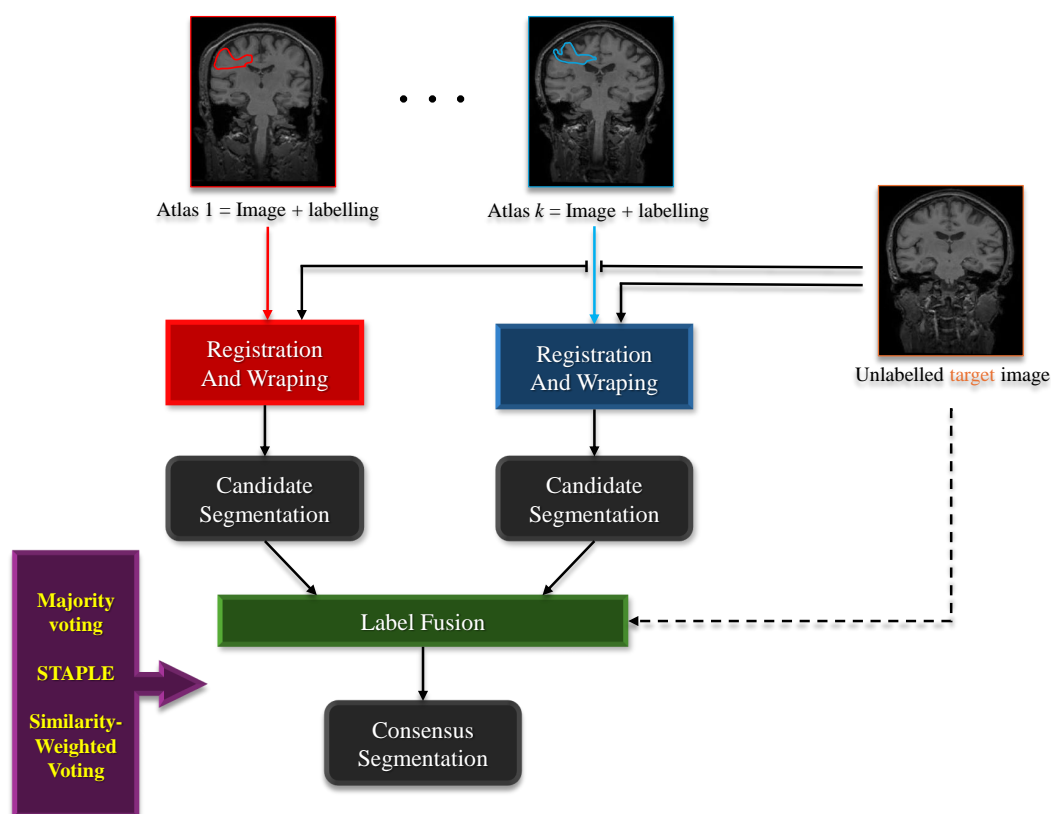


Figure 2.1: In multi-atlas segmentation two or k number of atlas images with their corresponding labelling are each registered and warped to the target image. This results in multiple candidate segmentations, which are then combined through label fusion to produce a final consensus segmentation. Between the label fusion and the final consensus segmentation, methods such as majority voting, STAPLE or similarity-weighted voting can be employed to combine the candidate segmentations [14].

2.2.1 Multi-Atlas Propagation with Enhanced Registration

MAPER is a robust and accurate method for whole-brain segmentation based on multi-atlas label propagation combined with *decision fusion*. Decision fusion refers to the process of combining multiple candidate segmentations—each derived by registering a different atlas to the target image—into a single, unified segmentation (Figure 2.1). Originally introduced by Heckemann et al. (2006) [10], MAPER works by registering a set of manually segmented brain atlases to a target MR image using high-dimensional non-rigid registration. The anatomical labels from each atlas are propagated to the target space, and a consensus segmentation is derived via voxel-wise vote-rule decision fusion.

In an enhancement described by Heckemann et al. (2010) [12], the registration process was modified to include tissue-class information during the initial stages. Specifically, probabilistic tissue maps (e.g., grey matter, white matter, cerebrospinal fluid/CSF) obtained through automatic classification are used to improve coarse alignment between atlas and target. This tissue-informed registration, followed by conventional intensity-based fine registration, significantly improves segmentation robustness, especially in persons with pathological changes such as *ventriculomegaly*. The enhanced pipeline preserves high segmentation performance across healthy and diseased brains while increasing applicability to heterogeneous datasets such as those from aging or Alzheimer’s disease populations.

Performance evaluations demonstrated that MAPER achieves high agreement with manual reference segmentations as measured by overlap indices (e.g. Jaccard coefficient; see Section 2.3) for numerous brain structures and shows substantial improvements in segmenting enlarged ventricles compared to intensity-only methods. Notably, the method achieved a success rate of 9/9 for cases with enlarged ventricles, compared to only 4/9 with the baseline approach.

2.3 Evaluation of Segmentation Algorithm

The most commonly used measures for evaluating segmentation performance are based on the overlap between the automated segmentation and a reference segmentation—typically expert-generated manual delineations—which serve as the closest available surrogate since true anatomical ground truth does not exist for in-vivo MR image segmentation (Figure 2.2).

One such metric is the *Jaccard index*, also known as the *Intersection over Union (IoU)*. It quantifies the similarity between two sets by dividing the number of overlapping elements by the number of elements in the union of the sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2.3.1)$$

Here, A denotes the set of voxels in the automated segmentation and B the set in the reference segmentation. The Jaccard index ranges from 0 (no overlap) to 1 (perfect overlap). It penalizes both false positives and false negatives, making it a robust and interpretable metric for segmentation quality.

In the context of anatomical segmentation, the Jaccard index is first computed for each anatomical region in a given target test image, resulting in a set of regional scores $\{JC_1, JC_2, \dots, JC_n\}$ (Figure 2.3). These scores quantify the segmentation performance

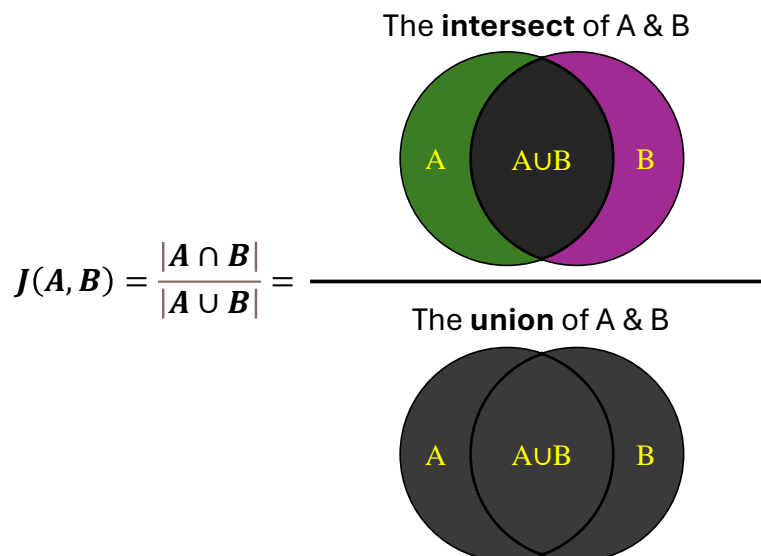


Figure 2.2: Jaccard index $J(A,B)$, a commonly used similarity metric for evaluating segmentation overlap. It is defined as the ratio between the size of the intersection $|A \cap B|$ and the size of the union $|A \cup B|$ of two sets (A) and (B). The intersection represents the common elements between the two segmentations, while the union includes all elements present in either segmentation.

for each region. To obtain a summary measure for that test image, the regional Jaccard scores are averaged into a individual-level mean Jaccard Index:

$$\overline{JC}_i = \frac{1}{n} \sum_{k=1}^n JC_k \quad (2.3.2)$$

where n is the number of anatomical regions and i denotes the target individual. To evaluate the overall performance of the segmentation algorithm, these individual-level means are further averaged across all test targets:

$$\overline{JC}(\text{fn}) = \frac{1}{N} \sum_{i=1}^N \overline{JC}_i \quad (2.3.3)$$

where N is the number of target images and fn represents the number of atlases used in the segmentation. This hierarchical averaging procedure produces a single summary measure of segmentation performance as a function of atlas count. However, it does not retain information about regional variability within individuals or consistency across participants; such variability must be assessed separately using dispersion metrics, such as standard deviation.

