NeuroPathPredict (NPP): A data-driven paradigm to map the distribution of Alzheimer's disease neuropathology.

Raghav Madan

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

John H. Gennari, Co-Chair

Paul K. Crane, Co-Chair

Thomas J. Grabowski

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

University of Washington

Abstract

NeuroPathPredict (NPP): A data-driven paradigm
to map the distribution of Alzheimer's disease neuropathology.

Raghav Madan

Co-Chair of the Supervisory Committee:

John H. Gennari
Biomedical Informatics and Medical Education

Paul K. Crane
General Internal Medicine

Alzheimer's disease (AD) is characterized by the progressive accumulation of misfolded proteins, primarily tau neurofibrillary tangles, amyloid-β plaques, and TDP-43 inclusions, across various brain networks. Despite a century of histopathological insight, quantitative understanding of how these pathologies spatially unfold remains limited. Classical frameworks such as Braak and Thal staging distilled sparse regional observations into ordinal categories, yielding reproducible heuristics for disease progression but obscuring fine-scale gradients and inter-individual heterogeneity. Modern imaging and network-based diffusion models have extended these ideas to the living brain; however, they remain constrained by coarse parcellations, strong mechanistic assumptions, and the lack of direct calibration against histological ground truth. The field, therefore, lacks a rigorous, data-driven framework that can translate quantitative neuropathology into spatially continuous, anatomically interpretable maps across the brain.

To address this gap, I developed **NeuroPathPredict (NPP)**, a modular, open-source system that integrates quantitative histopathology with high-resolution neuroimaging and spatial statistics. NPP comprises three primary pipelines.

(1) **QNPtoVox** transforms Halo-derived, tile-wise tau burden measurements into voxel-level maps co-registered to the MNI ICBM 2009b template via ex vivo MRI, preserving anatomical orientation and enabling cross-participant comparability.

(2) The **Integrated-Brain Information System (I-BIS)** creates a multilayer “brain GIS” by unifying various structural, functional, and vascular atlases into a shared 0.5 mm volumetric grid. It produces thousands of biologically relevant covariates, such as distances, densities, and neighborhood features, which help contextualize each voxel regarding white-matter tracts, networks, and tissue types.

(3) The **NPP modeling framework** combines these data using universal kriging with external drift, a geostatistical method that links anatomy-informed predictors with residual spatial autocorrelation to infer continuous tau fields from sparse observations, providing both predictions and spatially explicit results uncertainty.

Applied to ten donors from the Adult Changes in Thought (ACT) study, NPP demonstrates that anatomically enriched models outperform non-spatial baselines and recover mesoscale gradients of tau burden consistent with known vulnerability patterns along association tracts and functional networks. The results show that spatial autocorrelation persists beyond measured covariates, validating the use of kriging in brain space and underscoring the value of integrated anatomical context. More broadly, NPP establishes a reproducible computational framework that transforms post-mortem histology into standardized, voxel-wise maps suitable for cross-modal validation with MRI and PET, for testing theories of selective vulnerability, and for modeling the spatial interplay of multiple pathologies. By integrating digital pathology, neuroimaging, and spatial statistics, this work enhances the ability to reconstruct, predict, and ultimately comprehend the brain-wide spatial dynamics of neurodegeneration.

Acknowledgements

I would like to express my deepest gratitude to my dissertation committee (**Dr. Paul K. Crane, Dr. John H. Gennari, Dr. Caitlin S. Latimer, Dr. Thomas J. Grabowski, Dr. Seo-Eun Choi,** and **Dr. Eardi Lila**) for their invaluable guidance, mentorship, and encouragement throughout this work.

My two primary mentors, Dr. Paul Crane and Dr. John Gennari, have been instrumental to both my scientific development and professional growth. Dr. Crane has been a steadfast mentor whose intellectual rigor, integrity, and deep sense of purpose have profoundly shaped the trajectory of this dissertation. His ability to connect biological meaning with statistical inference taught me how to think critically and creatively about data. Dr. Gennari's mentorship in biomedical informatics has been equally transformative. His emphasis on reproducibility, computational clarity, and thoughtful design has defined my approach to building data-driven systems. I am deeply indebted to both for their generosity, patience, and belief in this project.

I am also sincerely grateful to Dr. Caitlin Latimer, whose expertise in neuropathology and precision in tissue-based science anchored this work in biological reality, and to Dr. Thomas Grabowski for his vision in bridging neuropathology with imaging and clinical neuroscience. I thank Dr. Seo-Eun Choi for her guidance in statistical modeling and interpretation, and Dr. Eardi Lila for his valuable insights into spatial methods and model generalizability.

This research was supported by the **Adult Changes in Thought (ACT) Study**, funded through the National Institute of Health U19 grant (AG066567). I am deeply grateful to the **ACT participants and their families** for their generous donations of brain tissue, which made this work possible. I extend my sincere thanks to the **ACT Neuropathology Core** and the **ACT Neuroimaging Core** at the University of Washington and Kaiser Permanente Washington for providing access to data, tissue resources, and technical expertise that were essential to this dissertation.

I would also like to thank the **Department of Biomedical Informatics and Medical Education (BIME) at UW Medicine** for fostering an environment that encourages interdisciplinary collaboration and scientific curiosity. The department's mentorship, technical community, and commitment to translational research provided the intellectual home in which this work evolved.

Finally, I am deeply thankful to my family, friends, and colleagues for their unwavering support throughout this journey. Their encouragement, understanding, and belief in me sustained every stage of this work. To those who offered time, perspective, or simply kindness when it mattered most, thank you. This dissertation is as much a reflection of their support as it is of my own effort.

Table of Contents

ACKNOWLEDGEMENTS **1**

CHAPTER 1. INTRODUCTION AND OVERVIEW **6**

1.1 Overview of Alzheimer's disease 6
1.2 Motivation and Objectives 7
1.3 Scope and Outline 9

CHAPTER 2. BACKGROUND **11**

2.1 Alzheimer's Disease and Neuropathology 11
2.2 Spatial Modeling of Neuropathology 13
2.2.1 A brief history of spread models in AD *14*
2.2.2 Limitations of prior spread models *16*
2.2.3 What GIS has already solved: from LUR to UK *17*
2.3 Advances in QNP, Neuroimaging, and the Adult Changes in Thought study 20
2.3.1 Advancements in digital histopathology *20*
2.3.2 Advancements in neuroimaging *21*
2.3.3 The Adult Changes in Thought (ACT) study *22*
2.4 Summary 25

CHAPTER 3. QUANTITATIVE NEUROPATHOLOGY TO MNI VOXEL COORDINATES **26**

3.1 Introduction 26
3.2 Data input 27
3.2.1 Halo extract *28*
3.2.2 Gross tissue images *31*
3.2.3 Ex vivo MRI *32*
3.3 QNPtoVox pipeline overview 33
3.3.1 Orchestrating the pipeline *33*
3.3.2 Configuring the environment *36*
3.3.3 Configuration files *37*
3.4 QNP and coordinate extraction 38
3.5 Preprocessing ex vivo MRI 40
3.6 Co-location (QNP and ex vivo MRI) 42
3.7 Voxel transformation and kernel application 43
3.8 MNI co-registration 46

CHAPTER 4. INTEGRATED-BRAIN INFORMATION SYSTEM (I-BIS) **49**

4.1 Introduction 49
4.2 Input Data 51
4.2.1 Standard volumetric space selection *51*
4.2.2 Atlas Selection *54*
4.3 I-BIS Pipeline Overview 68
4.3.1 Orchestrating the Pipeline *68*
4.3.2 Configuring the Environment *71*
4.3.3 Configuration Files *71*

4.4 ROI Voxel Extraction 72
4.4.1 Order of Operations and the Upsampling Artifact 73
4.4.2 Individual Variable Extraction 74
4.4.3 Upsampling and Co-registration to MNI 2009b 75
4.5 Buffer-Zone Analysis 76
4.5.1 Implementation in I-BIS 77
4.6 Absolute Distance Analysis 79
4.7 Point-Based Covariate Extraction and Data Consolidation 80
4.7.1 Implementation of Point-Based Covariate Extraction 80
4.7.2 Data consolidation 81

CHAPTER 5: NPP UK MODEL 83

5.1 Introduction 83
5.2 NPP UK model pipeline setup 84
5.2.1 Orchestrating the pipeline 85
5.2.2. Setting up the environment 85
5.2.3. Configuration files 86
5.3 Preprocessing: Merging Data into a Design Matrix 87
5.3.1 Handling missing data 88
5.3.2 Data standardization and transformation 88
5.3.3 Merging and final validation 88
5.4 CV Elastic Net: Dimensionality Reduction 89
5.4.1 Elastic Net Overview 89
5.4.2 Elastic Net in NPP UK model 90
5.5 Feature Selection 93
5.5.1 Factor analysis in NPP UK model 93
5.5.2 Selecting primary predictors for UK 97
5.6 Universal Kriging: Spatial Modeling and Prediction 98
5.6.1 Introduction to UK for NPP UK model 98
5.6.2 Components of 04_universal_kriging.R script 101
5.7 Saving Results and Visualizations 106
5.7.1 Saving predictions 106
5.7.2 Visualization Directory 108
5.7.3 Logging and Outputs 109

CHAPTER 6. RESULTS AND ANALYSIS 110

6.1 Results for participant 6966 110
6.1.1 Primary predictors 113
6.1.2 Universal Kriging (UK) 115
6.2 All 10 participants modeled independently 120
6.2.1 Block split analysis 120
6.3 Significant predictor analysis 127

CHAPTER 7. DISCUSSION & CONCLUSION 132

7.1 Interpreting results 132
7.1.1 Spatial block cross-validation and prediction distance. 132

7.1.2 White-matter tracts and functional networks *134*
7.1.3 Robustness across participants and the role of data "density" *134*
7.1.4 Variogram characteristics *135*
7.1.5 Added value of spatial modelling *136*
7.1.6 Block inclusion patterns and edge effects *136*
7.2 Future Work (Methods) 137
7.2.1 Regional expansion and brain-wide analysis *137*
7.2.2 Model development beyond UK *138*
7.2.3 Spatial omics integration (SEA-AD) *139*
7.2.4 Enriching I-BIS covariates *139*
7.2.5 Pipeline engineering and scalability *140*
7.2.6 Validation and benchmarking *140*
7.3 Generalizability and future applications 140
7.3.1 Multi-pathology interaction mapping (Tau × Aβ × TDP-43) *141*
7.3.2 Reproducing and extending Braak staging at scale *142*
7.3.3 Subtype-specific trajectories and cross-modal subtypes *142*
7.3.4 Data-efficient corpora for diffusion and neural networks *143*
7.4 Contributions 145
7.4.1 Mapping digital histopathology to neuroimaging coordinates *145*
7.4.2 An anatomically derived way to study spatial patterns in the brain *146*
7.4.3 Universal kriging in brain space *147*
7.4.4 Three open-source, semi-automated pipelines *148*
7.5 Limitations 149
7.5.1 Cohort size and sampling scope *150*
7.5.2 Spatial resolution and scale mismatch *150*
7.5.3 Finite correlation distance and stationarity in UK *151*
7.5.4 Automation, throughput, and computational controls *151*
7.6 Conclusion 152
BIBLIOGRAPHY 156
INDEX OF FIGURES 165
INDEX OF TABLES 169

Chapter 1. Introduction and Overview

1.1 Overview of Alzheimer's disease

Alzheimer's disease (AD) is a devastating neurodegenerative disease that currently affects millions of people worldwide. The disease is characterized by a progressive decline in cognitive function, which eventually leads to complete dependence on caregivers and ultimately to death. AD adds a significant burden on patients, society, and the economy, with estimated costs exceeding $360 billion per year in the United States alone[1]. With an increasingly aging population, the prevalence of AD is projected to rise steeply, highlighting the urgent need for effective treatments. Until recently, treatment options provided only symptomatic relief and had no significant effect on disease progression[2]. The complexity of AD pathology, with multiple proteinopathies and heterogeneity in clinical presentation, further complicates the development of effective treatments. There is an urgent need to gain a deeper understanding of the underlying mechanisms of AD to develop effective interventions.

Mounting evidence over the decades has shown a strong association between AD and widespread depositions of beta-amyloid (Aβ) plaques and Tau neurofibrillary tangles (NFTs)[3]. In addition, TDP-43 inclusions, which were earlier linked with amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration (FTLD), have been found in many AD cases[4]. These proteinopathies follow overlapping spatial spread patterns. Evidence suggests Aβ starts in association cortices, eventually spreading to the neocortex and allocortex. The Tau cascade is believed to be triggered by Aβ in the entorhinal cortex, spreading to limbic areas and finally reaching the neocortex[5]. Similarly, distinct patterns of spread have been found for TDP-43[6], with a limbic-predominant and neocortex-predominant pattern. All three proteinopathies form a dynamic system whose underlying mechanistic pathways are still not fully understood.

Tau, amongst beta-amyloid and TDP-43, is closely associated with neurodegeneration and focal clinical deficits experienced by AD patients[7]. Accordingly, in this dissertation, I focus on Tau as the primary neuropathological target, developing spatial mapping and predictive modeling methods around its distribution. The seminal papers by Braak & Braak[8,9] outlined a 6-stage tau progression framework that has been a cornerstone in AD research since. New studies with much larger sample sizes and the latest neuroimaging and neuropathology techniques have suggested a heterogeneous spatiotemporal spread of Tau falling into distinct trajectories[10,11]. Neuropathological subtypes have been shown to follow significantly different longitudinal cognitive trajectories[12]. Traditionally, AD was thought to universally involve a gradually progressing late-life syndrome with predominant memory impairment, followed by a cascade of cognitive impairments. However, several

syndromes have been identified with either linguistic, executive, behavioral, or visuospatial cognition deficit as the primary impairment, with AD as the underlying primary pathology[13,14] . In addition to these varied clinical presentations, there are spatial variations that have been reported in radiological and pathological presentations. The underlying mechanisms contributing to this heterogeneous rather than homogeneous expression of AD are not well understood. One possible common thread is the co-localization of tau and neurodegeneration in impacted regions. It has been conjectured that these distribution differences could be attributed to individual differences in brain organization, regional vulnerabilities, or intrinsic properties of disease pathology[13].

A major challenge in the field is that a tiny proportion of the brain is directly assessed in even the most comprehensive neuropathological evaluations for AD pathology. One rule of thumb is that a standard neuropathology workup includes direct observation under the microscope of 1/10,000th of the brain. Specific brain segments are stained for AD neuropathology findings and evaluated. As these neuropathologies are distributed brain-wide, there is a massive amount of missing structural information about their location and distribution.

1.2 Motivation and Objectives

I begin from a simple premise. The premise is that the spatial distribution of neuropathological proteins, their location, sequence, and preferred anatomical pathways, encodes information essential for understanding Alzheimer's disease. The goal is to transform sparse, heterogeneous quantitative histopathology into brain-wide, interpretable maps that can be analyzed with the same spatial logic, long available to neuroimaging. In this dissertation, I focus on tau as the primary target pathology and develop a pipeline that co-locates digital histology with ex vivo MRI in MNI[15] (Montreal Neuroimaging Institute) (a standard brain coordinate template), engineers biologically interpretable voxel-level covariates, and predicts pathology fields with calibrated uncertainty (Figure 1.2.1).

To meet this goal, I designed an end-to-end workflow composed of three components: (i) QNPtoVox, which maps Halo[16]-derived quantitative histopathology to MNI while preserving intensity and orientation; (ii) the Integrated-Brain Information System (I-BIS), which assembles a reusable covariate library and generates point values, distances, and buffer-density features across multiple concentric radii; and (iii) NeuroPathPredict (NPP) modeling stack, which is centered on universal kriging with external drift to recover continuous pathology surfaces from limited samples. Together, these components create a standardized, reproducible foundation for voxel-wise analysis across donors and studies.

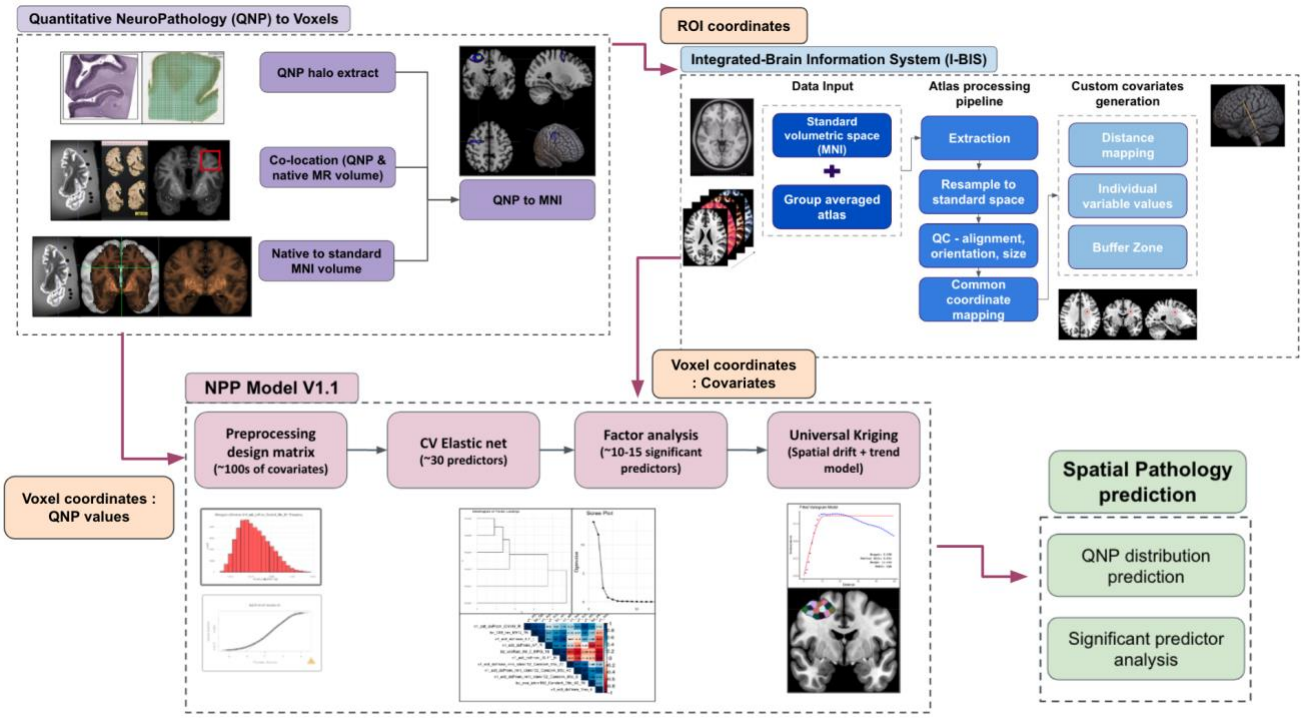


Figure 1.2.1: *NeuroPathPredict (NPP) pipeline: QNPtoVox mapping, I-BIS covariate engineering, and universal-kriging based spatial prediction of tau pathology.*

Motivated by this framework, I set three objectives. First, establish a robust co-location procedure that renders quantitative histology commensurate across donors and stains in a common MNI coordinate system. Second, build an extensible covariate library that embeds each voxel in anatomical and physiological context, white-matter tracts, functional networks, vasculature, and cortical geometry, without sacrificing interpretability. Third, given co-located pathology and contextual covariates, recover spatially coherent tau fields from sparse observations, and explain predictions in terms that engage current theories of AD progression.

Methodologically, I adapt ideas from geospatial science[17]: when measurements are sparse, models that harness spatial autocorrelation can interpolate fields credibly and reveal structure. NPP uses universal kriging with external drift to integrate co-located tau observations and I-BIS covariates, predicting pathology in the vast expanse of unobserved tissue, while indicating where predictions are reliable. In the proof-of-concept application to the middle frontal gyrus (ten donors), I compute hundreds of covariates in MNI space and evaluate model performance against non-spatial baselines, emphasizing generalizability and interpretability.

The practical objective is a deliverable toolkit and dataset: open, standardized code; precomputed, shareable MNI-space covariates; and validated modeling procedures that produce continuous tau maps. Strategically, the objective is a foundation that can scale beyond one region, one histological stain, or one disease stage, enabling comparative studies, hypothesis testing about anatomical pathways of spread, and, ultimately, integration with multimodal biomarkers to advance mechanistic understanding of Alzheimer's disease.

1.3 Scope and Outline

Building on section 1.2, I narrow the proof-of-concept to tau (% stained for AT8 antibody for phosphor-tau) in the middle frontal gyrus (MFG) using quantitative digital histopathology from ten donors co-located to MNI (Montreal Neuroimaging template) space. This scope lets me stress-test the end-to-end workflow on real tissue while keeping the problem tractable for careful validation and iteration. I treat amyloid and TDP-43, additional regions, and longitudinal progression as out of scope for this dissertation; the intent is to establish a rigorous template that can be extended later. Within this setting, I evaluate predictive performance and uncertainty using spatial block cross-validation and report RMSE, MAE, and R^2, benchmarking against non-spatial baselines. I prioritize reproducibility (standardized inputs, versioned code, precomputed MNI-space covariates) and interpretability (anatomically grounded features, explicit variogram modeling).

Next, I briefly outline how each chapter advances this agenda, from contextual background and data harmonization through covariate engineering and spatial modeling to results, limitations, and implications.

Chapter 2 (Background) lays the conceptual groundwork. I summarize the literature on tau/amyloid progression, classical staging, and PET-era adaptations, against their constraints. Then I develop the argument for treating neuropathology as a spatial field that must be modeled with anatomically interpretable covariates. I also preview the ACT cohort as the empirical setting in which this research can be executed.

In **Chapter 3 (QNPtoVox)**, I transform slide-level quantitative histopathology into analysis-ready data by anchoring sections with ex vivo MRI, resolving scale and distortion, and mapping measurements into MNI 2009b so that every voxel value is traceable to tissue.

In **Chapter 4 (I-BIS)**, I enrich those voxels with standardized anatomical context, functional networks, white-matter tracts, vascular territories, cortical geometry, and

distance transforms, thereby constructing a “brain GIS” that yields a clean, interpretable design matrix for modeling.

Chapter 5 (NPP UK model) presents the modeling workflow. I specify an anatomy-informed drift and fit a universal-kriging style spatial model, validate with spatial block cross-validation, and report per-voxel uncertainty to support calibrated, brain-wide inference.

Chapter 6 (Results) delivers the empirical evidence: I quantify predictive accuracy and calibration, examine which anatomical features most strongly explain tau burden, compare against alternative approaches and available in-vivo measures, and assess robustness across participants.

Finally, in **Chapter 7 (Discussion and Conclusion),** I close by integrating the biological and methodological lessons, acknowledging limitations, and outlining how the framework scales to additional regions, markers, and cohorts. I also argue that the artifacts produced here, standard-space histology volumes, the I-BIS atlas stack, and a reproducible NPP pipeline, constitute durable infrastructure for future studies and translation.

Chapter 2. Background

I use this chapter to establish the empirical and technical foundations that motivate the NeuroPathPredict (NPP) framework. I first synthesize what is known about the spatiotemporal unfolding of Alzheimer's disease (AD) pathology, emphasizing tau and amyloid staging, and make explicit the historical limitations in how these processes have been studied and mapped. I then trace the evolution of spatial modeling, from reaction–diffusion and connectome-based spread models to a GIS-inspired, universal-kriging framework, clarifying assumptions, evidence, and gaps. Finally, I show how digital quantitative neuropathology, together with a common neuroimaging coordinate system and modern atlases, supplies the measurements and spatial reference needed to build brain-wide predictive models for pathology in unsampled tissue.

2.1 Alzheimer's Disease and Neuropathology

Several decades of scientific research converged on the idea that misfolded proteins can seed further misfolding and move between cells, offering a mechanistic rationale for orderly, large-scale spread. Early unifying arguments framed neurodegenerative diseases through the lens of prion biology, emphasizing templated conformational change and intercellular transfer as drivers of progression[18]. Subsequent reviews synthesized experimental evidence that intraneuronal aggregates cross membranes, propagate trans-synaptically, and thereby advance pathology in a "prion-like" manner, explicitly extending these principles to tau and amyloid-β (Aβ)[19]. This framework explains why pathology appears in connected territories and sets expectations for stage-like topographies even before one specifies anatomical schemes.

Classical neuropathology research further mapped tau pathology into ordered stages that unfold across specific cortico-limbic circuits**.** In their landmark work, Braak and colleagues used whole-hemisphere histology to show that neurofibrillary tangles (NFTs) follow a six-stage topography: trans entorhinal (I–II), limbic (III–IV), and widespread neocortical involvement (V–VI)[8,9]. Subsequent elaborations described the same sequence with AT8 immunohistochemistry and detailed layer-specific patterns at each stage, reinforcing the canonical limbic-to-neocortical spread of tau[20]. Amyloid-β (Aβ) deposition follows a different but similarly stereotyped sequence described by Thal phases[21,22]**.** Thal et al. reported five phases beginning in the neocortex, then the allocortex, followed by the diencephalon/basal forebrain, select brainstem nuclei, and ultimately the cerebellum, establishing a hierarchical, regionally ordered scheme for Aβ distribution. Reviews of amyloid biology and staging corroborate this rostro-caudal sequence and place Thal phases alongside tau staging as core elements of neuropathologic assessment[23]. Consensus NIA–AA guidelines embed both frameworks, tau NFT stages, and Aβ phases

into standardized autopsy workflows that define Alzheimer's disease neuropathologic change (ADNC)[24].

PET imaging emerged as another way to study the spread and staging of neuropathologies. With tau-selective PET tracers, researchers adapted Braak's postmortem regions to in vivo "PET-Braak" staging, showing that higher stage assignments relate to worse cognition, greater Aβ burden, and medial temporal atrophy, thus conferring clinical relevance to staging in the living brain[25]. Methodological studies compared regional and whole-brain tau-PET measures and found that medial temporal and inferior/middle temporal uptake captured key clinical and imaging correlates, supporting the utility of PET-Braak schemas[26]. Large multi-center analyses further modeled natural history with tau-PET, observing that progression beyond mid-Braak levels occurs chiefly in the presence of Aβ, thereby anchoring stage transitions to amyloidosis[27]. Importantly, population-scale work with >1,600 individuals revealed heterogeneity: tau does not follow a single path but rather at least four distinct spatiotemporal trajectories, reframing "one-size-fits-all" staging as an average over divergent patterns[11]. These PET-era insights, both validation and heterogeneity, motivate modeling approaches that respect individual differences while preserving comparability across cohorts[28].

Despite the decisive role that classical staging has played in shaping our understanding of AD, I argue that the field now needs a synthesis that leverages the microscale truth of histology while escaping the methodological ceilings of PET-era staging. Histology remains the gold standard for establishing the presence and cellular localization of pathology, yet routine protocols examine a defined set of regions rather than the entire brain, which means that most tissue is never directly observed, and any spatial inference necessarily proceeds from sparse samples[29]. The staging systems that distilled these sparse observations into ordered progressions, Braak stages for tau and Thal phases for amyloid, were historically necessary and enormously influential, but by design they compress a continuous, heterogeneous distribution of pathology into a small number of ordinal categories defined over coarse regions, thereby obscuring fine-scale gradients that may matter for spatial prediction and mechanistic interpretation.

PET-based staging brought a transformative step: it moved staging into the living brain, enabling longitudinal observation and clinicopathologic correlations at scale. Tau PET "Braak-like" schemas are robust predictors of cognitive decline and disease progression, validating the overall logic of ordered progression in vivo[30]. However, PET-era staging inherits technical sensitivities: first-generation tau tracers such as ^{18}F-flortaucipir show off-target binding in regions such as the basal ganglia and choroid plexus[31,32], complicating quantification and visual reads; methodological choices around SUVR

reference regions, partial-volume correction, and pipeline details introduce method variance across sites[33]; and several autopsy-anchored studies and reviews have questioned early-stage sensitivity, particularly at Braak I–IV[34]. Together, these issues are exactly the kinds of in-vivo ambiguities that a histology-anchored framework can help complement.

Histology has limitations that modern pipelines must address explicitly. Beyond the basic sparsity of sampled tissue mandated by standard protocols, cross-site heterogeneity in tissue handling, staining, and interpretation persists even under consensus guidelines, as documented in multisite assessments and historical standardization efforts[35,36]. And because histology sections are acquired and analyzed at a cellular scale, reconciling those measurements with millimetric neuroimaging coordinates requires deliberate choices about scale, smoothing, and registration to avoid erasing biologically meaningful structure, challenges widely noted in the MRI–histology co-registration literature (e.g., slice-thickness mismatch, tissue shrinkage, non-linear distortion)[37,38].

My pipeline design makes these trade-offs transparent. I first generate voxel-wise, standard-space pathology volumes from quantitative histology (QNPtoVox), then embed those voxels in a rich anatomical context via an MNI-aligned atlas stack, and finally fit spatial models that predict AT8 intensity (percentage positively stained using AT8 antibody for phosphor-tau) in unobserved tissue, thereby preserving histology's ground-truth specificity while exploiting imaging's whole-brain coordinate system. This systematic workflow mitigates the structural weaknesses of both traditions: it goes beyond region-coarse, ordinal staging; calibrates and contextualizes PET findings against histological ground truth (including suspected off-target or partial-volume effects); and produces continuous, reproducible brain maps that are comparable across donors and studies[39].

2.2 Spatial Modeling of Neuropathology

In the previous section, I outlined the prion-like propagation and the staging traditions that have organized AD neuropathology for three decades. Here, I deepen the focus by reviewing the history of spatial models of AD pathology, making explicit their assumptions, findings, and limitations. Then, by analogy to methods from geographic information science (GIS), I assert that a data-first, universal kriging strategy can convert sparse histology into dense, region-wide maps with quantified uncertainty and biologically interpretable covariates. This section sets up the enabling advances in quantitative digital histopathology and neuroimaging (section 2.3) that make my approach feasible.

2.2.1 A brief history of spread models in AD

1. Reaction–diffusion and nucleation–growth models (1990s–2000s)

Early mathematical work adapted nucleated polymerization and reaction–diffusion ideas to misfolded proteins, modeling local seeding, templated growth, and diffusion/clearance in continuous media. These frameworks captured sigmoidal kinetics and wave-like spread but were largely spatially homogeneous (i.e., they did not encode brain anatomy beyond generic diffusion) and operated at coarse spatial scales[40,41].

- Assumptions- Isotropic diffusion with constant coefficients; uniform production/clearance; no explicit brain connectivity.

- Findings- Reproduced generic spatiotemporal patterns and offered mechanistic interpretations of sigmoidal accumulation.

- Limitations- Ignored the brain's architecture and heterogeneity; difficult to test against region-wise imaging or histology.

2. Network diffusion on the structural connectome (2010s)

A major step was to embed propagation in the human connectome. The Network Diffusion Model (NDM) treats pathology as a linear diffusion process over the white-matter graph: the graph Laplacian governs spread from initial seeds to connected regions[42]. With longitudinal imaging, NDM predicted atrophy/metabolism trajectories in AD[43], and extensions localized probable sources of spread[44]. Parallel epidemic-spreading models (ESM) added nonlinear production/clearance and saturation[45]. Network-based models reframed AD as a systems-level disease whose geography respects large-scale connectivity, moving beyond purely regional staging.

- Assumptions-
    - A fixed, group-average structural connectome approximates individual pathways.
    - Linear (NDM) or compartmental (ESM) kinetics with spatially homogeneous parameters.
    - Seeds are sparse; propagation is constrained primarily by white-matter connectivity.

- Findings-
    - Structural connectivity explains a non-trivial fraction of where pathology and atrophy appear next, especially for tau-related changes[43] .
    - For amyloid-β, simple Euclidean proximity sometimes rivals or exceeds connectomic constraints in animal models, suggesting protein-specific spread regimes[46].
    - Adding clearance/susceptibility terms improves fit and offers a unified language for misfolded proteins.[45]

- Limitations-
    - Models propagate on a static, group-average connectome, making results sensitive to parcellation and tractography choices and potentially misrepresenting participant-specific pathways.
    - The homogeneous, globally applied kinetics downplay regional heterogeneity, vascular/CSF dynamics, cytoarchitecture, and cell-type composition, which might modulate tau or Aβ burden.

