DEVELOPING INTEGRATED TOOLS FOR TERABYTE-SCALE IMAGE PRE-PROCESSING, REGISTRATION, AND VISUALIZATION

by
Vikram Chandrashekhar

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
September 2021

Abstract

Quantifying terabyte-scale multi-modal human and animal imaging data requires scalable analysis tools. We developed CloudReg, an open-source, automatic, terabyte-scale, cloud-based image analysis pipeline that pre-processes and registers cross-modal volumetric datasets with artifacts via spatially-varying polynomial intensity transform. CloudReg accurately registers the following datasets to their respective atlases: *in vivo* human and *ex vivo* macaque brain magnetic resonance imaging, *ex vivo* mouse brain micro-computed tomography, and cleared murine brain light-sheet microscopy.

Thesis Committee Members

Joshua T Vogelstein, PhD (Primary advisor)

Jeremias Sulam, PhD

Daniel J Tward, PhD

Acknowledgements

I would like to thank my advisor Joshua Vogelstein for providing tremendous amounts of guidance and insight since the day I joined his lab. Josh has catalyzed my growth as a scientist and a person. I am grateful and honored to have been able to work with him and my other advisors, Daniel Tward and Jeremias Sulam, without whom my work would be impossible.

I would also like to thank our collaborators at Stanford and the NIH for their willingness to provide data and insight, allowing us to develop novel methods. In particular, I would like to acknowledge Ailey Crow at Stanford and Jared Rosenblum at the NIH for their significant support.

Lastly, I would like to thank my family, particularly my parents and brother, who have been a consistent source of support over my many years in Baltimore. None of this would be possible without their unconditional support.

Contents

List of Tables

List of Figures

1 Introduction

Modern imaging methods can generate intact, whole brain data from a variety of modalities including magnetic resonance imaging (MRI), computed tomography (CT), and light-sheet microscopy (LSM) of cleared tissue samples. Each of these methods provides specific information about an individual sample based on the physical principles of the technique, and also produces artifacts unique to each technique. MRI can provide detailed anatomic or functional information but can be limited by intensity inhomogeneity due to magnetic field bias.[1] CT can provide detailed anatomic information but can be limited by radiodensity artifacts.[2] LSM, in combination with tissue clearing methods, can provide anatomic, functional, and molecular information at subcellular resolution,[3] but can be limited by intensity inhomogeneity due to microscope optics.

Clearing methods including CLARITY (Clear Lipid-exchanged Anatomically Rigid Imaging/immunostaining-compatible Tissue Hydrogel),[3] SHIELD (Stabilization to Harsh conditions via Intramolecular Epoxide Linkages to prevent Degradation),[4] and iDISCO (immunolabeling-enabled three-Dimensional Imaging of Solvent-Cleared Organs)[5] can generate terabytes of data per sample.[6] High-resolution, multi-field-of-view (mFOV) datasets require pre-processing to remove artifacts, stitching into a complete volume, registration to a reference atlas, and visualization in order to perform quantitative analyses.[7,8]

Each of these image processing steps presents unique challenges. First, aligning and stitching every FOV acquired into a complete volume requires significant compute power and is time-intensive. Second, pre-processing imaging data requires correcting artifacts unique to each modality and sample such as intensity inhomogeneity in LSM and MRI. Third, registration methods are frequently intra-modal, have manual components, and are limited by artifacts

introduced by specimen preparation and imaging.[9,10] Finally, visualization of these terabyte-scale datasets on a local machine is compute-intensive, slow, and expensive.[11]

To address these challenges, we present CloudReg, an automatic, cross-modal, cloud-based pipeline consisting of local and global intensity correction,[12,13] alignment and stitching,[8] image registration with nonlinear methods,[14,15] and interactive online visualization through Neuroglancer (https://github.com/google/neuroglancer).[16] We specifically developed algorithms for distributed local intensity correction and cross-modal registration while leveraging existing state-of-the-art, open-source tools.[12,13] We applied CloudReg to various datasets including *in vivo* human brain MRI,[17] *ex vivo* macaque brain MRI,[18,19] *ex vivo* in situ mouse brain micro-CT,[20,21] and LSM-imaged cleared mouse and rat brains.[3,4,5]

2 Methods

Cleared brain specimen preparation and image acquisition

CloudReg was developed on multiple brain imaging modalities including a variety of clearing methods, all imaged with LSM. Whole mouse and rat brains were optically cleared using CLARITY, SHIELD, or iDISCO as previously described.[3,4,5] Autofluorescence image volumes were acquired using either a CLARITY-Optimized Light-sheet Microscope (COLM)[6] or a LaVision UltraMicroscope II (Miltenyi Biotec, Bergisch Gladbach, Germany). We used the autofluorescence channel to register, and then applied the resulting transformation to any additional channels, but any/all channels could be run through CloudReg. The autofluorescence channel of an LSM image is the background natural fluorescence present in the sample that is not associated with any artificially added fluorophores. The COLM imaged whole mouse and rat brains with voxel size 0.585x0.585x5.0 μm^3, 1.46x1.46x5.0 μm^3, or 2.9x2.9x5.0 μm^3 resulting in terabytes of data per brain for these higher resolution samples. The LaVision UltraMicroscope II was used to acquire a whole mouse brain in a single z-stack at 5.16 μm isotropic resolution.

Ex vivo in situ mouse brain micro-computed tomography

The intact mouse head was imaged via micro-CT following terminal vascular polymer perfusion as previously described.[20] Soft tissue including brain was visualized by immersing the sample in phosphotungstic acid (PTA) before micro-CT as previously described.[21]

In vivo human brain magnetic resonance imaging

Human brain MRI data was obtained from the MRICloud atlas set as previously described.[17]

Upload raw data to cloud storage

To run CloudReg, raw data is made web-accessible, for example, by uploading to cloud storage. We use Amazon Web Services (AWS) Simple Storage Service (S3) for our cloud storage services. The mFOV raw data was stored as a 2D TIFF (Tagged Image File Format) series for each column in the image volume. mFOV data was organized in the COLM acquisition format,[6] but any format can be used with minor modifications. The single-FOV (sFOV) raw data was stored as a 2D TIFF series where each slice in the image was saved in a separate TIFF. The raw data was uploaded to S3 using the awscli Python package (https://github.com/aws/aws-cli).

Run CloudReg

CloudReg consists of two user-facing Python scripts that we developed along with other utility functions (available here: https://cloudreg.neurodata.io). All scripts referred to herein are available in this repository. Our two user-facing scripts automatically start and stop a cloud server after running a series of computations on that server. The first script requires a cloud server with attached solid-state drives (SSD) and the second script requires a server with sufficient memory for the given data. The first script requires SSDs since the computation steps are input/output (I/O) bound, whereas the second script requires more memory since it needs to apply transformations to the high resolution data. An AWS cloud server is called an Elastic Compute Cloud (EC2) instance. The first script (run_colm_pipeline_ec2.py) automatically starts an r5d-type EC2 instance (an instance that contains attached SSD storage) and performs the following steps (Figure 2.1 A-B): 1) transfers raw data from cloud storage, 2) corrects local intensity, 3) stitches the data into a three-dimensional volume, 4) corrects global intensity, and 5) downsamples and uploads stitched and preprocessed data to cloud storage for visualization. The second script