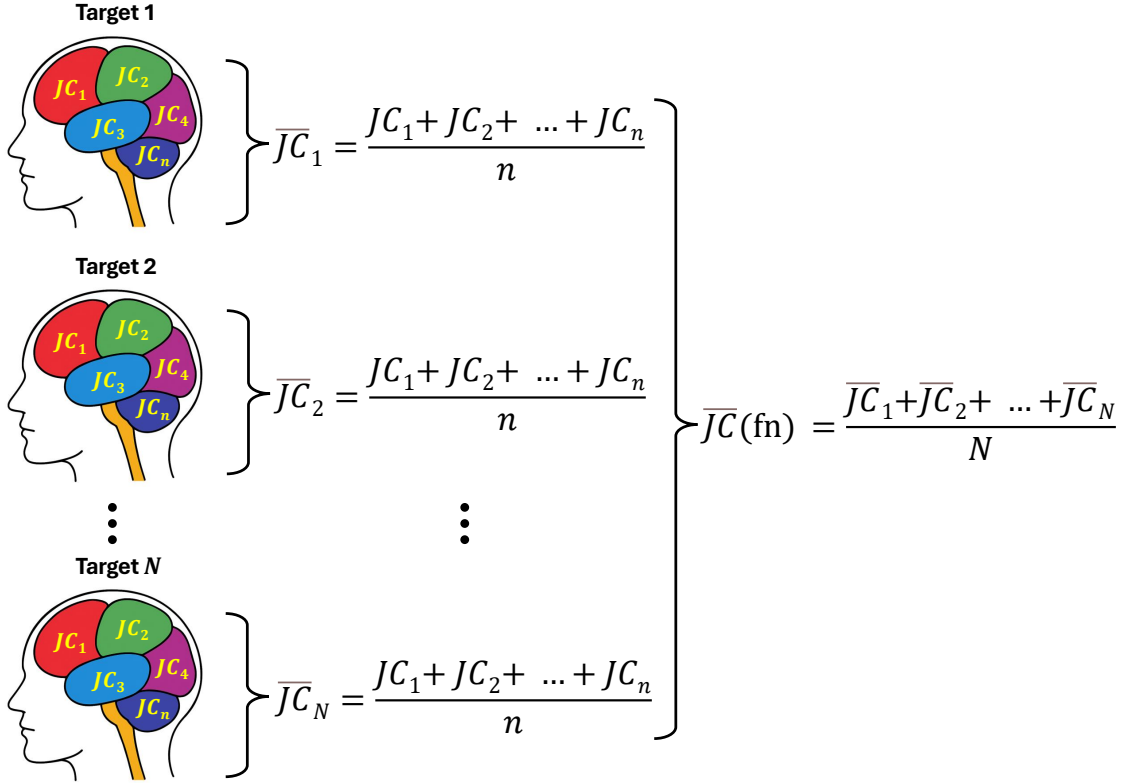


Figure 2.3: A hierarchical averaging process for the Jaccard Index. Individual region scores (JC_1 to JC_n) are first averaged per target individual, resulting in individual-level means (\overline{JC}_1 to \overline{JC}_N). These are then averaged across all participants to obtain the overall mean Jaccard Index $\overline{JC}(\text{fn})$ for a given number of atlases fn .

2.3.1 Benchmark performance & Synthesizing MR-images

Benchmarking segmentation algorithms using clinical datasets provides a realistic and clinically relevant means of evaluating performance. Unlike synthetic data, which is generated under idealized and controlled conditions, clinical MR-images contain natural anatomical variability, imaging artifacts and heterogeneity across individuals, scanners and acquisition protocols. When annotated datasets are available—i.e. clinical MR-images accompanied by expert manual segmentations—they can serve as a trusted reference for performance evaluation. Automated segmentations can then be quantitatively compared to these references using standard similarity metrics such as the Jaccard Index.

By applying the segmentation algorithm to a clinical dataset with known reference labels, and aggregating performance metrics across multiple brain regions and individuals, researchers can establish a benchmark that reflects the method's performance and robustness in real-world data. This process helps identify strengths and weaknesses of the algorithm, informs parameter tuning, and enables fair comparisons across studies or against other segmentation approaches.

However, clinical datasets are extremely rare and costly to acquire. Furthermore, expert manual segmentations, often used as references, inherently deviate from the unknowable true anatomical ground truth by an unknown degree, limiting their reliability for systematic testing and evaluation. To address this, synthetic MR-images can be generated to simulate specific anatomical or imaging scenarios, while preserving known ground truth

2.3. Evaluation of Segmentation Algorithm

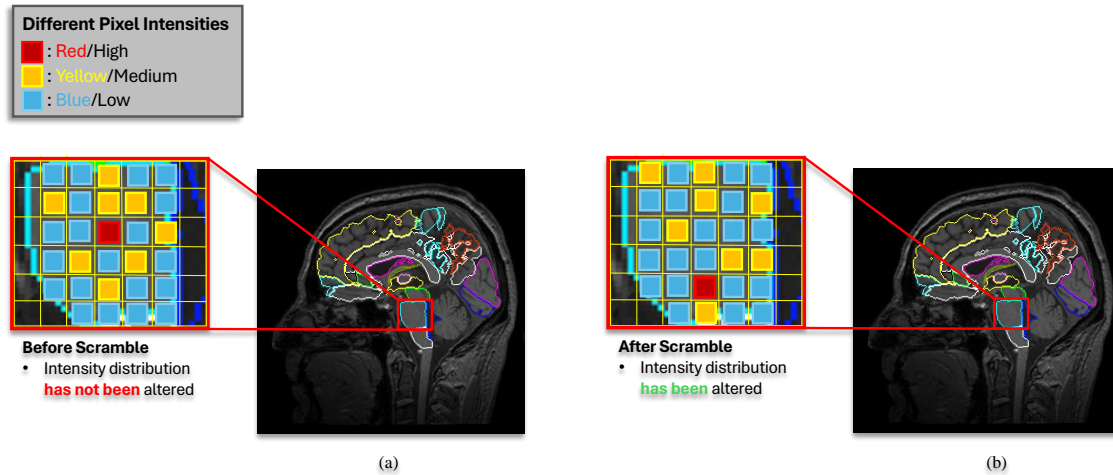


Figure 2.4: Illustration of intensity scrambling within a segmented brain region. (a) Before scrambling, the original spatial organization of intensity values is preserved. (b) After scrambling, the same intensity values have been randomly redistributed within the region, altering the spatial layout while maintaining the original intensity distribution. The resulting image constitutes a synthetic MR-image, which can be used to evaluate the sensitivity of segmentation algorithms to spatial arrangement.

structures [15]. One example involves randomly permuting the voxel intensities within each segmented region of a brain MR-image (called *scramble* in this work), see Figure 2.4. This preserves the overall regional shape while eliminating any internal tissue structure, which can be useful for evaluating whether a segmentation algorithm relies on spatial cues or intensity patterns. Other modifications might include applying regional smoothing, intensity scaling, local deformations to simulate specific imaging conditions or anatomical variability.

3

Method

3.1 Participants and MRI dataset

The MRI datasets used in this study were the IXI and Hammers atlas databases, obtained from the Brain Development repository [15]. All images for both datasets were provided in NIfTI format, with only T1-weighted MR-images included in the analysis. T1-weighted images were selected because they provide strong intensity contrast between distinct brain tissues, which is essential for accurate structural segmentation. The IXI dataset was released under the Creative Commons CC BY-SA 3.0 license.

The IXI dataset comprised approximately 600 MR-images acquired from healthy adult volunteers. Data collection was carried out across three hospitals in London: Hammersmith Hospital, where a Philips 3T scanner was used; Guy’s Hospital, employing a Philips 1.5T scanner; and the Institute of Psychiatry, utilizing a GE 1.5T scanner. The imaging protocol for each volunteer included T1-weighted, T2-weighted, proton density (PD)-weighted sequences, magnetic resonance angiography (MRA) and diffusion-weighted imaging with 15 directions [16].

The Hammers dataset consisted of 30 healthy volunteers (15 men and 15 women) aged between 20 and 54 years (median age 30.5 years). Scanning was performed at the National Society for Epilepsy MRI Unit in Buckinghamshire, UK. MRI acquisition was carried out using an inversion recovery followed by a fast spoiled gradient-recalled echo (FSPGR) sequence in the coronal plane, producing T1-weighted 3D volumes. The imaging parameters were as follows: echo time (TE) of 4.2 ms, repetition time (TR) of 15.5 ms, inversion time (TI) of 450 ms and a flip angle of 20°. The dataset comprised 124 slices, each with a thickness of 1.5 mm, covering a field of view (FOV) of 18 × 24 cm with a matrix size of 192 × 256 and a number of excitations (NEX) equal to 1 [10][17].

Each MR-image in both the IXI and Hammers datasets was accompanied by anatomical segmentations and corresponding tissue classification maps for white matter, gray matter, and CSF, resulting in a segmentation of the brain into 482 distinct regions.

3.2 Generation of Synthetic Ground Truth Images

The creation of five different types of synthetic ground truth images based on the IXI and Hammers datasets (only T1-weighted images) was performed using the R Project programming environment, together with the biomedical image processing library ANTsR, an R interface to the Advanced Normalization Tools (ANTs) framework that enables advanced medical image registration, segmentation, and analysis. In this project, ANTsR was used to extract voxel intensities from the MR-images with 482 segmentation regions.