3. Data-driven ordering and disease-progression modeling (late 2010s–2020s)

A complementary thread inferred temporal orderings directly from cohort data, without prescribing a transport law. Event-based models (EBM) estimate a probabilistic sequence in which regional measures transition from “normal” to “abnormal,” and assign individuals a stage on that sequence. Oxtoby et al. (2017)[47] applied EBM to anatomical brain connectivity and recovered an ordered pattern of disconnection with subject-level staging. The methodology has since been consolidated as a practical protocol, covering mixture modeling of normal/abnormal distributions, maximum-likelihood sequence estimation with uncertainty, and individual staging, for use across modalities[48].

- Assumptions-
    - Monotonic change of biomarkers/regions from normal to abnormal.
    - Conditional independence given disease stage, enabling tractable likelihoods.
    - Inference over cross-sectional or short-term longitudinal data.

- Findings-
    - Group-level sequences recover biologically plausible orders (e.g., early involvement of default-mode/medial parietal connectivity).
    - Individual-level probabilistic staging places each participant on the inferred sequence and enables downstream association with cognition.

- Limitations-
    - Region-averaged inputs; limited voxel-wise resolution.
    - Implicit spatial dependence rather than explicit modeling.
    - Sensitive to cohort composition and measurement pipelines.

2.2.2 Limitations of prior spread models

I build on the insights from modeling approaches discussed above, but depart methodologically for these four limitations:

(1) Scale mismatch with histological truth.

Most spread models operate at parcel granularity (i.e., dozens to a few hundred cortical/subcortical ROIs), whereas the ground truth for neuropathology is cellular-scale histology. Collapsing heterogeneous tissue into parcel averages obscures laminar and perivascular gradients that plausibly govern tau accumulation. As a result, such models are difficult to calibrate directly against histology and can only be validated indirectly via region-level PET or atrophy.

(2) Strong structural assumptions and equifinality.

Network diffusion model and Epidemic spread model specify how the pathology must move (diffusion/epidemic on a fixed graph with homogeneous parameters). In practice, multiple seed sets and parameter choices can fit the same cross-sectional patterns (equifinality), and subject-to-subject variability in connectivity further complicates inference. Mechanistic conclusions, therefore, rest on assumptions that are hard to verify and can be fragile to parcellation, tractography method, and seed selection.

(3) Limited use of rich anatomical context.

Connectivity is only one lens on vulnerability. Vascular territories, functional networks, gyral–sulcal geometry, distance to CSF spaces, and white-matter tract density all potentially shape exposure, clearance, or stress, yet they rarely appear explicitly in mechanistic spread models. When they do, they enter as post hoc correlates rather than as formal predictors with effect sizes and uncertainty.

(4) Comparability and uncertainty.

Outputs are often model states (e.g., time, concentration) in parcellated spaces

that are not directly comparable to voxel-wise maps across donors. Uncertainty quantification is typically global (parameter/posterior intervals), not spatially resolved per voxel/region, limiting downstream statistical use.

These limitations argue for a data-first spatial model that (i) honors histology as the target signal, (ii) uses anatomy-informed covariates rather than hard-coding a transport law, and (iii) produces voxel-wise predictions with calibrated uncertainty in a common template space.

2.2.3 What GIS has already solved: from LUR to UK

When I step back, the problem looks familiar, not from neurology, but from environmental exposure modeling based on geography. A handful of air-quality monitors provide local insights into NOx or $PM_{2.5}$ levels, whereas most street corners lack data monitoring. Urban scientists learned long ago how to turn those scattered numbers into a city-wide exposure map: they lean on what is known about the city itself, the roads and factories, the river and the ridge, the morning wind, and on the fact that nearby places tend to be similar. That is the essence of modern environmental exposure modeling: use interpretable spatial covariates to explain the mean, and let a residual spatial process borrow strength from neighbors[17,49,50].

There have been parallel observations in neuroimaging. The brain is a multilayered map, where coherent structure emerges only when we honor scales (from cellular laminae to large-scale networks) and reference frames (common coordinates and landmarks)[51]. I adopt that view throughout this dissertation. Histology tiles are our monitors. White-matter tracts, functional networks, ventricles and vessels, and cortical topography are our "roads," "districts," "shorelines," and "terrain." The goal remains the same: to predict a continuous field, specifically tau burden, at every location using covariates that carry biological meaning.

Land-use regression: the first layer of the story

Land-use regression (LUR) is the main tool in environmental epidemiology: it models pollutant levels at each monitoring site based on various spatial features, such as proximity to major roads, traffic volume, land cover, elevation, and weather conditions. These variables are carefully selected and regularized to prevent overfitting[52,53]. LUR scales well and yields interpretable coefficients ("each 100 m closer to a highway adds X μg/m³). Its limitation is also clear: residual spatial dependence almost always persists because no covariate stack fully captures how air mixes through streets and courtyards[49,54].

In brain terms, a LUR-only strategy would regress tau on neuroanatomical features, e.g., tract density, proximity to the default mode network, and distance to the lateral ventricle. That alone is valuable (we learn which features matter), but tau still exhibits mesoscale autocorrelation that such a regression will leave on the table. This is precisely the kind of multiscale, reference-anchored structure we must encode explicitly.

Universal kriging: adding the second layer

Geostatistics supplies the missing piece with universal kriging (UK): decompose the field into a drift (mean) explained by covariates and a residual spatial process governed by a variogram [55–59]. Formally, for location s,

$$Z(\mathbf{s}) = \underbrace{\mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta}}_{\text{drift (covariates)}} + \underbrace{\varepsilon(\mathbf{s})}_{\text{residual GP}}, \qquad \gamma(\mathbf{h}) = \tfrac{1}{2}\,\mathrm{Var}[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})],$$

where γ is the variogram parameterized by a range (correlation distance), sill (spatial variance), and nugget (microscale variation/measurement error). UK then produces best linear unbiased predictions (BLUPs) at any target location by weighting nearby residuals according to γ, and critically, returns a per-pixel (/voxel) prediction variance alongside the estimate.

Large multi-city projects such as MESA Air repeatedly found that UK (or its close cousins like regression-kriging) outperforms pure LUR when spatial structure persists after adjusting for covariates, yielding sharper maps and honest uncertainty[17,49,50].

Translating the blueprint to the brain

The analogy is not superficial; each GIS element has a precise neuroanatomical counterpart, exactly the multilayer-map logic of Zaleshina & Zaleshin (2017):

- Monitors → histology tiles: A sparse, irregular set of tissue samples with quantitative AT8 burden.

- Land-use layers → I-BIS layers**:** Distance to highways/industrial land-use/elevation becomes distance to ventricles or major vessels, tract density/proximity, functional network membership, vascular maps, cortical depth/curvature, gyral–sulcal position, and macro-parcellation labels.

- Exposure field → tau field**:** A continuous scalar defined everywhere in the brain volume, expected to be autocorrelated at mesoscales.

- City map → MNI 2009b: A common coordinate system (the reference frame) where features and outcomes align, enabling pooling and cross-study comparison.

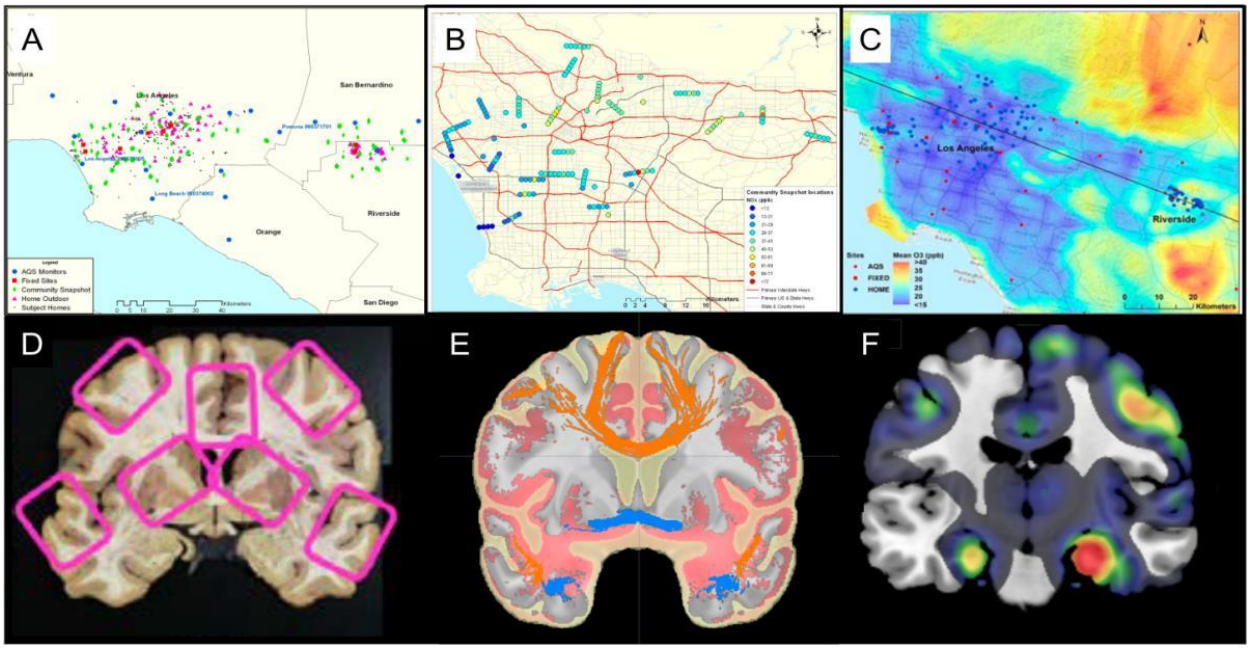


Figure 2.2.1: *From city maps to brain maps. Panels A–C[17] illustrate the land-use/universal-kriging blueprint: sparse monitors (A), contextual covariates such as roads/topography (B), and the kriged continuous NOx surface (C). Panels D–F translate that logic to neuropathology: NIA–AA diagnostic sampling sites as sparse measurements (D)[35], a brain GIS in MNI space with tracts, networks, vasculature, and geometry as drift covariates (E), and a potential voxel-wise disease map (F[60]).*

GIS taught us how to turn a few trustworthy measurements into a map you can use. By importing Universal Kriging into the brain, where our "monitors" are histology tiles and our "land-use" layers are neuroanatomy arranged as a multilayered map, I can produce continuous, tau pathology maps that respect biology, scale across donors, and speak directly to selective vulnerability. In section 2.3, I explain why this is now feasible: digital quantitative histopathology provides dense, reproducible measurements; MNI 2009b and modern atlases offer the coordinate framework and covariates; and the ACT cohort supplies the standardized tissue and imaging datasets needed to construct these maps at scale.

2.3 Advances in QNP, Neuroimaging, and the Adult Changes in Thought study

I use this section to assert that two parallel streams of progress enable this research. First, only recently, digital histopathology has become truly quantitative at the whole-slide scale. And second, neuroimaging has matured a common coordinate framework with rich atlases, enabling the fusion of gold-standard tissue evidence with brain-wide modeling in a rigorous, reproducible way. I first trace the methodological arc in digital histopathology that moved the field from expert, region-sampled reads to high-throughput, quantitative measures on entire slides. I then turn to neuroimaging, specifically, the MNI 2009b template[61] space and the ecosystem of atlases and PET pipelines that make multimodal data interoperable at the voxel level. I close by situating this dissertation in the Adult Changes in Thought (ACT) cohort[62], whose population-based design, tissue repository, and imaging resources supply the empirical substrate for the chapters that follow.

2.3.1 Advancements in digital histopathology

Classical neuropathology emphasized the reliability of diagnosis via semi-quantitative scoring on a limited set of blocks and stains, an approach codified in the NIA–AA[29] autopsy framework and its summary of AD neuropathologic change. That framework remains the field's backbone, but by design it summarizes spatially sparse observations. Over the past decade, whole-slide imaging (WSI)[63] has been validated as non-inferior to light microscopy for primary diagnosis across broad surgical pathology workloads, and major regulators have cleared WSI systems for clinical use, changes that catalyzed the shift from glass to pixels and opened the door to computational measurement at scale.

On top of WSI, software ecosystems now convert visual impressions into reproducible, continuous measurements. QuPath[64] is a widely used, open-source platform that provides batch processing, scripting, and object-based analytics designed for whole-slide workflows and has been adopted across biomedical research. In parallel, commercial platforms such as the HALO[16] package offer ready-to-use modules for area quantification, cell detection, and spatial analysis that have been applied in neurodegenerative disease research, including tau. Together, these tools have accelerated quantitative pipelines for AD neuropathology, including recent studies that benchmark automated measures of Aβ and AT8-positive tau against expert scoring[65].

At the marker level, phosphorylated-tau immunohistochemistry with AT8 remains central to quantifying neurofibrillary pathology. Early and subsequent epitope-mapping studies show that AT8 recognizes double phosphorylation at Ser202/Thr205[66], with additional evidence that nearby phosphorylation at Ser208[67] can enhance binding. These

biochemical details explain AT8's sensitivity to pre-tangles and tangles and justify its continued use as a quantitative readout in digital pipelines.

Together, validated whole slide imaging, mature analysis software, and marker-level clarity let me treat each slide as a spatially indexed measurement field, not a handful of ordinal grades. In practical terms, I can compute tile-wise or object-wise tau burden (e.g., percent area positive) and standardize outputs across batches, capabilities that are prerequisites for brain-wide modeling.

2.3.2 Advancements in neuroimaging

On the imaging side, community templates developed by the Montreal Neuroimaging Institute (MNI) in the MNI ICBM 2009[15] series supply a high-resolution, widely adopted coordinate system for pooling and comparing multimodal data. In particular, the MNI 2009b nonlinear templates (available at 0.5 mm isotropic resolution) make it feasible to align histology-derived volumes, ex vivo MRI, and in vivo imaging to a shared grid, enabling voxel-wise analyses and cross-study comparability. In this dissertation, I register quantitative histology to MNI 2009b to ensure that tissue-anchored maps can be analyzed with the same tools commonly used for MRI and PET.

Rich, population-derived atlases now annotate that common space with neurobiologically meaningful structure. For example, the Yeo 7/17-network[68] cortical parcellations provide large-scale functional systems derived from 1,000 participants, while statistical arterial atlases[69] situate voxels in a vascular supply context relevant to selective vulnerability. Diffusion MRI templates and tractography atlases built from HCP-1065[70] participants add white-matter information at scale. In my modeling, these atlases become voxel-wise covariates, biologically interpretable features that can help explain and predict where tau accumulates.

Crucially, tau-PET pipelines have been adapted explicitly to MNI-space frameworks. Following Schwarz et al. (2016)[71], many groups compute PET-Braak stages by summarizing tracer uptake in MNI-defined Braak ROIs; subsequent multi-center and longitudinal work shows that PET-Braak staging tracks clinical severity and prognoses cognitive decline. The practical implication is straightforward: mapping tissue-anchored tau burden to MNI space allows direct, apples-to-apples comparison with in-vivo PET staging, one coordinate system for living and post-mortem data. A robust common space with atlas context converts quantitative histology into brain-wide, voxel-wise maps amenable to statistical learning, spatial inference, and cross-modal validation. This is the substrate on which I build the predictive models in later chapters.

2.3.3 The Adult Changes in Thought (ACT) study

The ACT study[62,72] is a long-running, population-based cohort embedded in Kaiser Permanente Washington. Participants (≥65 years) who are free of dementia enroll, receive serial cognitive assessments, and many consent to brain donation, creating a clinico-pathologic resource designed for community representativeness and autopsy-anchored inference. Over three decades, ACT has built a platform that integrates longitudinal clinical data, neuropathology, and, more recently, advanced imaging to support research on aging and dementia. In addition to medical records and study visits, ACT study curates genetics/biomarkers, neuroimaging, and life-course exposures (Figure 2.3.1). For this dissertation, I draw primarily on two corpora: **(i)** the ACT Neuropathology/Tissue Repository[73] and **(ii)** ACT's post-mortem neuroimaging resources. I summarize the cohort and datasets I use again at the end of this chapter to anchor the empirical analyses that follow.

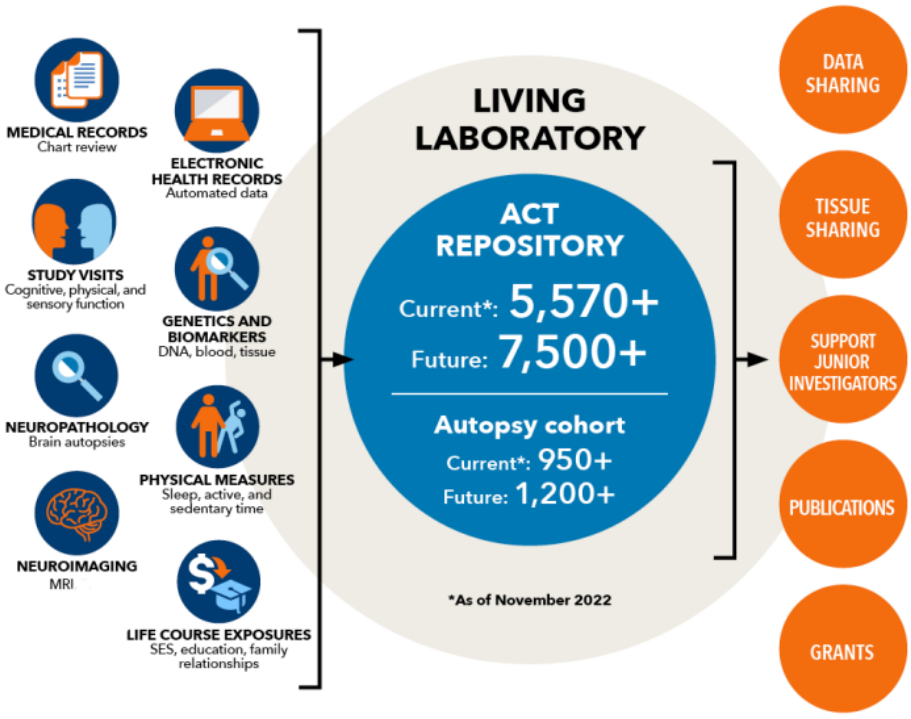


Figure 2.3.1. *The ACT[62] "Living Laboratory": An integrated clinico-pathologic and neuroimaging repository enabling brain-wide spatial modeling. Image from ACT study's website.*

ACT's autopsy pipeline is designed to satisfy consensus diagnostic criteria while enabling quantitative and imaging-informed analyses. The protocol specifies systematic fixation,

ex vivo MRI with specimen stabilization, circuit-based slab photography and sampling, standardized immunohistochemistry (IHC), optional multiplex quantification, and digital whole-slide analysis, each step documented for cross-case comparability[74].

After rapid procurement and gross documentation, whole brains or hemibrains are immersed in 10% neutral-buffered formalin for extended fixation (typically ≥2 weeks), with scheduled fixative changes. Specimens are then embedded in agarose to stabilize orientation and minimize deformation and air trapping during imaging and sectioning, procedures that improve downstream staining and facilitate accurate spatial registration.

Agarose-embedded specimens are scanned on a clinical 3T MRI with high-resolution 3D sequences (e.g., FLAIR/GRE). For downstream pipelines, a reverse-contrast ("pseudo-T1") image is derived and processed with standard neuroimaging tools (e.g., FreeSurfer-style segmentation), yielding a distortion-controlled volume that becomes the anatomical scaffold for slab photos, blockface localization, and slide-to-MRI registration (Figure 2.3.2). Reviews and method papers corroborate the translational value and practicality of ex vivo MRI for bridging histology and neuroimaging.

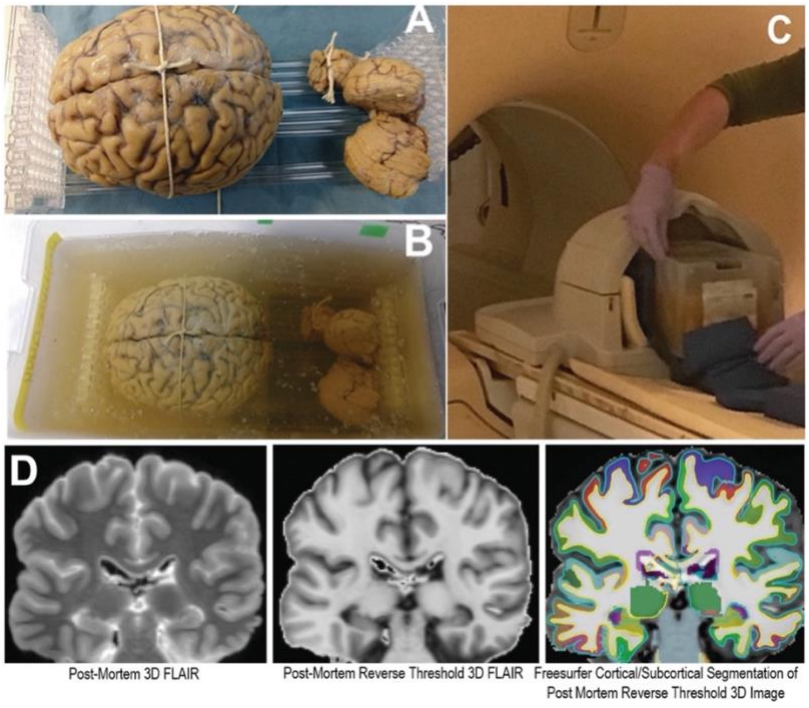


Figure 2.3.2. *Ex vivo MRI workflow in ACT[74]: (A) fixed brain, (B) agarose embedding, (C) specimen positioned for acquisition, and (D) post-mortem 3D FLAIR, reverse-contrast image, and representative cortical/subcortical segmentation. Image from Latimer et. al. 2023.*

Following MRI, the agarose block is sectioned into uniform coronal slabs on a calibrated slicer (ACT uses 4 mm), laid out anterior-to-posterior, and photographed with a fixed-height rig on a scale-referenced background (Figure 2.3.3). Diagnostic sampling required by NIA–AA guidelines is then complemented by imaging-guided circuit sampling that targets structures or lesions noted on ex vivo MRI. This dual approach preserves classical reads while enriching for spatially informative blocks that connect histology to the volumetric reference.

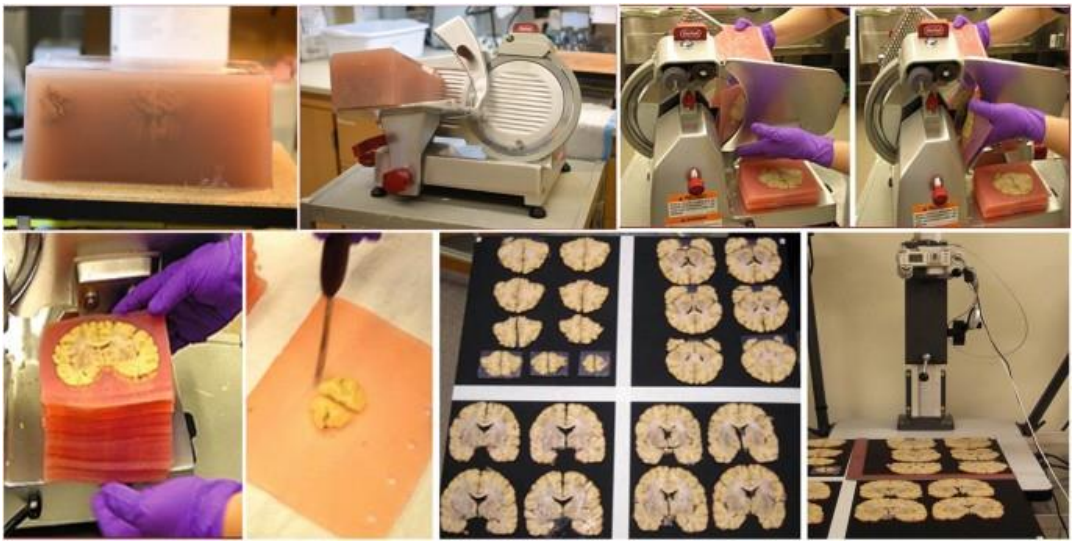


Figure 2.3.3. *Coronal slabbing and photographic documentation[74]: uniform 4 mm sections on a calibrated slicer, targeted block isolation, and slab mosaics photographed with a fixed camera rig for scale fidelity and alignment to MRI. Image from Latimer et. al. 2023.*

Formalin-fixed paraffin-embedded (FFPE) blocks support a standard panel of stains and IHC, including AT8 for phosphorylated tau and Aβ immunostaining. Where appropriate, ACT augments slide-based reads with Luminex-based multiplex assays performed on FFPE sections, providing quantitative measures of Aβ and p-tau. Slides are scanned as whole-slide images (WSI) and analyzed in dedicated software (e.g., HALO) to derive continuous burden metrics such as percent-area-positive. Because each block inherits coordinates from the slab photos, and slabs are registered to ex vivo MRI, tile-level measurements can be mapped back into volumetric space, enabling the voxel-wise reconstructions I use in this dissertation.

ACT's Neuropathology Core maintains a tissue and data repository with standardized datasets available to qualified investigators, alongside parallel program resources (clinical, imaging, and omics). This infrastructure is explicitly organized for quantitative reuse and cross-study integration, conditions that make ACT a strong substrate for MNI-aligned, atlas-informed spatial modeling.

2.4 Summary

In this chapter, I discussed the conceptual scaffolding for this dissertation and used it to arrive at a concrete modeling strategy. I began in section 2.1 by situating tau and amyloid within three decades of Alzheimer's disease neuropathology research, moving from prion-like propagation to the classical Braak/Thal staging systems and their PET-era adaptations. I made explicit the structural constraints that have accompanied those traditions. I then turned in section 2.2 to the history of spatial modeling itself: from early reaction–diffusion and nucleation–growth ideas through connectome-based network diffusion and epidemic-spreading models, and finally to a GIS-inspired, data-first approach that treats pathology as a spatial field to be predicted with interpretable covariates and calibrated uncertainty. I closed in section 2.3 by showing that recent technical progress, whole-slide digital pathology and quantitative image analysis, a mature common coordinate framework (MNI 2009b) with rich atlases, and standardized post-mortem protocols in the Adult Changes in Thought (ACT) study, render this approach practically achievable.

Taken together, these literatures crystallize the central problem I aim to solve. Routine autopsies directly observe only a tiny fraction of the brain and summarize it with region-level stages, leaving most tissue unmeasured and most spatial structure unmodeled. Mechanistic spread models, while generative and insightful, live at parcel scale and rely on strong assumptions about transport that are difficult to verify against cellular-scale histology. They typically do not yield voxel-based mapping that can support rigorous comparisons. Histology, imaging, and atlas context also rarely co-reside in a single coordinate system, which frustrates interpretability, comparability, and the use of biologically meaningful covariates. PET staging and multi-site histology add method variance that further complicates cross-participant analyses. The scientific opportunity, therefore, is to preserve histology's ground-truth while using a common imaging space and anatomy-aware statistics to infer the unobserved brain.

Chapter 3. Quantitative Neuropathology to MNI voxel coordinates

3.1 Introduction

A central aim of the NeuroPathPredict (NPP) pipeline is to bridge quantitative digital histopathology with biologically meaningful imaging phenotypes. In Alzheimer's disease research, standard neuro-histopathological workflows[75] often yield spatially limited tissue sampling, restricting the interpretability and scalability of quantitative histopathology data. To address this limitation and extend the utility of such data, one promising strategy is to spatially co-locate quantitative neuropathological (QNP) measures with whole-brain neuroimaging frameworks.

To that end, I developed **QNPtoVox**, a preprocessing pipeline that aligns voxel-level tau burden data, quantified as percent area stained for tau pathology using Halo software, with the MNI152 standard brain volume. This transformation enables harmonized, voxel-based analyses across participants, facilitating spatial modeling, inter-participant comparison, and integration with other neuroimaging modalities. Importantly, QNPtoVox serves as a foundational component of NPP but is also broadly extensible to other histological markers, datasets, and imaging workflows.

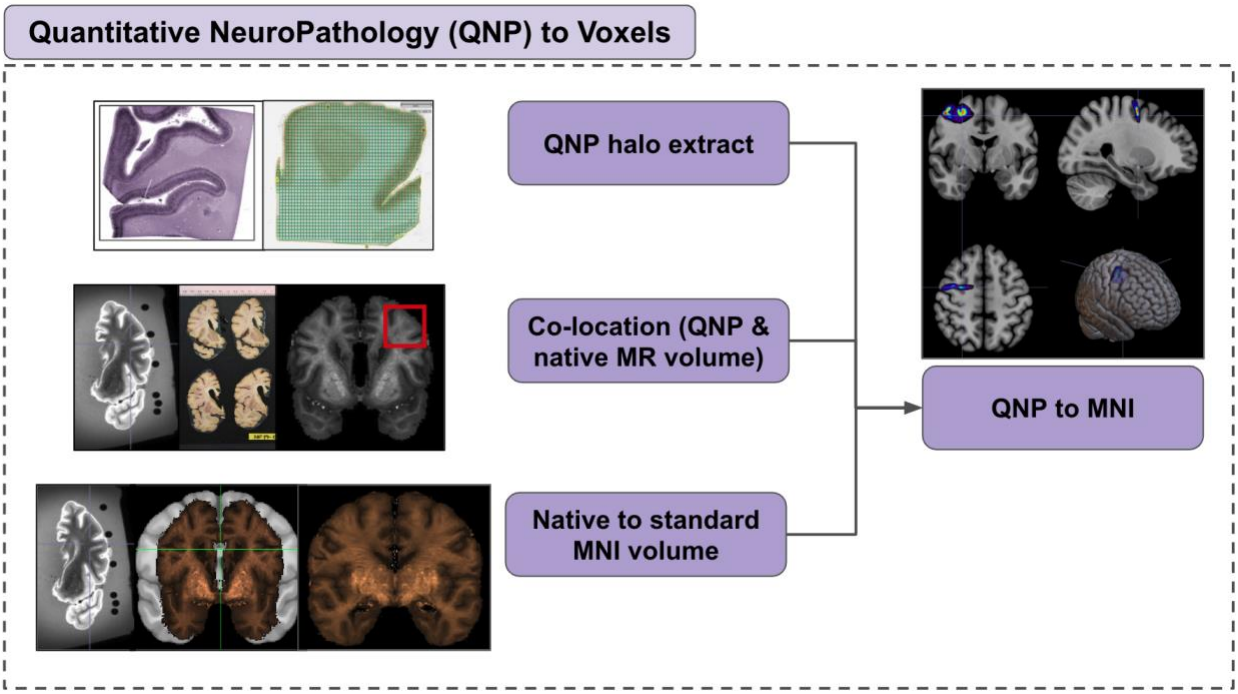


Figure 3.1.1: *Overview of the QNPtoVox pipeline. This schematic outlines the key steps in transforming QNP data, extracted from Halo-stained histological slides, into voxel representations*

in MNI space, enabling spatial integration with ex vivo and standard neuroimaging volumes for use in the NPP pipeline.

This chapter outlines the design and implementation of the QNPtoVox pipeline (Figure 3.1.1), which operationalizes the spatial transformation of quantitative histopathology into MNI-aligned voxel coordinates. It begins with a description of the core data inputs: Halo-derived pathology, gross tissue images, and ex vivo MRI scans. The chapter then details how these diverse modalities are organized and processed through a modular pipeline framework, including orchestration logic, environment setup, and participant-specific configuration files.