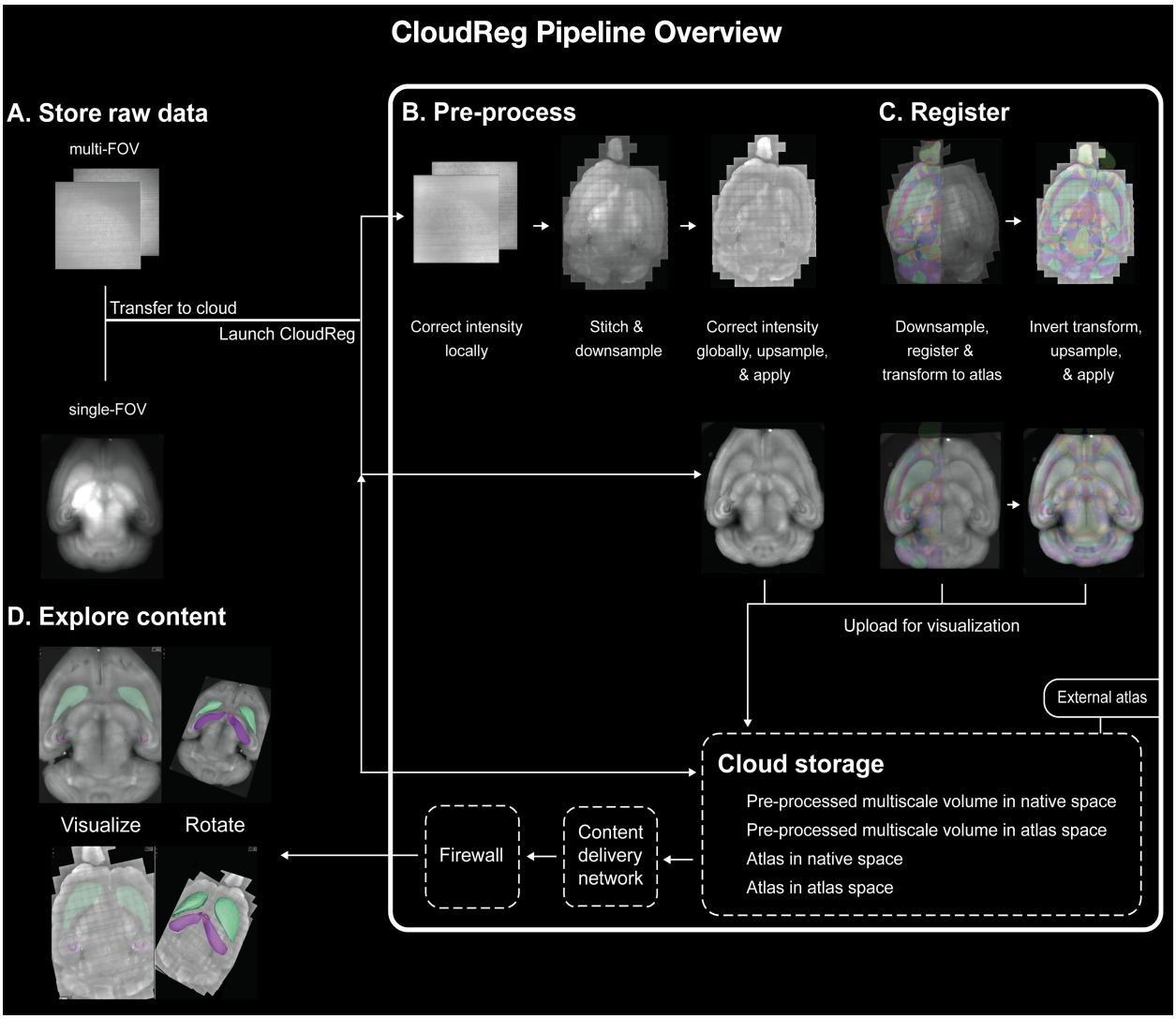

Figure 2.1: CloudReg pipeline schematic, example outputs at each step. A: Store raw data. Raw data can be either multi-field-of-view (mFOV) or single-field-of-view (sFOV). After raw data is generated, it is transferred to the cloud and CloudReg is launched. B: Pre-process. All of the following steps are performed on a cloud compute server that has access to the raw data. First, local intensity inhomogeneity is corrected per-FOV using an algorithm we developed. Next, mFOV data is aligned and stitched into a complete volume using Terastitcher; if the data is sFOV, this step is not performed. Then, global intensity inhomogeneity is corrected on the stitched whole brain by computing an intensity correction on the downsampled data, upsampling the result, and applying it at native resolution; this step applies to both mFOV and sFOV data. The pre-processed data is then downsampled and uploaded to cloud storage for visualization. C: Register. A downsampled version of the pre-processed data is registered to the ARA CCFv3 and the computed transformations are saved. These transformations are invertible so the atlas can be transformed to the data space and the data can be transformed to the atlas space. The transformations are upsampled and applied to the ARA anatomic parcellations, and input data and are uploaded to cloud storage for visualization through Neuroglancer. To maintain privacy of imaging data, visualization is restricted to authorized users by using a content delivery network and firewall. D: Explore content. Left column shows ARA parcellations transformed to the s- and mFOV data based on the computed transformations from the registration. Right column shows 2D axial slice from the pre-processed input data and 3D rendering of Caudoputamen and Hippocampal regions selected from the transformed ARA visualized in Neuroglancer. ARA CCFv3, Allen Reference Atlas Common Coordinate Framework version 3.

(run_registration_ec2.py) automatically starts an r5-type EC2 instance (an instance with high ratio

of memory to processing cores) and performs the following steps (Figure 2.1 C-D): 1) downloads downsampled, preprocessed data and the reference atlas data from cloud storage, 2) registers input data to the provided atlas, 3) transforms atlas anatomic parcellations to input data space and input data to atlas space, and 4) uploads transformed data to cloud storage for visualization. For example, for mouse brain data we use the Allen Reference Atlas (ARA) Common Coordinate Framework Version 3 (CCFv3), but any reference atlas or sample can be used.

Transfer raw data from cloud storage

After the cloud server is started, available SSDs must be formatted and mounted onto the server for use. We developed a bash script (mount_combined_ssds.sh) to do this automatically for any EC2 instance with SSD storage available. The raw data is then downloaded from cloud storage onto these SSDs. Data download is parallelized across all available cores to speed up the process via a Python script (download_raw_data.py).

Correct local intensity

Our local intensity correction can be applied to any mFOV samples with intensity inhomogeneity, though we developed our local intensity correction on LSM-imaged CLARITY data. For mFOV data, there are two intensity correction steps: 1) local correction per FOV and 2) global correction on the stitched image volume. A single intensity correction step is performed on on the whole image volume of sFOV data. To correct the per FOV intensity inhomogeneity, we developed an algorithm to estimate a multiplicative intensity correction directly from the data. Our algorithm begins with sub-sampling the mFOV raw data such that FOVs are uniformly sampled in voxel space in all three dimensions. Uniform sampling in voxel space uses the number of voxels in each dimension rather than their size in physical units. The amount of subsampling is a

configurable parameter; increased subsampling can speed up the rest of the computation but may provide a less accurate estimate of the intensity correction. We use a subsampling factor of two. Next, the mean is computed across these subsampled FOVs such that the resulting mean is a 2D image. The multiplicative intensity correction is then estimated by applying the N4 (Nick's Nonuniform Nonparametric intensity Normalization) bias correction algorithm to the computed mean image.[12] This algorithm is available as a Python script (correct_raw_data.py). Example results from this local intensity correction are shown in Figure 2.2.