For the first and simplest case, the extracted voxel matrix for each brain region was spatially rearranged by randomly permuting the voxel positions within the segmentation mask using the `sample()` function, resulting in a *scrambled* image (termed *scramble*). This procedure preserved the original intensity distribution but disrupted the spatial order of voxels within each region. The permuted intensities were then reassigned to their original anatomical locations, thus maintaining the outer shape of each region while scrambling its internal spatial structure (see Figure 2.4 in subsection 2.3.1). For the *smoothed* image (*smooth*), the voxel intensities within each brain region were homogenized by computing the median intensity using the `median()` function. This median value was then assigned to all voxels within the corresponding region, effectively unifying the intensity within the label while preserving anatomical boundaries (see Figure 3.2.1). With the smoothed images, an

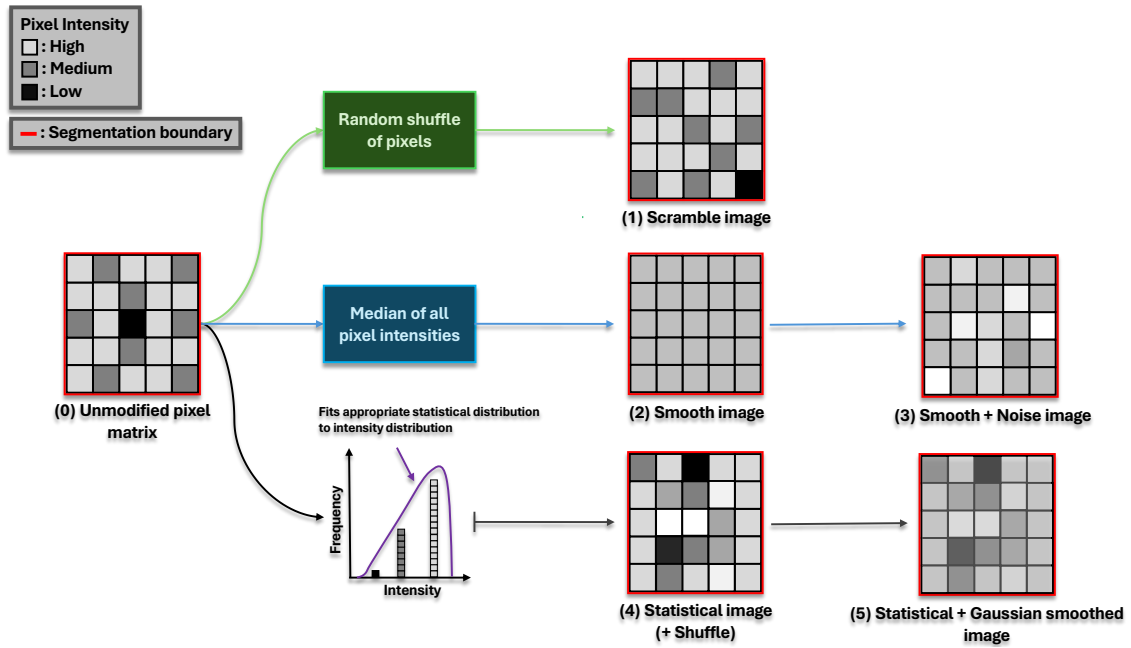


Figure 3.2.1: Five different types of synthetic MR-images were generated from the (0) original unmodified data: (1) scrambled images produced by randomly rearranging the spatial positions of voxel intensities within each region; (2) smoothed images generated by replacing all voxel intensities within the segmented region with the median intensity value; (3) smoothed images with added Rician noise; (4) statistical images created by fitting an appropriate statistical distribution to each region; (5) statistically generated images further processed with gaussian smoothing. The red outline surrounding each pixel matrix indicates the boundary of the segmented region.

additional step was implemented to generate a third set of synthetic images by adding *Rician noise*. Rician noise was specifically chosen because it better modeled the noise characteristics observed in real MR images, more precisely than Gaussian noise [15]. The mean background noise level, $\bar{\sigma}_b$, was estimated by automatically computing the mean intensity within four predefined regions of interest (ROI) located in the corners of a central coronal slice. The resulting values were then averaged to yield a robust estimate of the overall background noise level.

Based on the estimated background noise level $\bar{\sigma}_b$, Rician noise was simulated across the entire synthetic brain volume. Two independent zero-mean Gaussian noise components were generated using `rnorm(length(image), mean = 0, sd = $\bar{\sigma}_b$)`, each with standard deviation $\bar{\sigma}_b$. These components were denoted by n_r and n_i , representing the real and imaginary noise terms, respectively. A magnitude transformation was then applied to the smoothed synthetic image as follows:

$$I_{\text{noisy}} = \sqrt{(I + n_r)^2 + n_i^2}, \quad (3.2.1)$$

where I denotes the voxel intensity matrix of the smoothed image. The resulting I_{noisy} represents the new voxel intensities with simulated Rician noise. These values were then inserted into the corresponding locations within the brain mask, leaving anatomical boundaries unaffected. This simulation approach follows established conventions, such as those implemented in the BrainWeb project [15].

To generate the fourth type of synthetic image, referred to as the *statistical* image (*stat*), the intensity distribution within each anatomical brain region was analyzed individually. For each region, all voxel intensities were extracted and candidate probability distributions were fitted to the data. The set of candidate distributions included the Normal, Log-normal, Gamma, Weibull, Inverse Gaussian, Laplace, Beta prime, Cauchy, Exponential, Generalized Error, Gumbel, Inverse Gamma, Inverse Weibull, Kumaraswamy, Log-logistic, Logistic, Logit-normal, Nakagami, Pareto, Power distribution, Rayleigh, Skew Generalized Error, Skew Normal, Skew Student-t, Student-t, and Uniform distributions. For each fitted model, the Akaike Information Criterion (AIC) was calculated to evaluate the goodness of fit. The distribution yielding the lowest AIC value was selected as the best-fitting model for that region. To ensure reliable statistical modeling, this procedure was only applied to regions containing at least 20 distinct voxel intensities, as smaller regions were deemed insufficient for robust distribution fitting. For each eligible region, new synthetic voxel intensities were randomly generated from the selected distribution to match the number of voxels in the original region. This approach ensured that the synthetic image preserved the overall anatomical structure while introducing realistic, region-specific intensity variability based on statistical modeling (Figure 3.2.1).

The fifth and final synthetic image, referred to as the *statistical smooth* image (*statsmooth*) was generated by applying spatial smoothing to the statistical synthetic image. The smoothing was performed using the `seg Maths` tool from the NiftySeg software package, with a Gaussian smoothing kernel of standard deviation $\sigma = 2$ mm [18].

Specifically, the entire statistical MR image was smoothed by applying the `-smo 0.5` option, which convolved the image with a three-dimensional Gaussian filter of 2 mm standard deviation. This process moderated local intensity variations by smoothing transitions between adjacent brain regions, each characterized by their own statistical intensity distributions that could differ markedly. Although some pixel intensities increased overall, the smoothing filter reduced sharp intensity differences at regional boundaries, resulting in more gradual and natural transitions. This preserved the region-specific statistical intensity profiles while improving spatial coherence, producing softer voxel-level gradients and less distinct regional boundaries compared to the unsmoothed statistical image.

The five types of synthetic brain MR image described above are summarized and visualized in Figure 3.2.2. The figure illustrates the original unmodified MR-image (0) alongside each of the generated variants: the *scramble* (1), *smooth* (2), *smoothnoise*

(3), *stat* (4) and *statsmooth* (5). Each synthetic image or *benchmark* highlights different aspects of synthetic variation, enabling systematic benchmarking of segmentation algorithm performance under controlled intensity alterations.

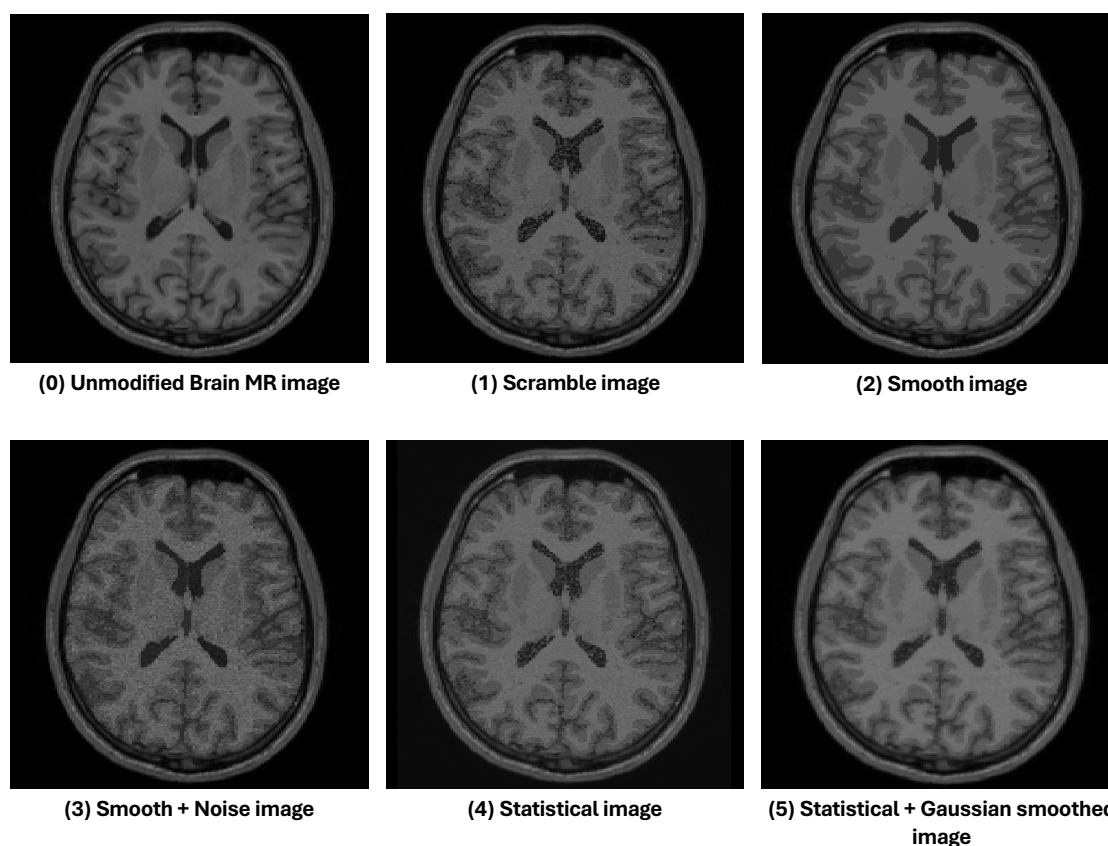


Figure 3.2.2: Examples of the original unmodified brain MR-image (0) and five generated synthetic images: *scramble* (1), *smooth* (2), *smoothnoise* (3), *stat* (4) and *statsmooth* (5) from the same volunteer. All images are displayed in the axial plane (horizontal slice through the brain).