Next, the chapter covers the extraction of QNP values and their associated spatial coordinates, followed by the preprocessing of ex vivo MRI to facilitate downstream alignment. The co-location module integrates these coordinate systems through a series of transformation and registration steps, enabling the mapping of histological data into a voxel space. Voxel transformation techniques, including resolution matching and kernel-based smoothing, are described in detail. Finally, the mapped volumes are co-registered to the MNI152 template space, yielding standardized outputs that support cross-participant comparisons and integration with broader neuroimaging analyses. Together, these steps establish a reproducible and extensible framework for spatially informed modeling of neuropathology.

3.2 Data input

The data used in this pipeline originates from the Adult Changes in Thought (ACT) study, which was discussed in Chapter 2. Within the ACT study, the Neuropathology Core provides high-resolution digital histopathology images from postmortem brain tissue, including quantitative measures of tau pathology, and pictures of gross brain tissue sliced coronally at 4 mm thickness. Complementing this, the Neuroimaging Core collects and processes ex vivo MRI scans from the same brains, enabling high-fidelity structural imaging. Together, these two cores offer a rich, multimodal dataset that supports the spatial integration of histological and imaging data, serving as the foundation for the QNPtoVox pipeline. Each of these data modalities is discussed in the following sub-sections.

The data used in QNPtoVox is derived from post-mortem work-up on 10 brains generously donated by ACT study participants. In the standardized workflow protocol[74], either the entire brain or one hemisphere is embedded in agarose. The brain is then scanned using 3T ex vivo MRI at 0.8mm isotropic voxel resolution to capture macroscopic structure with minimal distortion. After imaging, the brain is sectioned coronally into 4 mm-thick slabs. From these slabs, a standardized set of tissue blocks are obtained using a

punch tool according to NIA-AA (National Institute on Aging—Alzheimer's Association) standards[75]. Tissue slabs are photographed both before and after taking tissue block samples. These blocks are paraffin-embedded, sectioned, and stained using both routine and immunohistochemical protocols, including AT8 for phosphorylated tau. Digitized histological slides are analyzed quantitatively using software such as Halo[16] (Indica Labs).

For this dissertation, the group of ten participants was selected based on predefined inclusion criteria developed in consultation with the ACT study's Neuropathology and Neuroimaging Core teams. Participants were required to have Braak stage IV to VI pathology, ensuring moderate to severe tau deposition. The anatomical region of interest was limited to the Middle Frontal Gyrus (MFG), and all tissue samples were stained with the AT8 antibody specific for phosphorylated tau. Additionally, only cases with both ex vivo MRI scans and digitized histological slides from the same hemisphere were included to enable precise spatial alignment.

3.2.1 Halo extract

Once the tissue blocks were sectioned, stained, and scanned, they were processed for quantitative analysis using Halo software. While whole-slide quantification is typically performed using standard Halo modules, a customized workflow was developed in collaboration with the ACT Neuropathology Core to meet the specific spatial modeling requirements of the NPP pipeline. In this workflow, a 0.5 mm × 0.5 mm grid was overlaid onto each tissue section (Figure 3.2.1), approximating the 0.5 mm isometric voxel resolution of the MNI 2009 brain template. These grid tiles were arranged in relation to one another across the tissue section but did not conform to a standardized 3D orientation that would allow direct mapping to MNI coordinates (Figure 3.2.2).

To preserve relative spatial structure, the X and Y coordinates of each tile's vertices were extracted in the 2D plane of the histological image. Each tile was treated as a closed polygon, with five vertices recorded in microns, the first and last representing the same point to denote closure. An example of this polygon-based representation is shown in Figure 3.2.3B. Concurrently, quantitative tau pathology was assessed using Halo's Area Quantification module, which calculated the percent area stained with AT8 antibody for each tile (Figure 3.2.3A). This dual extraction of spatial geometry and staining intensity enables downstream voxel alignment and statistical modeling.

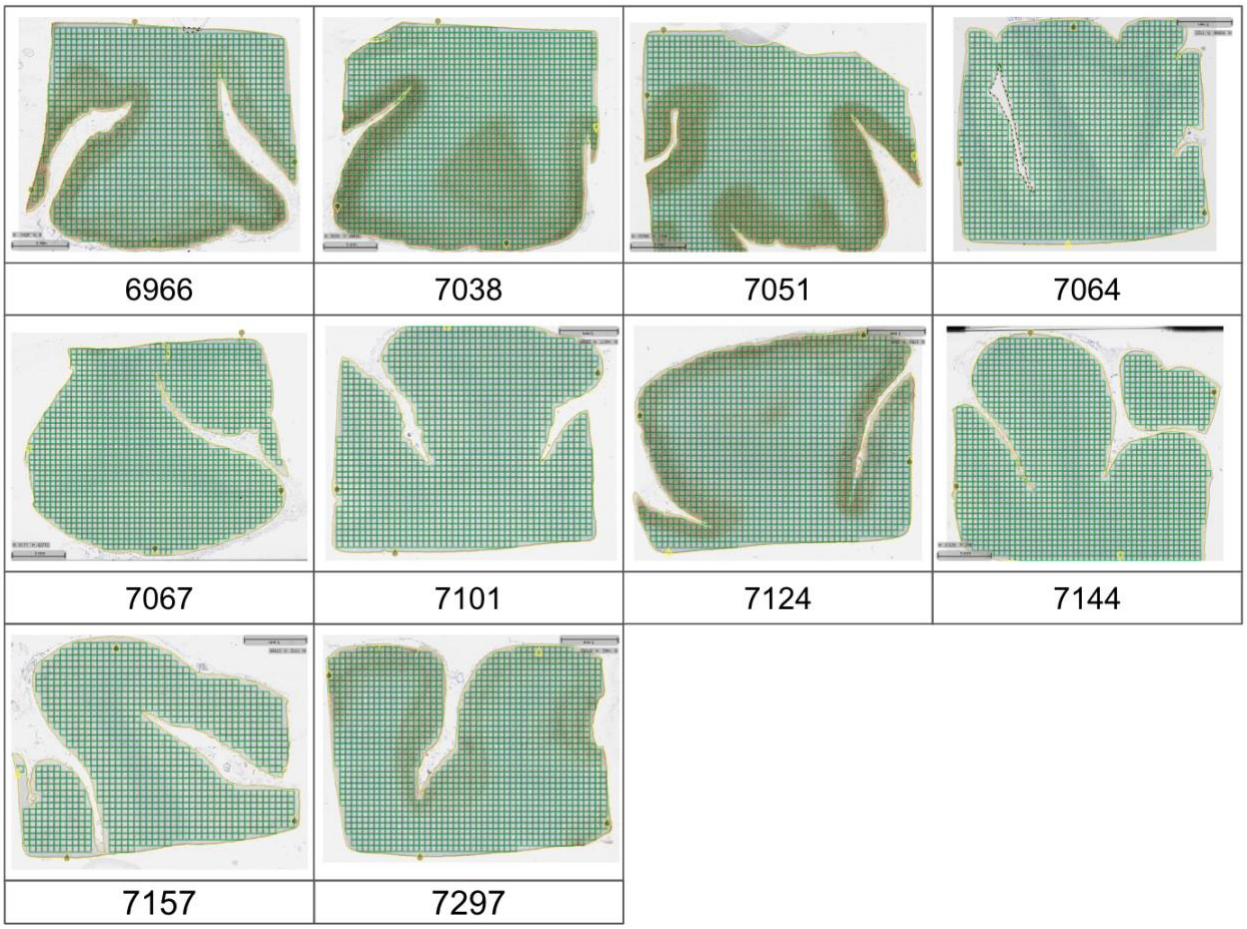


Figure 3.2.1: *Grid overlay of 0.5 × 0.5 mm tiles across Middle Frontal Gyrus (MFG) tissue sections from all 10 participants. The grid facilitates quantitative analysis of tau pathology and spatial alignment with voxel-based imaging data. Participant IDs are shown below each section.*

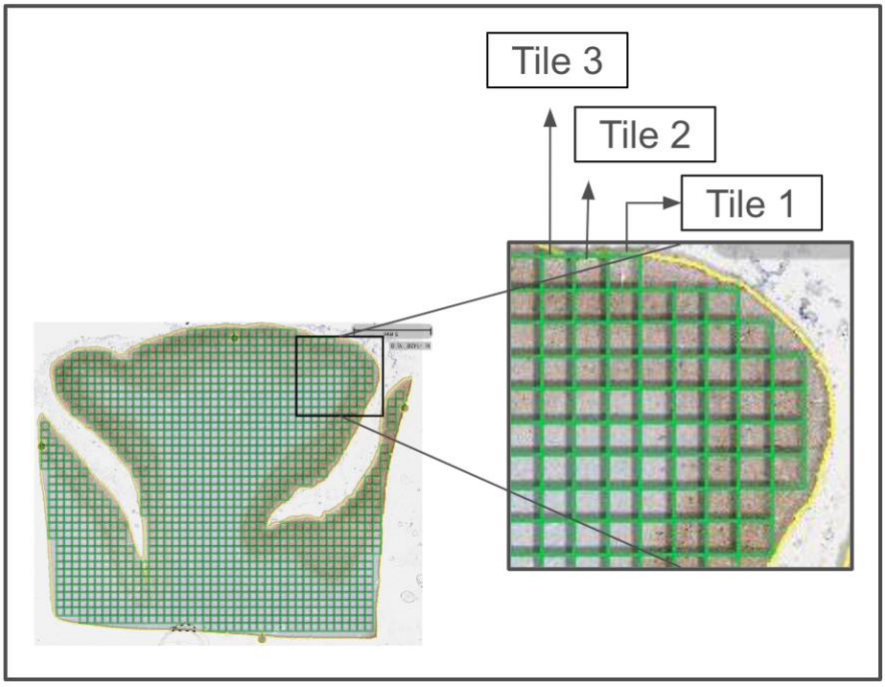


Figure 3.2.2: *Overlay of 0.5 × 0.5 mm analysis grid on a digitized histological section of the MFG for one participant. The magnified inset highlights the relative spatial arrangement and indexing of individual tiles used for coordinate extraction and quantitative analysis.*

1	Image Tag	Analysis Region	% AT8 Positive Tissue
2855	7038-A2-AT8.svs	Tile 1	3.22263
2856	7038-A2-AT8.svs	Tile 2	0.968064
2857	7038-A2-AT8.svs	Tile 3	1.91409
2858	7038-A2-AT8.svs	Tile 4	1.01487
2859	7038-A2-AT8.svs	Tile 5	1.050945
2860	7038-A2-AT8.svs	Tile 6	1.270973
2861	7038-A2-AT8.svs	Tile 7	0.861112
2862	7038-A2-AT8.svs	Tile 8	0.694421
2863	7038-A2-AT8.svs	Tile 9	0.753835
2864	7038-A2-AT8.svs	Tile 10	0.726456

```
<Annotation Name="Tile 11" Visible
  <Regions>
    <Region Type="Polygon" HasEnd
      <Vertices>
        <V X="23659" Y="1816" />
        <V X="23659" Y="2806" />
        <V X="24649" Y="2806" />
        <V X="24649" Y="1816" />
        <V X="23659" Y="1816" />
<Annotation Name="Tile 12" Visibl
  <Regions>
    <Region Type="Polygon" HasEnd
      <Vertices>
        <V X="24649" Y="1816" />
        <V X="24649" Y="2806" />
        <V X="25639" Y="2806" />
        <V X="25639" Y="1816" />
        <V X="24649" Y="1816" />
```

***Fi*gure 3.2.3:** *Example of grid tile-based quantification workflow. The left panel shows percent AT8-positive staining per tile as extracted from Halo's area quantification module. The right panel displays the corresponding XML annotation data, where each tile is defined by five polygon vertices in micron-scale X-Y coordinates.*

3.2.2 Gross tissue images

In accordance with the standardized protocol described by Latimer et al. (2022), whole brains are immersion-fixed in 10% neutral buffered formalin (NBF) for a minimum of two weeks to ensure adequate preservation of macroscopic and microscopic structures. Following fixation, brains are embedded in agarose gel to stabilize the tissue for uniform sectioning. The agarose-embedded brain is then coronally sectioned into 4 mm-thick slabs using a custom slicing apparatus designed to preserve anatomical fidelity across specimens. Each slab is carefully laid out on a black background and photographed using a high-resolution imaging setup with calibrated metric rulers placed alongside for scale and orientation.

For this dissertation, gross images were obtained from the ACT Neuropathology Core for each of the 10 selected participants. These gross tissue images (Figure 3.2.4) provide critical anatomical context throughout the QNPtoVox pipeline. While they are not used for direct image registration, they serve as an important visual reference for validating the anatomical location and orientation of histological tissue blocks. They also support the detection of tissue deformation or sampling inconsistencies that may arise during downstream processing. By bridging the macroscopic tissue geometry seen in ex vivo MRI with the localized sampling of histological analysis, gross tissue photographs enable greater interpretability and quality control in spatial modeling and voxel-based reconstruction.

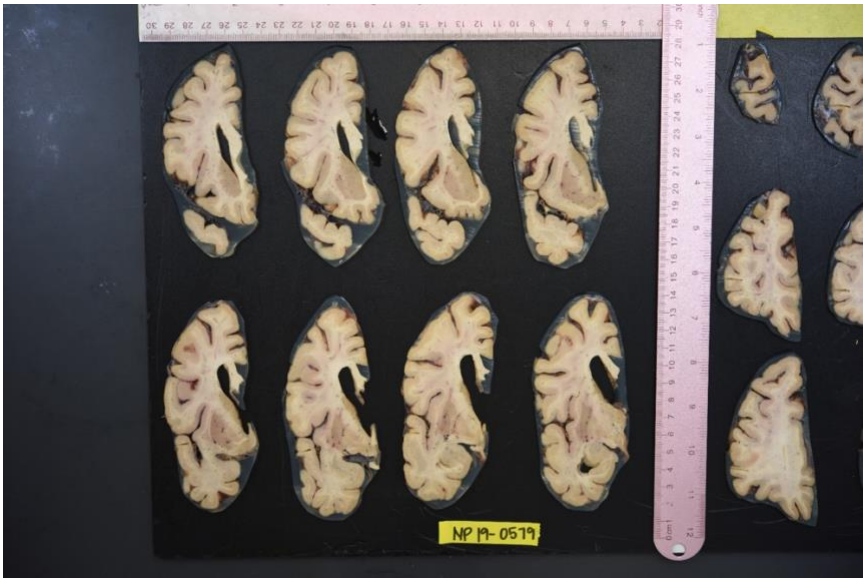


Figure 3.2.4: *Gross tissue image of fixed coronal brain slabs used for QNPtoVox processing. These 4 mm sections, photographed post-fixation and agarose embedding, provide anatomical reference for validating histological sampling and spatial alignment within the NPP pipeline.*

3.2.3 Ex vivo MRI

High-resolution postmortem MRI plays a central role in the QNPtoVox pipeline by providing a volumetric anatomical reference that bridges histological data with standard neuroimaging frameworks. Ex vivo MRI scans are acquired using protocols[74] specifically adapted for postmortem tissue, typically yielding 800 µm isotropic resolution. These scans provide exceptional structural contrast without the motion artifacts seen in vivo, enabling precise delineation of cortical and subcortical boundaries. For the NPP pipeline, these MRI volumes serve two key functions: (1) they act as a spatial scaffold onto which histological data can be co-located, and (2) they facilitate registration to a standard neuroimaging template (MNI152) to allow voxel-based modeling and cross-participant comparisons.

For each participant, ex vivo MRI was performed on a single fixed hemisphere, as illustrated in the left panel of Figure 3.2.5. These high-resolution FLAIR volumes served as the anatomical basis for downstream spatial alignment. To enable bilateral processing and compatibility with standard neuroimaging tools, the single-hemisphere image was mirrored across the midsagittal plane to generate a synthetic whole-brain volume. The mirrored volume, shown in the right panel of Figure 3.2.5, preserves anatomical symmetry and allows for complete downstream whole-brain workflows. This synthetic bilateral brain forms the structural backbone for spatial registration and voxel-based modeling within the QNPtoVox framework.

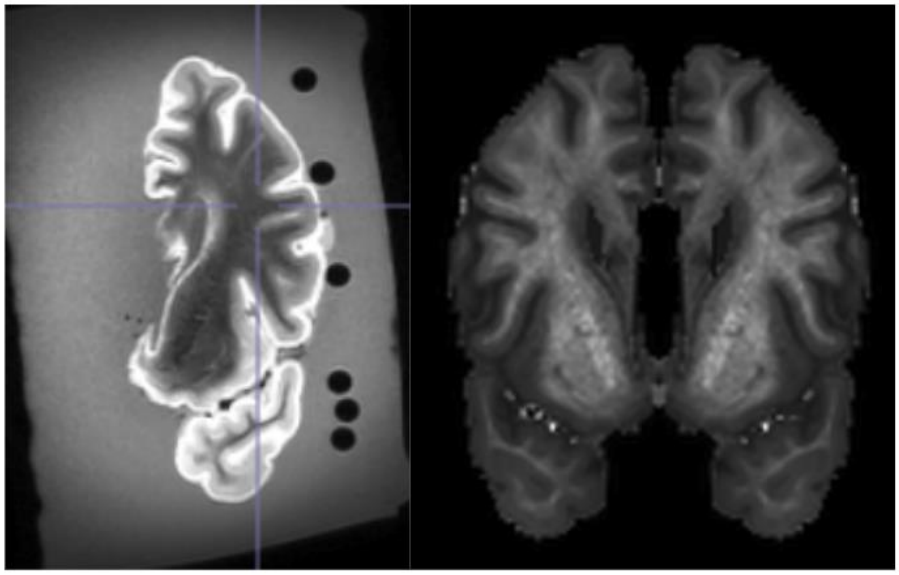


Figure 3.2.5: *The left panel shows the acquired postmortem FLAIR image of a single hemisphere embedded in agarose. The right panel displays the mirrored synthetic whole-brain volume generated originally for FreeSurfer-based surface reconstruction and spatial normalization.*

3.3 QNPtoVox pipeline overview

The QNPtoVox pipeline is a modular processing framework designed to transform QNP data into voxel-aligned representations in MNI space. The pipeline integrates participant-specific histopathological annotations with postmortem MRI data through a sequence of preprocessing, alignment, and transformation steps. It is implemented in *Python* and *R*, with support from *FreeSurfer*[76], *FSLeyes*[77], and external neuroimaging tools. The structure is fully configurable, allowing for batch processing of multiple participants with participant-specific customization.

The pipeline consists of three core processing stages. First, the input ex vivo MRI volumes (in *MGZ* format) are converted to *NIfTI*, reoriented to MNI space, and manually upsampled to 0.5 mm isotropic resolution using *FSLeyes*[77]. Second, these upsampled volumes are sliced at 0.5 mm intervals from anterior to posterior using a custom *R* script. This produces a set of standardized slice images aligned with the native orientation of Halo histopathology tiles. Third, spatial coordinates and percent-stained area values are extracted from Halo annotations and mapped onto these brain slices. The output of this stage is a set of voxel-level representations of pathology, organized for integration with volumetric imaging data.

Each step of the pipeline is driven by a simple configuration system using `.txt` files, which specify participants, file paths, transformation offsets, and processing parameters. Logs are generated for each run, and intermediate outputs are organized in participant-specific directories. The pipeline can be executed in whole or stepwise mode, allowing for iterative development and manual intervention (e.g., upsampling or block alignment) where needed. The subsequent subsections describe the orchestration logic, environment setup, and configuration structure in detail.

3.3.1 Orchestrating the pipeline

I orchestrated the QNPtoVox pipeline using a command-line interface built around the *run_qnp_pipeline.py* script, which serves as the main controller for executing one or more processing stages across selected participants. The pipeline is designed for modular execution, allowing me to run individual steps, such as upsample, slice, or extract, either independently or as a sequence. Each stage follows a consistent input-output convention and is implemented within *pipeline_steps.py*, while supporting functions for logging, validation, and file handling are defined in *pipeline_utils.py*.

The pipeline directory is structured as follows:

```
QNPtoVox/
├── config/
  ├── pipeline_config.txt          # Main configuration file
  ├── participant_list.txt                 # List of participants to process
  └── manual_coordinates.txt       # List of manually identified X,Y, Z coordinates
├── scripts/
  ├── run_qnp_pipeline.py          # Main pipeline script
  ├── pipeline_utils.py            # Utility functions
  ├── pipeline_steps.py            # Step implementations
  └── virtualmeatslicerNative.R    # R script for brain slicing
├── Input/
  ├── exvivo_transformed/          # MGZ files for each participant
  └── Halo_extract/
    └── Annotations/               # Halo annotation files
├── output/                        # Pipeline outputs
  └── XXXX/                        # Participant-specific outputs
    ├── XXXX_upsampled/            # Upsampled NIfTI files
    ├── XXXX_slices/               # Brain slice images
    └── XXXX_coordinates/          # Extracted coordinates and AT8 values
├── logs/                          # Pipeline logs
├── requirements.txt               # Python dependencies
├── setup.sh                       # Setup the environment
├── manual_upsampling_instructions.md        # Setup the environment
├── manual_alignment_instructions.md         # Setup the environment
└── README.md                      # Readme file
```

To initiate the pipeline, I used a lightweight configuration file (*pipeline_config.txt*) that defines all critical runtime parameters, including directory paths, processing options, and participant-specific settings. The pipeline is designed to be highly controllable through command-line flags, allowing fine-grained execution tailored to both development and batch-processing needs. For instance, the user can run only the upsampling step for specific participants using:

```
python3 scripts/run_qnp_pipeline.py --participants 6966 7038 --steps upsample
```

Alternatively, running the full pipeline for all configured participants requires a simple call:

```
python3 scripts/run_qnp_pipeline.py
```

The script also includes a '--help' flag that provides instructions on how to use the pipeline:

```
python3 run_qnp_pipeline.py --help
usage: run_qnp_pipeline.py [-h] [--info] [--validate-only]
                           [--steps {upsample,slice,extract,transform,kernel,mni} ]
                           [--participants PARTICIPANTS [PARTICIPANTS ...]] [--force] [--dry-run] [--verbose]
                           [--config CONFIG]
options:
  -h, --help            show this help message and exit
  --info                Show pipeline status and exit
  --validate-only       Only validate inputs and exit
  --steps               {upsample,slice,extract,transform,kernel,mni}]
                        Pipeline steps to run (default: all steps)
  --participants PARTICIPANTS [PARTICIPANTS ...]
                        Participant IDs to process (default: all participants from config)
  --force               Overwrite existing output files
  --dry-run             Show what would be done without executing
  --verbose, -v         Enable verbose logging
  --config CONFIG       Path to configuration file

Usage Examples:
    # Run full pipeline
    python3 scripts/run_qnp_pipeline.py

    # Run specific steps
    python3 scripts/run_qnp_pipeline.py --steps slice extract transform

    # Process specific participants
    python3 scripts/run_qnp_pipeline.py --participants 6966 7038 --steps slice

    # Debug mode with dry run
    python3 scripts/run_qnp_pipeline.py --steps slice --dry-run --verbose --participants 6966

    # Force overwrite existing files
    python3 scripts/run_qnp_pipeline.py --steps slice --force

    # Show status and validate inputs
    python3 scripts/run_qnp_pipeline.py --info
```

python3 scripts/run_qnp_pipeline.py --validate-only

Additional flags such as *--dry-run, --force, --validate-only*, and *--info* provide further control over pipeline execution. These options enable testing pipeline logic without executing commands, overwriting existing outputs if needed, validating inputs before running, and inspecting the current processing state of all participants.

Given that some key steps, such as upsampling/reorientation and block alignment, require manual visual inspection using GUI tools like *FSLeyes*, I developed step-by-step instructional guides (*manual_upsampling_instructions.md* and *manual_alignment_instructions.md*). When the pipeline encounters a missing intermediate file that must be generated manually (e.g., an aligned block mask), it would automatically halt execution, inform the user of the required manual step, and direct them to the corresponding guide. This integration ensures that manual interventions are well-documented, reproducible, and seamlessly embedded within the broader automated workflow.

Each participant's outputs are stored in a standardized folder hierarchy under *output/*, with dedicated subdirectories for upsampled images, brain slices, coordinates, transformation masks, and final MNI-aligned volumes. All processing events are logged to the *logs/* directory with time stamps, error messages, and stepwise status reports, facilitating debugging and full auditability of the pipeline's execution history.

3.3.2 Configuring the environment

To ensure reproducibility and compatibility across systems, I configured the QNPtoVox pipeline to rely on a minimal set of core tools and dependencies. These include *Python* (version 3.7 or higher), *R* (with specific imaging packages), and *FreeSurfe*r for format conversions. Environment setup was managed through a shell script (*setup.sh*) that automated most of the initial checks and directory creation.

This script performed several essential tasks. It made the Python script executable, verified the presence of required directories (output/, logs/), and checked for key dependencies. Specifically, it tested whether core Python libraries such as *nibabel*, *pandas*, and *numpy* were installed; if not, it prompted installation via:

pip install -r requirements.txt

It also confirmed the availability of R and the necessary R packages (*RNifti* and *png*) used in the slicing stage. These were installed using:

install.packages(c("RNifti", "png"))

FreeSurfer was required to convert *MGZ* files to *NIfTI* format during the initial preprocessing step. The setup script checked for the *mri_convert* utility and set the *FREESURFER_HOME* environment variable and sourced the appropriate setup script if it was missing:

```
export FREESURFER_HOME=/path/to/freesurfer
source $FREESURFER_HOME/SetUpFreeSurfer.sh
```

While most of the pipeline is executed through command-line scripts, two steps, upsampling and alignment, require manual intervention through the *FSLeyes GUI*. To support this, I ensured that *FSL* (and *FSLeyes*) were installed and correctly linked to the MNI 2009b template, which was used as the reference space throughout the pipeline. By validating all software dependencies early and organizing setup within a single script, I minimized compatibility issues and streamlined reproducibility across sessions and systems.

3.3.3 Configuration files

To ensure flexibility and participant-level customization, I implemented a lightweight configuration system for the QNPtoVox pipeline. All runtime parameters, paths, and participant-specific settings are stored in plain-text .txt files, following a simple key-value format. This modular structure made it easy to adjust parameters across different pipeline stages without altering the core codebase.

The main configuration file, *pipeline_config.txt*, includes general information about the pipeline, such as its name, version, and a description of its function. More importantly, it defines the relative paths to input and output directories, the list of participants to process, and the expected file formats (e.g., *MGZ* file names and Halo annotation suffixes). This file also specifies the parameters governing each processing stage. For example:

- Upsampling: output resolution (set to 0.5 mm isotropic), interpolation method.
- Slicing: slice thickness (0.5 mm), image orientation (anterior to posterior), normalization flags.
- Coordinate extraction: scaling factors, tile sampling intervals, and outlier filtering options.

- Registration: MNI template path and interpolation settings.

In addition to the main config file, I maintained a participant list file and a separate manual coordinate file (*manual_coordinates.txt*). The latter stored manually identified offsets (X, Y, Z) for mapping histological tile data to the approximate MNI position. These values were determined through visual inspection of brain slices and used during the transformation stage.

Finally, the configuration files supported several quality control parameters (e.g., minimum number of coordinates, thresholding rules) and execution settings such as log levels, parallelization, and worker limits. This level of control allowed me to adapt the pipeline to both development and batch-processing environments while preserving reproducibility.

3.4 QNP and coordinate extraction

In order to spatially map tau pathology for each participant, I first needed to extract structured coordinates and quantitative values from Halo-derived histological annotations. These annotations, stored as XML files, contain polygon definitions for each 0.5 × 0.5 mm grid tile overlaid on the stained tissue sections. Each polygon represents a spatial unit for which AT8 percent area has been computed, and these tile definitions form the foundation for voxel-based analysis in QNPtoVox. Because the original data are organized in 2D slide space and lack direct voxel alignment, this step was essential to bridge the histological and imaging domains. This step utilized *'--extract'* step flag to execute.

Figure 3.2.3 shows an example of data inputs used for this step. Briefly, they include:

- Annotation XML Files - Located in *Input/Halo_extract/Annotations/,* these files contain vector definitions of polygonal regions (tiles) and their associated labels.

- Summary CSV File - *Input/Halo_extract/Summary Analysis(in).csv* provides the % AT8-positive tissue per tile, matching tile labels in the XML annotations.

- Configuration Parameters - These include scaling factors (*coordinate_scale_factor*), sampling interval, and annotation suffixes specified in *pipeline_config.txt*.

The main processing is implemented in the *CoordinateExtractionStep* class in *pipeline_steps.py* and performs the following:

1. Parse the XML annotation file
    - Skips annotations labeled as "Layer 1" (non-informative overlays).
    - Extracts the X and Y coordinates of all vertices for each polygon (tile).

2. Extract AT8 values from the Summary CSV
    - Matches the Image Tag and Analysis Region fields to participant-specific entries.
    - Filters rows where Analysis Region starts with "Tile".
    - Stores % AT8 Positive Tissue per tile.

3. Save Raw Coordinates
    - Writes a CSV file *{participant}_tile_coord.csv* with raw polygon coordinate vertices.

4. Save Raw AT8 Data
    - Writes *{participant}_AT8.csv* with tile IDs and associated AT8 percentages.

5. Merge and Process Data
    - Combines tile centroid coordinates (converted from polygon vertices) and AT8 values.
    - Saves final merged file *{participant}_tile_proc.csv*, which is used in the transformation step.

6. Logging and Error Handling
    - Checks for missing files and records parsing errors.
    - Exits if no coordinates or AT8 values are found.

Finally, each participant's processed coordinate files are saved in the *output/{participant}/{participant}_coordinates/* directory:

- *XXXX_tile_coord.csv:* Raw XY polygon vertices (micron scale).

- *XXXX_AT8.csv*: Matched tile-wise AT8 values (% positive area).

- *XXXX_tile_proc.csv:* Combined and processed coordinate-AT8 dataset, used in the transformation step.

3.5 Preprocessing ex vivo MRI

To enable spatial alignment between histopathology-derived coordinates and brain anatomy, each participant's ex vivo MRI scan is first processed into a form that matches the resolution and orientation of the MNI 2009b volume. This preprocessing step ensures that all downstream spatial operations, such as slicing, coordinate transformation, and registration to MNI space, are grounded in a common, high-resolution anatomical reference.

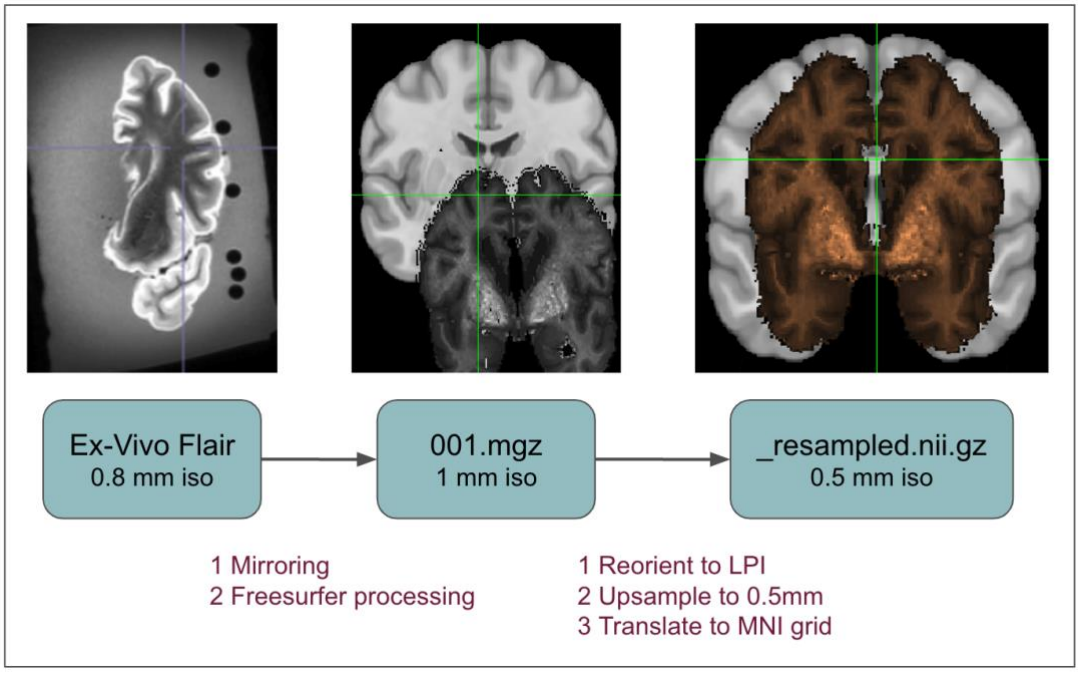


Figure 3.5.1: *Preprocessing workflow for ex vivo MRI volumes. Starting with a participant's original FLAIR scan (left), the image is mirrored and converted to FreeSurfer's .mgz format (center), followed by reorientation to LPI, upsampling to 0.5 mm isotropic resolution, and translation to the MNI grid (right). The final output serves as the anatomical reference for voxel-level integration with QNP data.*

For each participant, the input to this stage is a postmortem MRI scan of a single hemisphere in FreeSurfer's .mgz format, located in *Input/exvivo_transformed/{participant}/001.mgz*. These scans are derived from fixed, agarose-embedded hemispheres and typically offer high tissue contrast, especially in FLAIR-weighted sequences. However, they are not directly usable for spatial slicing or coordinate registration because of their native resolution, orientation, and format.