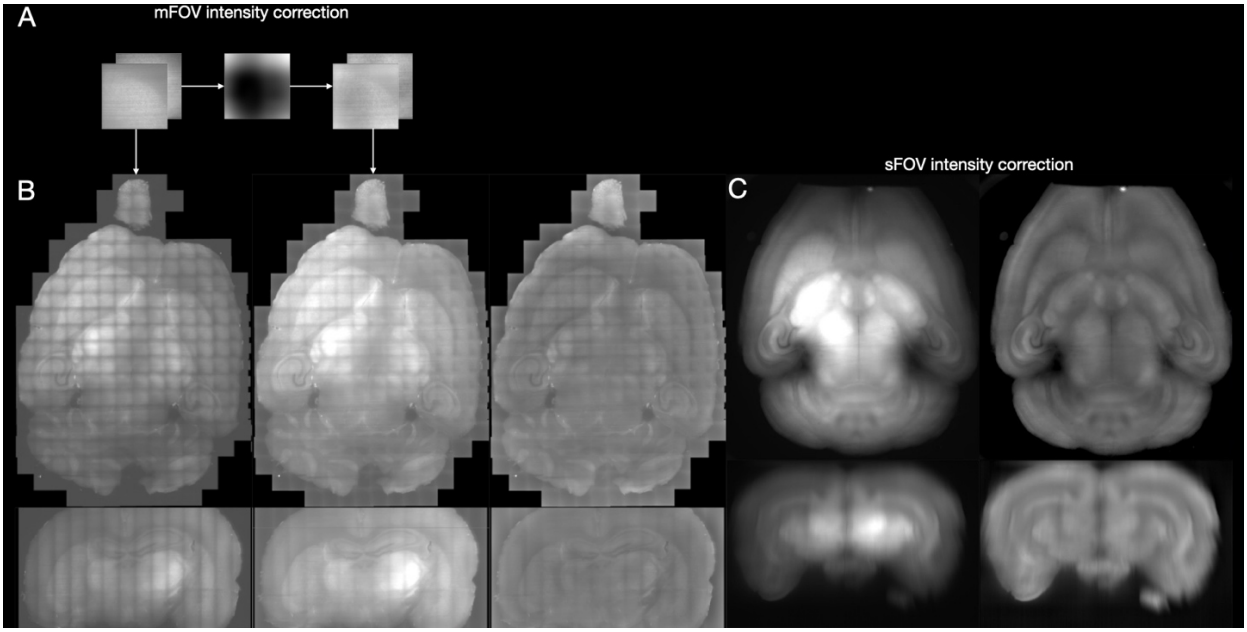

Figure 2.2: ***Intensity correction on mFOV and sFOV data.*** *A: Intensity correction per FOV. The left images show raw data from a mFOV dataset. The middle image shows the computed local multiplicative intensity correction that is applied across every FOV from the raw mFOV data. The right images show the FOVs after applying intensity correction. B: Effect of intensity correction on stitched mFOV data. The left column shows a single axial (top) and coronal (bottom) slice from the complete stitched volume from the raw data in A. The middle column shows the same axial (top) and coronal (bottom) slices after per-FOV correction. The right column shows the same slices after global intensity correction. C: Effect of intensity correction on sFOV data. Axial (top) and coronal (bottom) slices of the raw sFOV data before (left) and after (right) global intensity correction. FOV, field-of-view; mFOV, multi-field-of-view; sFOV, single-field-of-view.*

Stitch mFOV data into a volume

Stitching mFOV data into a volume was performed using Terastitcher,[8] an open-source software for stitching terascale images. Terastitcher uses a maximum intensity projection, normalized cross correlation (MIP-NCC) method to align each of the individual 2D FOVs in the

3D volume. Generating accurate stitching results requires each 2D FOV to have some overlap with neighboring FOVs. The stitched image volume is saved as a 2D TIFF series on the SSDs of the r5d-type EC2 instance. To accelerate this process, we adapted Python scripts from Terastitcher to parallelize stitching across all available cores subject to memory constraints. To use these scripts with our pipeline, we had to first convert them to Python 3 (paraconverter.py and parastitcher.py). Functionality in these scripts is wrapped and available in a Python script (stitching.py).

Upload

Stitched data is uploaded to cloud storage, for example S3, in a highly parallelized fashion using cloud-volume (https://github.com/seung-lab/cloud-volume), an open-source Python package that can write Neuroglancer-compatible data to cloud storage. We have made contributions to cloud-volume to provide additional compression methods (https://github.com/seung-lab/cloud-volume/pull/291). The stitched data is concurrently downsampled using a package called tinybrain (https://github.com/seung-lab/tinybrain) for quicker visualization and uploaded to cloud storage. We implemented this parallelized upload procedure to leverage as many cores as are available subject to memory constraints (create_precomputed_volume.py).

Correct global intensity

To correct intensity inhomogeneity in the stitched image volume, we applied the N4 bias correction algorithm to a downsampled version of the whole volume in 3D.[12] The downsampled intensity correction produced is then upsampled and applied to the native resolution data, which is then uploaded back to cloud storage in Neuroglancer precomputed format on a slice-by-slice

basis (correct_stitched_data.py). Example results from this global intensity correction are shown in Figure 2.2.

Visualize with Neuroglancer

Once the data is uploaded to cloud storage, we use Neuroglancer to view that data in a web browser. Neuroglancer stores the state of the visualizer in a JSON format which is appended to the URL. When visualizing many samples at once, this URL can exceed the maximum allowed length of URLs. To enable sharing of this URL, we implemented a URL shortener using a custom JSON-state (Javascript Object Notation) server with our Neuroglancer deployment (https://viz.neurodata.io). To enable visualization of multi-channel LSM data, we added differential channel coloring that has been merged into the Neuroglancer core code base.[16] We also set up a Content Delivery Network through AWS called CloudFront on top of our S3 bucket to enable HTTP/2 to speed up visualization on Neuroglancer. We also added Internet Protocol (IP)-address-based restriction to ensure limited access to private data using AWS Web Application Firewall (WAF). Example visualization is shown in Figure 2.3