3.3 Experimental Setup

3.3.1 Synthetic Images as Atlases

Following the generation of synthetic atlases, a quantitative evaluation was conducted to assess their segmentation performance relative to conventional atlases with MAPER onepad version. For each of the 30 Hammers target images, MAPER was applied using one of three atlas sets: *conventional*, *scramble* and *stat* each consisting of 40 MR-images from the IXI dataset. The segmentation performance was quantified using the mean *JC* computed across all anatomical regions in each target image.

To enable statistical comparison between methods, the results were compiled into paired datasets corresponding to the same target segmented under different atlas conditions. Three pairwise comparisons were evaluated: *conventional* vs. *scramble*, *conventional* vs. *stat*, and *scramble* vs. *stat*. For each pair, a two-sided paired *t*-test was performed to determine whether the mean segmentation performance differed significantly between methods. The distribution of segmentation performance was further visualized using violin

plot, illustrating method-specific variability across targets and supporting the interpretation of the statistical results.

3.3.2 Leave-One-Out Cross-Validation on Synthetic Images

The second experiment involved performing leave-one-out cross-validation (LOOCV), as illustrated in Figure 3.3.1, using 32 randomly selected MR-images from the IXI dataset. This analysis was carried out separately for each of the three atlas sets: conventional images, scramble and stat synthetic images. In the LOOCV procedure, one image was iteratively held out as the target while the remaining 31 images were used as atlases for segmentation. This process was repeated 32 times, ensuring that each image was used exactly once as the target. To statistically compare the performance of these methods, three

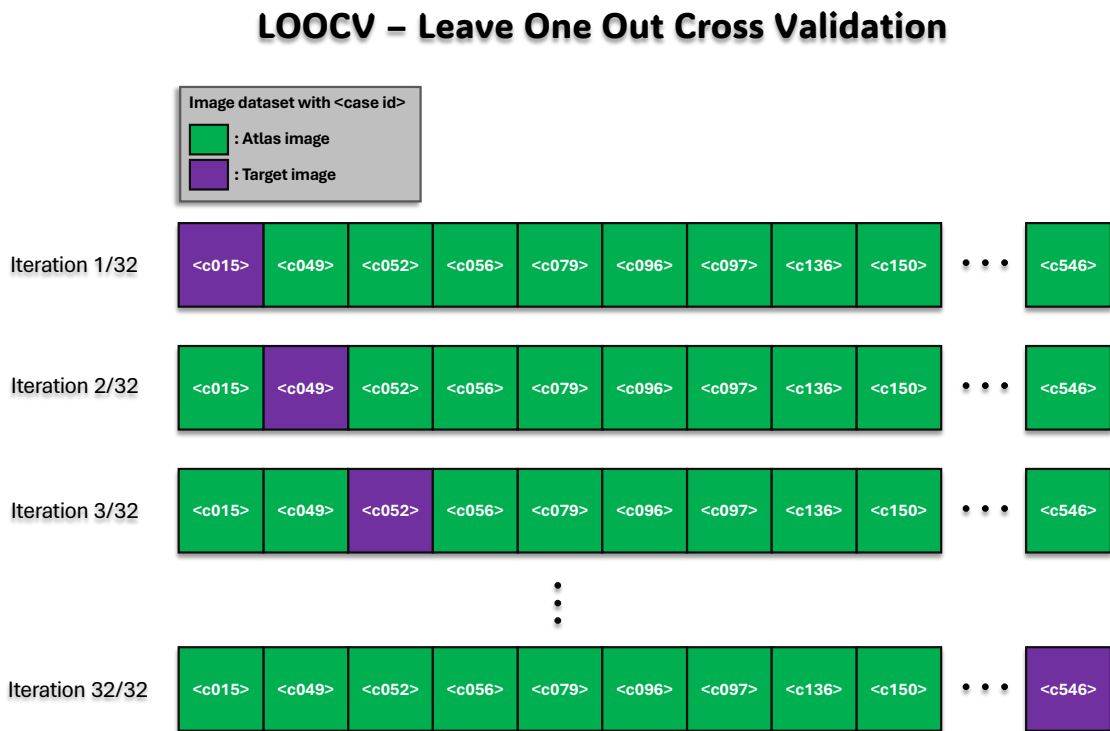


Figure 3.3.1: The LOOCV used with MAPER. In each iteration, one image from the dataset was designated as the target image (purple), while the remaining images served as atlases (green). This procedure was repeated 32 times, ensuring that each image was used once as a target and 31 times as part of the atlas set. Each volunteer is identified by a case ID displayed within angle brackets (e.g., <c015>).

pairwise comparisons were conducted: conventional vs. scramble, conventional vs. stat, and scramble vs. scramble. For each comparison, a two-sided paired t -test was used to assess whether the mean segmentation performance differed significantly between methods. The statistical results were tabulated and included t -values, degrees of freedom, p -values, and indicators of statistical significance at the $\alpha = 0.05$ level. To complement the statistical analysis, the distribution of mean JC scores across all 32 LOOCV targets was visualized using violin plot. Each method was plotted separately, with individual data points shown to reflect per-target variability.

3.3.3 Benchmarking the Ground Truth Benchmark

To benchmark the synthetic targets, or, in other words, the ground-truth benchmark, two separate atlas setups were created: one using 30 Hammers images and the other using 32 IXI images. MAPER was then applied to segment a set of 14 target images, which included synthetic variants generated through `scramble`, `smooth`, `smoothnoise` (smoothing combined with Rician noise), `stat` and `statsmooth`. The conventional (non-synthetic) images were also segmented to serve as a reference, ensuring that the synthetic data exhibited segmentation behaviour consistent with that of conventional target images.

To further investigate the influence of atlas set size on segmentation performance, the experiment proceeded with a fusion-based analysis in which the number of atlases (denoted by fn) was systematically varied. For the Hammers dataset, the number of atlases ranged from three to 29, as the case using all 30 had already been evaluated. For the IXI dataset, the atlas count ranged from three to 31, complementing the existing result with 32 atlases. For each target image, segmentation was performed using randomly sampled subsets of atlas images and label fusion was carried out using majority voting. To ensure reproducibility in the atlas selection process, a fixed-seed pseudorandom number generator was used to control the .

Although both datasets were processed using identical procedures, only the results related to the IXI dataset were considered for benchmarking the ground truth benchmark. The Hammers dataset was included for methodological validation, as a form of internal consistency control, to verify that the segmentation behaves in line with expectations. Since the MAPER method was originally developed and tuned using the Hammers dataset, it was expected to perform optimally on that data. Consequently, the Hammers results were not included in the primary analysis, to avoid potential bias. The IXI dataset, which did not play a role in the MAPER development process, was therefore used as the basis for assessing segmentation performance in an unbiased setting.

Following each segmentation, the mean JC across all 120 brain regions was computed per target image. These values were then averaged across the seven targets to obtain an overall mean JC per fn . This final mean was plotted against the number of fused atlases, separately for the conventional target images and each type of synthetic target image.

To model the relationship between segmentation performance and the number of fused atlases, a parametric function was fitted to the mean JC data points. Specifically, the following model was used, as introduced by Heckemann et al. (2006) [10]:

$$\overline{JC}(fn) = 1 - a - \frac{b}{\sqrt{fn}} \quad (3.3.1)$$

where $\overline{JC}(fn)$ denotes the average of the individual mean scores \overline{JC}_i , computed across all $i = 1, \dots, N$ target individuals segmented with a set of fn fused atlases (see Figure 2.3 in Section 2.3). Parameters a and b represent the asymptotic error floor and the convergence rate, respectively. This formulation captures the commonly observed diminishing returns in multi-atlas segmentation as more atlases are fused.

The model fitting was conducted separately for each combination of source dataset (Hammers or IXI) and synthetic image type (`smooth`, `smoothnoise`, `stat` and `statsmooth`). Aggregated $\overline{JC}(fn)$ values were used to fit the model using nonlinear least squares (NLS)

estimation. Constrained initial parameter values of $a = 0.9$ and $b = 0.1$, and the optimization was limited to a maximum of 100 iterations to prevent infinite loops in case of non-convergence; this limit was never reached. The resulting estimates of a and b were then extracted and stored for comparative analysis across image types and datasets. To evaluate the variability and stability of the estimated convergence parameter b , a non-parametric *bootstrapping* procedure was subsequently performed. For each combination of source dataset (Hammers and IXI) and synthetic image type (scramble, smooth, smoothnoise, stat and statsmooth), a total of 1000 bootstrap iterations were carried out [19]. In each iteration k , the \overline{JC}_i values for the seven target images, covering atlas fusion from $fn = 3$ to $fn = n$ (where $n = 32$ for IXI and $n = 30$ for Hammers), were randomly resampled from the original dataset. The same \overline{JC}_i value for a given target image could be selected multiple times within a fn -group (see Figure 3.3.2). The model defined in Eq.