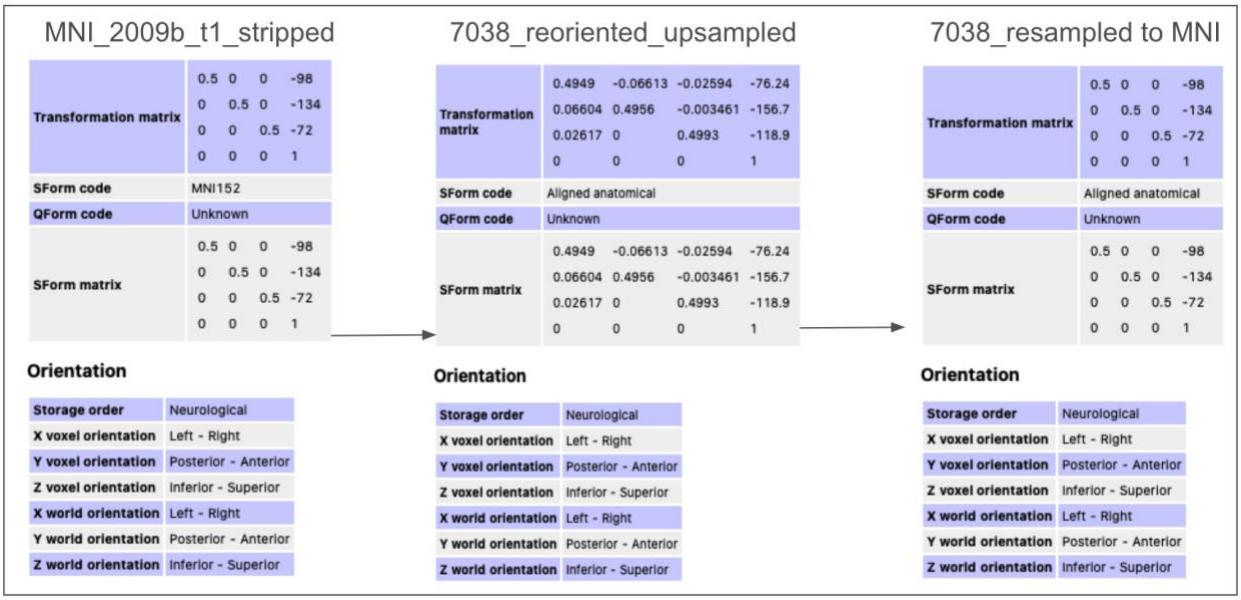


Figure 3.5.2: *Preprocessing example of ex vivo MRI for participant 7038. The left panel shows the orientation and transformation matrix of the MNI2009b template used as a reference. The middle panel displays the manually reoriented ex vivo MRI scan in anatomical space. The right panel shows the final upsampled volume aligned to MNI space, ensuring consistent voxel orientation and spatial reference across the QNPtoVox pipeline.*

To standardize these scans, I implemented a manual preprocessing step using *FSLeyes*, as described in the guide *manual_upsampling_instructions.md*. This involves three key actions:

- Format conversion: The *.mgz file* is converted to compressed *NIfTI* format using *FreeSurfer's mri_convert*[76].

- Reorientation: The *NIfTI* file is manually aligned to the *MNI 2009b template* in RAS orientation using the reorientation tools in *FSLeyes*[77].

- Upsampling: The reoriented image is resampled to 0.5 mm isotropic resolution to match the spatial scale of Halo-derived pathology tiles.

After completing these steps, I saved the processed file as *{participant}_001_up_re.nii.gz* in the *output/{participant}/{participant}_upsampled/* directory. This file serves as the anatomical base for brain slicing (Section 3.6), transformation (Section 3.7), and final MNI registration (Section 3.8).

To ensure quality and consistency, the pipeline checks for the existence of this upsampled file before allowing downstream steps to proceed. If the file is missing, the user is prompted to complete the manual steps, with instructions and filename

conventions clearly outlined in the documentation. While this manual step introduces some subjectivity, it is necessary to ensure visual alignment with MNI space and accommodate inter-participant variability in tissue geometry. The output of this step, a reoriented, 0.5 mm isotropic NIfTI volume, serves as the standardized anatomical reference for all subsequent spatial operations in QNPtoVox.

3.6 Co-location (QNP and ex vivo MRI)

Following the extraction of histological tile coordinates and the preprocessing of ex vivo MRI, the next step in the pipeline involves aligning 2D Halo-derived pathology data with the corresponding 3D anatomy of the participant's brain. This co-location step is essential for expressing quantitative neuropathology in anatomical space, enabling voxel-level modeling and downstream registration.

To begin, the upsampled MRI volume is sliced into a series of 0.5 mm coronal sections using a custom ***R script*** (*virtualmeatslicerNative.R*), called through the main script with the flag, *'--slice'*. These slices span the anterior-to-posterior extent of the hemisphere and are saved as PNG images in participant-specific directories. This slicing step serves both computational and anatomical purposes: it creates a slice-wise reference system that aligns with the resolution of Halo-derived pathology tiles and supports the manual identification of the coronal plane corresponding to the histological sampling region.

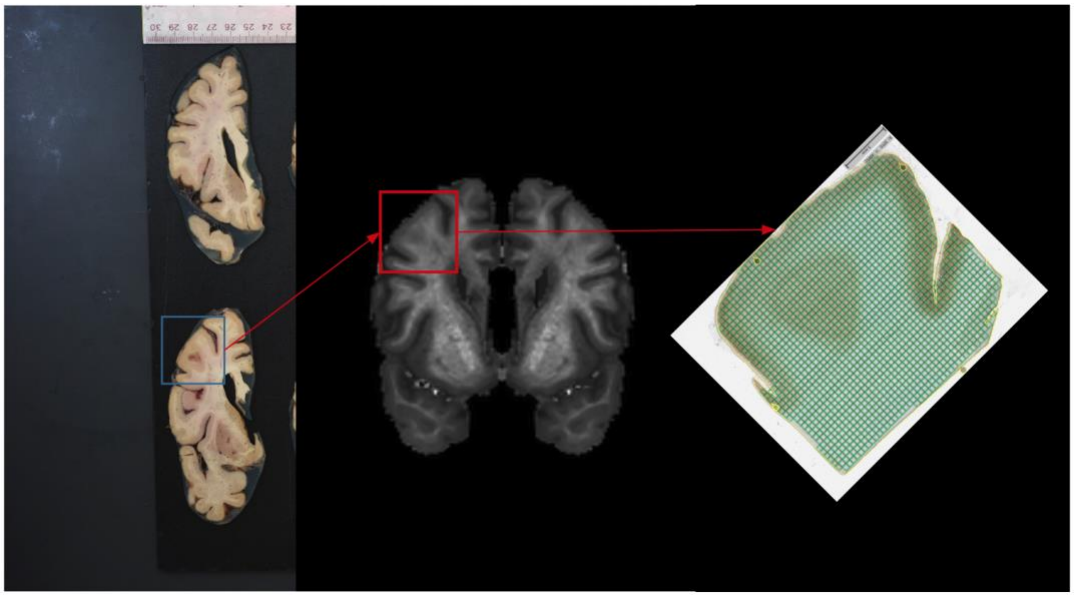


Figure 3.6.1: *Co-location of quantitative neuropathology with ex vivo MRI for participant 7038. The left panel shows the gross tissue slab used for histological sampling. The middle panel displays a coronal slice from the participant's upsampled ex vivo MRI volume. The right panel shows the corresponding Halo-annotated tissue section overlaid with a 0.5 × 0.5 mm tile grid. Red*

arrows illustrate the anatomical alignment process used to identify the Y-plane and apply coordinate offsets for accurate spatial mapping.

After generating these slices, I consulted with a neuroanatomist to determine the approximate Y-plane (coronal slice index) that best matched the gross tissue sampling location for each participant. The expert reviewed both the histological images and the MRI-derived slices, identifying the slice that corresponded most closely in terms of anatomical landmarks and tissue shape (see Figure 3.6.1). This slice index, along with estimated X and Z offsets based on anatomical alignment, was manually entered into the *manual_coordinates.txt* file for each participant.

With these participant-specific spatial offsets in hand, the pipeline proceeds to map the 2D tile coordinates (from *tile_proc.csv*) into 3D voxel space. Each tile's centroid is shifted by the provided X and Z offsets and placed at the manually identified Y-slice to localize the anatomical plane corresponding to the histological sample. The output is a set of transformed voxel coordinates and corresponding AT8 values, which preserve the anatomical distribution of pathology for each participant.

The transformed data are saved as:

- *XXXX_transformed_coords.csv:* A table of 3D voxel-space coordinates and associated AT8 intensities.

This co-location procedure effectively bridges the 2D histological grid and the 3D brain volume, anchoring QNP data within participant-specific anatomy. By incorporating expert-identified Y-slice positions and manually tuned offsets, it ensures anatomically valid placement of each data point, setting the stage for volumetric mask generation, smoothing, and eventual registration to standard space.

3.7 Voxel transformation and kernel application

After determining the 3D spatial coordinates of each histological tile through the co-location step, the next stage in the QNPtoVox pipeline involved transforming these discrete data points into a volumetric mask representing pathology in the participant's native brain space. This transformation was executed in three key phases: mask creation, manual alignment, and kernel smoothing (see figure 3.7.1).

The first step was to generate a 3D binary mask using the transformed tile coordinates stored in **transformed_coords.csv*. Each row in this file contained the X, Y, and Z voxel-space coordinates of a single tile's centroid along with its corresponding AT8 value. I defined a cubic block, 8 voxels thick (corresponding to 4mm thick coronal tissue blocks),

centered at each coordinate, and filled it with the tile's Percentage AT8 positive value. This operation effectively produced a voxel volume in which each block represented a histological tile's spatial footprint and signal strength.

The output of this operation was saved as:

1. **_QNP_AT8_mask_block.nii.gz*

However, this volume, although in voxel space, was not yet anatomically aligned with precision. Due to variability in coordinate estimation, a manual adjustment was necessary to bring the pathology mask into proper anatomical correspondence with the participant's upsampled ex vivo MRI.

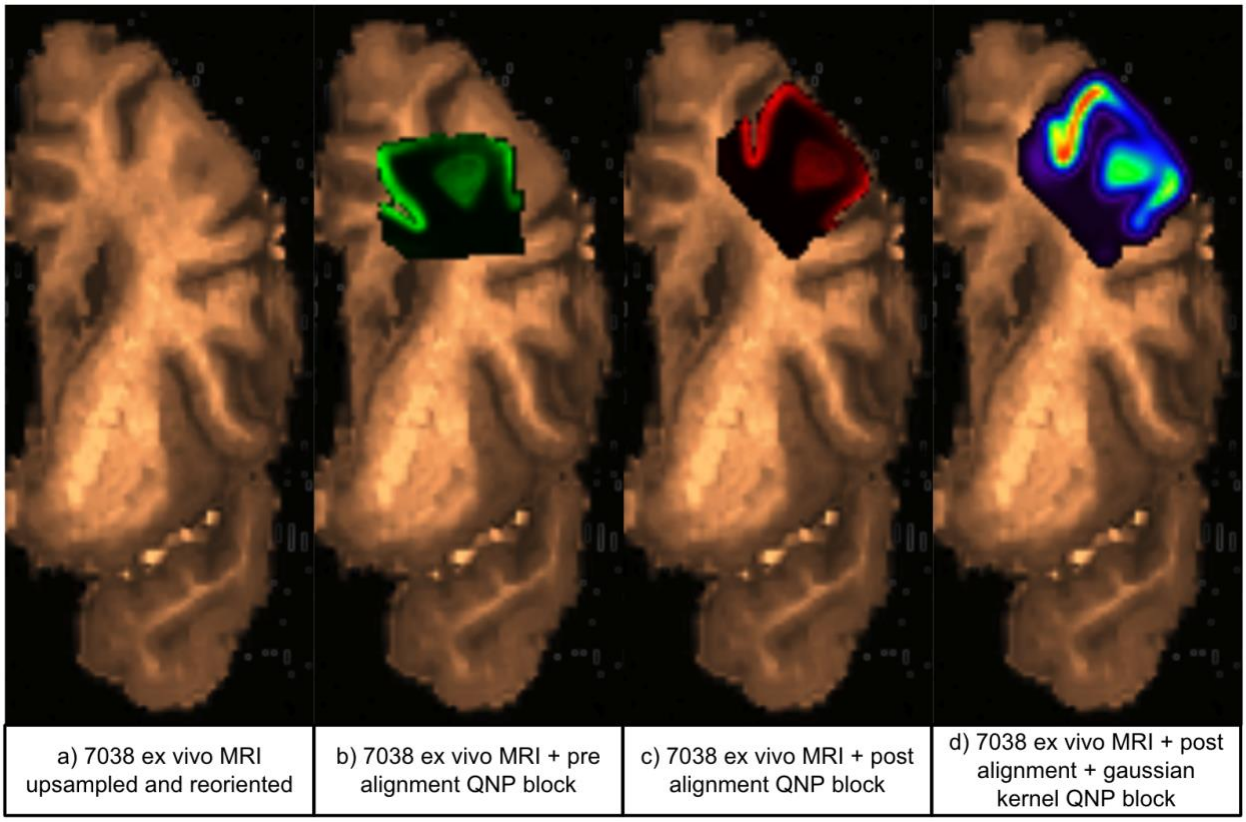


Figure 3.7.1: *Transformation of QNP block to voxel space for participant 7038. (a) Upsampled and reoriented ex vivo MRI in native space. (b) Initial placement of the QNP block based on transformed coordinates prior to alignment. (c) Manually aligned QNP block adjusted to match anatomical boundaries using FSLeyes. (d) Final smoothed pathology map after application of a 2 mm Gaussian kernel, illustrating local diffusion and anatomical conformity.*

Using *FSLeyes*, I loaded both the mask block and the corresponding upsampled brain volume *(*_001_up_re.nii.gz*) and performed manual alignment. This involved translating and rotating the mask to visually match anatomical boundaries. To guide consistency, I followed the procedure outlined in *manual_alignment_instructions.md*. Once alignment was satisfactory, the adjusted volume was saved as:

2. *_QNP_AT8_mask_block_aligned.nii.gz

The final phase involved applying a Gaussian kernel to the manually aligned mask. This step addressed three key challenges. First, there is inherent variability in the anatomical location of tissue sampling within and across participants, even within the same labeled region, such as the MFG. The smoothing operation helps account for these small shifts by distributing the pathological signal locally, making group-level comparisons more robust. Second, histopathological signals in the brain are not confined to sharply bounded cubes; they diffuse through tissue in a gradient-like fashion. Third, hard-edged binary masks are analytically limiting in neuroimaging workflows, as they introduce edge artifacts and reduce the generalizability of voxel-wise analyses.

To mitigate these issues, I applied a 3D Gaussian smoothing kernel with a 2 mm full width at half-maximum (FWHM) using *scipy.ndimage.gaussian_filter*[78]. This operation diffused the AT8 values over neighboring voxels, producing a soft, continuous map of pathology density. I then thresholded the resulting smoothed volume at 0.01 to generate a binary support mask, which was multiplied with the smoothed data to remove edge artifacts and restrict the volume to biologically meaningful signal. This ensured that only AT8 positive values of 1% or greater were modeled, thereby boosting signal intensity.

The final output was saved as:

3. *_QNP_AT8_smoothed_sig2.nii.gz

All files were stored in the participant's transformation directory *(output/{participant}/{participant}_transformation/).* This step produced a smooth, anatomically aligned, and quantitatively rich 3D map of tau pathology, ready for registration to MNI space and voxel-level statistical modeling. By integrating volumetric transformation, visual alignment, and principled smoothing, this stage provided the essential bridge between histological specificity and neuroimaging compatibility.

3.8 MNI co-registration

To enable group-level comparisons, spatial modeling, and integration with standardized neuroimaging datasets, all QNP-derived pathology volumes must ultimately be transformed into a common anatomical framework. In this pipeline, as discussed earlier, I chose the MNI ICBM 2009b nonlinear symmetric (0.5 mm)[15] template as the target space and used the Advanced Normalization Tools (ANTs)[79] suite to perform image registration. ANTs was selected for its proven accuracy in nonlinear deformable registration, particularly in postmortem and anatomically variable brains, such as those in the ACT cohort.

This transformation is critical for harmonizing pathology data across participants. Without registration to the MNI space, voxel-level comparisons would be anatomically inconsistent, and group-wise analysis would be confounded by inter-individual differences in brain shape, size, and orientation. ANTs provides the flexibility and precision needed to address these challenges while preserving the anatomical localization of smoothed pathology signals.

To support reproducibility and accessibility, I included setup support for ANTs within the pipeline. A shell script (*install_ants.sh*) located in the *scripts/ANTs/* directory offers two installation options:

- A Homebrew-based installation for macOS users with package manager access.
- An interactive installation mode that walks users through the manual setup process on any Unix-based system.

Once ANTs is installed, the registration procedure proceeds in two primary stages for each participant (see Figure 3.8.1):

Step 1: Native ex vivo MRI to MNI Registration

The first step involves registering the participant's native-space ex vivo MRI volume to the MNI template. Specifically, I used the reoriented and upsampled file (**_001_up_re.nii.gz*) as the moving image, and the 0.5 mm MNI ICBM 2009b template (*mni_icbm152_t1_nlin_sym_09b_hires_stripped.nii.gz*) as the fixed image.

This process is executed using ***antsRegistration***, configured to include:

- Rigid transformation for initial alignment of orientation and position

- Affine transformation to account for scaling and shearing
- Nonlinear SyN (Symmetric Normalization) transformation to accommodate local anatomical deformations.

The result is a set of transformation parameters:

- **_to_MNI0GenericAffine.mat*: Affine matrix
- **to_MNI1Warp.nii.gz*: Forward nonlinear warp field
- **to_MNI0Warp.nii.gz*: Initial warp estimate (usually zero field)

These files are saved in the participant's **_registration/* ***directory*** and are reused in the subsequent transformation step.

Step 2: Warping the QNP Volume to MNI Space

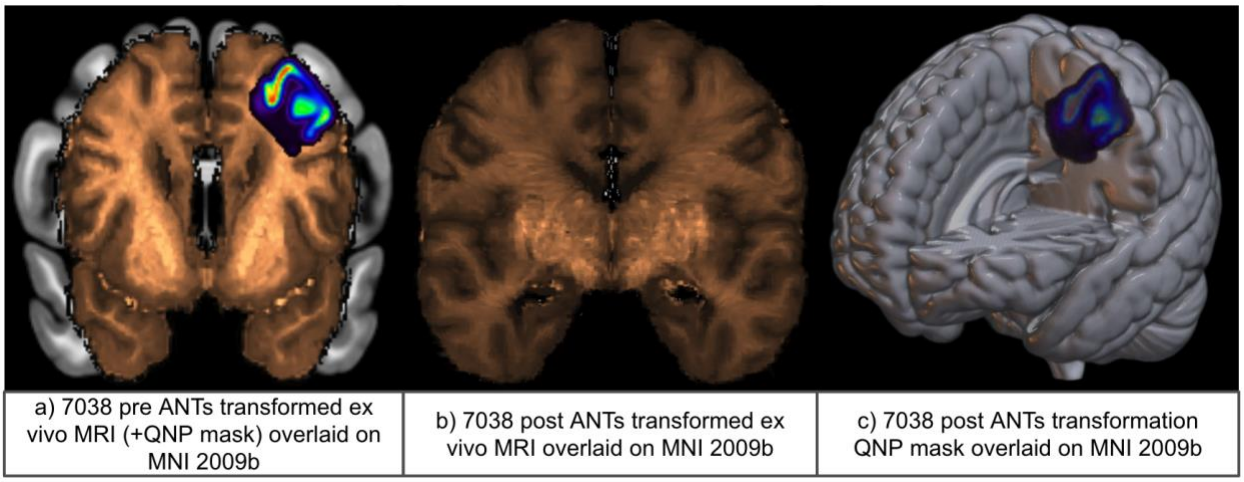
Once the registration parameters have been computed, the second step involves applying these transformations to the participant's smoothed pathology volume *(*_QNP_AT8_smoothed_sig2.nii.gz*). Using the ***antsApplyTransforms*** function, I warped the QNP volume into MNI space with the following options:

- *--reference-image*: the MNI template
- *--input*: the native space smoothed QNP mask
- *--transform:* affine + warp (in correct order)
- **--***interpolation:* linear (to preserve smoothed intensity values)

The transformed volume was saved as:

4. **_QNP_AT8_smoothed_sig2_MNI.nii.gz*

This file represents the fully standardized version of each participant's pathology map, aligned voxel-for-voxel with other participants and MNI-based neuroimaging resources. It is stored in the **_final_results/* directory.



a) 7038 pre ANTs transformed ex vivo MRI (+QNP mask) overlaid on MNI 2009b	b) 7038 post ANTs transformed ex vivo MRI overlaid on MNI 2009b	c) 7038 post ANTs transformation QNP mask overlaid on MNI 2009b

Figure 3.8.1: *Registration of participant 7038's ex vivo MRI and QNP volume to MNI space using ANTs. (a) Pre-ANTs transformation view showing the native-space ex vivo MRI overlaid with the aligned QNP mask. (b) Post-ANTs transformation of the ex vivo MRI, nonlinearly warped and aligned to the MNI2009b template. (c) Final smoothed QNP mask registered to MNI space and visualized on the MNI2009b cortical surface.*

Once the smoothed QNP volumes were successfully registered to MNI space, the final operation involved extracting voxel-level information for downstream spatial modeling. For each participant, I parsed the registered QNP volume (**_QNP_AT8_smoothed_sig2_MNI.nii.gz*) to extract non-zero voxel coordinates (in MNI space) and their corresponding AT8 values. This conversion produced a standardized voxel-wise representation of QNP across participants, each row encoding an (X, Y, Z) coordinate and a continuous-valued % positively stained for AT8 value.

These extracted coordinates served two critical functions in the next phase of the analysis. First, they were used as inputs to the I-BIS framework to generate spatially informed covariates. Second, the combination of voxel coordinates and AT8 pathology values formed the input dataset for the Universal Kriging modeling phase. The quality of kriging predictions depends on accurate spatial alignment and the richness of input covariates, both of which are supported by the MNI co-registration step.

Together, these operations complete the QNPtoVox pipeline, transforming raw histopathology into a fully spatial, voxel-wise dataset aligned with a standard neuroimaging framework. This enables scalable, participant-normalized modeling of neuropathology and forms the analytical foundation for predictive disease mapping in Alzheimer's disease.

Chapter 4. Integrated-Brain Information System (I-BIS)

4.1 Introduction

I developed the **Integrated-Brain Information System (I-BIS)** as a modular and extensible framework to bring together multimodal neuroanatomical resources into a single, high-resolution reference space. My goal was to generate spatial covariates systematically for modeling neuropathology, and within the NeuroPathPredict (NPP) pipeline, I-BIS provides the anatomical context needed to enrich voxel-level quantitative neuropathology (QNP) data with descriptors from multiple curated brain atlases.

Integrating multiple atlases in a common volumetric space is not new, but existing tools often limit flexibility. Many come with fixed atlas sets, predefined outputs, or are tailored to narrow applications. For example, Lead-DBS[80] offers advanced registration and visualization capabilities but focuses on deep brain stimulation planning, not large-scale, custom covariate generation. I designed I-BIS to be atlas-agnostic and highly configurable, allowing me to swap, update, or expand the atlas library with minimal effort while maintaining strict spatial consistency and reproducibility.

I drew inspiration from geographic information systems (GIS), where various geospatial layers, including topography, infrastructure, and environmental maps, are integrated within a common coordinate framework for multi-factor spatial modeling. I approached brain mapping in the same way. Each atlas in I-BIS is a “spatial layer” that I preprocess to match the resolution, dimensions, and orientation of the MNI ICBM 2009b[81] template. This ensures that every voxel in my analysis is enriched with precise anatomical and functional context.

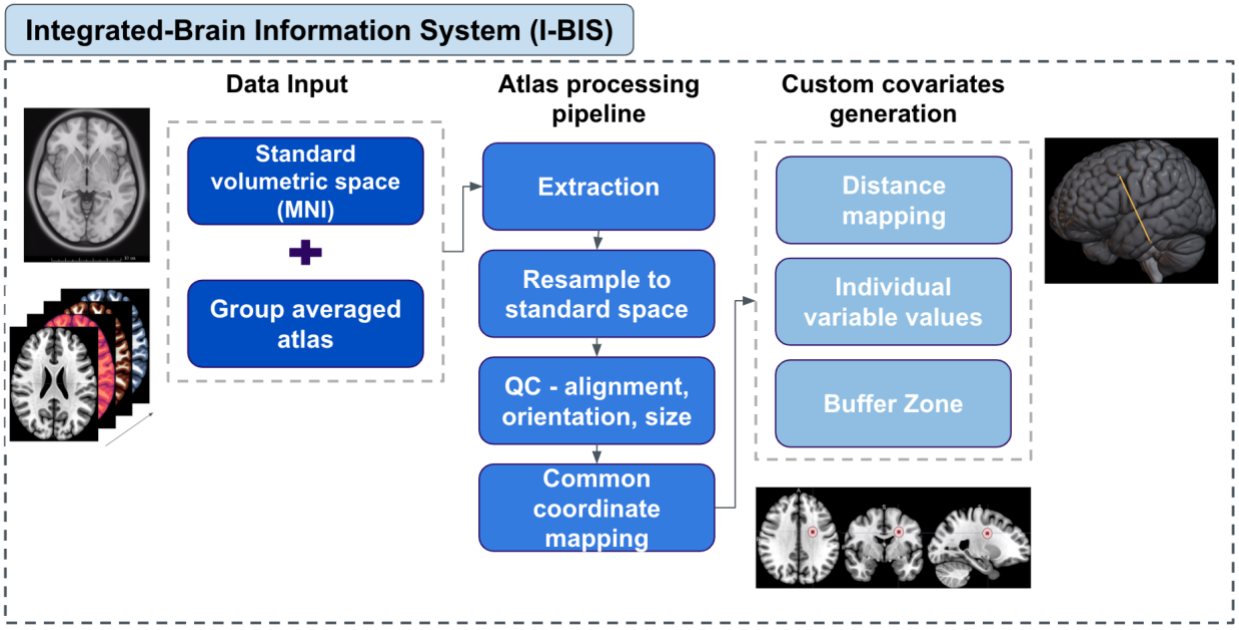


Figure 4.1.1: *Overview of the I-BIS pipeline. This schematic outlines the key steps of data input, pre-processing atlases, and generating custom covariates.*

I structured I-BIS into three main stages (Figure 4.1.1):

1. **Data Input** – I select a standard volumetric space (MNI ICBM 2009b in this work) and load selected group-averaged atlases that capture structural, functional, or microstructural properties.

2. **Atlas Processing Pipeline** – I extract the relevant atlas maps, upsample them to the target resolution, check alignment, orientation, and size, and bring them into a common coordinate framework.

3. **Custom Covariate Generation** – I produce analytical variables in three main categories:

   a. **Distance mapping** – computing the Euclidean distance from each voxel to specific anatomical features or network boundaries.

   b. **Individual variable values** – assigning atlas-defined properties (e.g., tract probability, functional network membership) directly to voxels.

   c. **Buffer zones** – defining probabilistic or fixed-radius neighborhoods around target regions.

By converting static atlas maps into structured covariates, I make anatomical and functional information directly usable in predictive models. In this dissertation, I pair these covariates with QNPtoVox (Chapter 3) outputs to build predictor sets for spatial models such as universal kriging. Beyond this work, I-BIS is ready to scale and adapt for other neuroimaging and neurodegenerative disease mapping projects where anatomical context is critical.

4.2 Input Data

4.2.1 Standard volumetric space selection

In neuroimaging, there is no single universal reference brain space. Instead, several widely recognized volumetric templates have been developed, each with its own anatomical basis, resolution, and intended applications. Among the most influential are the Talairach space[82], the Montreal Neurological Institute (MNI) space[15], and the Human Brain Project (HBP) BigBrain template[83]. In the subsections that follow, I describe each of these three reference spaces in detail, outlining their origins, technical specifications, advantages, and limitations. My goal in this comparison is to evaluate their suitability for integrating multiple brain atlases and generating covariates within I-BIS, and to select the one that offers the best balance between anatomical fidelity, atlas compatibility, and practical usability.

Talairach Space

The Talairach coordinate system, introduced in 1988, was an early milestone in stereotaxic neuroanatomy. It is based on a single post-mortem brain, scaled, and partitioned to fit a fixed anterior-posterior aligned grid. While it provided the first widely used 3D reference for neurosurgery and functional localization, its reliance on a single brain introduces anatomical bias. The original brain's proportions, such as its unusually small cerebellum, do not match population averages, leading to spatial mismatches of up to 5–10 mm in peripheral regions when compared to modern MRI templates. From a neuropathology mapping perspective, Talairach space is problematic. Most modern atlases are no longer developed in Talairach coordinates, and converting between Talairach and other spaces requires additional transforms that can propagate error. For I-BIS, where I aim to integrate multiple high-resolution, population-based atlases, these limitations made Talairach unsuitable.

MNI (ICBM) Standard Space

The MNI framework improves on Talairach by averaging MRI scans from many individuals, thereby capturing a more representative anatomy. Over time, the MNI

templates have evolved from lower-resolution linear averages to the high-resolution, nonlinear MNI ICBM 2009 series. The version I selected MNI ICBM 2009b, is a nonlinear average of 152 brains at 0.5 mm isotropic resolution, with multiple MRI contrasts (T1, T2, PD). In addition, tissue probability maps for gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) are available in the 2009a and 2009c versions at 1mm isotropic resolution (Figure 4.2.1). This resolution is twice as fine as the older 1 mm MNI templates, making it possible to map small subcortical structures and narrow sulci with precision. Because MNI space is population-based, it minimizes single-participant bias and achieves residual cross-participant misalignments of only a few millimeters after nonlinear registration, well within acceptable limits for voxel-based modeling. Crucially, the majority of modern structural, functional, and connectivity atlases are defined directly in MNI coordinates or can be transformed to them with minimal preprocessing. This widespread compatibility greatly simplifies integration. For I-BIS, this means I can combine diverse atlases, cytoarchitectonic, vascular, white matter tract, and functional network maps, without bespoke registration pipelines for each.