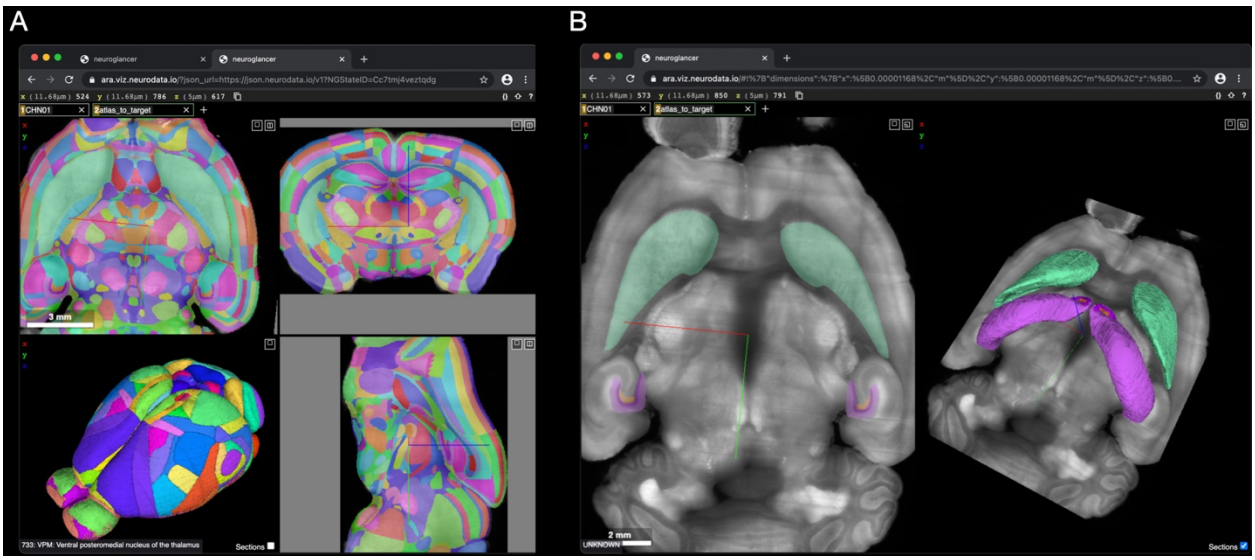

Figure 2.3 Interactive web-based visualization with Neuroglancer. A: CLARITY mouse brain with ARA parcellations overlaid. LSM-imaged CLARITY-cleared mouse brain with the ARA CCFv3 anatomic parcellations registered and

transformed to the input sample is shown. A 3D rendering of the resulting ARA brain nuclei segmentation based on our registration is shown in the bottom left quadrant. This data is being visualized in a web browser and is served from cloud storage. B: CLARITY mouse brain with selected ARA regions. The left side shows the sample from A with only Caudoputamen and Hippocampal regions selected from the ARA. The right side shows those regions rendered in 3D and overlaid on the raw data. CLARITY, Clear Lipid-exchanged Anatomically Rigid Imaging/immunostaining-compatible Tissue Hydrogel; ARA CCFv3, Allen Reference Atlas Common Coordinate Framework version 3; LSM, Light-Sheet Microscopy.

Register to a reference atlas

CloudReg computes affine and nonlinear transformations using a modified version of the Expectation-Maximization Large Deformation Diffeomorphic Metric Mapping (EM-LDDMM) registration algorithm.[15] Our modified EM-LDDMM enables cross-modal registration by estimating spatially-varying polynomial transformations of the atlas intensity to match our input data intensities. Per-voxel error signals in our intensity transform allow detection of artifacts and missing tissue. These concepts are combined within an EM framework where deformation parameters and polynomial coefficients are updated iteratively.

Below is a description of each variable necessary to specify an objective function to be optimized for image registration.

Name	Definition	To be optimized
x	A point in 3D space describing the location of a voxel.	
$J(x)$	The input image which is a real-number-valued function of x	Input; Fixed parameter
$I(x)$	The atlas image which is a real-valued function of x	Input; Fixed parameter
$v(x)$	A velocity field which is a 3D vector-valued function of x and time	Yes
$\varphi(x)$	A position field which is a 3D vector-valued function of x. It includes a component found from integrating $v(x)$ over time and an affine component.	Yes
$\hat{I}(x)$	$(1, I, I^2, I^3)^{\mathrm{T}} \circ \varphi^{-1}(x)$	Yes
$c(x)$	A 4D vector-valued function of x representing the coefficients of a 3rd order polynomial contrast transform at each voxel in $J(x)$	Yes
$w(x)$	A real-number-valued function of x taking values between 0 and 1 representing the posterior probability that a voxel in $J(x)$ corresponds to some voxel in I as opposed to missing tissue or artifact	Yes
σ_M	A positive real number representing the standard deviation of the noise in the image J and a weighting of the matching term in our objective function. To calculate w we assume background and artifact has twice and five times the standard deviation, respectively.	Fixed parameter; unitless; default is standard deviation of J
σ_R	A positive real number representing the weighting of the regularization of v in our objective function.	Fixed parameter; unitless; default is 10,000
σ_C	A positive real number representing the weighting of the regularization of c in our objective function	Fixed parameter; unitless; default is 5
a	A characteristic length scale for regularizing v	Fixed parameter; microns; default is 500
a'	A characteristic length scale for regularizing c	Fixed parameter; microns; default is 750
$L_a/L_{a'}$	A highpass differential operator for encouraging spatial smoothness in regularization equal to $(1 + a^2\Delta)^2$ where Δ is the Laplacian.	
id	Identity operator	

Table 2.1: Variables in our registration objective function defined along with default parameters and optimization conditions.

The objective function we minimize is

$$\underbrace{\frac{1}{2\sigma_M^2}\int |c^T(x)\hat{I}(x) - J(x)|^2 w(x)dx}_{\text{Matching accuracy}} + \underbrace{\frac{1}{2\sigma_R^2}\int_0^1 \int |L_a v_t(x)|^2 dxdt}_{\text{Deformation regularization}} + \underbrace{\frac{1}{2\sigma_C^2}\int |L_{a'}c(x)|^2 dx}_{\text{contrast regularization}}$$

Note that $c^T \hat{I}$ is the atlas deformed and intensity-transformed to the input data. The highpass operator used in regularization of v and c were initially described in Beg et al.[14] w is a set of weights estimated using the EM algorithm, designed to downweight voxels that contain missing tissue or large artifacts. For a given w, the cost is optimized over v, c, and affine parameters using gradient descent. This is the maximization step of the EM algorithm. For the expectation step, w is updated using Gaussian mixture modeling (GMM) which depends on the value σ_M. The procedure is described in more detail in Tward et al.[15] Our contribution in this work is twofold: (1) we introduce including c as a function of space rather than a constant, and (2) we add a contrast regularization term in the objective function.

CloudReg can be run on any two image volumes that have correspondence. We optimized aspects of the pipeline for mouse and rat, whole-brain, LSM-imaged CLARITY data. Specifically, registration is performed on a downsampled version of the input data, at 100 μm. The resulting transformations are upsampled and applied to the atlas to transform it to the input data coordinate space and the input data to the atlas space. The resulting transformations generated by EM-LDDMM are smooth and can be upsampled with linear interpolation,[15] enabling visualization of the registered atlas at the native resolution of our observed image. Transformations are stored on cloud storage with transformed images stored in Neuroglancer precomputed format for visualization. Our modified EM-LDDMM is available in CloudReg as a MATLAB script (map_multiscale_nonuniform_v02_mouse_gauss_newton.m) and exists in a fork of scikit-image, an open-source python image analysis toolkit (https://github.com/scikit-image/scikit-image/pull/4390). The python script remains four-fold slower than the MATLAB script largely due to the relative computational efficiency of 3D interpolation.

Spatially varying polynomial intensity transform

Since the objective function we optimize (see equation above) is quadratic in c, we can solve for c by solving a linear system of equations, given by

$$\frac{1}{\sigma_C^2} L_{a'} L_{a'} c + \frac{1}{\sigma_M^2} \hat{I}\hat{I}^T w c - \frac{1}{\sigma_M^2} \hat{I} w J = 0$$

In the above expression we use the fact that L is self-adjoint. Since the first two terms on the left correspond to positive semidefinite operations, we can rewrite the problem as:

$$\mathrm{B} = \frac{\hat{\mathrm{I}}^{\mathrm{T}}\sqrt{\mathrm{w}}}{\sigma_M} \qquad \mathrm{A} = \frac{L_{a'}}{\sigma_C} \qquad \mathrm{b} = \frac{\hat{I} w J}{\sigma_M^2}$$

$$AAc + B^T B c - b = 0$$

To improve conditioning, and using the fact that A is easily inverted in the Fourier domain because it is diagonalized by the Fourier transform, we use an elimination approach, letting $y = Ac$, and solving the system:

$$(id + A^{-1} B^T B A^{-1}) y - A^{-1} b = 0$$

by subtracting a constant multiplied by the residual (the left-hand side of the above equation) at each step (*i.e.,* gradient descent for the corresponding least squares problem). The code to solve this system and compute the coefficients c is available in a MATLAB script (estimate_coeffs_3d.m). An example registration demonstrating the spatially-varying intensity transform is shown in Figure 2.4.