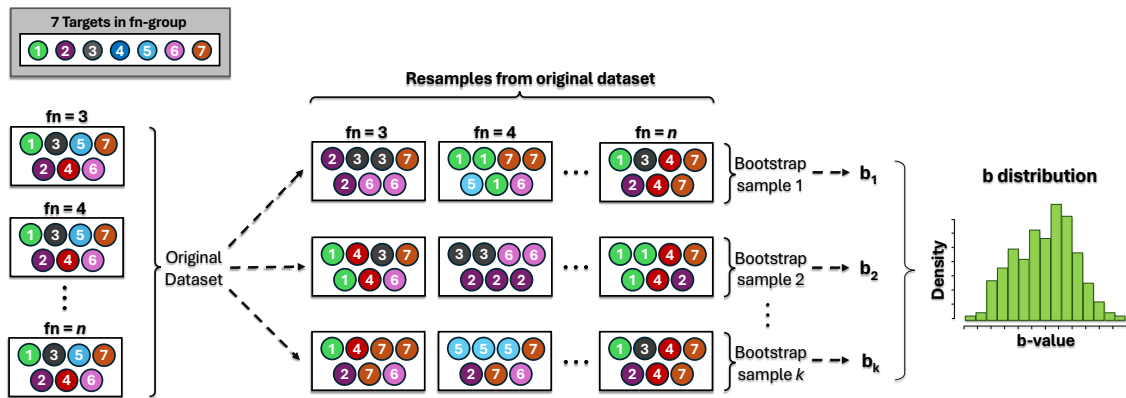


Figure 3.3.2: The original dataset of seven synthetic target in each fusion group fn n was resampled to generate multiple bootstrap samples. For each sample, the segmentation model was fitted and the parameter b was estimated. This procedure was repeated 1000 times for each combination of dataset and synthetic image type, resulting in distributions of b -values used for statistical comparison and confidence interval estimation.

(3.3.1) was then re-fitted to each bootstrap sample using NLS, while keeping the same initial parameter values and convergence criteria as in the primary fitting procedure. For every bootstrap iteration, a single estimate of the parameter b_k was extracted, resulting in a distribution of 1000 b -values per synthetic type and dataset. From these distributions, the mean b -value and corresponding 95% confidence intervals were computed using the 2.5th and 97.5th percentiles. These summary statistics were then used to quantify uncertainty and to enable comparison of convergence behaviour across different synthetic targets.

In addition, pairwise comparisons between synthesis types were performed within each dataset using independent two-sample t-tests applied to the bootstrapped b -distributions. A threshold of $p < 0.05$ was used to determine statistical significance. This allowed for a formal evaluation of whether the convergence profiles differed significantly between the synthesis types. To visualize the variability in the b -estimates, violin plots were generated for each combination of source and target type. These plots provided a compact and interpretable summary of the distributional properties of the fitted b -values, highlighting both central tendency and spread.

3.3.4 Example Application of Ground Truth Benchmark

As an example application of the ground-truth benchmark, a new, hitherto untested development branch of the MAPER project called `no-onepad` was compared to the standard software version residing in the `master` branch. The modification consisted of removing a preprocessing step in which the extracted brain image was "padded" with voxels of value 1, forming a three-voxel-thick shell completely surrounding the brain surface, where background (zero-valued voxels) would otherwise be. This modification had originally served a purpose that has since been addressed by other modifications to the MAPER code. The `no-onepad` branch was created by the developer on the suspicion that this obsolete step was now damaging accuracy, rather than improving it. Dismissing or confirming this suspicion became a test case for the ground-truth benchmark.

For this purpose, the synthetic target benchmarks that had performed best in the benchmarking with bootstrapping was chosen as the basis for evaluating the performance difference between the two MAPER versions. This ensured that the comparison was conducted using the most discriminative and sensitive (higher convergence rate b) target configuration identified in prior analyses. Both MAPER versions were applied to the same set of target images, using identical processing pipelines apart from the presence or absence of brain surface padding.

To ensure that the comparison would be statistically well-powered, the optimal number of target images was determined prior to the experiment through a *power analysis*. This was performed with the following steps, for each of the seven synthetic IXI images. The mean Jaccard index was calculated separately for the two versions, with 32 randomly selected IXI atlases. Using $\overline{JC}_i^{(M)}$ and $\overline{JC}_i^{(N)}$ to denote the mean Jaccard index for the i -th target under the `master` (M) and `no-onepad` (N) configurations, respectively, the paired difference was defined as:

$$\vec{\Delta} = \overline{JC}_i^{(N)} - \overline{JC}_i^{(M)}, \quad i = 1, 2, 3, \dots, 7. \quad (3.3.2)$$

The vector $\vec{\Delta}$, containing the pairwise differences in mean JC , was used to compute the empirical standard deviation $s_{\Delta} = \text{sd}(\vec{\Delta})$. To estimate the number of target images required to detect a difference of $\delta = 0.02$ between the segmentation methods, this value of δ was combined with the computed s_{Δ} in a two-sided paired-sample power analysis (since both MAPER versions were applied to the same targets, a paired design was employed). The calculation assumed a significance level of $\alpha = 0.05$ and a statistical power of $1 - \beta = 0.80$. The required sample size n_s was obtained using the following equation:

$$1 - \beta = \Phi \left(\frac{\delta \sqrt{n_s}}{s_{\Delta}} - z_{1-\alpha/2} \right), \quad (3.3.3)$$

where Φ is the standard normal cumulative distribution function and $z_{1-\alpha/2}$ is the critical value for a two-sided test. This approach ensured that the sample size estimation was grounded in the empirically observed variability of the paired mean JC differences, providing a more accurate and context-specific basis for planning the final comparative analysis. Once the optimal number of targets n_s had been determined, the MAPER algorithm was executed using both the `master` and `no-onepad` configurations across n_s synthetic target images. The resulting segmentations were then subjected to a comparative analysis. The distribution of mean JC at fusion level $\text{fn} = 32$ was visualized using a violin plot with one point per target and coloured connecting lines to indicate the paired relationship between the two MAPER versions.

4

Results

4.1 Synthetic Images as Atlases

The Figure 4.1.1 shows the distribution of mean Jaccard indices obtained for each target when using conventional, scramble, and stat atlas images. The conventional method resulted in the highest segmentation accuracy, with higher central tendency (median and mean) and similar variability compared to the synthetic variants. The stat yielded intermediate performance, while the scrambled method consistently produced the lowest scores. All pairwise comparisons yielded statistically significant differences at the

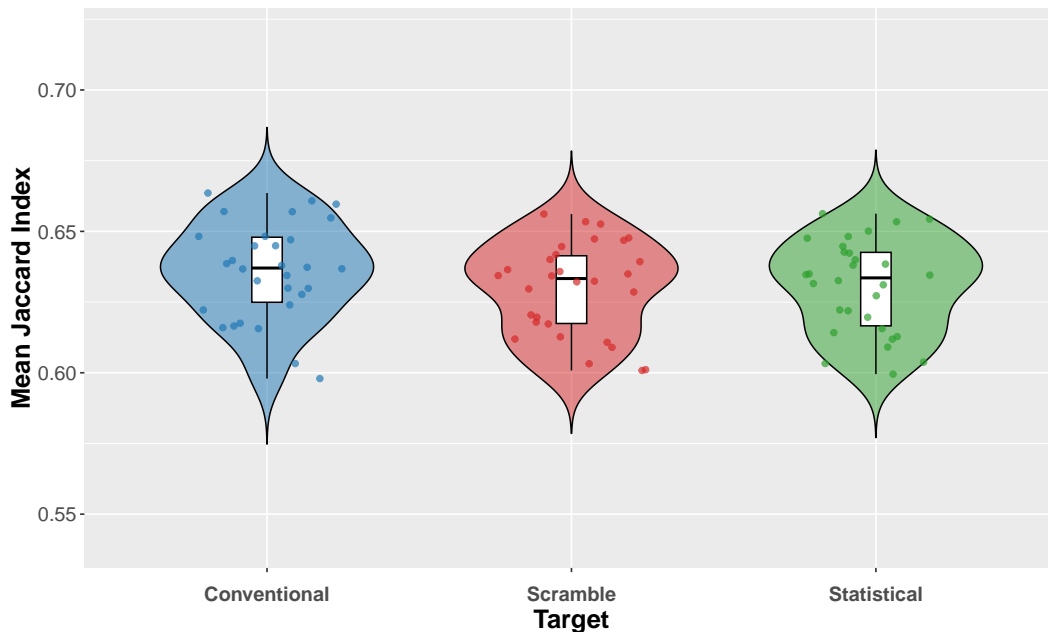


Figure 4.1.1: Distribution of mean Jaccard indices for segmentations using conventional, scramble and stat atlases with $n = 30$ targets from the Hammers dataset. Each point represents one mean Jaccard \overline{JC}_i index from an target image. The conventional atlas outperforms both scramble and stat atlases.

$\alpha = 0.05$ level as seen in Table 4.1.1. The conventional atlas performed significantly better than both the scramble ($t = 7.70$, $p = 1.73 \times 10^{-8}$) and stat atlas ($t = 7.25$, $p = 5.47 \times 10^{-8}$). The statistical atlas outperformed the scramble atlas ($t = -3.02$, $p = 0.0052$). A positive t-value indicates that the statistical atlas had a higher mean Jaccard index (i.e., greater overlap with the whole brain of the reference segmentation), while a negative t-value indicates that the scramble atlas performed better.

Table 4.1.1: Paired t -test results that compared mean Jaccard indices between atlas images (source: 30 Hammers targets). All pairwise comparisons yielded statistically significant differences at the $\alpha = 0.05$ level. Here, $t > 0$ indicated that the first method achieved a higher mean Jaccard index, whereas $t < 0$ meant the second method performed better.

Comparison	t -value	df	p -value	Significant
Conventional vs. Scramble	7.70	29	1.73×10^{-8}	Yes
Conventional vs. Stat	7.25	29	5.47×10^{-8}	Yes
Scramble vs. Stat	-3.02	29	5.20×10^{-3}	Yes

4.2 Leave-One-Out Cross-Validation on Synthetic Images

The distribution of mean Jaccard indices obtained through leave-one-out cross-validation for three target types conventional, scramble, and stat is presented in Figure 4.2.1. The performance of these targets was compared using paired t -tests across 32 targets from the IXI dataset.