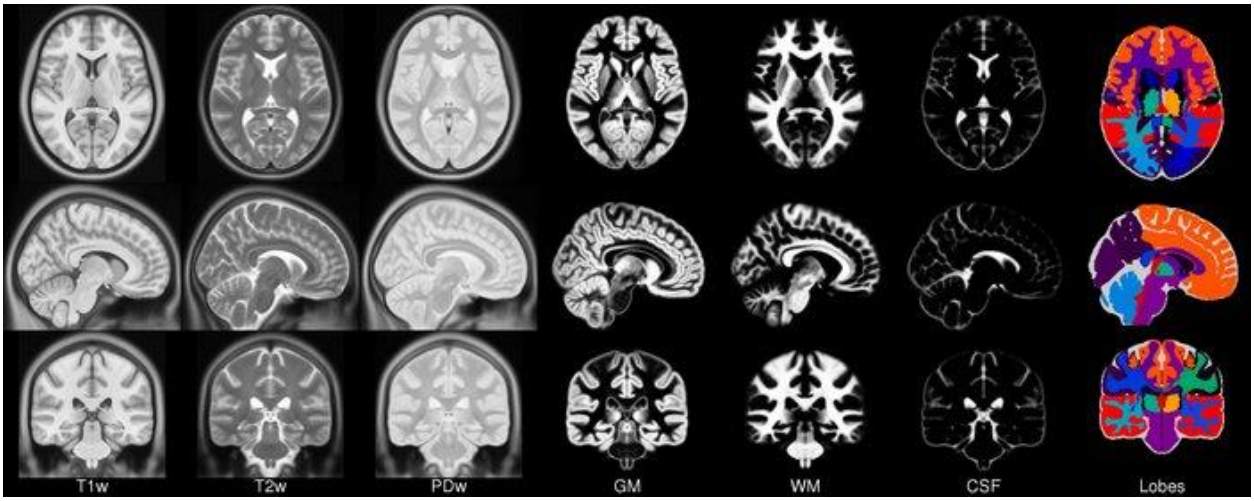


Figure 4.2.1: *The MNI ICBM 2009c nonlinear symmetric template[15] displayed across multiple contrasts and tissue segmentations. Columns show: T1-weighted (T1w), T2-weighted (T2w), and proton density-weighted (PDw) volumes; segmented gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) probability maps; and a lobar parcellation map. Axial and sagittal views are shown for each modality. The template's 0.5 mm isotropic resolution and population-averaged anatomy provide high anatomical fidelity and broad compatibility with structural, functional, and microstructural atlases used in I-BIS.*

HBP BigBrain Template

The BigBrain dataset, part of the HBP, is a histology-based 3D reconstruction of a single human brain at an extraordinary 20 µm isotropic resolution. It offers unprecedented

microstructural detail, including cortical lamination and subnuclear organization. In principle, such detail could support neuropathology mapping at a near-cellular scale. However, BigBrain shares some of Talairach's core limitations: it is based on a single donor brain, which means it does not reflect inter-individual variability, and normalizing other brains to it is challenging. The dataset is also computationally enormous, making it unwieldy for large-scale voxel-based modeling. While transformations between BigBrain and MNI exist, most atlases are not natively defined in BigBrain space, which reduces their practical utility for a multi-atlas integration pipeline like I-BIS.

Why I chose MNI ICBM 2009b?

After weighing spatial accuracy, anatomical representativeness, resolution, compatibility, and computational feasibility, I selected MNI ICBM 2009b as the standard volumetric space for I-BIS. This template is a nonlinear average of 152 individual brains at 0.5 mm isotropic resolution, offering double the linear resolution of the earlier 1 mm MNI templates. The higher resolution enables finer anatomical delineation while remaining computationally manageable for large-scale voxel-based modeling. A key advantage of the 2009b template is its rich multimodal information; it includes T1-weighted (T1w), T2-weighted (T2w), and proton density-weighted (PDw) volumes at 0.5 mm, along with tissue probability maps for gray matter, white matter, and cerebrospinal fluid. These modalities are directly useful for preprocessing steps such as segmentation, tissue classification, and structural boundary identification.

From a workflow perspective, using MNI ICBM 2009b means I can normalize participant data with well-validated pipelines in SPM, FSL, ANTs, and other standard neuroimaging tools, because this template is simply a high-fidelity evolution of the widely adopted MNI152 space. It is also backward compatible with atlases defined on earlier MNI versions, minimizing the need for additional transformations. Atlas compatibility was a decisive factor. Many modern parcellations are defined directly in the 2009b space, including the Allen Human Brain Atlas 3D parcellation, which was drawn on the 2009b nonlinear symmetric template. Specialized subcortical atlases, high-resolution connectivity maps from the Human Connectome Project (HCP), and functional network definitions are all available in MNI space and can be applied to the 2009b template with minimal adjustment. Compared to Talairach, where virtually no new atlases are developed, MNI 2009b offers vastly broader coverage, supporting integration of cytoarchitectonic maps, connectivity-based parcellations, and vascular territories in a single framework.

The computational manageability of MNI 2009b was also important. At ~10 million voxels, it is orders of magnitude lighter than the HBP BigBrain dataset, which contains billions of

voxels and requires specialized processing resources. While BigBrain's extreme 20 μm resolution is unmatched, most MRI data (and most pathology features visible on MRI) are on the order of 1 mm. For my purposes, 0.5 mm is more than sufficient to capture relevant features without introducing unnecessary complexity. In essence, MNI ICBM 2009b provides the highest-fidelity macroscopic representation of the human brain that remains fully compatible with mainstream neuroimaging workflows. By anchoring I-BIS to this template, I ensure that all atlas-derived covariates align seamlessly with QNPtoVox outputs, enabling accurate, reproducible, and scalable spatial modeling.

4.2.2 Atlas Selection

Brain atlases provide standardized spatial frameworks that allow me to define anatomical and functional boundaries consistently across individuals and studies. Their development has paralleled advances in neuroimaging and computational methods. Early atlases were largely histological, constructed from postmortem brain slices to delineate structures at the microscopic level. While these maps offered exquisite cellular detail, they had limited generalizability because they were based on single brains and could not account for normal inter-individual variation.

The advent of MRI-based atlases transformed the field by enabling scalable, reproducible mapping of brain anatomy in vivo. Templates such as the MNI ICBM 2009b Nonlinear Symmetric standard space have become widely adopted reference frameworks, allowing diverse datasets to be co-registered into a common coordinate system. Over time, atlases have expanded beyond purely anatomical definitions to include functional parcellations, structural connectivity maps, cytoarchitectonic boundaries, vascular territories, neurotransmitter distributions, and task activation patterns. This multimodal evolution has greatly increased their utility for both basic and clinical neuroscience research[84].

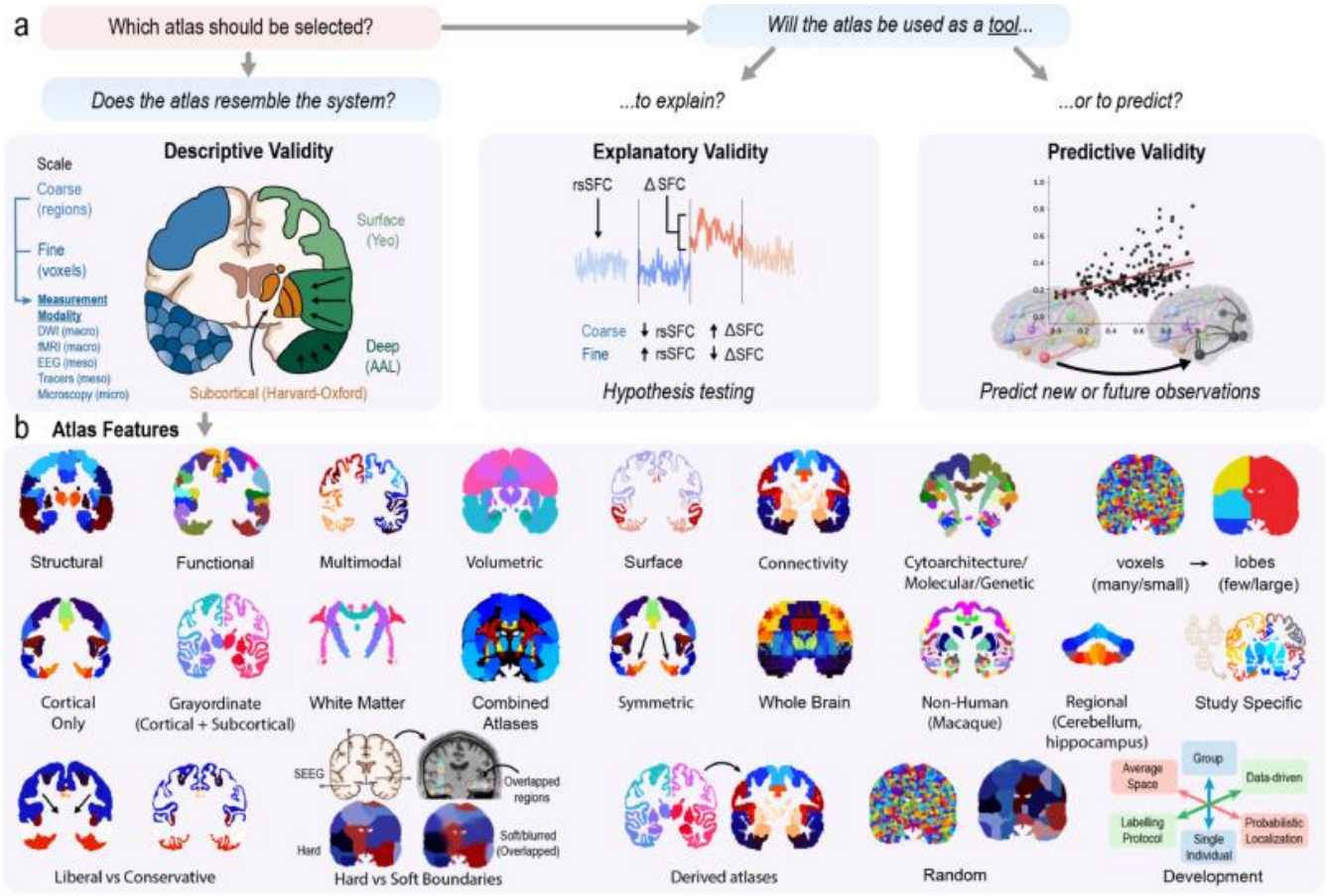


Figure 4.2.2: *Adapted from Revell et al. (2022). Framework for selecting brain atlases based on validity dimensions (descriptive, explanatory, predictive) and atlas features (structural, functional, multimodal, connectivity, cytoarchitectonic, boundary type, symmetry, coverage, source population). This framework guided the selection of atlases for I-BIS to ensure diverse coverage of anatomical and functional features relevant to spatial covariate generation.*

Atlas selection should be informed by both purpose and features. Revell et al. (2022) propose three broad validity dimensions that can guide selection (Figure 4.2.2):

- **Descriptive validity** – how well the atlas represents the system of interest (e.g., anatomical fidelity, parcellation granularity).

- **Explanatory validity** – suitability for testing hypotheses about relationships between regions or networks (e.g., structure–function correspondence).

- **Predictive validity** – utility for forecasting new or future observations (e.g., disease progression modeling).

Beneath these validity considerations lies a diverse set of atlas features, also illustrated in Figure 4.2.2. These include differences in:

- **Domain** (structural, functional, multimodal, connectivity-based, cytoarchitectonic).
- **Representation** (volumetric vs. surface-based, symmetric vs. asymmetric, whole-brain vs. regional).
- **Boundary characteristics** (hard vs. soft, probabilistic vs. deterministic).
- **Population source** (human vs. non-human, developmental vs. adult).
- **Coverage scale** (lobar, voxels, fine-grained subfields).

Along with considerations listed above, one of my core objectives in developing I-BIS is to integrate a curated set of atlases into a high-resolution, MNI-registered coordinate framework that collectively spans relevant validity types and features. In an ideal scenario, I-BIS would include representative atlases from all major categories, structural segmentation, functional networks, connectivity, vascular territories, and fine-grained subfields. In practice, I selected atlases based on four explicit criteria:

1. **Anatomical significance** – relevance to neuropathology modeling in Alzheimer's disease and related disorders.
2. **Open-source availability** – ensuring reproducibility and accessibility for other researchers.
3. **Prior co-registration to MNI space** – minimizing transformation error and preserving spatial fidelity.
4. **Base voxel resolution** – matching or exceeding the 0.5–1 mm range suitable for integration into MNI ICBM 2009b.

Table 4.2.1 below lists the atlases included in I-BIS. Each row corresponds to a distinct *type* of atlas, such as tissue segmentation, white matter tractography, vascular territories, macrostructural parcellations, cortical functional networks, cerebellar functional networks, and fine-grained subfield segmentation. For each type, I selected one atlas that best met my criteria after comparing it to alternative atlases in the same category. In the subsections that follow, I will detail the rationale for each selection, describing the

alternatives considered, their strengths and weaknesses, and why the final choice was most suitable for integration into I-BIS.

Atlas type (Category)	Selected atlas	Anatomical variable of Interest	Base coordinate system	Base voxel resolution
Tissue Segmentation	MNI152 2009c[15]	Grey matter, white matter, CSF	MNI152_2009c	1 mm
White matter tractography	HCP1065[70]	87 White matter tracts	MNI152_2009a	1 mm
Vascular maps	Mouches & Forkert[69]	Arterial Map	MNI152_2009b	0.5 mm
Macrostructural parcellation	CerebrA[85]	102 Regional parcellation	MNI152_2009c	1 mm
Cortical functional networks	Yeo17[68]	17 Cortical functional networks	MNI152_2009c	1 mm
Cerebellar functional networks	Buckner17[86]	17 Cerebellum functional networks	MNI152_2009c	1 mm
Fine grained HC subfields	EugenioHC[87]	14 Hippocampal subfields	MNI152_2009b	0.5 mm

Table 4.2.1: *List of atlases included in I-BIS, their corresponding anatomical variables of interest, base coordinate system, and base voxel resolution.*

Tissue Segmentation (MNI 2009c)[15]

Tissue segmentation maps are a foundational component of I-BIS because they define the spatial distribution of grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF) throughout the brain. These features are directly relevant to NPP modeling: they allow me to assess whether a voxel's tissue type influences its vulnerability to pathology, quantify distances to GM–WM boundaries (which may mark anatomical transition zones for disease spread), and apply tissue-specific density functions during spatial modeling. Without an accurate tissue segmentation layer, many structural covariates used in NPP would be either absent or prone to registration error.

I compared two main candidates:

1. MNI ICBM 152 (2009 series) tissue probability maps
2. BigBrain Project histology-derived segmentations

As discussed in Section 4.2.1, while BigBrain offers exceptional microstructural detail, it suffers from single-participant bias, limited direct compatibility with other MNI-space

atlases, and high computational cost. For tissue segmentation, these drawbacks outweigh the benefits of its extreme resolution, particularly since my modeling operates at the macroscopic scale (0.5–1 mm).

In addition, although my primary I-BIS reference space is MNI ICBM 2009b Nonlinear Symmetric, I selected MNI ICBM 2009c for the tissue segmentation layer. The distinction between these two templates lies in symmetry and available derivatives:

- 2009b – nonlinear average of 152 participants, symmetric template (left–right mirrored), available at 0.5 mm resolution.

- 2009c – nonlinear average of 152 participants, asymmetric template preserving actual population-level asymmetries, with high-quality probabilistic GM, WM, and CSF maps provided at 1 mm resolution.

For tissue segmentation, 2009c includes validated probabilistic tissue maps (see Figure 4.2.1) widely used in neuroimaging workflows, ensuring strong compatibility with standard preprocessing pipelines, hence the choice.

White Matter Tractography (HCP1065[70])

White matter tractography (WMT) atlases are critical to I-BIS because they represent the brain's structural connectivity pathways, which can serve as potential routes for trans-synaptic or axonal spread of neuropathology. For the WMT layer in I-BIS, my aim was to include a high-quality atlas that provides comprehensive coverage of major association, commissural, and projection fiber pathways in MNI space. This layer supports covariates describing a voxel's probability of belonging to a specific tract, distance to tract, or proximity to tract boundaries, features that are relevant for modeling pathology spread along axonal projections.

I compared two main candidates:

1. HCP1065 tractography atlas[70] – generated from diffusion MRI data from 1,065 participants in the Human Connectome Project.

2. JHU ICBM-DTI-81 white matter labels and tracts[88] – derived from DTI data of 81 participants in the ICBM dataset.

The JHU ICBM-DTI-81 atlas has been a standard in neuroimaging for over a decade, offering manually labeled major white matter tracts and probability maps. However, it has some limitations relative to more recent resources:

- Lower spatial resolution – original voxel size is 1 mm, with less anatomical precision at tract boundaries.

- Fewer tracts – covers a limited set of large, well-defined pathways.

- Older acquisition and modeling – based on single-shell DTI and an earlier probabilistic tractography pipeline, which can underrepresent crossing fibers.

By contrast, the HCP1065 atlas benefits from:

- Large sample size – diffusion MRI from over 1,000 healthy adults, increasing population representativeness.

- Higher quality acquisition – multi-shell, high b-value diffusion data with submillimeter resolution, improving modeling of crossing and branching fibers.

- Greater tract coverage – includes 87 named tracts spanning association, commissural, and projection systems.

- Improved anatomical fidelity – derived using advanced tractography algorithms that produce sharper tract boundaries and more complete coverage of smaller pathways.

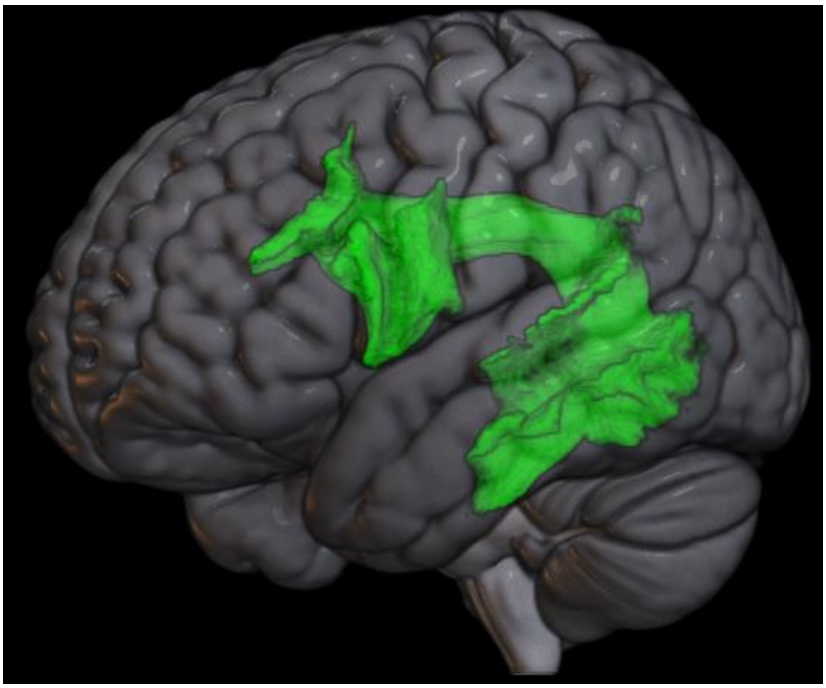


Figure 4.2.3: *Representative white matter tract, AF_L (Arcuate Fasciculus_L) derived from HCP1065 and placed in the MNI 2009b brain.*

The HCP1065 atlas meets all four of my selection criteria. It is anatomically significant for neurodegeneration modeling because it offers detailed mapping of both short-range and long-range pathways potentially involved in pathology propagation. It is open source, natively registered to MNI space (2009a), and provides tracts at 1 mm resolution, which is computationally manageable while still anatomically detailed.

Vascular Maps (Forkert & Mouches)[69]

Vascular atlases capture the spatial distribution of the brain's blood supply and drainage systems, providing a reference for understanding how perfusion patterns may influence the localization and spread of neuropathology. In I-BIS, vascular territory information supports covariates such as arterial territory membership, distance to vascular boundaries, and location within watershed zones. These covariates are particularly relevant to NPP, as both hypoperfusion and vascular architecture can interact with disease processes like Alzheimer's pathology or co-occurring cerebrovascular disease.

Ideally, both arterial and venous atlases would be incorporated into I-BIS to comprehensively capture vascular influences. Arterial maps define supply territories, while venous maps define drainage pathways; both could contribute complementary covariates. However, for this proof-of-concept implementation, I chose to focus on the arterial system to reduce complexity and prioritize a vascular layer that is directly relevant to many well-established neuropathological mechanisms (e.g., selective vulnerability in arterial watershed zones, ischemia-related pathology patterns).

Candidates considered:

1. Forkert & Mouches (2019) – Statistical cerebral artery atlas
2. Huck et al. (2019) – VENAT venous vasculature atlas

Forkert & Mouches (2019) created a statistical atlas of cerebral arteries using time-of-flight MR angiography (TOF-MRA) from 544 healthy participants. They automatically segmented the cerebral arteries in each participant, measured vessel radii, and then nonlinearly registered these maps to MNI space. The aggregated data produced volumetric images encoding the probability of an artery at each voxel, along with the mean and variability of vessel radii across the population. The atlas, released in 2019, has 0.5 mm isotropic resolution in MNI ICBM 2009b space. It covers the large and medium arteries of the brain, including ACA, MCA, and PCA branches, providing probabilistic occurrence maps that highlight regions of higher vessel density. This detailed probabilistic approach allows not only for the identification of supply territories but also for quantifying uncertainty and inter-individual variability in vascular anatomy.

Huck et al. (2019) developed the VENAT atlas, a complementary venous vasculature map, using 7 Tesla MRI quantitative susceptibility mapping (QSM) from 20 young adults. They segmented cortical and deep veins, measured average vessel diameter (~0.84 mm) and curvature, and combined these into a high-resolution 0.6 mm isotropic average venous anatomy atlas in stereotaxic space. VENAT allows visualization of the venous network, including the superior sagittal sinus and major cortical veins, and provides a normative reference for venous structure location and variability.

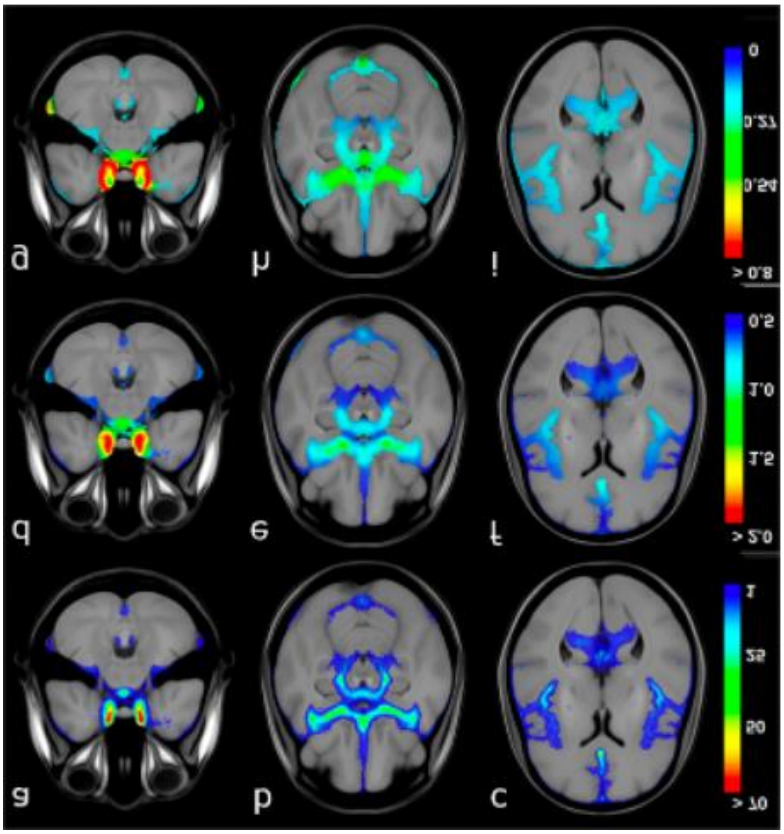


Figure 4.2.4: *Forkert & Mouches (2019)[69] probabilistic cerebral artery atlas. Color intensity reflects the occurrence probability of major arteries, including ACA, MCA, and PCA branches. The atlas also encodes mean vessel radius, providing a high-resolution reference for defining arterial territories.*

Given my modeling goals in NPP, where arterial territories are a well-established explanatory factor for spatial pathology patterns, the Forkert & Mouches arterial atlas offers a more directly relevant and statistically robust foundation. It captures arterial supply zones that correspond closely to known perfusion territories and potential hypoperfusion-related vulnerability patterns. While the Huck et al. VENAT atlas is valuable for venous system mapping and could complement the arterial atlas in future I-BIS iterations. The Forkert & Mouches atlas provides the most practical and informative arterial reference for this phase of the project.

Macrostructural Parcellation (CerebrA)[85]

Macrostructural parcellations divide the brain into anatomically meaningful regions based on gross morphological landmarks, such as gyri, sulci, and subcortical boundaries. Unlike tissue segmentation, which assigns each voxel a probability of belonging to a broad tissue class (GM, WM, or CSF), macrostructural parcellations operate at a higher level of organization: they delineate specific cortical and subcortical structures within the grey matter. While tissue segmentation tells us *what type* of brain tissue a voxel contains, macrostructural parcellation tells us *where in the brain's anatomical hierarchy* that voxel resides.

This distinction is critical for I-BIS and NPP. Regional identity is a powerful explanatory variable in neuropathological modeling; certain regions are more vulnerable to pathology due to differences in cytoarchitecture, connectivity, and metabolic demand. Macrostructural boundaries also provide a scaffold for summarizing or constraining voxel-based data, enabling region-based statistics, distance-to-boundary calculations, and region-specific covariates.

I considered the following candidate atlases for macrostructural parcellation:

1. CerebrA[85] – A 102-region parcellation derived from the Mindboggle-101 dataset, manually corrected to improve anatomical accuracy, and aligned to MNI ICBM 2009c space.

2. HCP-MMP1[89] – A surface-based parcellation of 180 cortical areas per hemisphere, defined using multimodal data (cortical architecture, function, connectivity, and topography), later mapped to MNI volumetric space.

3. Destrieux atlas[90] – A FreeSurfer-based cortical parcellation defining ~148 regions using sulcal and gyral anatomy, available in both surface and volumetric forms.

Between the three macrostructural parcellations considered, each presented distinct strengths and limitations. CerebrA struck an optimal balance between anatomical fidelity and computational manageability. Its volumetric boundaries, manually refined by expert neuroanatomists, corrected misalignments and label errors common in automated segmentations. The resulting 102-region framework provided enough granularity for region-level covariates without introducing unnecessary complexity. Crucially, CerebrA is already aligned to the MNI ICBM 2009c space, enabling seamless integration with other atlas layers in I-BIS.

In contrast, HCP-MMP1 delivered the highest cortical resolution, with 180 regions per hemisphere defined using a rich combination of architectural, functional, and connectivity-based features. While this fine-scale detail is valuable for specialized cortical analyses, it imposes significant computational demands in large-scale voxel-based modeling. Furthermore, its native surface-based format requires volumetric conversion, a process prone to inaccuracies in deep sulci and subcortical areas. In addition, upon manual inspection, HCP-MMP1 did not have sufficient cortical ribbon coverage as shown in Figure 4.2.5 (highlighted via the red and yellow circles).

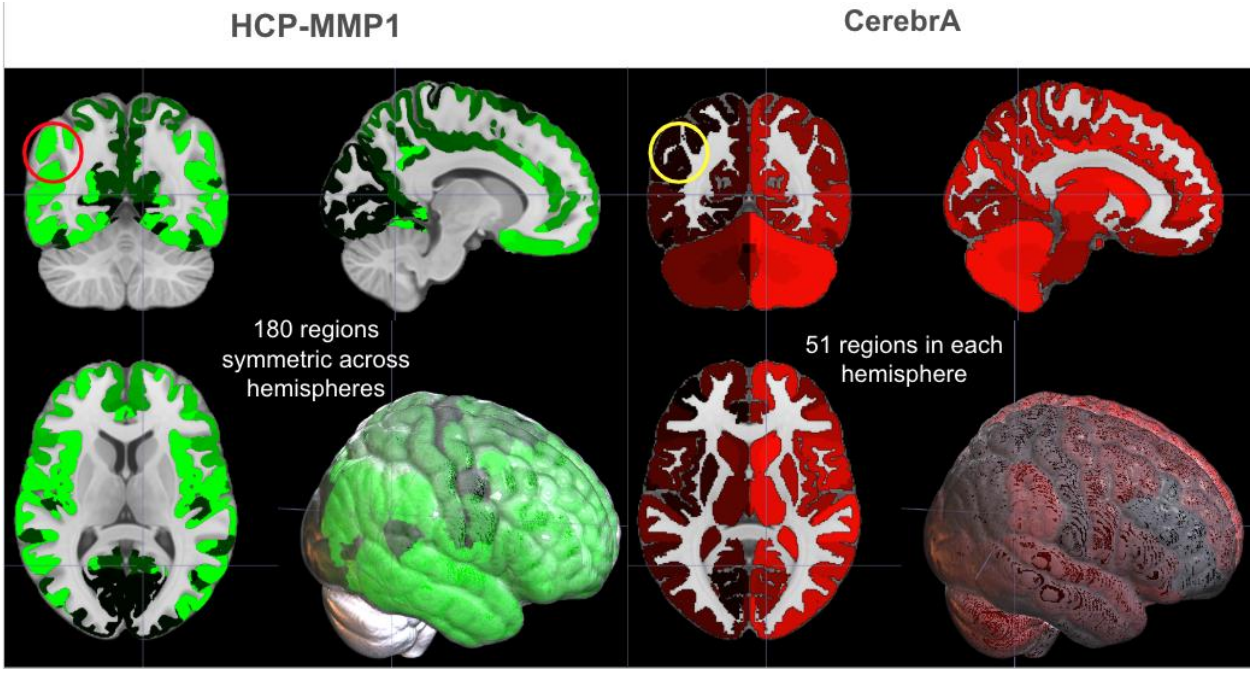


Figure 4.2.5. *Comparison of cortical coverage across macrostructural parcellations. HCP-MMP1 (left) divides the cortex into 180 regions per hemisphere, offering fine-scale parcellation with multimodal definitions. CerebrA (right) defines 51 regions per hemisphere, manually corrected for anatomical accuracy and aligned to MNI space. The figure illustrates the difference in cortical coverage and regional granularity between the two atlases.*

Destrieux occupied a middle ground in terms of granularity, with roughly 148 cortical regions based on widely recognized sulcal and gyral landmarks. However, it too is primarily surface-based, making it susceptible to similar volumetric mapping artifacts as HCP-MMP1. Moreover, it lacked the targeted volumetric refinements found in CerebrA.

For the purposes of I-BIS, I required a parcellation that was volumetric from the outset, anatomically precise, computationally manageable, and applicable to both cortical and subcortical regions within the same framework. CerebrA met all of these criteria, making it the most suitable choice for this proof-of-concept implementation, while HCP-MMP1

and Destrieux, though valuable in their own contexts, posed unnecessary integration and complexity challenges for the current scope.

Cortical Functional Networks (Yeo17)[68]

Functional network parcellations group cortical regions based on patterns of synchronized neural activity, typically derived from resting-state fMRI. Unlike macrostructural parcellations, which follow anatomical landmarks, functional network maps reflect the brain's intrinsic connectivity architecture, identifying sets of regions that co-activate during specific cognitive, sensory, or motor states.

In I-BIS, including a functional network layer enables covariates that quantify a voxel's network membership, its distance to network boundaries, or its proximity to high-connectivity hubs. These covariates allow NPP to model pathology not only as a function of anatomy or tissue class but also as a function of functional network architecture, for example, to test hypotheses about trans-synaptic propagation or preferential vulnerability of specific networks in Alzheimer's disease (e.g., the default mode network).