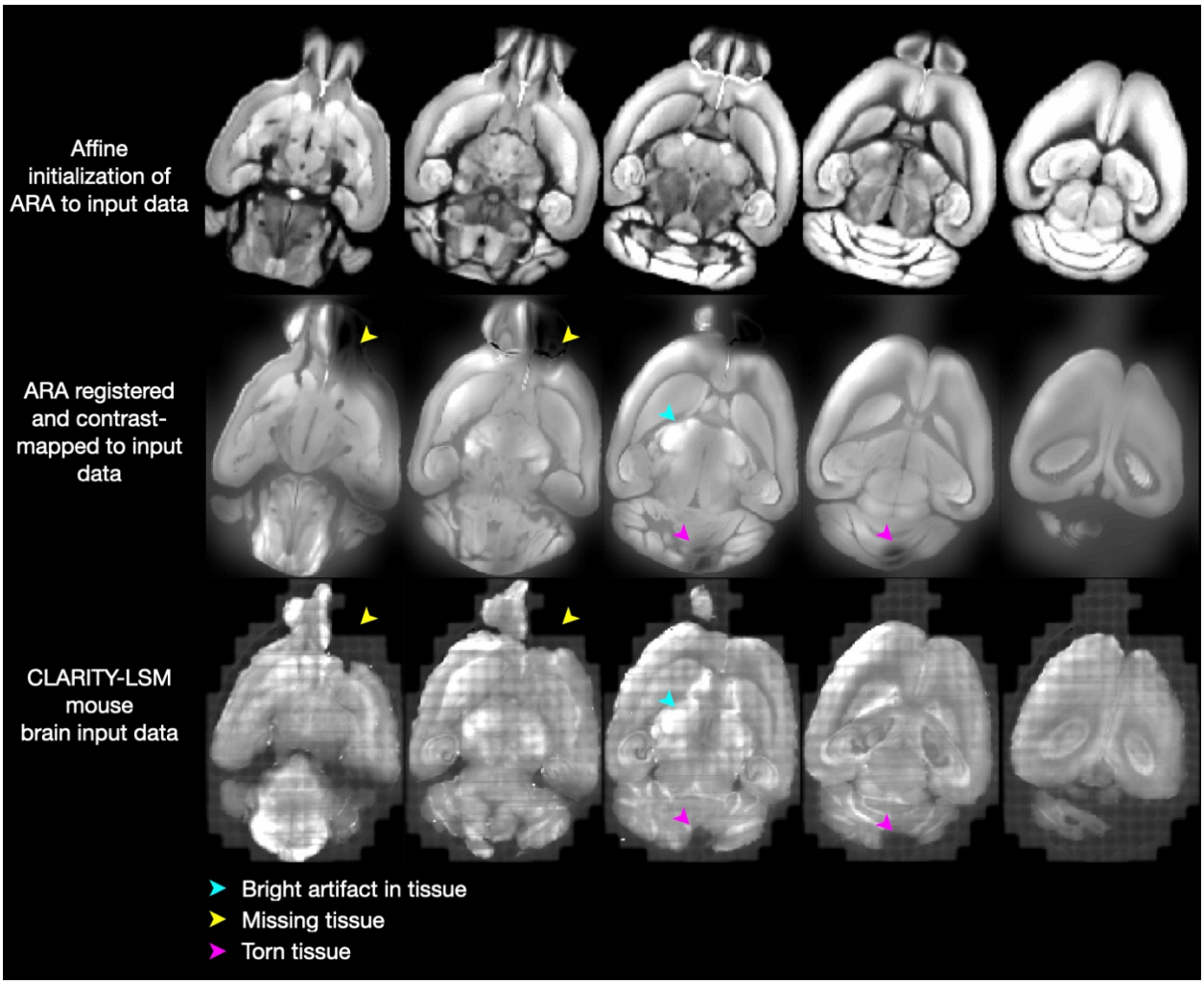

Figure 2.4: Registration and contrast-mapping tissue with artifacts. Top row shows an affine initialization aligning the ARA CCFv3 raw data to our input data generated from LSM-imaged CLARITY cleared whole mouse brain in the bottom row. The middle row shows the ARA transformed to match the input data with a spatially varying, cubic polynomial contrast transform applied. Arrowheads indicate regions of the input tissue that contained artifacts, missing data, or torn tissue that were mapped onto the atlas using our spatially varying, cubic contrast transform, despite these significant artifacts. The bottom row contains our input data with arrowheads corresponding to the middle row. ARA CCFv3, Allen Reference Atlas Common Coordinate Framework version 3. LSM, Light-Sheet Microscopy; CLARITY, Clear Lipid-exchanged Anatomically Rigid Imaging/immunostaining-compatible Tissue Hydrogel.

Determine registration accuracy

To determine registration accuracy, an expert placed 19 landmarks shown in Figure 2.5 on both the ARA and three input LSM-imaged cleared tissue samples. In areas of significant deformation or damaged tissue, landmarks were not placed. The transformations computed in the registration are applied to the landmarks placed on the LSM-imaged cleared tissue data to transform them to the landmarks placed on ARA data. The Euclidean distance between the LSM-

imaged cleared tissue transformed points and corresponding ARA points is computed and reported in microns in Table 2.1. We developed a Python script that computes registration accuracy given two Neuroglancer links with landmarks and computed transformations (registration_accuracy.py).