All pairwise comparisons revealed statistically significant differences at the $\alpha = 0.05$ level (see Table 4.2.1). Conventional targets significantly outperformed scramble ($t = 3.33$, $p = 0.0022$), as did stat targets ($t = -2.10$, $p = 0.0442$).

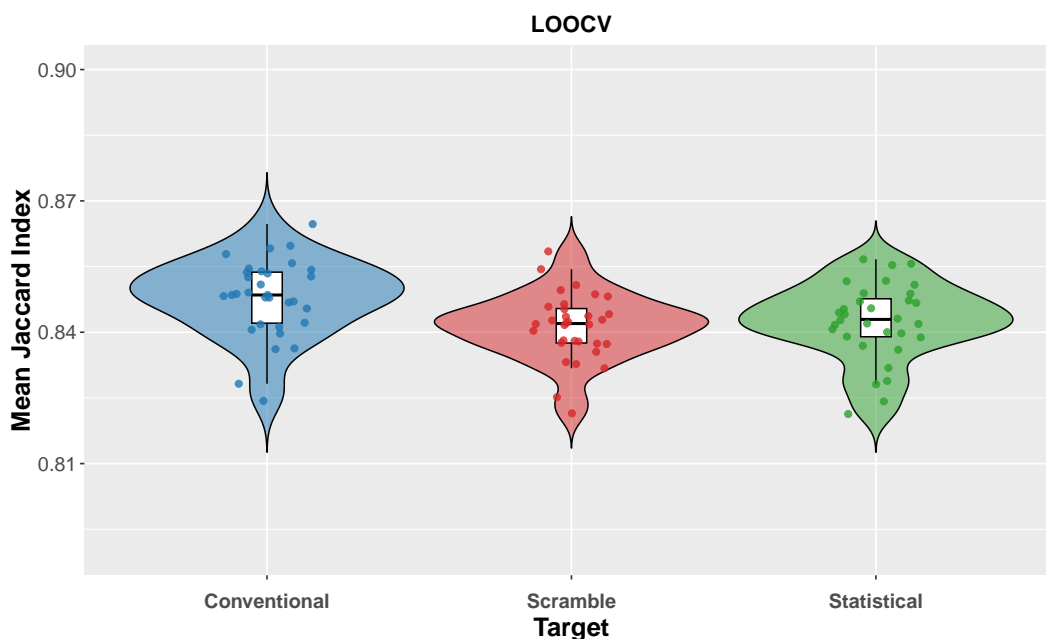


Figure 4.2.1: Mean Jaccard indices from leave-one-out cross-validation across 32 targets in the IXI dataset, comparing segmentation performance for three target types: conventional, scramble and stat. Higher values indicate greater overlap with the reference segmentation. The results show that both conventional and statistical targets outperform scramble targets, with conventional targets achieving the highest overall performance.

A higher Jaccard index indicates greater overlap with the reference segmentation, and a positive t -value means the first target had a higher mean value. Therefore, the negative t -value in the scramble vs. stat comparison confirms that stat targets yielded better segmentation results. Additionally, conventional targets also outperformed stat ones ($t = 2.52, p = 0.0171$), suggesting that while both performed well, conventional targets offered slightly better accuracy under LOOCV setup.

Table 4.2.1: Paired t -test results comparing mean Jaccard indices from leave-one-out cross-validation (source: 32 IXI targets). All comparisons showed statistically significant differences at $\alpha = 0.05$. A positive t -value indicated superior performance of the first target, whereas a negative t -value reflected better results for the second target.

Comparison	t -value	df	p -value	Significant
Conventional vs. Scramble	3.33	31	0.0022	Yes
Conventional vs. Stat	2.52	31	0.0171	Yes
Scramble vs. Stat	-2.10	31	0.0442	Yes

4.3 Benchmarking the Ground Truth Benchmark

Figure 4.3.1 presents both the observed mean Jaccard index values \overline{JC}_i (dotted lines) and the corresponding fitted model curves $\overline{JC}(\text{fn})$ (solid lines) for the five ground truth benchmarks (see observed b in Table 4.3.2). The fitted trends highlight systematic differences in convergence behaviour, with certain target types reaching higher accuracy levels more rapidly as the number of fused atlases increases. Figure 4.3.2 shows the distribution of

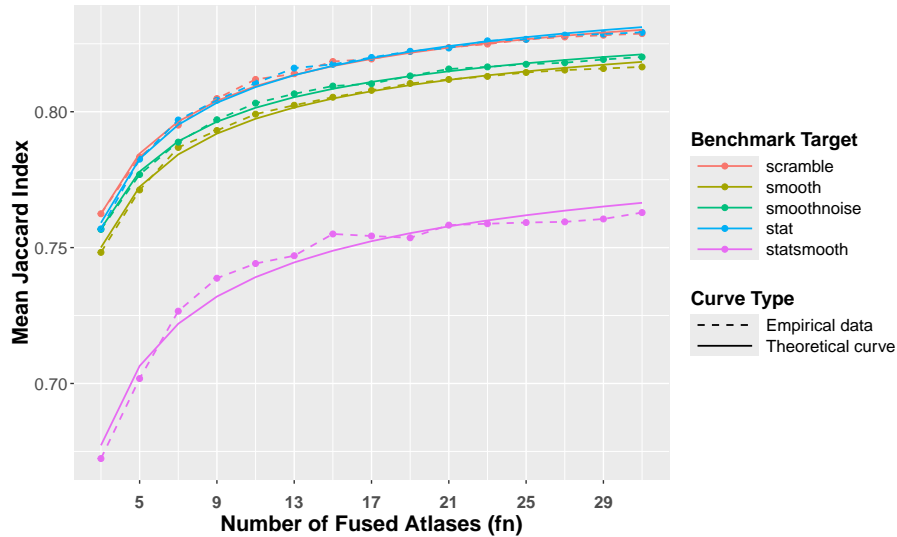


Figure 4.3.1: Mean Jaccard index as a function of the number of fused atlases (fn) for five types of synthetic MR-images: scramble, smooth, smoothnoise, stat and statsmooth. Dashed lines represent \overline{JC}_i (empirical data), while solid lines show the fitted model estimates $\overline{JC}(\text{fn})$ (theoretical curve) for each ground truth benchmark.

1000 bootstrapped b -values across the five synthetic target types. These distributions reflect differences in the rate at which the Jaccard index converges as the number of fused atlases increases.

4.3. Benchmarking the Ground Truth Benchmark

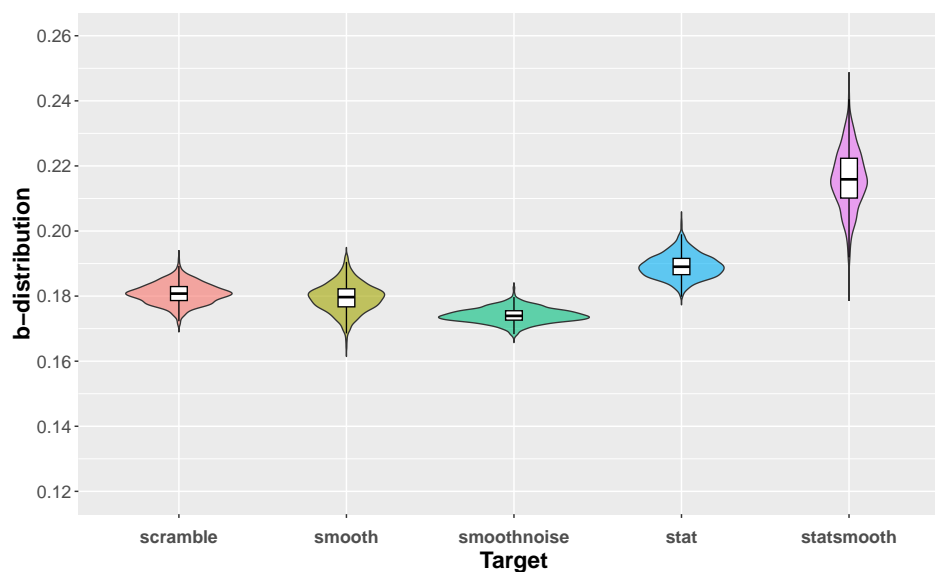


Figure 4.3.2: Distribution of 1000 bootstrapped b -values for five synthetic target types (32 IXI targets). *Statsmooth* showed the highest b -values with greater spread, suggesting faster but more variable convergence. *Smoothnoise* had the lowest values with the narrowest spread, indicating slower but more consistent convergence. The remaining types showed intermediate b -values with relatively compact distributions.

Among the target types, *statsmooth* exhibits the highest median and the broadest spread of b -values, suggesting more rapid convergence. In contrast, *smoothnoise* presents the narrowest distribution and the lowest b -values, indicative of a more gradual convergence pattern. The remaining target types — *scramble*, *smooth*, and *stat* — lie between these extremes and show relatively compact distributions.

These visual differences are statistically supported by the results in Table 4.3.1, where all pairwise t -tests between target methods revealed significant differences at the $\alpha = 0.05$ level. This confirms that the choice of synthetic target has a statistically significant effect on the convergence characteristics of the segmentation model. Further quantitative insight is

Table 4.3.1: Pairwise t -tests of b -values between synthetic target methods (source: 32 IXI targets). Reported values include the t -value, p -value, and whether the result was statistically significant at $\alpha = 0.05$.