For the functional network, I considered the following atlases:

1. Yeo et al. (2011)[68] – Defined using resting-state fMRI data from 1,000 participants, available in two resolutions: a coarse 7-network model (broad functional divisions) and a finer 17-network model (subdivisions of the 7-network set). Provided in both "tight" and "liberal" masks in MNI152 volumetric space.

2. Dworetsky et al. (2021) [91]– Developed using a multimodal, multi-dataset approach, incorporating both resting-state and task-based connectivity. Provides refined cortical parcellations that align functional networks with structural and cytoarchitectonic boundaries.

Parcellations in Nonlinear MNI152 Volume Space

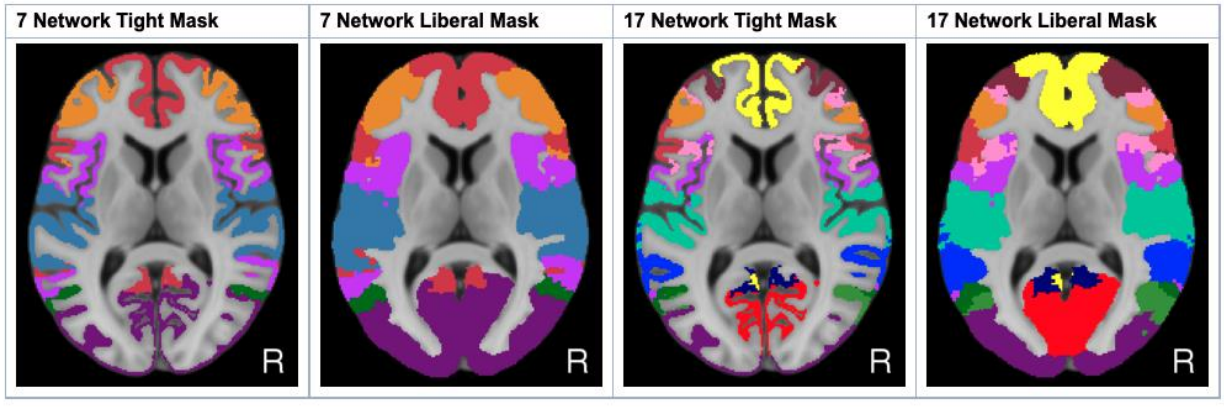


Figure 4.2.6. Yeo et al. (2011)[68] cortical functional networks. *The 7-network and 17-network parcellations are shown with both "tight" and "liberal" masks in nonlinear MNI152 volumetric space. Tight masks (left in each pair) include only high-probability voxels for each network, while liberal masks (right) extend boundaries to include lower-probability voxels. The 17-network version was selected for I-BIS due to its finer functional subdivisions and enhanced utility for network-based neuropathology modeling.*

Between the options considered, the Yeo et al. (2011) atlas offered clear advantages for I-BIS. The original Yeo parcellation is available in both a 7-network and a 17-network version. The 7-network configuration captures broad-scale divisions of cortical function, such as visual, somatomotor, default mode, limbic, dorsal attention, ventral attention, and frontoparietal control, and is well suited for coarse functional analyses. However, its broad boundaries risk obscuring functionally distinct subdivisions that may be relevant for modeling disease spread. The 17-network version retains the overall large-scale structure but further subdivides major networks into finer, functionally coherent components. This greater granularity increases the potential for detecting network-specific associations with pathology, while still maintaining interpretability and computational tractability for voxel-based modeling.

When compared with the more recent Dworetsky et al. (2021) atlas, which integrates resting-state and task-based connectivity with structural and cytoarchitectonic boundaries, the Yeo atlas remains more widely recognized and universally supported across neuroimaging pipelines. Its exceptionally large sample size of 1,000 participants ensures strong generalizability, and the availability of both "tight" and "liberal" volumetric masks in MNI space facilitates flexible inclusion criteria for voxels at network boundaries. For these reasons, I adopted the Yeo 17-network parcellation for I-BIS, as it provides the optimal balance between anatomical coverage, functional specificity, and cross-study comparability.

Cerebellar functional and HC subfield atlases: Buckner 17[86] and Eugenio HC[87]

In addition to cortical functional networks, I-BIS incorporates specialized network and subfield maps for the cerebellum and hippocampus to broaden its anatomical and functional coverage. The Buckner 17-network cerebellar atlas was derived using resting-state fMRI connectivity to cortical networks, revealing that the cerebellum contains topographically organized regions aligned with the same functional networks identified in the cerebral cortex. This atlas enables voxel-based mapping of cerebellar regions to their corresponding cortical functional networks, supporting analyses of cortico-cerebellar loops and their role in distributed brain function (Figure 4.2.7). In I-BIS, this opens the

possibility of extending network-based covariates to subcortical structures with well-established functional coupling to the neocortex.

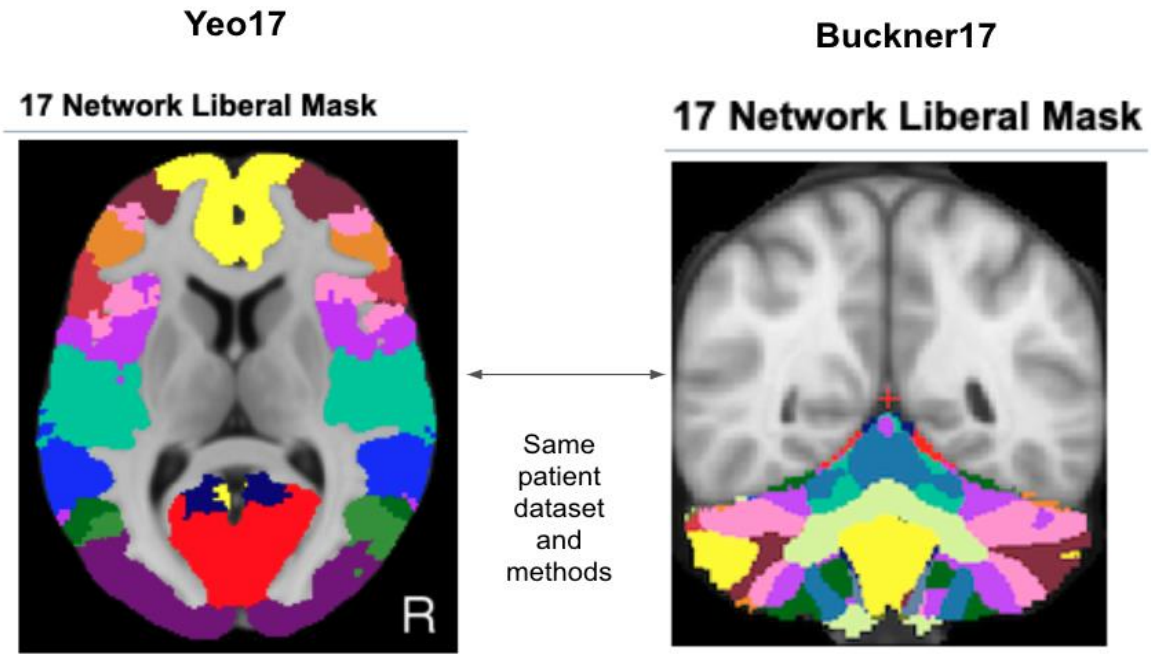


Figure 4.2.7. Yeo17 cortical functional networks and Buckner17 cerebellar networks. *The Yeo 17-network liberal mask (left) maps cortical functional organization into 17 distinct networks based on resting-state connectivity from 1,000 participants. The Buckner 17-network liberal mask (right) applies the same patient dataset and connectivity-based methodology to the cerebellum, enabling each cerebellar region to be assigned to its corresponding cortical network.*

The Eugenio HC pipeline provides a detailed parcellation of the hippocampus into 14 subfields, integrating structural MRI delineations with prior histological and cytoarchitectonic knowledge. Importantly for I-BIS integration, I applied the published Eugenio HC pipeline to the MNI ICBM 2009b template to extract these hippocampal subfields directly in the standard space used for other I-BIS layers (see Figure 4.2.8). This ensured that the subfield labels were fully co-registered with the MNI 2009b coordinate framework, enabling seamless use alongside other volumetric atlases. The resulting segmentation allows for subfield-specific covariates that capture potential differences in connectivity, cytoarchitecture, and selective vulnerability, particularly relevant for neurodegenerative diseases such as Alzheimer's, where certain hippocampal subfields are disproportionately affected.

Both the Buckner cerebellar atlas and Eugenio HC hippocampal subfield atlas were included in I-BIS because the initial plan for NPP was to model neuropathology across multiple brain regions, including the hippocampus and cerebellum, rather than focusing

exclusively on the middle frontal gyrus (MFG). In the current proof-of-concept scope, these atlases are not directly used in modeling; however, they remain strategically important. As I-BIS is extended to whole-brain modeling, these specialized maps will enable region-specific and network-level covariates beyond the neocortex, enhancing the capacity to investigate distributed pathology patterns across the brain.

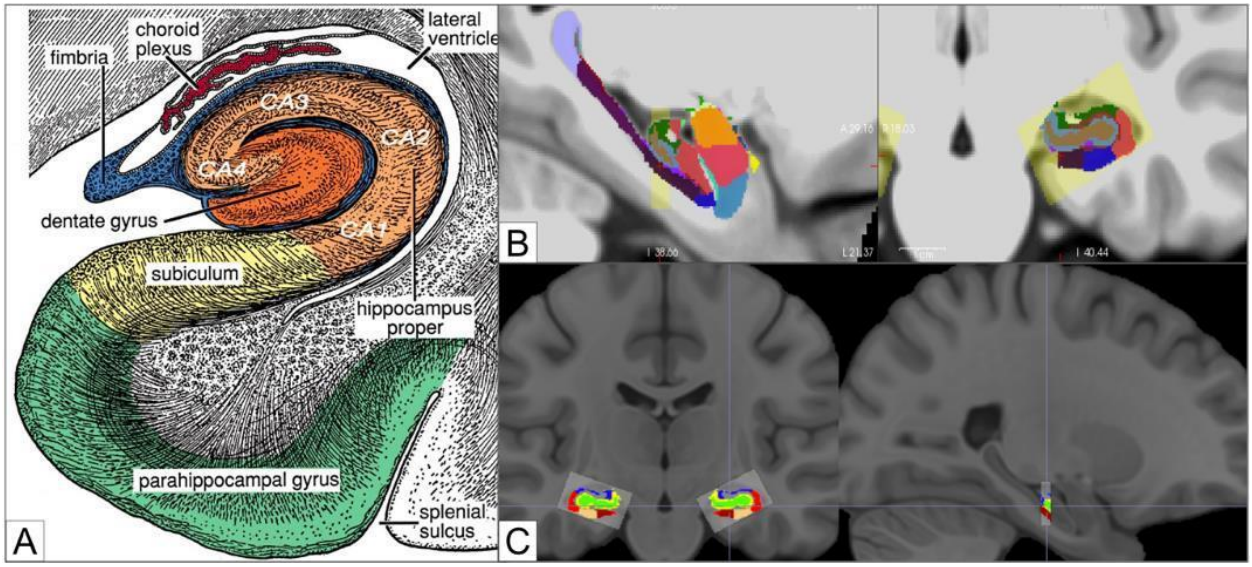


Figure 4.2.8. Eugenio HC hippocampal subfield atlas applied to MNI ICBM 2009b. *Panel A: schematic representation of hippocampal anatomy, showing subfields including CA1–CA4, dentate gyrus, subiculum, and hippocampal proper. Panels B and C: volumetric representation of the 14 hippocampal subfields generated by applying the Eugenio HC pipeline to the MNI 2009b template, ensuring direct alignment with other I-BIS atlas layers. This enables subfield-specific covariates for modeling differential vulnerability patterns in neurodegenerative diseases.*

In summary, the atlas selection for I-BIS was guided by the need to balance anatomical fidelity, functional relevance, computational manageability, and compatibility within a unified MNI standard space. Each atlas was chosen to capture a distinct and complementary aspect of brain organization, from tissue segmentation and vascular maps to macrostructural anatomy, cortical and cerebellar functional networks, and hippocampal subfields. While some atlases, such as those for the cerebellum and hippocampus, are not directly applied in the current proof-of-concept scope of NPP, their inclusion ensures that I-BIS remains extensible to whole-brain modeling in future work. This layered atlas framework provides a flexible yet standardized foundation for generating diverse covariates, enabling the modeling of neuropathology through multiple anatomical, vascular, and functional dimensions. The selections made here lay the groundwork for a scalable system that can adapt to progressively broader datasets and research questions.

4.3 I-BIS Pipeline Overview

The I-BIS pipeline operationalizes the extraction of imaging-based covariates from structural MRI and associated ROI masks into a reproducible, modular workflow. Built in *Python 3.7+* and structured for open-source distribution, the pipeline adheres to the reproducibility, transparency, and scalability requirements defined in the broader NPP framework.

At its core, I-BIS consolidates four primary processing stages: ROI voxel extraction, buffer zone analysis, variable extraction including distance-based covariates, and data consolidation, each implemented as a standalone script but orchestrated through a unified execution controller, *run_ibis_pipeline.py*. The design permits both end-to-end execution and selective, stepwise runs, allowing researchers to isolate and debug specific components without re-running the entire pipeline.

This chapter begins by outlining the core structural elements of the I-BIS pipeline: how the workflow is orchestrated through a unified controller script (Section 4.3.1), how the computing environment is configured and validated to ensure reproducibility (Section 4.3.2), and how parameterization is managed via a modular YAML configuration system (Section 4.3.3).

The subsequent sections (4.4–4.7) present each major processing stage in sequence, explaining both the technical implementation ("how") and the scientific rationale ("why"). Section 4.4 describes ROI voxel extraction, in which raw atlas volumes are parsed to isolate regions of interest, extract variables of interest (e.g., tissue class intensities, coordinate locations), and preprocess these for downstream covariate engineering. Section 4.5 covers buffer-zone analysis, where spatial neighborhoods around ROIs are defined to capture local tissue context and vascular proximity. Section 4.6 focuses on variable extraction, applying methods such as Euclidean distance transforms to generate biologically meaningful covariates. Section 4.7 concludes with data consolidation, merging outputs from all prior stages into standardized, machine-learning-ready covariate tables. Together, these steps transform raw atlas data into structured, spatially contextualized variables that can be directly integrated into the NeuroPathPredict modeling framework described in Chapter 5.

4.3.1 Orchestrating the Pipeline

The entry point for all processing is the *run_ibis_pipeline.py* script, located in the repository's root directory. This script is the "conductor" of the I-BIS workflow, sequentially

or selectively invoking the four major processing stages while ensuring all prerequisites are satisfied. The pipeline directory is structured as follows:

```
V1/IBIS/
├── README.md                          # Main documentation
├── requirements.txt                   # Python dependencies
├── run_ibis_pipeline.py               # Main pipeline execution script
├── config/                            # Configuration files
│   ├── pipeline_config.yaml           # Main pipeline configuration
│   ├── roi_config.yaml                # ROI-specific settings
│   └── buffer_zone_config.yaml        # Buffer zone parameters
├── scripts/                           # Pipeline components
│   ├── utils.py                       # Shared utility functions
│   ├── roi_extraction.py              # ROI extraction module
│   ├── buffer_zone.py                 # Buffer zone analysis module
│   ├── variable_extraction.py         # Variable extraction module
│   └── data_consolidation.py          # Data consolidation module
├── input/                             # Input data directory
│   ├── coordinates/                   # Coordinate files
│   ├── images/                        # NIfTI image files
│   ├── masks/                         # ROI mask files
│   ├── QNP_vox_coords/                # QNP coordinate files
│   ├── EDT/                           # EDT directory
│   └── Var/                           # Variable input files
└── output/                            # Output directory
    ├── roi/                           # ROI extraction results
    ├── buffer_zone/                   # Buffer zone analysis results
    ├── variables/                     # Variable extraction results
    │   ├── edt/                       # EDT extraction results
    │   └── var/                       # Variable extraction results
    ├── consolidated/                  # Consolidated datasets
    └── logs/                          # Pipeline logs
```

There are two execution modes-

- Full pipeline mode – If no step restriction is given, the script runs all four stages in the standard order: ROI voxel extraction → buffer zone analysis → variable extraction → data consolidation.

- Step-restricted mode – The --steps CLI flag allows execution of a subset of stages, e.g.:

python run_ibis_pipeline.py --config pipeline_config.yaml --steps roi_extraction buffer_zone

This is particularly useful when adjusting parameters for one step (such as buffer radius) without repeating earlier or later stages. There are several core control mechanisms in place-

- Argument parsing – The script uses Python's *argparse* to capture the --*config* path, *--steps list*, and optional overrides (e.g., *--log_level* DEBUG).

- Configuration loading – The specified YAML config is parsed, and any referenced sub-configs (e.g., *roi_config.yaml, buffer_zone_config.yaml*) are loaded and validated.

- Output directory preparation – The pipeline ensures that all expected directories (e.g., */output/roi_extraction, /output/buffer_zone, /output/variable_extraction, /output/consolidation, /logs*) exist, creating them if necessary to prevent write errors.

- Logging initialization – A timestamped log file is created under /logs. All parameter values, start/end times for each stage, and warnings/errors are recorded.

- Sequential stage execution – For each selected stage, the corresponding script (e.g., *roi_extraction.py*) is imported and executed with parameters from the configuration. Runtime exceptions are caught and logged without aborting the entire run unless explicitly configured to do so.

- Completion report – Upon finishing, the script outputs a summary of run time per stage and the location of all generated outputs.

A centralized controller script ensures consistent execution order, parameter sharing, and file-naming conventions across stages. It also supports iterative refinement, an essential requirement for research workflows where parameters are tuned based on intermediate outputs. The logging and directory management mirror the robustness approach in QNPtoVox (Ch. 3.3.1) but are adapted for I-BIS's multi-atlas, multi-ROI context.

4.3.2 Configuring the Environment

Ensuring reproducibility in neuroimaging analysis requires a controlled and well-documented software environment. I-BIS formalizes this through dependency management, validation scripts, and environment capture. All required Python packages and their minimum versions are listed in *requirements.txt*, which covers:

- Neuroimaging I/O and manipulation: nibabel, nilearn for loading, manipulating, and saving NIfTI volumes.
- Numerical computing: numpy, scipy for array operations and distance transforms.
- Data wrangling: pandas for tabular covariate storage.
- Machine learning and statistics: scikit-learn for feature scaling and transformations.
- Configuration parsing: PyYAML for reading YAML files.
- Visualization and logging: matplotlib, seaborn, tqdm for plots, progress bars, and log messages.

These requirements can be installed simply by running *pip install -r requirements.txt*.

Without version-locked dependencies, even minor library updates can alter numerical outputs (e.g., changes in interpolation defaults in Nilearn or SciPy). Explicit environment capture ensures that published analyses can be rerun exactly, a key criterion for open and reproducible science.

4.3.3 Configuration Files

I-BIS uses a YAML-based configuration system to control all aspects of execution, avoiding the need to modify source code between runs. Config files reside in the */config* directory.The following three files control the execution of I-BIS-

1.*pipeline_config.yaml*

This global file specifies:

- Input/output root directories (absolute or relative paths).
- Logging settings (level, file name patterns).

- Execution order and which atlas layers to include.
- Default parameters for each stage (e.g., whether to overwrite existing outputs).

2.*roi_config.yaml*

Defines ROI extraction-specific parameters, including:

- ROI name and expected NIfTI mask file pattern.
- Column names for coordinates and intensity in the extracted CSVs.
- Options for voxel filtering (e.g., minimum intensity thresholds).

3.*buffer_zone_config.yaml*

Controls buffer-zone creation, specifying:

- Radii to be generated (in mm).
- Overlap rules between multiple ROIs.
- Binary vs. probabilistic masking options.

Each stage of *run_ibis_pipeline.py* reads only the parameters it needs from the relevant YAML, ensuring that unrelated settings do not cause conflicts. This modular approach allows, for example, the buffer zone radius to be tuned without touching ROI extraction parameters. Centralizing parameters in human-readable files supports transparency, reproducibility, and version control. It also allows for experimental branching different YAML configs can represent different study designs without altering the underlying codebase.

4.4 ROI Voxel Extraction

In neuroimaging workflows such as I-BIS, the accuracy of downstream modeling depends on the precision of voxel-level data extracted from reference atlases. This step is foundational because it bridges the high-level atlas definitions discussed in Section 4.2 with the covariate engineering steps that follow. Specifically, ROI voxel extraction in I-BIS involves (i) isolating and saving individual atlas labels as independent volumetric masks, (ii) extracting voxel coordinates and associated intensity or classification values, and (iii) preprocessing these outputs to ensure compatibility with the spatial framework used in this research, the MNI ICBM 2009b nonlinear symmetric template at 0.5 mm isotropic resolution. From an operational perspective, Section 4.4 serves three roles within the pipeline:

1. Transforming complex multi-label atlases into analyzable components by separating each anatomical or functional label into its own NIfTI file.

2. Extracting spatially explicit quantitative data (coordinates + variable values) from each separated variable of interest for later transformation into covariates.

3. Standardizing spatial properties via upsampling and co-registration, ensuring interoperability between datasets and accurate voxel-to-voxel correspondence across atlases.

4.4.1 Order of Operations and the Upsampling Artifact

An early implementation decision concerned the order of variable extraction relative to spatial transformations. Two possible workflows existed:

- **Approach A:** Upsample and co-register the full atlas first, then extract variables.
- **Approach B:** Extract variables first, then apply upsampling and co-registration to each variable map individually.

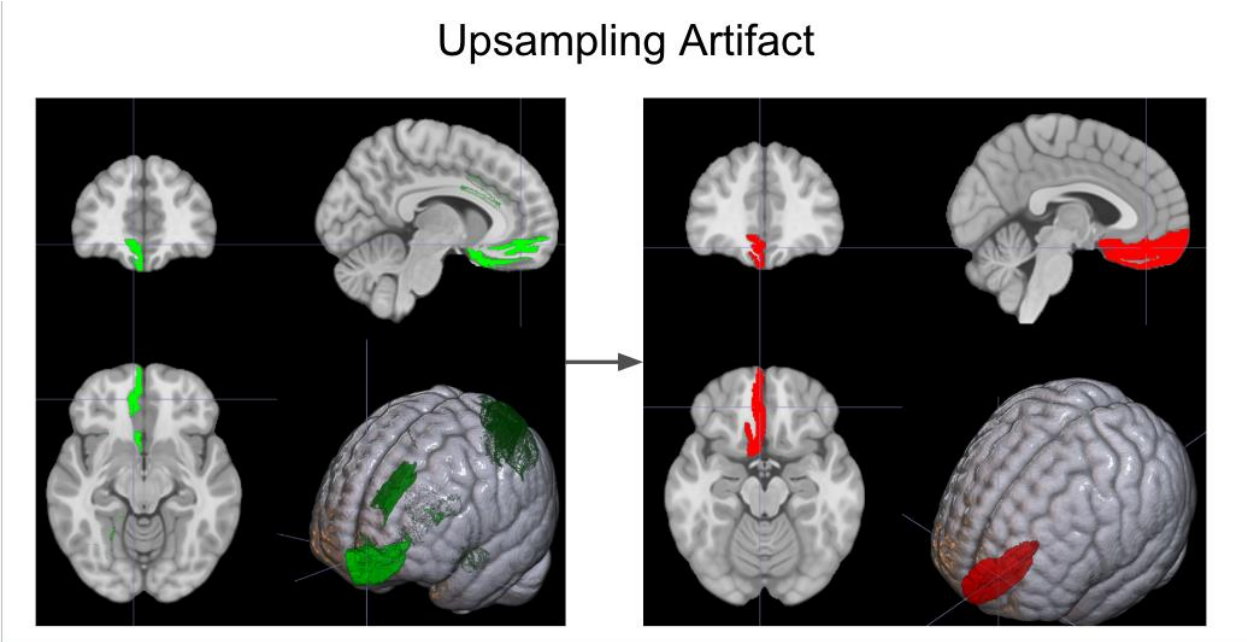


***Figure 4.4.1**:* Impact of operation order on ROI boundary quality. *Left: Variable maps generated by first upsampling/co-registering multi-label atlases, then extracting variables—showing interpolation-induced graininess. Right: Maps generated by first extracting binary label masks, then upsampling/co-registering—yielding crisp boundaries and anatomically faithful representations.*

Initial testing of Approach A revealed a critical artifact (Figure 4.4.1, left panel). Interpolation during upsampling introduced “fuzzy” label boundaries, leading to discontinuous or grainy variable maps, particularly in thin or oblique structures. These artifacts arose because multi-label volumes, when interpolated, produce intermediate voxel intensities that do not correspond to valid anatomical labels.

Switching to Approach B, extracting individual binary label masks first and then upsampling and co-registering each mask, resolved the problem (Figure 4.4.1, right panel). This sequence preserved crisp regional boundaries and prevented mislabeling, producing seamless variable maps registered to MNI 2009b. This change, while computationally heavier, was critical for ensuring the fidelity of voxel-level anatomical and functional definitions.

4.4.2 Individual Variable Extraction

Using the final workflow order (Approach B), a total of 232 variables were extracted from the atlases selected in Section 4.2. These include:

- Tissue segmentations: gray matter, white matter, cerebrospinal fluid
- White matter tracts: 87 probabilistic tracts from HCP1065
- Regional parcellations: 102 regions from CerebrA
- Functional networks: 17 cortical (Yeo17), 17 cerebellar (Buckner17)
- Vasculature metrics: 6 arterial territory probability and radius variables from Forkert & Mouches

Extraction proceeded in two scripted steps:

Step 1: Atlas Label Separation

Each atlas volume was decomposed into binary NIfTI masks, one per label value. The separation script, *roi_extraction.py* in combination with atlas-specific utilities (e.g., *separate_labels_CerebrA.py*), identified all unique nonzero label integers in the atlas file. For each label:

- Voxels matching the label were assigned a value of 1; all others were set to 0.

- The binary mask was saved to disk as *<atlas>_<label>.nii.gz* in the pipeline's *roi_masks/* directory.

This step, implemented with Nibabel, ensured that each variable of interest was spatially isolated, enabling clean variable extraction without cross-label contamination.

Step 2: Coordinate and Value Extraction

For each binary mask, the pipeline iterated through the 3D voxel grid to extract:

- (x, y, z) voxel coordinates in atlas space
- Voxel intensity values (from the original atlas or associated quantitative map)

The output was stored in per-ROI CSV files with columns for x, y, z, and value. These files formed the initial covariate dataset for each label and were placed in the *roi_data/* directory. Storing this data in tabular format allowed for flexible downstream transformations, such as adding Euclidean distance metrics or density calculations, without re-accessing the volumetric files.

4.4.3 Upsampling and Co-registration to MNI 2009b

Following variable extraction, each ROI mask was transformed to the MNI 2009b nonlinear symmetric space at 0.5 mm isotropic resolution. This step combined resampling (changing the voxel size and grid) and co-registration (aligning coordinate systems) in a single interpolation pass (see Figure 4.4.2):

- Upsampling addressed differences in voxel size between source atlases (often 1 mm or 0.6 mm) and the target 0.5 mm grid, using spline interpolation for binary masks to preserve regional boundaries.

- Rigid transformation corrected for translational and rotational offsets.

- Nonlinear warping accounted for local anatomical differences between the source atlas and the MNI 2009b template, if any.

All transformations were implemented in Python using Nibabel[92] and Nilearn[93] (for image resampling and affine transformations), with configuration parameters defined in *pipeline_config.yaml*. Outputs were stored in *roi_masks_mni/* and logged with their applied transform matrices for reproducibility.

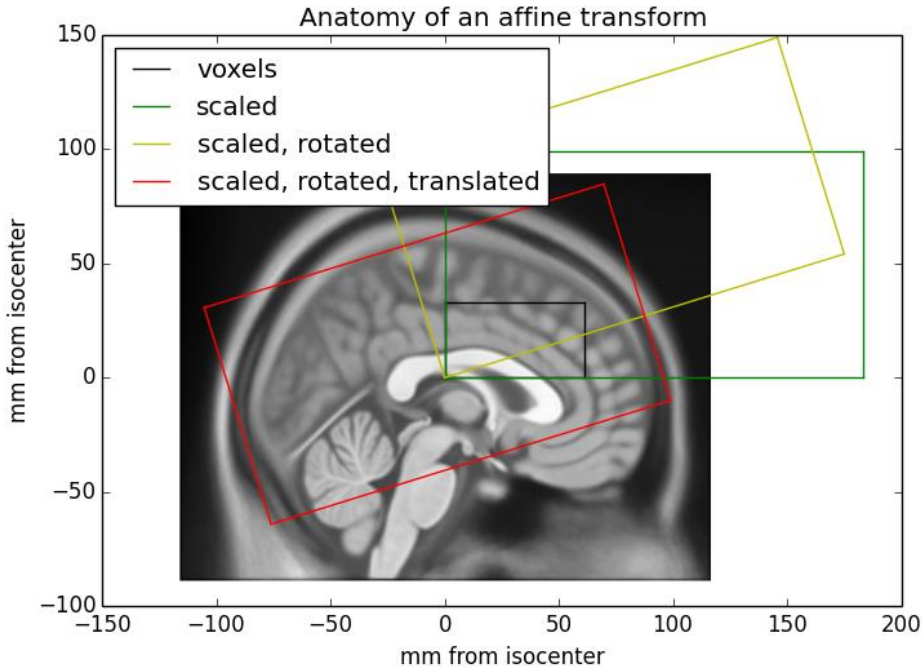


Figure 4.4.2: Anatomy of an affine transform[94]. *Example illustration showing how affine transformations adjust voxel coordinates during image registration. The transformation sequence includes scaling (green), rotation (yellow), and translation (red), relative to the original voxel grid (black). These operations correct for differences in orientation, position, and scale between source images and the MNI 2009b template, forming the first stage of atlas co-registration in the I-BIS pipeline.*

By completing ROI extraction before spatial transformations, I-BIS ensures that each variable map retains anatomically valid boundaries, is accurately positioned in MNI space, and is ready for covariate engineering in Sections 4.5–4.7.

4.5 Buffer-Zone Analysis

One of the core premises introduced in Chapter 2 is that the brain can be conceptualized as a spatially organized, multi-layered map, analogous to a geographic information system (GIS) that integrates terrain, infrastructure, and demographic layers. In GIS-based environmental modeling, buffer zones are routinely used to quantify the local context surrounding a point of interest, such as measuring population density within a given radius or road length within a set distance. In I-BIS, we extend this analogy to neuroimaging: for each region of interest (ROI) voxel, we calculate spatial context variables derived from brain atlases within concentric “buffer” radii. This approach encodes not only the voxel’s immediate anatomical identity but also the neighboring tissue composition, vascular

distribution, and network membership, features that can capture spatial dependencies relevant to neuropathological spread.

Spatial autocorrelation is a hallmark of many neuropathological processes. For example, tau pathology often extends contiguously along axonal projections, and ischemic damage is frequently confined to well-defined vascular territories. Buffer-zone covariates operationalize this by quantifying the proportion of specific atlas-defined features (e.g., white matter tracts, arterial territories, functional networks) in the surrounding neighborhood of a voxel. The choice of multiple radii, small (1–5 mm), medium (5–10 mm), and large (10–15 mm), mirrors the multi-scale strategy used in land-use regression[17], enabling the model to learn associations that manifest at different spatial extents. This ensures that both highly localized and broader contextual influences are represented in the covariate set.

4.5.1 Implementation in I-BIS

The buffer-zone step is implemented in *buffer_zone.py*, which is invoked when the *--buffer-zone* flag is set in *run_ibis_pipeline.py*. Parameters controlling this step are defined in *buffer_zone_config.yaml*, which specifies:

- radii: list of buffer distances in millimeters (e.g., [2, 5, 10, 15]).
- mask_dir: path to the preprocessed ROI masks in MNI 2009b space.
- atlas_dir: location of harmonized atlas volumes (output from Section 4.4).
- output_dir: destination for per-radius buffer maps and summary CSVs.