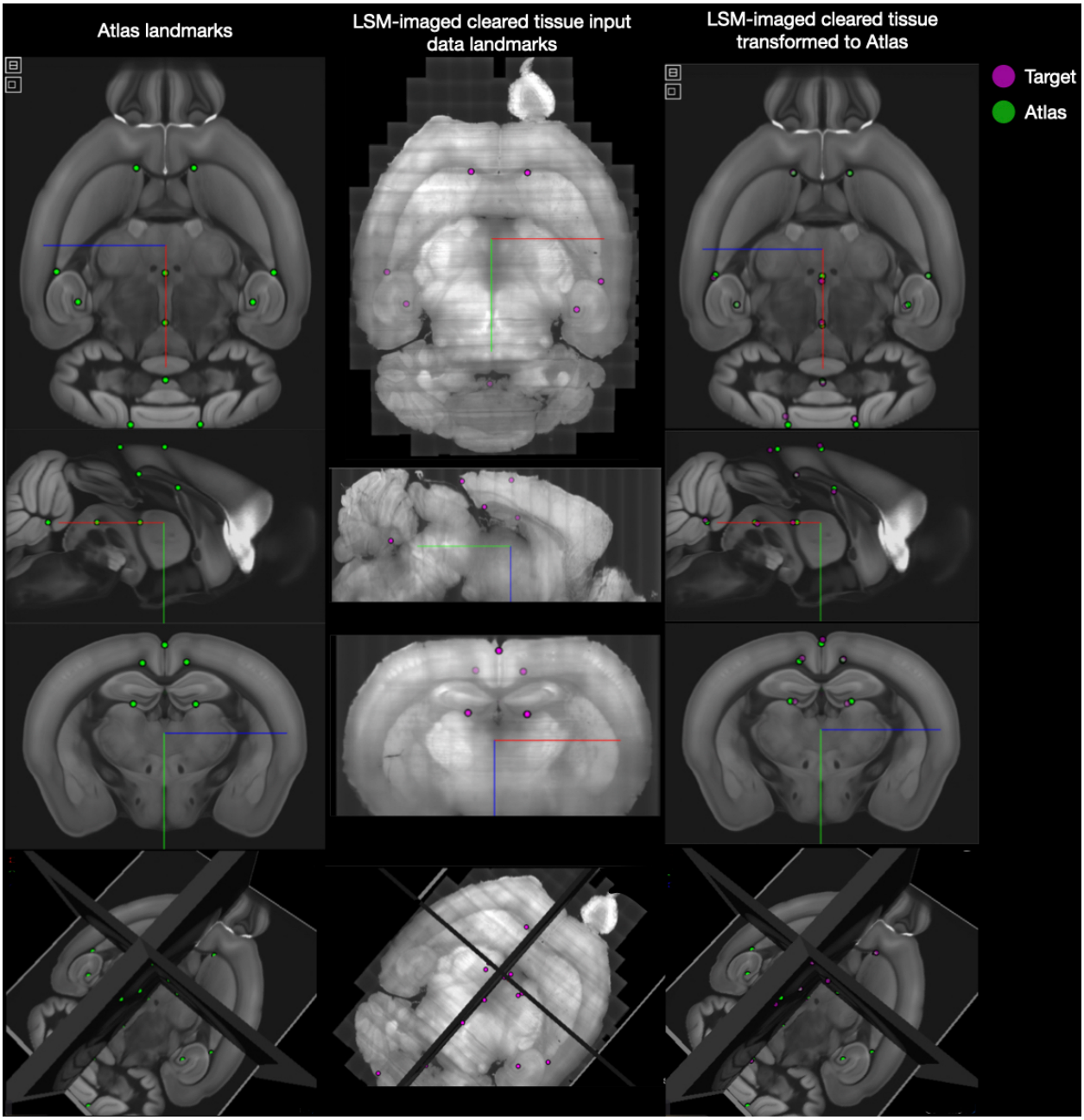

Figure 2.5: Landmarks placed for registration accuracy assessment. Left Column: ARA landmarks. Locations of landmarks for assessment of accuracy shown in axial (1st row), sagittal (2nd row), and coronal (3rd row) views; landmarks are also shown in 3D space (4th row). Middle Column: LSM-imaged CLARTY input data landmarks. Locations of landmarks for accuracy shown in same views. Right Column: Input data landmarks transformed to ARA.

Locations of CLARITY landmarks transformed to ARA shown in same views. CLARITY points are shown in magenta, ARA points are shown in green, and the overlap is grayscale. ARA, Allen Reference Atlas; CLARITY, Clear Lipid-exchanged Anatomically Rigid Imaging/immunostaining-compatible Tissue Hydrogel; LSM, Light-Sheet Microscopy.

Statistical analysis

We computed the mean and standard deviation of landmark error across all landmarks for each of our three LSM-imaged cleared tissue samples and computed means across landmarks by region including cortex, midbrain, and cerebellum.

Utility Functions

We have also implemented a Python script to downsample and upload a 3D image stack to the cloud for visualization with Neuroglancer (ingest_image_stack.py).

3 Results

Figure 2.1 shows an overview of the CloudReg pipeline. Data is uploaded to a web-accessible cloud storage provider; CloudReg is launched from a local machine (Figure 2.1A). The pipeline is run automatically in the cloud. CloudReg starts a cloud computing instance with sufficient RAM to perform pre-processing and registration, downloads raw data onto that server, corrects local intensity, stitches, corrects global intensity, and uploads pre-processed data to cloud storage for online visualization and analysis with Neuroglancer (Figure 2.1B). Next, registration is started by providing an affine initialization to roughly align an atlas to the input data and is automatically updated to include non-linear deformation via an expectation-maximization optimization process. Then, the atlas anatomic parcellations are automatically transformed to the input data at high resolution for visualization (Figure 2.1C). All the data is stored in the cloud, and then routed through a content delivery network and firewall to facilitate efficient and secure visualization via Neuroglancer (Figure 2.1D). The resulting data can be shared by sending a Universal Resource Locator (URL) and visualized from anywhere with a web browser and internet connection. Our deployment of Neuroglancer enables instant visualization of multi-channel, terabyte-scale datasets and provision of a shortened URL with one-click to share data views and analysis results (Figure 2.3).[16]

We initially developed CloudReg using high-resolution, LSM-imaged CLARITY mouse brain data[6] and used the Allen Reference Atlas (ARA) Common Coordinate Framework Version 3 (CCFv3) as the reference atlas (Figure 3.1, rows 1 and 2).[22]

Tissue clearing procedures and optics of LSM introduce sample-specific artifacts and intensity inhomogeneity in the imaged samples, which we aimed to correct with our mFOV-

based pre-processing algorithm. Intensity inhomogeneity manifests as a decay of intensity away from the center of each FOV and, for mFOV samples, the center of the stitched whole brain.