Comparison	t -value	p -value	Significant
Scramble vs Smooth	7	1.56e-12	Yes
Scramble vs Smoothnoise	55	<0.0001	Yes
Scramble vs Stat	-55	<0.0001	Yes
Scramble vs Statsmooth	-115	<0.0001	Yes
Smooth vs Smoothnoise	35	<0.0001	Yes
Smooth vs Stat	-53	<0.0001	Yes
Smooth vs Statsmooth	-114	<0.0001	Yes
Smoothnoise vs Stat	-112	<0.0001	Yes
Smoothnoise vs Statsmooth	-141	<0.0001	Yes

provided in Table 4.3.2, which lists the mean b -values and corresponding 95% confidence

intervals for each target type. Statsmooth has the highest mean b -value (0.216, 95% CI: [0.198, 0.234]) and smoothnoise, on the other hand, has the lowest mean b -value (0.174, 95% CI: [0.170, 0.178]).

Table 4.3.2: Observed and mean bootstrapped b -values with 95% confidence intervals for each synthetic target type (source: 32 IXI targets). Statsmooth shows the highest values among the different targets.

Target	Observed b	Mean b	95% CI Lower	95% CI Upper
Scramble	0.170	0.181	0.174	0.187
Smooth	0.172	0.180	0.170	0.188
Smoothnoise	0.161	0.174	0.170	0.178
Stat	0.181	0.189	0.183	0.197
Statsmooth	0.224	0.216	0.198	0.234

4.4 Example Application of Ground Truth Benchmark

A comparison between the master and no-onepad MAPER versions was conducted at fusion number $fn = 32$, using the synthetic target configuration based on statsmooth. Figure 4.4.1 shows the distribution of mean Jaccard indices \overline{JC}_i per target (19 targets) for each version. Each line connects paired values from the same target, visualizing within-individual differences.

A paired t -test was performed to evaluate whether the two versions differed in segmentation accuracy. The result indicated a statistically significant difference, with $t = -4.586$ and $p = 0.0002$ at a significance level of $\alpha = 0.05$.

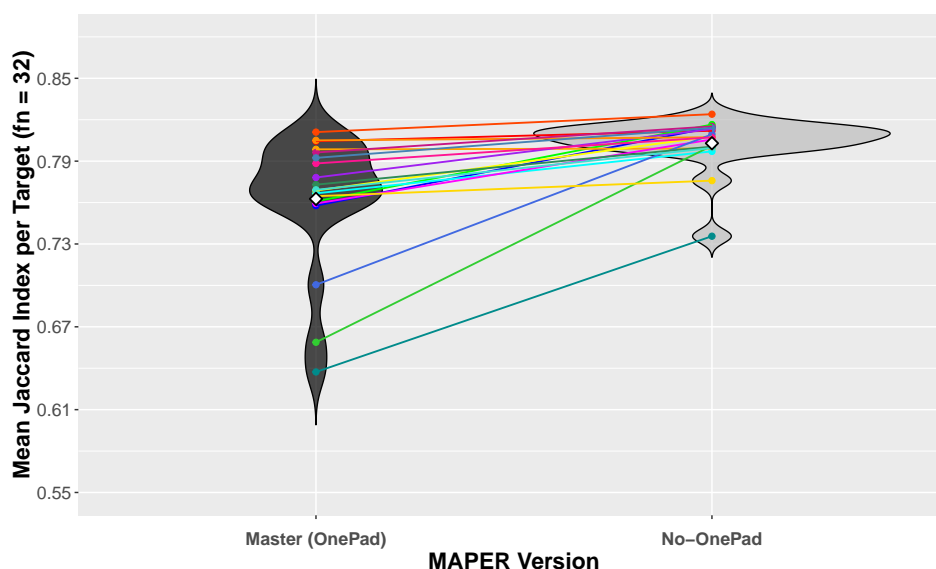


Figure 4.4.1: Mean Jaccard index per target image at $fn = 32$ for the master and no-onepad MAPER versions. Each point represents the mean Jaccard index for a single statsmooth target image, while the coloured lines connect paired values corresponding to the same individual across the different MAPER versions.

5

Discussion

5.1 Synthetic Images as Atlases

The first experiment evaluates whether synthetic MR images—specifically `scramble` and statistical variants—could serve as viable atlas inputs for automated segmentation using MAPER. While only these two types were tested, the results allow broader reflections on how different degrees of structural disruption in synthetic images may influence segmentation performance. This question is of practical relevance, as it explores whether synthetic MR-images can serve effectively as atlases in automated segmentation workflows. While the clinical need for generating new synthetic MR atlases on a regular basis may be limited, especially given the routine use of pre-prepared atlases in surgical planning and acute neurological assessments, the development of synthetic MR images as atlas inputs remains a promising avenue primarily for research applications. Synthetic atlases could provide valuable alternatives in research contexts where access to large-scale, annotated MR datasets is constrained by privacy, institutional availability, or demographic diversity. The utility of such synthetic atlases depends on their ability to achieve segmentation performance comparable to `conventional` atlases.

The results presented in Figure 4.1.1 show that `conventional` atlases consistently achieve the highest segmentation performance, while both synthetic variants perform significantly worse. `Scramble` atlases yield the lowest overall segmentation scores among the three atlas types, likely due to the substantial degradation of spatial coherence within each anatomical region. Although the voxelwise intensity histograms are preserved, the random permutation of spatial locations eliminates the gradients and textures required for effective non-rigid registration. Since MAPER relies heavily on spatial correspondence between atlas and target images, this internal disorganization hampers accurate alignment and label propagation. Nevertheless, the segmentations produced with `scramble` atlases remain structurally plausible in many cases, indicating that even heavily altered synthetic images retain a degree of anatomical utility. While their performance is statistically inferior to that of `conventional` atlases, `scramble` images still provide valuable insight into algorithmic robustness under spatially degraded conditions.

`Stat` atlases perform better than `scramble` variants, despite both types involving a random rearrangement of voxel spatial location within each anatomical region. However, `stat` atlases potentially gain an advantage over `scramble` through region-specific intensity modeling based on empirically fitted distributions. A possible explanation is that this likely provided enhanced contrast between anatomical regions and more informative cues for differentiating adjacent structures during segmentation. This is likely due to the amplification or attenuation of voxel intensities, depending on the fit to the selected statistical distribution, which in turn could contribute to improved contrast between regions.

Additionally, the fitted distributions provide a more structured intensity profile, whereas the original intensity distributions tend to be broader and less organized. Another possible explanation is that, since the `stat` images were generated by fitting distributions to each of the 120 anatomical regions with associated tissue classification, regions with fewer than 20 voxels were left unmodified because of insufficient data for reliable distribution fitting. This selective modification may have also contributed to the segmentation higher Jaccard index observed in Figure 4.1.1.

Quantitatively, Figure 4.1.1 showed that `conventional` atlases achieved a higher median Jaccard index (0.637) compared to `stat` (0.634) and `scramble` (0.633) variants. Although the absolute numerical differences appeared small, statistical tests confirmed their significance, reinforcing that segmentation pipelines are sensitive to subtle losses in image realism. This is particularly relevant in clinical contexts where segmentation quality can influence diagnostic or treatment decisions.

These findings also have implications for future development of synthetic benchmarks. While the `stat` images are not adequate substitutes for real data in precision-demanding segmentation pipelines, they may still be useful for controlled benchmarking, particularly when the aim is to isolate algorithmic sensitivity to spatial context or to simulate conditions of degraded image quality. Moreover, the framework used to create these synthetic images could be extended by incorporating more sophisticated spatial priors or by adding noise and partial volume effects to mimic clinical acquisition conditions more closely.

5.2 Leave-One-Out Cross-Validation on Synthetic Images

The second experiment uses leave-one-out cross-validation to evaluate the segmentation performance of `conventional`, `scramble`, and `stat` target images. Each of the 32 images in the IXI dataset was segmented using the remaining images of the same type as atlases, enabling an internally controlled assessment of how well each image type performs as a segmentation target under consistent conditions. Moreover, there are several reasons why LOOCV is particularly well-suited to this analysis. First, it isolates the performance of the segmentation algorithm from bias introduced by test data selection. Secondly, applying LOOCV across all three image types ensures that performance differences can be attributed to the image properties themselves, rather than to inconsistencies in the experimental setup. Lastly, LOOCV can help to assess whether synthetic images contain the anatomical and intensity characteristics required to function as meaningful test inputs for segmentation algorithms.

The results, presented in Figure 4.2.1 and Table 4.2.1, show that `conventional` targets achieve the highest segmentation scores, followed by `stat` and then `scramble` images. All differences between the target types are statistically significant. The fact that `stat` targets outperform `scramble` ones, despite both being synthetically generated and structurally altered, suggests that the underlying modeling strategy used to create the `stat` images may preserve properties that are beneficial to the automated segmentation process. The superior performance of `conventional` targets is expected, as these images maintain both voxelwise spatial structure and natural intensity distributions, allowing for accurate registration and majority voting during label fusion.

Additionally, the LOOCV results further offer valuable insight into MAPER’s behavior under varying image conditions. That MAPER performs best when using `conventional`

targets highlights its reliance on realistic anatomical structure and native intensity distributions for optimal registration and label propagation. However, the fact that `stat` targets result in significantly better performance than `scramble` ones demonstrates that MAPER can still extract and utilize higher-level regional intensity cues, even when voxel-level spatial coherence is disrupted. Although a significant difference is observed between the synthetic image types and the results obtained with `conventional` targets, MAPER still demonstrates robust performance, indicating its ability to operate effectively even when confronted with structurally or statistically altered input images.