I designed the core algorithm to proceed as follows:

1. Load ROI mask: The voxel coordinates of each ROI are read into memory.
2. Iterate over radii: For each specified radius, a spherical structuring element is generated using NumPy and SciPy morphological operations.
3. Mask dilation: The ROI mask is dilated by the spherical element, creating a “buffer mask” that includes all voxels within the target distance.
4. Atlas overlay: Each atlas volume is intersected with the buffer mask to compute:
    - Count of voxels belonging to each label within the buffer.

- Proportion of the buffer occupied by each label.

5. Store results: For each ROI and radius, a table is written containing buffer-level covariates. Outputs are stored in a radius-specific subdirectory under *output_dir*.

In cases where buffer zones extend beyond brain boundaries (e.g., near cortical surface or ventricles), the script masks out non-brain voxels using the MNI brain mask distributed with the atlas set. Overlaps between neighboring ROIs are preserved, reflecting the reality that spatial context is not exclusive, the same voxel can be in the buffer of multiple ROIs if they are spatially close.

For each ROI and buffer radius, the process generates:

- A binary NIfTI mask of the buffer zone.
- A CSV file summarizing label-wise counts and proportions from all input atlases.

These outputs feed directly into the variable extraction stage (Section 4.6), where they are merged with distance-based and point-specific covariates into the master covariate table for modeling (Chapter 5).
QC involves visually inspecting a subset of buffer NIfTI masks in FSLeyes or FreeView to confirm correct dilation distances and atlas overlay alignment. I also ensured reproducibility by:

- Version-locking spatial libraries (Nibabel, Nilearn, SciPy) via *requirements.txt*.
- Storing all intermediate buffer mask files for audit and re-analysis.
- Logging all parameters (radii, atlas list, voxel resolution) to the run log created by *run_ibis_pipeline.py*.

The buffer-zone analysis step in I-BIS translates the GIS-inspired principle of local spatial context into a neuroimaging-compatible framework. By systematically quantifying atlas-defined features within multiple radii around each ROI voxel, it captures anatomical, functional, and vascular neighborhood information that would be missed by point-based covariates alone. This multi-scale representation is particularly valuable for modeling neuropathological phenomena that exhibit both highly localized effects and broader network or territory-level patterns. The resulting buffer-derived metrics form one of the three major categories of engineered covariates in I-BIS, complementing distance-based and point-specific variables. In the next stage (Section 4.6), these outputs are integrated

with other spatial measures through variable extraction, producing a unified covariate table ready for downstream modeling.

4.6 Absolute Distance Analysis

Distance-based covariates quantify the proximity of each brain voxel to a spatial reference feature, serving an analogous role to distance-to-road or distance-to-source predictors in land-use regression (LUR) models. In environmental modeling, proximity often serves as a proxy for exposure intensity; for example, being closer to a highway may increase exposure to vehicular emissions. In the neuroanatomical context, the same principle applies; distance to a vascular structure, ventricle, cortical surface, or other landmark may influence local physiology or pathology. Proximity to a large vessel might modulate exposure to vascular perfusion patterns; nearness to a ventricle could relate to cerebrospinal fluid (CSF) dynamics or the spread of soluble factors. By embedding this allocentric spatial reference frame into our feature set, I-BIS situates each voxel in the broader geography of the brain.

Absolute distance metrics in I-BIS are computed using Euclidean Distance Transforms (EDT), a voxel-based method for calculating the shortest straight-line distance in three-dimensional space from each voxel in the ROI to the nearest voxel belonging to a reference structure. The procedure is implemented in *variable_extraction.py* when the *--distance-analysis* flag is set in *run_ibis_pipeline.py*.
The method for EDT calculation proceeds as follows:

1. Binary mask creation – For each atlas-defined label or structure of interest (e.g., major arteries from Forkert & Mouches atlas, cortical ribbons from CerebrA), a binary mask is generated in MNI 2009b space as mentioned in 4.4.

2. EDT computation – The mask is passed to an EDT function (via SciPy's[78] *ndimage.distance_transform_edt*), producing a distance map where each voxel's value encodes its minimum Euclidean distance to the structure.

3. ROI sampling – For each ROI voxel, the corresponding EDT value is extracted, giving a precise measure of proximity to the reference structure.

4. Multi-feature integration – This process is repeated for all selected atlas features, generating a set of absolute distance covariates that are stored alongside other covariate types (e.g., buffer-based, point-based).

This approach is computationally efficient, with EDT calculations scaling linearly with the size of the binary mask and no requirement for iterative search across voxels. For each ROI, the EDT analysis produces:

- Coordinate triplets (x, y, z) in MNI space.
- Distance metrics for each selected feature, expressed in voxel distance.
- A single CSV file per ROI or per participant, integrating distances for all features into one table.

Quality check involves random visual inspection of distance maps in FSLeyes to confirm smooth gradient fields and correct anatomical targeting. I ensured reproducibility by:

- Saving all binary masks and distance maps to disk for auditability.
- Recording the atlas source and label ID for each feature.
- Version-locking dependencies in *requirements.txt* to prevent numerical discrepancies across environments.

EDT-derived covariates provide a straightforward yet powerful means of embedding spatial context into voxel-based models. While buffer-zone analysis (Section 4.5) captures aggregated neighborhood information, absolute distances offer a direct measure of "nearness" that is invariant to ROI size or shape. The combination of both approaches allows I-BIS to represent spatial relationships at multiple levels of abstraction, enriching the final feature set for downstream modeling. In the next stage (Section 4.7), these distance metrics are merged with other covariates to form the consolidated dataset used in predictive modeling.

4.7 Point-Based Covariate Extraction and Data Consolidation

4.7.1 Implementation of Point-Based Covariate Extraction

While buffer-zone (Section 4.5) and absolute distance-based covariates (Section 4.6) quantify spatial relationships around or to reference features, point-based covariates directly encode the value of a feature *at the voxel location itself*. This approach parallels point sampling in GIS, where the value of an environmental layer (e.g., elevation, pollution concentration) is extracted at each coordinate of interest. In the neuroanatomical domain, this could correspond to reading the probability of being gray matter from a tissue segmentation map, the presence/absence of a functional network label, or the occurrence probability of a major artery. These pointwise measurements provide voxel-specific feature values without aggregation or transformation, preserving the highest level of spatial fidelity. When integrated with buffer- and distance-based features, point-based

covariates complete the triad of spatial descriptors in I-BIS, offering a direct, localized attribute to complement the broader spatial context.

Point-based covariates are computed in *variable_extraction.py* when the *--point-analysis* option is enabled in *run_ibis_pipeline.py*.

The workflow proceeds as follows:

1. ROI voxel list ingestion – The pipeline reads the coordinate list generated during ROI voxel extraction (Section 4.4).

2. Atlas sampling – For each selected atlas layer, the voxel intensity at each ROI coordinate is extracted using Nibabel/Nilearn functions (*nibabel.load* + array indexing in MNI 2009b space).

3. Value mapping – The extracted value could represent:
    - A binary membership mask (1 if voxel is part of the label, else 0).
    - An anatomical measurement (e.g., mean vessel radius from Forkert & Mouches atlas).

4. Data storage – Each sampled value is appended to the voxel's row in the intermediate feature table for that ROI.

Point-based sampling is computationally lightweight and requires no preprocessing beyond ensuring atlas alignment to MNI 2009b (performed in Section 4.4).

4.7.2 Data consolidation

The value of I-BIS lies not only in generating a variety of spatial covariates but in harmonizing them into a single, coherent dataset. Without systematic consolidation, the outputs of Sections 4.4–4.6 would remain fragmented: each atlas, feature type, and measurement scale stored in separate files with their own coordinate indexing and metadata. Data consolidation therefore serves as the bridge between spatial feature engineering and statistical modeling. It guarantees that every voxel in the dataset is represented by a complete, consistent set of covariates, enabling the NPP model (Chapter 5) to operate on a single, standardized input table.

Following the generation of buffer-zone, absolute distance, and point-based covariates, the *data_consolidation.py* module merges all feature tables into a unified dataset. This process includes:

1. File aggregation – The intermediate CSVs from Sections 4.4–4.6 are loaded into memory in batches (memory-safe processing is implemented to handle large voxel counts).

2. De-duplication – Coordinates are cross-checked to avoid repeated entries from overlapping processes.

3. NA handling – Missing values (e.g., outside atlas coverage) are replaced with zero or NaN based on the feature type, with an accompanying metadata log.

4. Feature indexing – Columns are systematically named to indicate the atlas source, feature type (buffer, distance, point), and measurement scale, where applicable.

5. Final merge – All covariates are horizontally concatenated into a single table indexed by (x, y, z) coordinates in MNI space.

The consolidated dataset is saved in both CSV and parquet formats for compatibility with downstream modeling scripts. The final I-BIS output file contains:

- Coordinate triplets (x, y, z) in MNI 2009b space.

- Covariate columns from:
    - Buffer-zone analysis (multi-radius, multi-atlas).
    - Absolute distance analysis (EDT-derived).
    - Point-based atlas sampling.

- Metadata fields (participant ID, atlas name, label description).

With the I-BIS pipeline complete, the resulting consolidated dataset encapsulates the spatial, anatomical, and functional context of each voxel in a unified, analysis-ready form. Every row now contains the MNI-standardized coordinates, engineered covariates from multiple atlas layers. This table is the keystone connecting the preprocessing framework of Chapter 4 to the predictive modeling framework of Chapter 5. In the next chapter, I describe how these spatially enriched features serve as inputs to the NPP UK model, enabling robust, voxel-level predictions of pathology distributions across the brain.

Chapter 5: NPP UK Model

5.1 Introduction

By incorporating advanced spatial statistics, the **NPP UK model** facilitates the prediction and visualization of neuropathological distributions within the brain, specifically addressing the spatial relationships between key regions of interest (ROIs) and underlying brain pathology. It is designed to be semi-automated, streamlining much of the analysis process for users while retaining manual input options that provide customizable control over the modeling parameters. This makes the pipeline not only accessible but also flexible, allowing users to tailor it to specific needs and datasets. Furthermore, the model generates detailed output data, including logs and intermediary results, which offer the option for fine-grained analysis and validation at every step of the process. In this chapter, I discuss the methodological framework for the NPP UK model pipeline. Results from applying this pipeline are presented in the next chapter, chapter 6.

At the heart of the model's functionality is the integration of Elastic Net regression and Factor Analysis to identify primary predictors from a high-dimensional dataset. By narrowing down hundreds of potential predictors to a manageable set of the top 30 significant features, and further refining these using factor analysis (which includes correlation analysis, cluster analysis, and scree plot analysis), the pipeline highlights the most relevant variables for predicting neuropathological changes. These variables are then analyzed, expanding the model's utility for understanding the spatial underpinnings of Alzheimer's disease and other neurodegenerative conditions.

The input to the model consists of two key elements: (1) quantitative neuropathology data in ROIs mapped to MNI voxel coordinates (Chapter 3), and (2) covariate data from I-BIS, also mapped to the MNI space (Chapter 4). These files serve as the foundation for the model, which processes them through a sequence of four major steps (See Figure 5.1.1): preprocessing (merging the inputs into a design matrix), feature selection via cross-validation Elastic Net, factor analysis to narrow the predictor set, and finally, Universal Kriging to predict spatial distributions.

The NPP UK model culminates in a spatial variogram model and external drift model (two components of UK) that together predict the spatial distribution of neuropathology across the ROIs. By applying these models to the neuroimaging and neuropathology data, the pipeline generates detailed predictions of where pathology is most likely to be within the brain, specifically in relation to the regions of interest, i.e., MFG in this case. This UK model pipeline is applied to the data derived from 10 donors of the ACT study, as was discussed in Chapter 2. This is the final summation of the NPP project, bringing

together spatial statistical models to provide insights into regional neuropathological distribution.

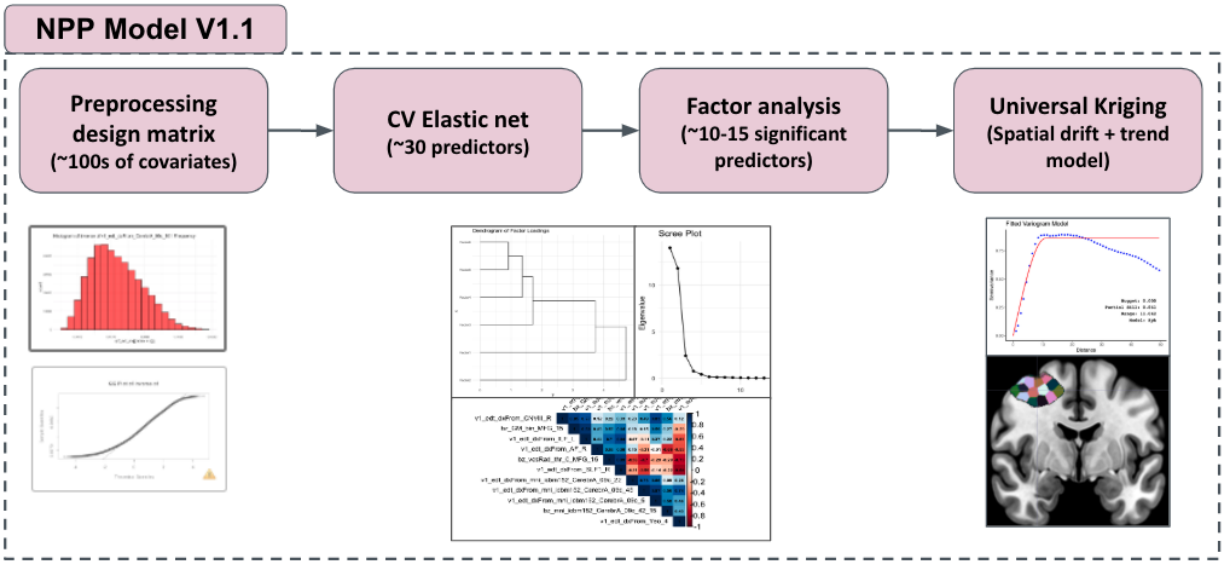


Figure 5.1.1: *This figure depicts the four primary components of the NPP UK model pipeline that enable the universal kriging modelling approach to predict the distribution of neuropathology. Preprocessing involves data standardization and transformation. CV elastic net and factor analysis are deployed to reduce covariate dimensionality, including cluster, scree, and correlation analyses. Finally, the UK model is applied with the help of a spatial variogram and leave-one-out block analysis.*

Next, I'll walk through the technical setup of this pipeline (section 5.2), preprocessing input data (section 5.3), applying cross-validated elastic net for dimensionality reduction (section 5.4), using factor analysis to select primary predictors (section 5.5), modeling spatial data with a two-part UK model (section 5.6), and finally saving results and visualizations from this pipeline (section 5.7). These sections mainly focus on the developed methodology for executing this pipeline. The results from the UK model are examined in more detail in the next chapter (Chapter 6).

5.2 NPP UK model pipeline setup

The pipeline is orchestrated by a series of interconnected R scripts and configuration files, which collectively facilitate the seamless processing of the data, from preprocessing through to spatial prediction. It is designed to be modular, allowing for flexibility and reproducibility in neuropathological modeling. At its core, the pipeline is executed through *run_pipeline.R*, with three critical components setting it up for execution: the *setup.R* script and the two configuration files: *input_variables.txt* and *subject_list.txt*. Each of these

components plays a pivotal role in structuring the pipeline and ensuring that data is processed systematically.

5.2.1 Orchestrating the pipeline

The *run_pipeline.R* script serves as the central driver of the entire NPP UK model pipeline. This script sequentially executes the various stages of the analysis, from data preprocessing to spatial prediction. The script ensures that all required steps are completed in the correct order, facilitating a smooth flow from one analysis stage to the next. Upon execution, *run_pipeline.R* first loads the configuration files and reads in the participant list (specified in *subject_list.txt*) to determine which data files need to be processed. It then proceeds through the following steps:

- Data Preprocessing: This step is performed by the *01_preprocessing.R* script, which validates, cleans, and standardizes the input data, merging them into a unified design matrix for downstream analysis. It then performs a covariate transformation and standardization, preparing the data for modeling.

- Elastic Net Regression: *02_elastic_net.R* applies Elastic Net to select primary predictors, ensuring a sparse yet accurate model by applying regularization to a high-dimensional dataset.

- Factor Analysis: The *03_factor_analysis.R* script is run to reduce dimensionality further, performing clustering and scree plot analysis to determine the most relevant predictors.

- Universal Kriging: Finally, the *04_universal_kriging.R* script applies Universal Kriging for spatial prediction, modeling the residuals, and adjusting for spatial autocorrelation.

Each of these four sub-components is described in detail in subsequent sections. The script ensures that all intermediate outputs are saved in a structured directory hierarchy, making it easy to track progress and access results at each stage. The modular design of the pipeline enables easy adaptation, allowing specific stages to be executed independently or reconfigured according to the user's needs.

5.2.2. Setting up the environment

Before any data processing begins, the *setup.R* script initializes the R environment, ensuring that all necessary libraries are loaded and all paths are correctly configured. It performs the following functions:

- Library Loading: Essential libraries for data manipulation (dplyr, tidyr), statistical modeling (caret, glmnet), spatial analysis (gstat, sp), and visualization (ggplot2, matplotlib) are loaded into the environment.

- Validation of Input Parameters: The script checks the input paths for raw data files, participant IDs, and configuration files, ensuring they are correctly specified and accessible. This validation step helps avoid errors during data processing by confirming that all necessary files are present before starting the analysis.

- Directory Setup: This step creates the necessary output directories for each analysis step (e.g., *output/preprocessing/, output/covariate_analysis/)* to store intermediate and final results. It ensures that outputs are organized according to the participant IDs and analysis stages, making it easy to manage and retrieve results.

By setting up the environment and verifying configurations before execution, the *setup.R* script ensures that the pipeline runs smoothly, minimizing potential errors related to file paths or missing dependencies. The script's error handling and logging features also provide a safeguard, ensuring that any issues are caught early and logged for troubleshooting.

5.2.3. Configuration files

The configuration files, *input_variables.txt* and *subject_list.txt*, are crucial for customizing the pipeline to different datasets and research needs. These files are read by the *run_pipeline.R* script to gather the necessary input data paths, model parameters, and participant information.

- *input_variables.txt* contains parameter settings that guide the pipeline's processing steps. It specifies the file paths for the input data (neuropathology data, coordinates, and covariates) and defines output directory locations for each pipeline stage (preprocessing, covariate analysis, Elastic Net, etc.). It also includes detailed settings for preprocessing (e.g., standardization, transformations for specific variables such as Percentage AT8 positive value and Euclidean distance transform (EDT)), as well as Elastic Net settings (e.g., alpha grid values, cross-validation parameters, and parallel processing configurations). Additionally, it outlines the configuration for Universal Kriging, specifying spatial settings such as the variogram model, nugget, range, and grid size for spatial prediction. The use of key-value pairs in this file enables easy modification of parameters, making the pipeline adaptable to various datasets or research scenarios. These parameters will be discussed in subsequent sections.

- *subject_list.txt* contains a list of participant IDs to be processed by the pipeline. Each line in this file corresponds to a single participant, ensuring that the pipeline knows which datasets to load and process. This file is essential when running the pipeline across multiple participants, as it automates the selection and batching of participant-specific data. The participant list is read by the pipeline, which then uses the IDs to access corresponding neuropathology data, ROI coordinates, and covariates stored in separate input directories. Lines beginning with "#" are treated as comments and ignored, allowing for easy documentation or exclusion of specific participants without modifying the entire list.

The modular design of this pipeline ensures that it can be easily extended or adapted for different research purposes, with clear separation of concerns between configuration, environment setup, and execution. Furthermore, the use of external configuration files makes the system flexible and accessible for future modifications or updates, supporting long-term scalability and reproducibility of results.

5.3 Preprocessing: Merging Data into a Design Matrix

The preprocessing step is crucial for preparing the input data for further analysis in the NPP UK Model pipeline. The goal of preprocessing is to transform input data into a unified design matrix, which serves as the foundation for subsequent modeling steps. This step involves several processes, including data cleaning, standardization, and integration of multiple data sources (i.e., neuropathology data, voxel coordinates, and covariate data).

The preprocessing module begins by loading two key input files for each participant. These include:

1. QNP Data (*MFG_QNP_vox_index_combined.csv*)**:** This file contains QNP data that includes intensity measures of pathology (e.g., tau pathology in Alzheimer's disease as measured by the AT8 stain) mapped to MNI voxel coordinates (Chapter 3).

2. Covariate Data (*Cov_all_consolidated_MFG_v1.csv*)**:** This file includes covariate data, derived from I-BIS (Chapter 4), and represents the structural and functional hierarchy of the human brain in the ROI.

These files are loaded into the pipeline and merged into a single design matrix, ensuring that each row corresponds to a unique combination of participant ID and voxel coordinates and each column represents a variable (e.g., neuropathology intensity, covariates). This unified matrix provides a comprehensive view of the data, ensuring that

all features are aligned in the same coordinate space (MNI 2009b), which facilitates the next steps of the analysis.

5.3.1 Handling missing data

Once the data are loaded, the preprocessing module ensures that the design matrix is clean and complete for analysis. Missing values are identified and addressed according to predefined criteria. Typically, missing data may be imputed or excluded based on the proportion of missing values in each variable. Additionally, zero-values, which may represent background tissue or non-brain regions, are either removed or flagged to prevent introducing noise into the subsequent analysis. The threshold for missing values is set carefully (e.g., 0.1), establishing a tolerance level for missing data, which ensures that only variables with a sufficient amount of data are retained for analysis.

5.3.2 Data standardization and transformation

Data standardization is a key step in preparing the design matrix for modeling. The goal of standardization is to ensure that all input variables are on a comparable scale, which is particularly important when dealing with high-dimensional datasets containing many covariates. In NPP, standardization is generally achieved by transforming the variables to have a mean of zero and a variance of one, using the z-scoring method. This ensures that no single covariate dominates the model simply due to its scale. Some covariates are transformed in a specific manner based on their distribution properties, which were analyzed separately outside the scope of this pipeline. For example:

- Cube-root transformation is applied to buffer zone (BZ) variables to normalize the distribution of values that may be skewed.
- Log transformation is applied to the Percentage AT8 positive value data to handle potential skewness and stabilize variance.
- Certain variables, such as Euclidean distance transform (EDT) variables, may not require transformation and are left unmodified.

The script also excludes specific columns from the standardization process (e.g., coordinates or participant IDs), which are necessary for the design matrix but do not require standardization.

5.3.3 Merging and final validation

The final step of preprocessing involves merging the cleaned, standardized, and transformed datasets into a single design matrix. This matrix contains all the relevant features for each voxel, including the spatial coordinates (x, y, z), neuropathology data,

and the covariates. The merged dataset is saved in a CSV format, making it accessible for subsequent modeling steps. The merged matrix is validated for consistency and accuracy. Intermediate results, such as cleaned and transformed datasets, are saved at each step. After preprocessing, the design matrix is ready for the following three stages of analysis.

5.4 CV Elastic Net: Dimensionality Reduction

Elastic Net regression is a powerful technique employed to perform dimensionality reduction by selecting the most relevant features from a high-dimensional dataset. In the context of the NPP UK model, Elastic Net serves a dual purpose: it reduces the number of predictors (covariates) while also addressing multicollinearity among highly correlated features. By utilizing both L1 (Lasso) and L2 (Ridge) regularization, Elastic Net effectively selects predictors that contribute most to the variance in the target variable, which in this case is typically related to phospho-Tau distribution. In this section, I explain what Elastic Net does, how it compares to Principal Component Analysis (PCA), and provide a detailed explanation of how the process is implemented in the *02_elastic_net.R* script.

5.4.1 Elastic Net Overview

Elastic Net is a regularized linear regression technique that combines the strengths of LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge regression. These methods, individually, apply penalties to the regression coefficients to promote sparsity (Lasso) or prevent overfitting by shrinking the coefficients towards zero (Ridge). Elastic Net is particularly useful when dealing with datasets where the number of predictors is large relative to the number of observations, or when predictors are highly correlated. In highly correlated datasets, it addresses one of the limitations of Lasso, which tends to select only one predictor from a group of highly correlated features, by incorporating a Ridge-like behavior that groups together correlated predictors and penalizes them collectively.

The regularization parameters in Elastic Net are controlled by two hyperparameters:

1. Alpha (α): Determines the balance between Lasso and Ridge. When α = 1, the model behaves like Lasso; when α = 0, it behaves like Ridge.

2. Lambda (λ): Controls the strength of the penalty, with larger values of λ leading to stronger regularization and thus fewer selected covariates.

I chose the Elastic Net approach, among multiple methods that help with dimensionality reduction, such as Lasso and PCA. Elastic Net's ability to balance between Lasso and

Ridge makes it well-suited for high-dimensional data with correlated predictors, ensuring that the model remains interpretable while avoiding overfitting. PCA transforms the original variables into a new set of uncorrelated components based on their variance. However, PCA does not provide interpretable coefficients in relation to the original variables, making it less useful for feature selection. Indeed, PCA can be considered a dimensionality reduction approach that does not attempt to omit any covariates. In contrast, Elastic Net regression retains the interpretability of the original features, selecting a subset of variables that directly contribute to predicting the target variable. Thus, Elastic Net not only reduces dimensionality but also facilitates a clearer understanding of the relationships between predictors and the target variable; hence, it was my prime choice for NPP.

5.4.2 Elastic Net in NPP UK model

The Elastic Net implementation in the NPP–UK pipeline is the gatekeeper that identifies the top 30 primary predictors used in subsequent factor analysis and Universal Kriging. I cap the selection at 30 for a pragmatic balance of tractability, interpretability, and coverage. At the upper end, pushing to 50 or 60 predictors would roughly double the dimensionality of the drift term and materially increase the computational burden of cross-validated Elastic Net, generalized least-squares estimation in the kriging drift, and per-voxel prediction, while also diluting interpretability when inspecting effects one by one. At the lower end (e.g., 15-20), the model risks discarding anatomically important covariates drawn from distinct I-BIS families. 30 therefore serves as the highest number I can responsibly manage through modeling and interpretation: it keeps the drift estimable and stable, keeps runtimes within available resources, and still affords sufficient breadth to capture the major sources of spatial variation.

Within this constraint, the *02_elastic_net.R* script standardizes the design matrix, performs cross-validated Elastic Net to shrink and decorrelate features, ranks non-zero coefficients by standardized magnitude, and passes the top 30 predictors forward. As in previous chapters, I illustrate the NPP workflow with data from participant 6966.

Step 1: Loading Transformed Data

The script starts by loading the transformed and standardized data (preprocessed in earlier steps), using the *load_transformed_data* function. This function ensures that the data are correctly formatted and contain the necessary covariates and target variable (e.g., percentage AT8 positive stained for phosphor-Tau) for model training. The data are loaded into a data table and checked for consistency, with logging output that informs the user about the dimensions and structure of the data.

Step 2: Preparing Data for Elastic Net

The *prepare_data* function is responsible for setting up the data for Elastic Net regression. The function separates the features (covariates) from the target variable, which is typically the percentage AT8 positive or another measure of neuropathological findings. The *prepare_data* function also ensures that the features are divided into training and testing sets. This separation enables the assessment of the model's generalizability by training it on one subset of the data and testing it on another.

Step 3: Cross-Validation and Model Training

The *train_elastic_net* function applies cross-validation to identify an optimal lambda (λ) parameter, which controls the penalty applied during training. The function uses the *cv.glmnet* function from the glmnet package to perform k-fold cross-validation, where the data is split into folds, and the model is trained and tested on different subsets to assess its performance. Internally, glmnet constructs a log-spaced λ path from a data-derived λ_max to λ_min and fits it via coordinate descent. *cv.glmnet* returns lambda.min (or lambda . 1se), the lowest-error choice (mean cross-validation error), which I adopt for the final fit.

For participant 6966, the model undergoes 10-fold cross-validation to select the optimal lambda value of 0.0408. This value is chosen because it minimizes the Mean Squared Error (MSE) of the model on the training set.

Step 4: Evaluating the Model

Once the best *cv.glmnet* model is selected, the function proceeds to calculate predictions for both the training and testing sets. The *train_elastic_net* function then evaluates the model's performance using various metrics, including mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2), to assess model accuracy and fit. These metrics help to quantify how well the Elastic Net model performs in predicting the target variable (e.g., neuropathological intensity) based on the selected covariates.

For participant 6966, the model's performance on the training set is as follows:

- MSE: 0.3664
- RMSE: 0.6053
- MAE: 0.4732
- R^2: 0.6341
- Adjusted R^2: 0.6340

These metrics indicate that the model explains approximately 63% of the variance in Percentage AT8 positive value for the training set. The model is then evaluated on the test set, with the following performance metrics:

- MSE: 0.3776
- RMSE: 0.6145
- MAE: 0.4801
- R^2: 0.6206
- Adjusted R^2: 0.6205

The test set results are consistent with the training set performance, suggesting that the model is generalizing well to unseen data.

Step 5: Feature Importance and Visualization

After model training, the script generates an importance ranking for the predictors based on the coefficients from the Elastic Net model. Features with non-zero coefficients are considered primary predictors, and the top 30 predictors are selected for further analysis. The *analyze_features* function produces a sorted list of features based on their absolute coefficient values, with the most important predictors listed at the top.

For participant 6966, the top 30 predictors are ranked based on their coefficient values. Examples of the most important predictors for this participant are:

1. v1_edt_dxFrom_CPT_F_R (Coefficient: 0.3242): Distance to Corticopontine Tract Frontal Right white matter tract.

2. v1_edt_dxFrom_Yeo_5 (Coefficient: -0.3091): Distance to Dorsal Attention A functional network.

3. bz_TR_S_R_7.5 (Coefficient: -0.3064): Density of Thalamic Radiation Superior Right white matter tract within a buffer zone of radius 7.5mm.

4. bz_SLF2_R_7.5 (Coefficient: -0.2662): Density of Superior Longitudinal Fasciculus2 Right white matter tract within a buffer zone of radius 7.5mm.

5. bz_CPT_F_R_7.5 (Coefficient: 0.2577): Density of Corticopontine Tract Frontal Right white matter tract within a buffer zone of radius 7.5mm.

These predictors represent the covariates with the strongest associations with percentage AT8 positive values for participant 6966. The Elastic Net model assigns each of these features a coefficient, indicating the strength and direction of its relationship with the

target variable. Positive coefficients indicate direct relationships, while negative coefficients indicate inverse relationships.

Step 6: Saving Results

Finally, the model results, including the trained model, performance metrics, and primary predictors, are saved to output files. These include CSV files for the top predictors, model summaries, and prediction results, which can be further reviewed and used for downstream analysis. Additionally, the model is saved as an RDS file (RDS is a storage format for R), ensuring that the model can be reloaded and reused without the need to retrain. Importantly, it also saves the file '*XXXX_predictors_to_review.txt' (XXXX corresponds to participant ID)*, which contains a list of the top 30 predictors, and is further used as an input for the next step of factor analysis.

5.5 Feature Selection

Following the cross-validated Elastic Net regression, the next step in the NPP pipeline focuses on feature selection and dimensionality reduction. By performing this dimensionality reduction, the model improves interpretability and reduces overfitting, enabling more reliable predictions. The results of this analysis help refine the predictor set before moving on to Universal Kriging for spatial predictions. The aim is to reduce the number of predictors from the top 30 selected by Elastic Net to 10-20 primary predictors. This reduction makes the model more interpretable while maintaining its predictive power.

This step is operationalized in the *03_factor_analysis.R* script. Although the script name suggests suitable inter-***factor analysis***, the implementation has since evolved to use Principal Components Analysis (PCA), which is more appropriate for achieving orthogonal, variance-preserving dimensionality reduction. The current pipeline retains the historical naming convention for backward compatibility, but future iterations of NPP will rename this module to reflect its function more accurately.