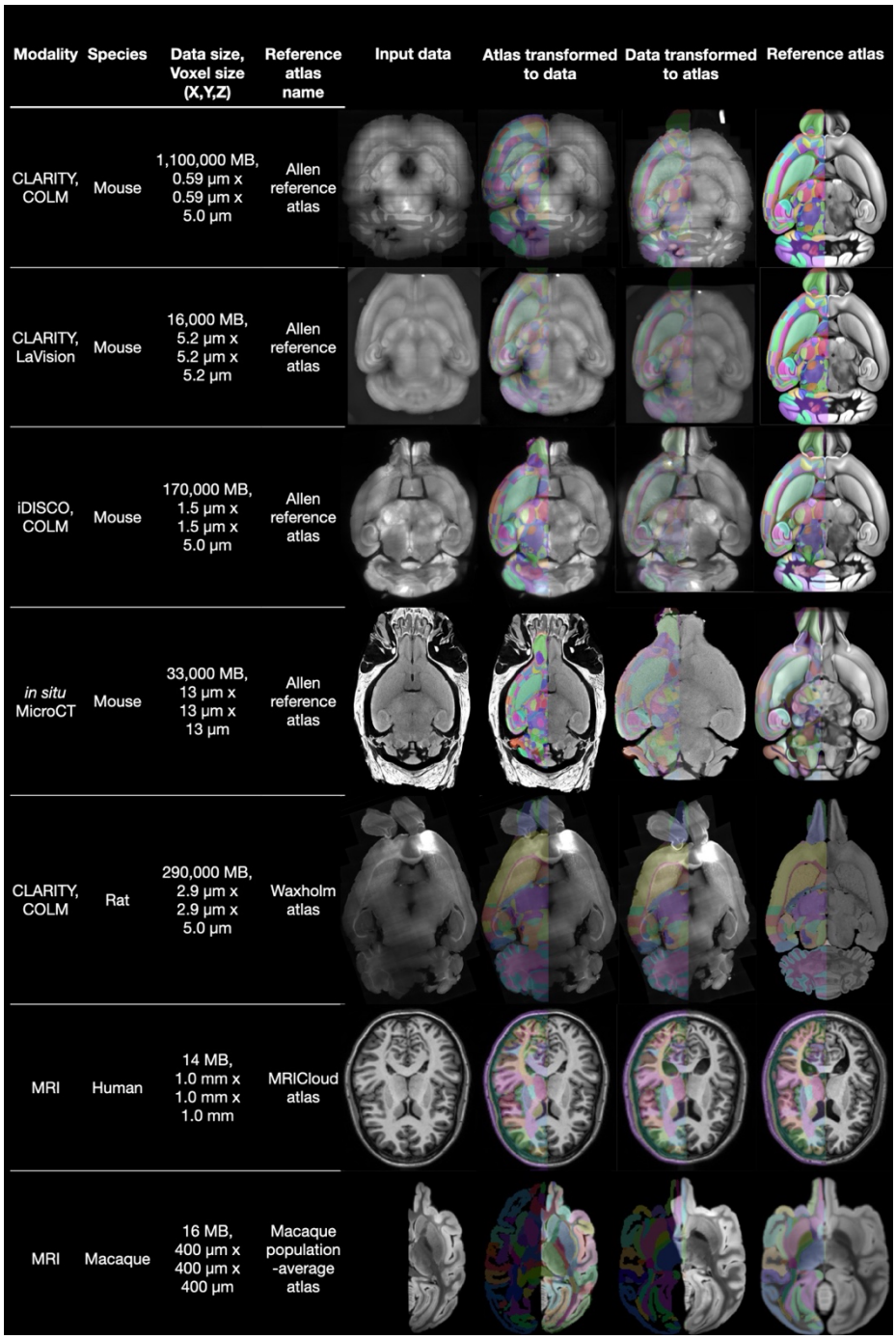

Figure 3.1: CloudReg pipeline registration outputs from multiple species imaged with various modalities. Each row demonstrates registration of either mouse, rat, macaque, or human brain imaging data to the corresponding atlas using CloudReg. The leftmost column of images shows the input data; the data from the autofluorescence channel is used for samples imaged with a light-sheet microscope (LSM). The rightmost column shows the atlas parcellations overlaid on one hemisphere of the atlas image data. The second and third columns show the respective atlas parcellations transformed to and overlaid on the original samples and vice-versa, respectively. CLARITY, Clear Lipid-exchanged Anatomically Rigid Imaging/immunostaining-compatible Tissue Hydrogel; COLM, CLARITY-Optimized Light-sheet Microscopy; GB, Gigabyte; iDISCO, immunolabeling-enabled three-dimensional imaging of solvent-cleared organs; MB, Megabyte; Micro-CT, Micro-Computed Tomography; TB, Terabyte.

This makes automatic, intensity-based registration a significant challenge. To minimize this per-

FOV artifact, we developed a parallelized intensity correction algorithm based on the hypothesis that the introduced intensity inhomogeneity is the same in each FOV. To efficiently compute this correction, we uniformly subsample the mFOV data in three dimensions, compute the mean across subsampled FOVs in parallel, and apply the N4 bias correction algorithm[12] to the resulting mean FOV (Figure 2.2A). Our intensity correction algorithm accounts for differences in tissue scattering from different clearing methods by estimating the intensity correction directly from the data. This pre-processed data is then automatically aligned and stitched using Terasticher, an open-source tool for stitching teravoxel microscopy datasets.[8] To minimize intensity inhomogeneity at the whole-brain scale, we apply the N4 bias correction algorithm to the whole stitched volume directly (Figure 2.2B-C).

The fully pre-processed sample, which can be acquired from a variety of modalities in a number of species, is then registered to a corresponding reference atlas. To enable registration of LSM-imaged tissue samples, we developed a spatially-varying polynomial intensity transform, expanding the scope of samples that can be automatically registered (Figure 3.1; Figure 3.2;). CloudReg computes affine and nonlinear transformations by building on the Expectation-Maximization Large Deformation Diffeomorphic Metric Mapping (EM-LDDMM) registration algorithm that we previously developed.[15] Our extension of EM-LDDMM built into CloudReg enables cross-modal registration of a diversity of brain volume samples with artifacts, tears, and deformations (Figure 2.4).

To assess registration accuracy, we used Target Registration Error (TRE) by computing the Euclidean distance between 19 landmarks (Figure 2.5) placed by experts on the ARA and our samples, where possible. TRE for samples 1, 2, and 3 was 2.27 ± 0.86 voxels, 2.42 ± 1.31 voxels, and 2.74 ± 0.88 voxels, respectively (Table 2.1). These voxels are 100 μm indicating the resolution at which the registration was performed. Measurements and summary statistics can be found in Table 2.1. Examining landmark error by brain region shows regional error differences,

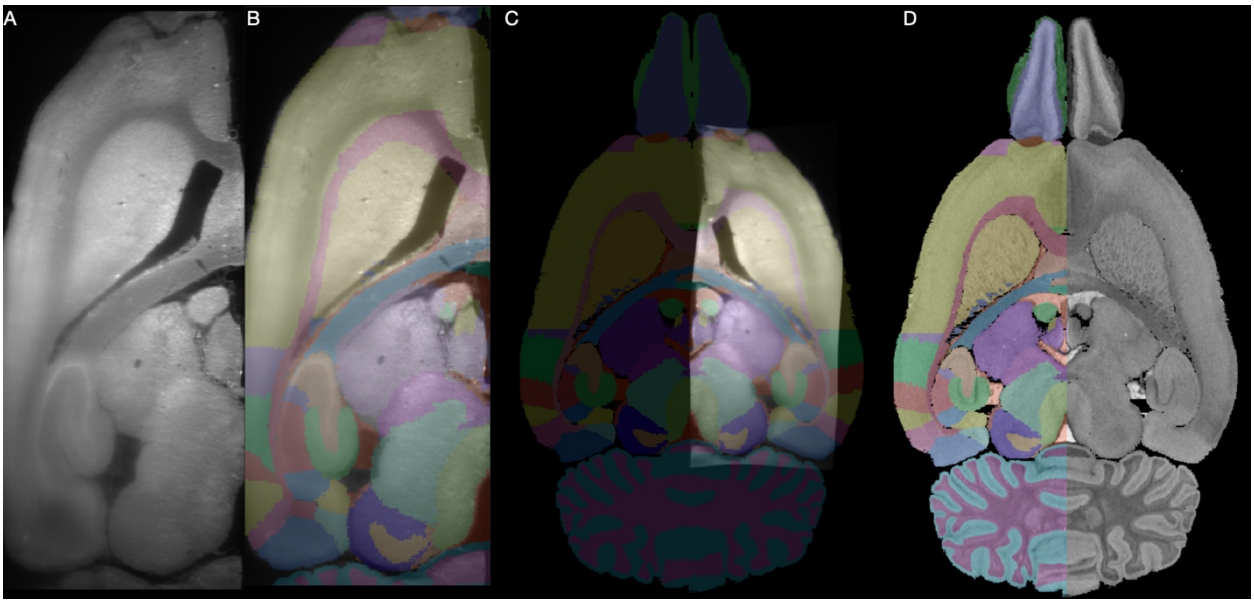

Figure 3.2: iDISCO rat hemisphere registered to Waxholm atlas.[23-25] A: Input data. iDISCO-cleared light-sheet microscopy-imaged rat brain region of interest. B: Waxholm atlas parcellations overlaid on input data. C: Input data transformed to Waxholm atlas. D: Waxholm atlas raw data with parcellations overlaid. iDISCO, immunolabeling-enabled three-dimensional imaging of solvent-cleared organs.

reflecting gross regional deformations. Mean landmark error for midbrain, cortical, and cerebellar regions across all samples are 1.47, 2.32, and 2.65 voxels, respectively.