5.3 Benchmarking the Ground Truth Benchmark

The third experiment evaluates how segmentation performance varies as a function of the number of fused atlases, ranging from three to 32 IXI-atlases, using five different types of synthetic images with 19 targets per synthetic type. The results were interpreted through the convergence model in Eq. (3.3.1), where the parameter b describes the rate at which segmentation performance approaches its asymptote; as the number of fused atlases increases. This means a higher b -value indicates faster convergence rate.

The empirical results demonstrate clear and consistent differences in convergence behaviour across the five synthetic image types. This is further supported by the statistical analysis, which confirms that all pairwise differences in b -values are significant (Table 4.3.1). Among the tested types, `statsmooth` stands out by exhibiting both the highest median and broadest distribution of b -values, as visualized in Figure 4.3.2. This indicates faster convergence overall, but also greater variability between target instances. More specifically, the 19 `statsmooth` target set exhibits higher fluctuation and inconsistencies in the mean Jaccard Index, which consequently leads to a spread in b -values in the bootstrapping setup. In contrast, `smoothnoise` yields the lowest and most compact distribution of b -values, suggesting a slower yet more consistent convergence pattern.

These findings are further supported by the bootstrapped distributions and 95% confidence intervals reported in Table 4.3.2. `Statsmooth` reaches the highest mean b -value (0.216), while `smoothnoise` reaches the lowest (0.174), with the other three types—`scramble`, `smooth`, and `stat`—occupying intermediate positions. The ordering and statistical separability of these values underline the sensitivity of segmentation convergence to structural and statistical properties of the synthetic target images.

Even though the mean Jaccard index as a function of `fn` is overall lower for `statsmooth` in Figure 4.3.1, the main objective in this context is to identify a b -value that reflects rapid convergence, as this enhances the sensitivity of the benchmark when used to distinguish between different segmentation models. From the benchmarking of the ground truth benchmark, `statsmooth` emerges as a particularly suitable choice for evaluating differences between automated segmentation algorithms, primarily due to its higher median convergence rate b . Currently, this remains the main practical application of the proposed framework. However, the methodology holds promise for broader applications beyond its current use. As such, there is potential to further expand the role of synthetic ground truth benchmarks, making them a valuable resource not only for comparing segmentation algorithms but also for addressing other methodological questions in the field. Due to time constraints and the computational cost associated with conducting large-scale experiments, other practical applications could not be explored within the scope of this work. Future studies are therefore encouraged to investigate additional use cases for ground truth

benchmarking, as well as to identify or design new types of synthetic images that may yield even higher b -values and thereby enhance sensitivity when comparing segmentation models.

5.4 Example Application of Ground Truth Benchmark

The comparative evaluation between the `master` and `no-onepad` versions of the MAPER algorithm provided a compelling demonstration of the utility of the synthetic ground truth benchmark. Using `statsmooth` targets (based on the results in Section 4.3) at the fusion level `fn = 32`, results indicated a statistically significant difference in segmentation accuracy (Figure 4.4.1), with the `no-onepad` consistently outperforming the `master` version, which includes a standard outer padding step. A paired t -test confirmed the reliability of this observation, yielding a test statistic of $t = -4.586$ and a corresponding p -value of 0.0002, well below the conventional significance threshold of $\alpha = 0.05$.

Based on the empirical findings from this benchmark-driven comparison, future development of the MAPER framework should proceed using the `no-onepad` configuration as the default. The better segmentation performance observed under this setup highlights its suitability as a baseline for further algorithmic refinement and evaluation.

6

Conclusion

This study investigates the use of synthetically generated MR images with known ground truth for assessing the performance and robustness of the automated brain segmentation algorithm MAPER. Several types of synthetic images were generated, each differing in spatial or intensity properties, and were tested in multiple experimental settings.

The results show that synthetic images, namely `scramble` and `stat`, both perform significantly worse as atlases than `conventional` unmodified MR-images when used for automated segmentation with MAPER. This indicates that `conventional` MR images remain the superior choice for segmentation tasks.

In the leave-one-out cross-validation, similar results were observed: MAPER performed better on `conventional` MR images than on `scramble` and `stat` ones, which is to be expected. However, the fact that synthetic images performed worse does not render them useless. On the contrary, the results provide valuable insight into the robustness and performance of MAPER, as well as its strengths and limitations.

In the experiment involving benchmarking of the ground truth benchmark, the synthetic image type `statsmooth` performed significantly better and yielded the highest convergence rate b among all synthetic types. This made `statsmooth` the most suitable benchmark for comparing the two MAPER versions.

In the application of the ground truth benchmark, the `statsmooth` benchmark revealed that the `no-onepad` version of MAPER is a better alternative than `master`, as the mean Jaccard index was consistently higher across all targets. This suggests that the `master` configuration should be replaced with the better-performing `no-onepad` version.

References

- [1] Atta, H. M. (1999). Edwin smith surgical papyrus: The oldest known surgical treatise. *American Surgeon*, 65(12), 1190–1192. <https://doi.org/10.1177/000313489906501222>.
- [2] Pozeg, Z. I., & Flamm, E. S. (2009). Vesalius and the 1543 epitome of his "de humani corporis fabrica librorum": A uniquely illuminated copy. *Papers of the Bibliographical Society of America*, 103(2), 199–220.
- [3] Judas, M., Ceganec, M., & Sedmak, G. (2012). Brodmann's map of the human cerebral cortex—or brodmann's maps? *Translational Neuroscience*, 3(1), 67–74. <https://doi.org/10.2478/s13380-012-0009-x>.
- [4] Talairach coordinates—an overview | sciencedirect topics [[Online; accessed 26 Jan. 2025]]. (2025).
- [5] Talairach, J., & Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system, an approach to cerebral imaging [[Online; accessed 26 Jan. 2025]].
- [6] Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. (1994a). Automatic 3d inter-subject registration of mr volumetric data in standardized talairach space. *Journal of Computer Assisted Tomography*, 18(2), 192–205. <https://pubmed.ncbi.nlm.nih.gov/8126267>
- [7] Mazziotta, J. C., Toga, A. W., Evans, A., & et al. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development: The international consortium for brain mapping (icbm). *NeuroImage*, 2(2), 89–101. <https://doi.org/10.1006/nimg.1995.1012>.
- [8] Iglesias, J. E., & Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1), 205–219. <https://doi.org/10.1016/j.media.2015.06.012>.
- [9] Gass, T., Székely, G., & Goksel, O. (2013). Semi-supervised segmentation using multiple segmentation hypotheses from a single atlas. In *Medical computer vision: Recognition techniques and applications in medical imaging* (pp. 29–37). https://doi.org/10.1007/978-3-642-36620-8_4.
- [10] Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., & Hammers, A. (2006). Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1), 115–126. <https://doi.org/10.1016/j.neuroimage.2006.05.061>.
- [11] Sanroma, G., Wu, G., Kim, M. K., & Shen, D. (2016). Multiple-atlas segmentation in medical imaging. In *Medical image recognition, segmentation and parsing* (pp. 231–257). <https://doi.org/10.1016/B978-0-12-802581-9.00011-1>.
- [12] Heckemann, R. A., Keihaninejad, S., Aljabar, P., Rueckert, D., & Hammers, A. (2010). Improving intersubject image registration using tissue-class information

- benefits robustness and accuracy of multi-atlas based anatomical segmentation. *NeuroImage*, 51(1), 221–227. <https://doi.org/10.1016/j.neuroimage.2010.01.072>.
- [13] Yaakub, S. N., Heckemann, R. A., Keller, S. S., McGinnity, C. J., Weber, B., Elger, C. E., & Hammers, A. (2020). On brain atlas choice and automatic segmentation methods: A comparison of maper and freesurfer using three atlas databases. *Scientific Reports*, 10(1), 2837. <https://doi.org/10.1038/s41598-020-57951-6>.
- [14] SciInstitute. (2012). Label fusion strategies for multi-atlas segmentation and group-wise correspondence in medical imaging [[Online; accessed 13 Apr. 2025]]. <https://www.youtube.com/watch?v=XB1XKj5QdDc>
- [15] Brainweb: Simulated brain database [[Online; accessed 23 Apr. 2025]]. (2006). <https://brainweb.bic.mni.mcgill.ca/brainweb>
- [16] Ixi dataset—brain development [[Online; accessed 25 Apr. 2025]]. (2025). <https://brain-development.org/ixi-dataset>
- [17] Brain atlases—brain development [[Online; accessed 25 Apr. 2025]]. (2025). <https://brain-development.org/brain-atlases>
- [18] Neuroimaging in python—pipelines and interfaces (nipy package) [[Online; accessed 18 May 2025]]. (2021). <https://nipy.readthedocs.io/en/1.1.0/index.html>
- [19] Annis, D. H. (2005). Permutation, parametric, and bootstrap tests of hypotheses. *Journal of the American Statistical Association*, 100(472), 1457–1458. <https://doi.org/10.1198/jasa.2005.s48>.
- [20] Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. (1994b). Automatic 3d inter-subject registration of mr volumetric data in standardized talairach space. *Journal of Computer Assisted Tomography*, 18(2), 192–205. <https://pubmed.ncbi.nlm.nih.gov/8126267>
- [21] Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., & Cuadra, M. B. (2011). A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*, 104(3), e158–e177. <https://doi.org/10.1016/j.cmpb.2011.07.015>.
- [22] Rohlfing, T., & Maurer, C. R. (2004). Multi-classifier framework for atlas-based image segmentation. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, II-206–II-213. <https://doi.org/10.1109/CVPR.2004.1315040>.

**SAHLGRENSKA ACADEMY
UNIVERSITY OF GOTHENBURG**

Gothenburg, Sweden

www.gu.se



UNIVERSITY OF GOTHENBURG

Sahlgrenska Academy

Gothenburg, Sweden 2025