Landmark names	Sample 1 (μm)	Sample 2 (μm)	Sample 3 (μm)	Category
1	n/a	160	184	cortex
2	172	142	458	cortex
3	201	262	336	cortex
4	236	279	280	cortex
5	85	252	293	cortex
6	183	320	312	cortex
7	315	200	252	cortex
8	358	244	146	cortex
9	210	290	312	cortex
10	207	153	304	cortex
11	257	282	170	cortex
12	267	321	237	cortex
13	178	37	131	cortex
14	192	135	165	cortex
15	106	176	71	midbrain
16	117	111	299	midbrain
17	223	106	n/a	cerebellum
18	n/a	216	n/a	cerebellum
19	367	384	n/a	cerebellum
MEAN	227	242	274	
STDEV	74	64	85	
MEAN (cortex)	220	220	256	232
STDEV	69	84	91	81
MEAN (cerebellum)	295	235	N/A	265
STDEV	102	140	N/A	121
MEAN (midbrain)	111	144	185	147
STDEV	8	46	161	71

Table 3.1: *Euclidean distance between corresponding pairs of manually placed landmarks in the LSM-imaged CLARITY and SHIELD cleared input data and ARA is given. Mean and standard deviation (STDEV) of the landmark error is given for each sample; the mean and standard deviation of landmark error by brain region is also given for each sample. LSM, Light-Sheet Microscopy; CLARITY, Clear Lipid-exchanged Anatomically Rigid Imaging/immunostaining-compatible Tissue Hydrogel; SHIELD, Stabilization to Harsh conditions via Intramolecular Epoxide Linkages to prevent Degradation; ARA, Allen Reference Atlas.*

4 Discussion

The region with the greatest error, the cerebellum, was most often subject to gross deformation including rotation and translation relative to the rest of the brain. By comparison, the midbrain, a relatively fixed structure in the brain and scan, had the lowest error. However, we found that the midbrain is most subject to the two artifacts typical of tissue clearing and LSM methods: intensity inhomogeneity and hydrogel-based deformation. Our TRE demonstrates that CloudReg handles artifacts typical of hydrogel-based tissue clearing methods including intensity inhomogeneity, hyperlocal structural deformation (nonuniform, micron-scale deviation from true anatomic position), and local missing tissue exceedingly well. CloudReg achieves this by relying on a rough affine initialization. Thus, in gross regions of the brain that have been displaced relative to the rest of the intact brain, the error will increase.

Registration is a crucial first step in analyzing a single or cohort of samples but can be more informative if combined with additional downstream analysis methods including cell and axon detection. A potential extension of our current work will be to accelerate the registration component of the code by leveraging our existing C++ implementations (https://github.com/InsightSoftwareConsortium/ITKNDReg).

CloudReg can correct intensity, align and stitch, register, and visualize terabyte-scale brain volumes with artifacts and tears. CloudReg is immediately applicable to brain volumes spanning a variety of species and imaging modalities including mouse, rat, monkey, and human brain imaging.

Bibliography

1. Zhuo, J. & Gullapalli, R.P. MR Artifacts, Safety, and Quality Control. *Radiographics* **26,** 275–297 (2006).
2. Barrett, J.F. & Keat, N. Artifacts in CT: Recognition and Avoidance. *Radiographics* **24,** 1679–1691 (2004).
3. Tomer, R., Ye, L., Hsueh, B. & Deisseroth, K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. *Nature Protocols* **9,** (2014).
4. Park, Y.-G. et al. Protection of tissue physicochemical properties using polyfunctional crosslinkers. *Nat. Biotechnol.* **37,** 73–83 (2018).
5. Renier, N. et al. iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell* **159,** (2014).
6. Tomer, R., Ye, L., Hsueh, B. & Deisseroth, K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. *Nat. Protoc.* **9**, (2014).
7. Chung, K. & Deisseroth, K. CLARITY for mapping the nervous system. *Nat. Methods* **10**, (2013).
8. Bria, A. & Iannello, G. TeraStitcher - a tool for fast automatic 3D-stitching of teravoxel-sized microscopy images. *BMC Bioinformatics* **13,** (2012).
9. Goubran, M. et al. Multimodal image registration and connectivity analysis for integration of connectomic data from microscopy to MRI. *Nature Communications* **10,** (2019).
10. Renier, N. et al. Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes. *Cell* **165,** (2016).
11. Husz, Z.L., Burton, N., Hill, B., Milayev, N. & Baldock, R.A. Web tools for large-scale 3D biological images and atlases. *BMC Bioinformatics* **13,** (2012).

12. Tustison, N. J. et al. N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29,** (2010).

13. Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* **17,** (1998).

14. Faisal Beg, M., Miller, M. I., Trouvé, A. & Younes, L. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. *Int. J. Comput. Vis.* **61,** (2005).

15. Tward, D. et al. Diffeomorphic Registration With Intensity Transformation and Missing Data: Application to 3D Digital Pathology of Alzheimer's Disease. *Front. Neurosci.* **14,** (2020).

16. Vogelstein, J. T. et al. A community-developed open-source computational ecosystem for big neuro data. *Nature Methods* **15,** (2018).

17. Mori, S. et. al. MRICloud: Delivering High-Throughput MRI Neuroinformatics as Cloud-Based Software as a Service. *Computing in Science & Engineering* **18,** (2016).

18. Feng, L et al. Population-averaged macaque brain atlas with high-resolution ex vivo DTI integrated into in vivo space. *Brain Struct Funct.* **222(9),** (2017).

19. Ratnanather JT et al. Cortico-cortical, cortico-striatal, and cortico-thalamic white matter fiber tracts generated in the macaque brain via dynamic programming. *Brain Connect.* **3(5),** (2013).

20. Rosenblum, JS et al. Neuraxial dysraphism in EPAS1-associated syndrome due to improper mesenchymal transition. *Neurol. Genet*. **6,** (2020).

21. Lesciotto, KM et al. Phosphotungstic acid-enhanced microCT: Optimized protocols for embryonic and early postnatal mice. *Dev. Dyn.* **249,** 573–585 (2020).

22. Wang, Q. et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181,** (2020).

23. Papp EA, Leergaard TB, Calabrese E, Johnson GA, Bjaalie JG. Waxholm Space atlas of the Sprague Dawley rat brain. *NeuroImage* **97**:374-386 (2014).

24. Kjonigsen LJ, Lillehaug S, Bjaalie JG, Witter MP, Leergaard TB. Waxholm Space atlas of the rat brain hippocampal region: Three-dimensional delineations based on magnetic resonance and diffusion tensor imaging. *NeuroImage* **108**:441–449, (2015).
25. Sergejeva M, Papp EA, Bakker R, Gaudnek MA, Okamura-Oho Y, Boline J, Bjaalie JG, Hess A. Anatomical landmarks for registration of experimental image data to volumetric rodent brain atlasing templates. *Journal of Neuroscience Methods* **240**:161-169, (2015).