5.5.1 Factor analysis in NPP UK model

Step 1: Loading Primary Predictors and Transformed Data

The *load_significant_predictors* function in the *03_factor_analysis.R* script loads the primary predictors identified from the Elastic Net step. For participant 6966, the Elastic Net regression identified a set of top predictors that include both Euclidean distance transforms (EDT) and buffer zone (BZ) variables. The list of top 30 predictors for participant 6966, as extracted in the previous step, is provided in the *6966_predictors_to_review.txt*. In addition to loading the predictors, ***the***

load_transformed_data *function* is used to load the preprocessed and standardized data. This data includes both the selected predictors and the target variable (Percentage AT8 positive value). This data serves as the input matrix for correlation analysis and PCA-based dimensionality reduction.

Step 2: Exploring Correlations Among Predictors

Before applying PCA, the correlation structure among the predictors is examined to assess redundancy and guide interpretation. The correlation matrix (Figure 5.5.1) visualizes the pairwise correlations among the 30 predictors, while the hierarchical clustering dendrogram (Figure 5.5.2) groups variables based on their similarity. These visualizations reveal clusters of highly correlated features, often buffer-zone variants of the same anatomical structure (e.g., multiple radii around a single tract or network).

Although originally included as part of the factor analysis module, these visualizations now primarily inform the Elastic Net step by highlighting multicollinearity within its selected features. They are retained here for diagnostic clarity and to illustrate the degree of overlap among predictors selected for PCA.

Step 3: PCA for Dimensionality Reduction

To mitigate multicollinearity and compress the correlated predictor space into a smaller set of independent dimensions, PCA is performed using the maximum-likelihood extraction method with Varimax rotation for numerical stability. While the earlier design of this module envisioned a classical factor analysis approach, subsequent testing showed that PCA better fits the purpose of this stage: achieving orthogonality among predictors rather than inferring latent constructs.

The scree plot (Figure 5.5.3) and eigenvalue criteria (eigenvalues > 1) were used to determine the number of components to retain. For participant 6966, four principal components captured approximately 91 % of the total variance, providing a compact and non-collinear representation of the original 30 predictors.

These components summarize the major variance directions in the data and serve as efficient, orthogonal inputs for subsequent modeling in the Universal Kriging step.

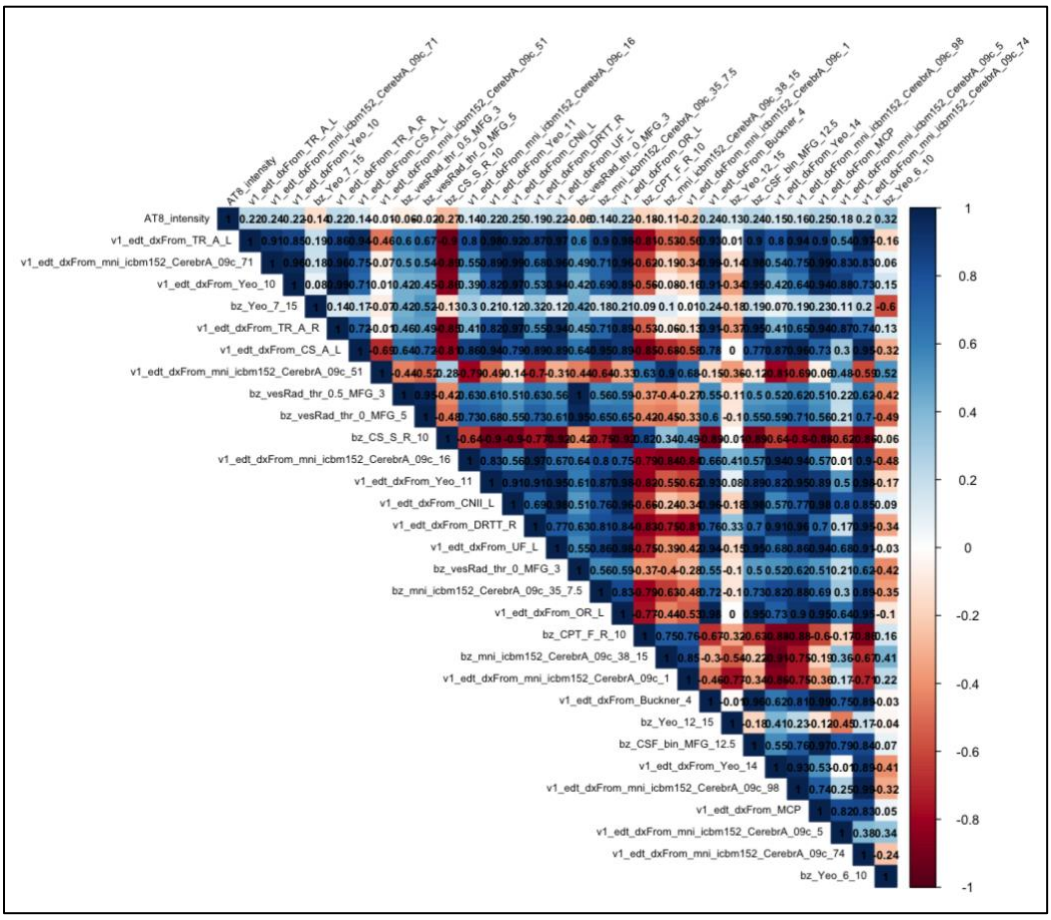


***Figure 5.5.1**: Correlation plot for participant 6966, representing the top 30 predictors that are selected from the CV-Elastic net model.*

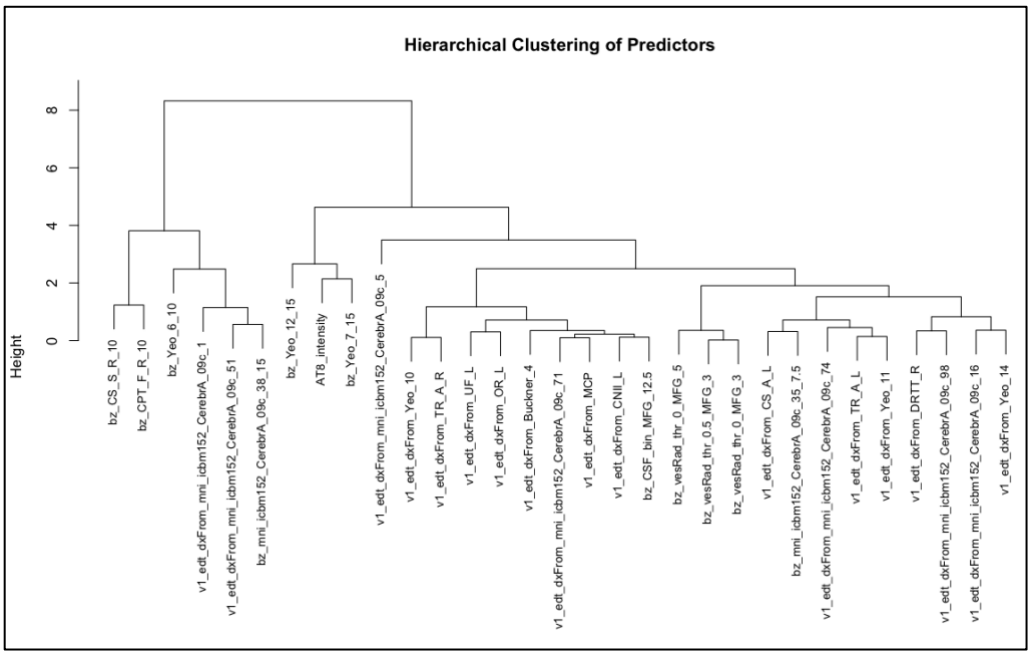


Figure 5.5.2: *Dendrogram plot for participant 6966, representing hierarchical clustering of the top 30 predictors that are selected from the CV-Elastic net model.*

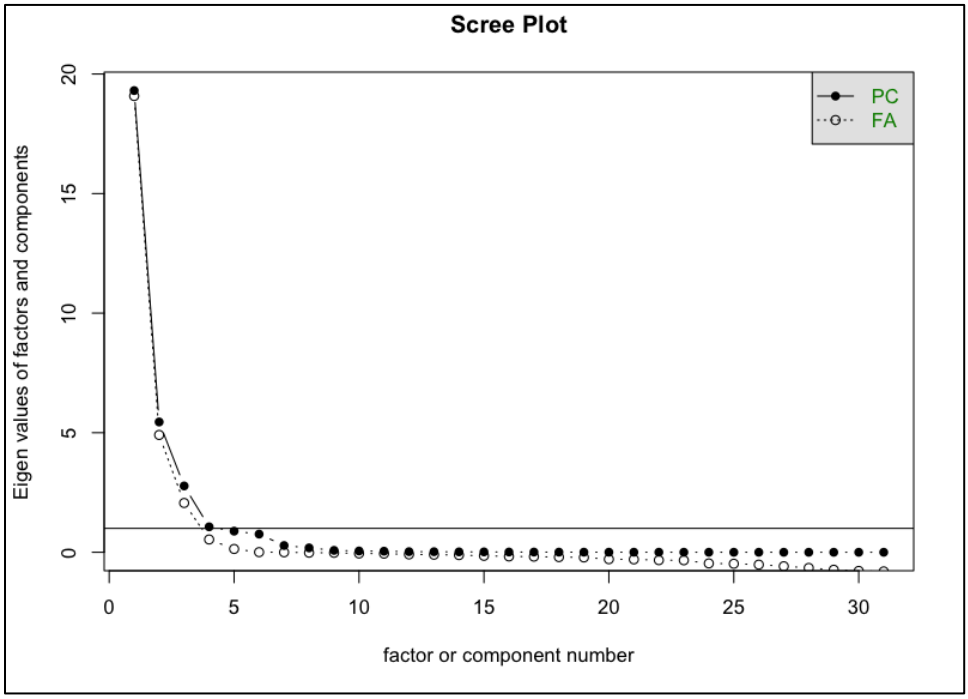


Figure 5.5.3: *A scree plot for participant 6966, depicting the minimum number of factors or components needed to predict optimum variance from 30 predictors that are selected from the CV-Elastic net model.*

Step 4: Saving Results

After PCA, the results are summarized and saved by the *03_factor_analysis.R* ***script*** into several output files:

- Factor/PCA Loadings: The loadings for each factor are saved in a CSV file (*6966_factor_loadings.csv*), showing the contribution of each original variable to the retained components (for transparency, not interpretation).
- Scree Plot: Saved as 6966_scree_plot.png, used to justify the choice of four components.
- PCA Summary: Recorded in 6966_factor_analysis_results.txt, including eigenvalues, variance explained, and model fit indices.

- Correlation and Dendrogram plots: Saved in the same directory to document inter-variable relationships before dimensionality reduction.

In future versions of NPP, these outputs will be reorganized under a revised module titled *"03_feature_selection.R"* to better reflect their purpose and avoid confusion with interpretive factor analysis.

5.5.2 Selecting primary predictors for UK

At this stage, the pipeline prompts the user to review the list of predictors and manually save only the ones they wish to use for further analysis as *XXXX_predictors_reviewed.txt*. If this file already exists in the output directory, it indicates that the list of primary predictors has been reviewed manually. To further facilitate the analysis of primary predictors, the script *03_factor_analysis.R* takes input *from XXXX_predictors_reviewed.txt* and creates a subfolder named *SigPred*, in which the shortlisted predictors are analyzed. Similar to primary factor analysis, this subsequent analysis plots a correlation matrix and a dendrogram. Additionally, it lists predictor pairs with high correlation ($|r| > 0.8$). This provides an extra layer of verification for the user to review if needed.

For participant 6966, after conducting factor analysis, I have narrowed down the list to 10 primary predictors from the initial 30. As shown in Figure 5.5.4, multicollinearity has been reduced. While we could further reduce predictors to eliminate high correlation pairs entirely, in this specific case, to ensure a sufficient number of predictors for subsequent steps, I chose to retain at least 14 primary predictors. This number may vary depending on the participant and their corresponding neuropathology.

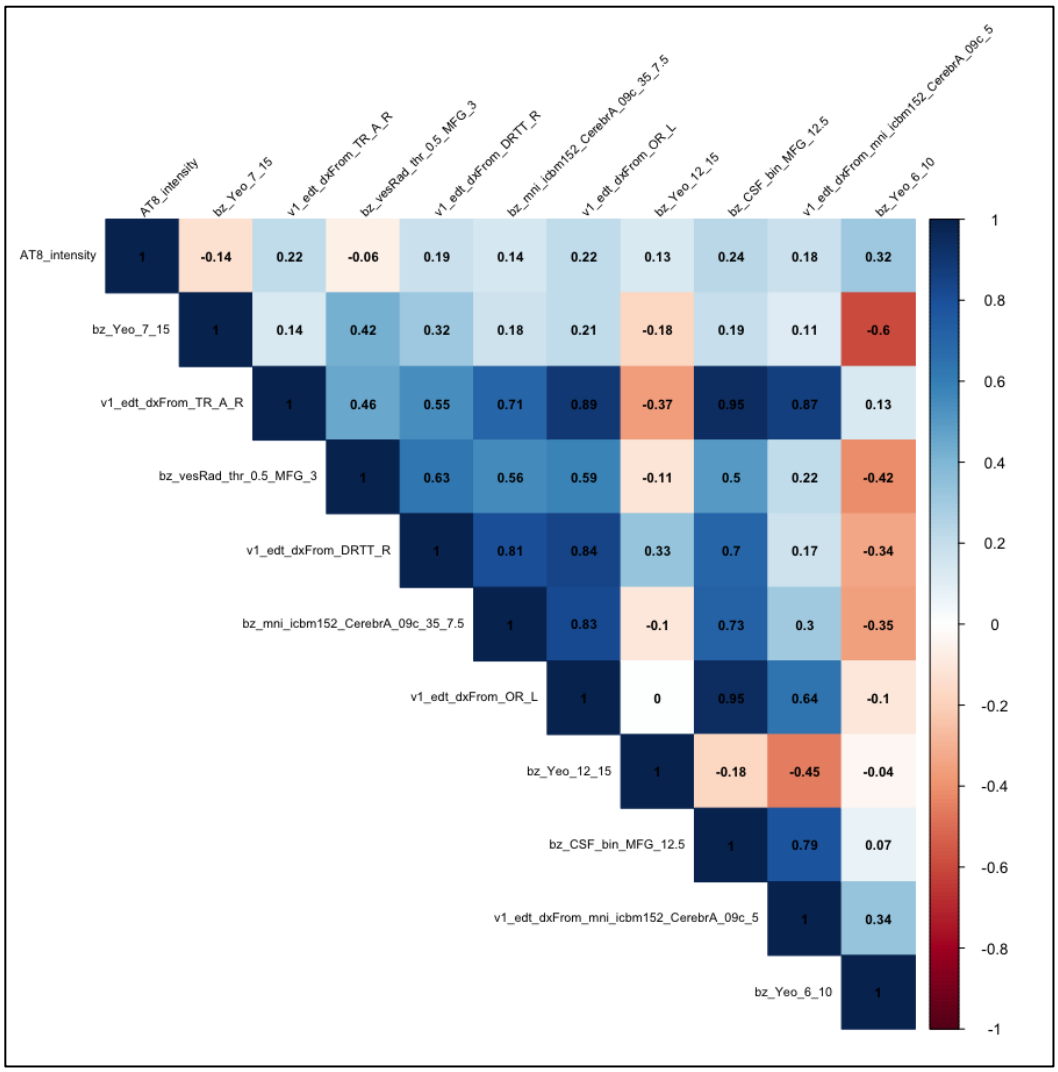


***Figure 5.5.4**: Correlation plot for participant 6966, representing primary predictors that are selected after factor analysis.*

5.6 Universal Kriging: Spatial Modeling and Prediction

5.6.1 Introduction to UK for NPP UK model

In this model, I deploy Universal Kriging (UK) to predict the spatial distribution of neuropathology (% AT8 positively stained for p-tau) across the brain. Specifically, like geospatial modeling[17] (Chapter 2), this process models the spatial spread of tau by considering the spatial autocorrelation of the data (i.e., the tendency of nearby regions to have similar neuropathological values) alongside external covariates that influence the distribution of neuropathology, such as neuroimaging-derived features from I-BIS.

The UK model incorporates both spatial drift and spatial residuals, where the drift represents the overall trend or pattern (derived from anatomical covariates) of neuropathology across the brain, and the residuals capture the local variations (spatial autocorrelation) in the data. This method is particularly valuable, as neuropathological proteins like tau do not always follow a simple pattern (prion-like) but are influenced by various dynamic factors such as disease progression, anatomical brain regions, and individual differences in age or genetic factors. By incorporating anatomical covariates into the model, UK offers a more flexible and accurate prediction of neuropathology distribution in the brain, as opposed to traditional methods like Ordinary Kriging, which would ignore these external influences.

Before I go into specific details, I will introduce key components of the Kriging Process-

1. The UK Formula

UK formula integrates both spatial coordinates and external anatomical covariates to predict the distribution of neuropathology. Specifically, the formula is as follows:

$$Log(AT8\ intensity) = \beta_0 \sum_{i=1}^{n} \beta_i\ * covariate_i + spatial\ residual$$

Where:

- **AT8 intensity** is the target variable (percentage AT8 positive value).
- **β_0** is the intercept.
- **β_i** are the coefficients for the anatomical covariates (/primary predictors).
- **Covariate$_i$** are the individual anatomical covariates.
- **Spatial residuals** represent the part of the variance that is not explained by the covariates and is partially captured through spatial autocorrelation (Variogram model).

2. Variogram:

The variogram is a core component of UK. It measures the spatial dependence of the data by calculating the semi-variance between pairs of points at different distances. The variogram helps in understanding how spatially autocorrelated the data are at various distances, which is crucial for predicting neuropathology in regions that have not been

directly sampled. The variogram component captures the stationary residuals, assuming constancy over a region.

For participant 6966, the empirical variogram is first computed by calculating the semi-variance between all pairs of voxel locations (X, Y, Z) in the preprocessed data. This provides an estimate of how similar or dissimilar the neuropathological values (Percentage AT8 positive value) are at different spatial separations.

The variogram is then fitted to a model, usually a Spherical, Exponential, or Gaussian model, to characterize how the spatial correlation decays with distance. After modeling all three types and comparing their fit for spatial semi-variance, I chose a Spherical model for all participants, with key parameters such as nugget, partial sill, and range being optimized based on the empirical variogram. Incidentally, the spherical model also geometrically aligns with spherical buffer zone covariates.

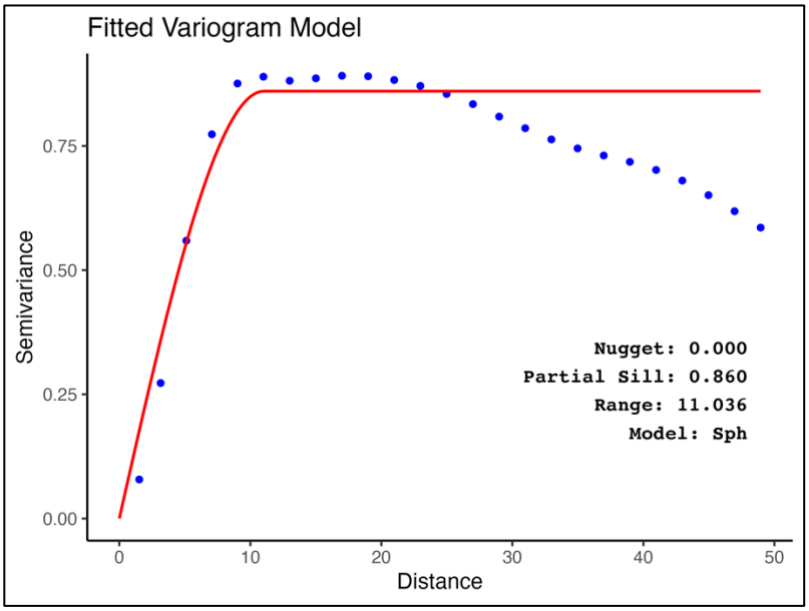


Figure 5.6.1: *An example of a fitted variogram that depicts the spatial autocorrelation of the percentage AT8 positive value distribution across voxels in MFG for participant 6966.*

Nugget: Represents the variance at very short distances (or noise), which is typically small but accounts for measurement error or unmodeled variations.

Partial Sill: Indicates the asymptotic value of the variogram, reflecting the total spatial variance that the model can explain.

Range: Defines the distance at which the spatial autocorrelation between data points effectively becomes zero, suggesting that beyond this distance, the data points are spatially independent.

3. Spatial Drift:

The spatial drift in UK refers to the deterministic, smoothly varying mean of the process that external covariates can explain. This drift accounts for systematic variations in the data that are not solely due to spatial proximity but are driven by other factors. For

example, in participant 6966, factors such as bz_CSF_bin_MFG_12.5 (density of CSF within a sphere of 12.5 mm) could explain variations in Tau distribution across the ROI, i.e., MFG.

By incorporating external covariates such as neuroimaging features (e.g., gray matter volume, functional connectivity), the model can predict how neuropathology varies due to these non-stationary influences. Non-stationary because they vary from one location to the other. This is important because neuropathological features like tau deposition do not follow a uniform spatial pattern but are influenced by a variety of biological and clinical factors.

5.6.2 Components of 04_universal_kriging.R script

In this subsection, I will explain the subfunctions that constitute the UK script, *04_universal_kriging.R, **the fourth and final step in the NPP UK model pipeline***. The script is modular and provides several key functions, each of which is essential for predicting and evaluating neuropathological distributions. I will break down each function and step involved in the process.

a. Preparing Spatial Data:

Before applying UK, it is essential to ensure that the data are in the correct spatial format. The *prepare_spatial_data* function is responsible for converting the data into a structure that the Kriging model can use. This step involves:

- Ensuring that all data columns are numeric: As Kriging requires numerical input, all covariates (e.g., neuropathology values, spatial coordinates) must be transformed into numeric values. This is especially important when dealing with data sources that may contain categorical values or non-numeric features.

- Extracting spatial coordinates (X, Y, Z): These coordinates represent the spatial locations of each data point, which are crucial for modeling spatial dependence. These coordinates are used to determine how the neuropathology values are related to the physical locations in the brain.

This function ensures that the data are correctly structured, with appropriate numeric values for both the response variable (e.g., neuropathology intensity) and the spatial coordinates.

b. Spatial Block Splitting:

In my approach to deploying UK, cross-validation is a crucial step for evaluating the model's performance. Instead of randomly splitting the data, spatial block splitting is used to partition the data into non-overlapping spatial blocks. The *create_spatial_blocks* function divides the data into spatial blocks based on the spatial coordinates (X, Y, Z) of the data points.

This method of splitting is crucial for geostatistical models**,** such as Kriging, as it ensures that the model's predictions are validated using spatially independent data. The blocks are used to assess how well the model generalizes by training the model on one set of blocks and testing it on others, preventing overfitting due to spatial autocorrelation. In NPP's case, the data correspond to a three-dimensional volume that can have varying shapes based on the tissue sampling protocol, the participant, and the region of interest (ROI). The tissue volume could be divided along any of the three axes (X, Y, Z). However, this choice has limitations. Splitting along the axes would likely lead to uneven block sizes based on the volume's geometry and orientation. Additionally, by choosing an axis to split, the UK model will only be applied to nearly two-dimensional objects, thus not utilizing the potential of deploying UK in a three-dimensional brain volume.

To account for these limitations, I decided to split the three-dimensional tissue volume into contiguous blocks of three dimensions. This was accomplished using a clustering algorithm, specifically K-means clustering. This algorithm splits the data into a user-defined number of blocks (e.g., 100 splits, 64 splits) and assigns each voxel a block number, which is then added to the *XXXX_uk_data.csv* file created earlier. These block assignments are then used to validate the model using a leave-one-out cross-validation approach that is described later in this section.

c. Loading UK Configuration

The next step is to load the participant-specific Universal Kriging configuration settings, which define key parameters for the Kriging process. The *Load_uk_config* function loads these parameters, found in the participant-specific UK configuration file (e.g., *XXXX_uk_config.txt*), including the following:

- Variogram Model: Specifies the theoretical model used to represent spatial autocorrelation in the data (e.g., spherical, exponential, gaussian). The spherical model is often chosen for neuroimaging data due to its ability to model spatial correlation at shorter distances.

- Range: The distance at which spatial correlation decays to near-zero. It defines the spatial extent over which the data points are expected to be correlated. (Figure 5.6.2)

- Nugget: The small-scale variance or noise that cannot be explained by spatial correlation, typically representing measurement error or short-range variability. (Figure 5.6.2)

- Partial Sill: Represents the portion of the total variance in the data that is explained by spatial autocorrelation. (Figure 5.6.2)

- Cutoff: The maximum distance used in the variogram calculation, limiting how far the model will consider spatial correlations.

- Width: The bin width for calculating the variogram.

- Grid Size: Defines the resolution of the prediction grid, determining the spacing between points where predictions will be made.

- Nmax: The maximum number of neighboring points used in the Kriging prediction for each point. (Figure 5.6.3)

- Nmin: The minimum number of neighbors required to make a valid prediction at a given location. (Figure 5.6.3)

- Maxdist: The maximum distance between points considered for Kriging predictions. (Figure 5.6.3)

- N_splits: Defines how many spatial splits (blocks) will be used for cross-validation. This parameter determines how the data will be partitioned for model evaluation.

- Cv_folds: The number of folds used in cross-validation. This determines how the spatial blocks will be divided during validation to assess the model's generalizability.

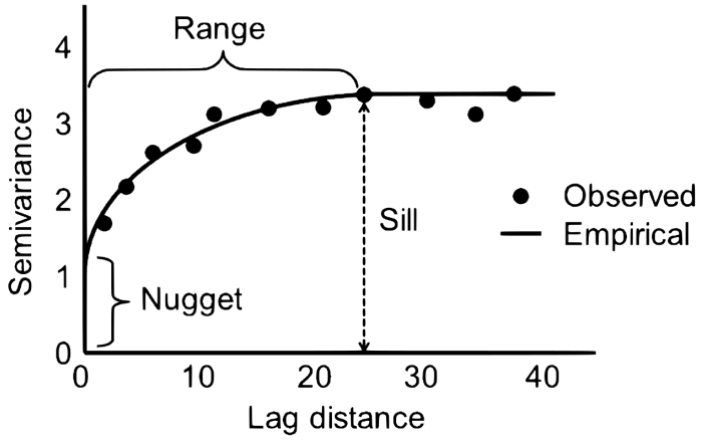


Figure 5.6.2: *A typical example of a semi-variogram showing different components*[95].

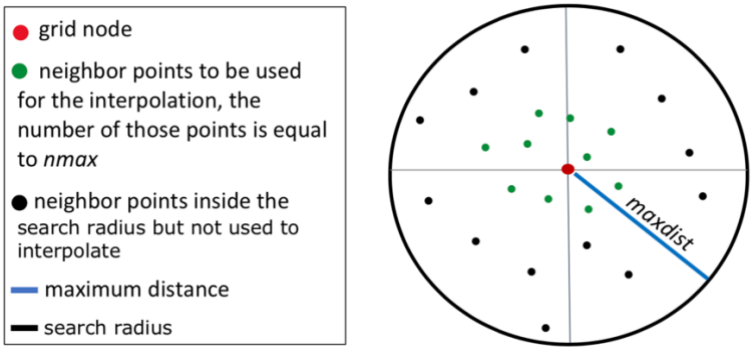


Figure 5.6.3: *An example depicting how nmax, nmin, and maxdist are used to tune the spatial variogram and the UK model.* [96]

d. Initial Variogram Plotting:

The variogram is a key concept in Kriging, representing the spatial correlation of data. The empirical variogram is computed from the observed data and visualized to assess the degree of spatial autocorrelation at different distances between data points.

The *plot_initial_variogram* function calculates the empirical variogram and generates a plot that shows how the semivariance (γ) differs as a function of distance between voxel pairs (X, Y, Z). The plot provides valuable insights into the spatial structure of the data, helping to determine whether a spherical, exponential, or gaussian model would be most suitable for fitting the variogram.

The empirical variogram plot (Figure 5.6.1) is saved as a PNG file for visual inspection, ensuring that the spatial correlation is well-captured and aligns with expectations. At this stage, the variogram is fitted using all the data, as the goal is to determine if the variogram

model specified through the configuration parameters is a reasonable fit for this set of data. If not, the configuration file is updated automatically to account for the appropriate model.

e. Fitting the Variogram:

Once the empirical variogram is generated, it is fitted to a theoretical variogram model. This step is performed using the *gstat package* in R, which provides the functionality to fit various types of variogram models to the data. The *fit_variogram* function takes the empirical variogram and fits it to a chosen spherical, exponential, or Gaussian model.

For example, suppose the empirical variogram suggests a spherical model. In that case, the function will fit the variogram to the spherical model and estimate the key parameters, including the nugget, partial sill, and range, that are used for UK model in the next step. These parameters are calculated for each time a variogram is fitted, even for the leave-one-out cross-validation step.

f. Fitting the Kriging Model:

Once the variogram is fitted, the next step is to apply the UK model. The model applied through the *perform_kriging* function incorporates both spatial autocorrelation (captured by the variogram) and external covariates (primary predictors), which represent the spatial drift. This function uses the formula described earlier in the section, the spatial coordinates, the model coordinates, and the variogram parameters to perform universal kriging.

g. Prediction Grid Creation:

A prediction grid is created over the region of interest (e.g., MFG) to predict neuropathology at unsampled locations. The *create_prediction_grid* function generates a 3D grid that covers the spatial domain, with the grid size defined in the configuration file. This grid is used as the basis for generating predictions across unsampled regions of the brain.

Once the grid is created, the UK model is applied to predict the neuropathology intensity at each grid location, using the fitted variogram model and the external covariates. These predictions provide a spatially continuous map of the neuropathology distribution.

h. Cross-Validation:

Cross-validation is an essential step in evaluating the performance of the UK model. As discussed earlier, instead of using random data splits, spatial block splitting ensures that

the training and testing data are spatially independent, preventing spatial autocorrelation from influencing model performance. Using these splits, the model is trained on a subset of blocks and validated on the remaining block. Performance metrics such as RMSE, R^2, and MAE are calculated to evaluate the model's ability to generalize to new spatial regions. I used various splits of 16, 32, 64, and 100 to assess the maximum prediction distance that is generalizable using this modeling approach. The results will be discussed in the next chapter.

5.7 Saving Results and Visualizations

After the Universal Kriging (UK) process has been completed, it is essential to save both the predictions and the associated diagnostic plots for future analysis, review, and visualization. This section outlines how the predictions and visualizations are saved, ensuring that all relevant data and results are securely stored in accessible formats. Additionally, it explains the role of logging in tracking progress and maintaining a record of the modeling process.

5.7.1 Saving predictions

Once the UK model has made predictions for the spatial distribution of neuropathology across the ROI, these predictions need to be saved in a suitable format for further analysis and interpretation. The predictions are stored in both CSV and NIfTI formats, providing flexibility in accessing and visualizing the results.

- CSV Format: The predictions and associated metrics (e.g., predicted values, residuals, spatial coordinates) are saved in a CSV file, providing an easily accessible table format for subsequent analysis. This allows users to review the predicted neuropathology values across different spatial coordinates (X, Y, Z) and analyze them using standard tools such as R, Excel, or Python.
    - For example, for each participant, a file such as *6966_cv_16_splits.csv* is saved. This file contains the predicted neuropathology intensity at each voxel location (X, Y, Z), the observed intensity, the predicted value (and any other relevant metrics (e.g., variance).
- NIfTI Format: To enable visualization in neuroimaging software, the blocks are also saved in NIfTI format. NIfTI files provide a standard format for storing 3D volumetric data, which can then be used to visualize the predicted neuropathology distribution as spatial maps.

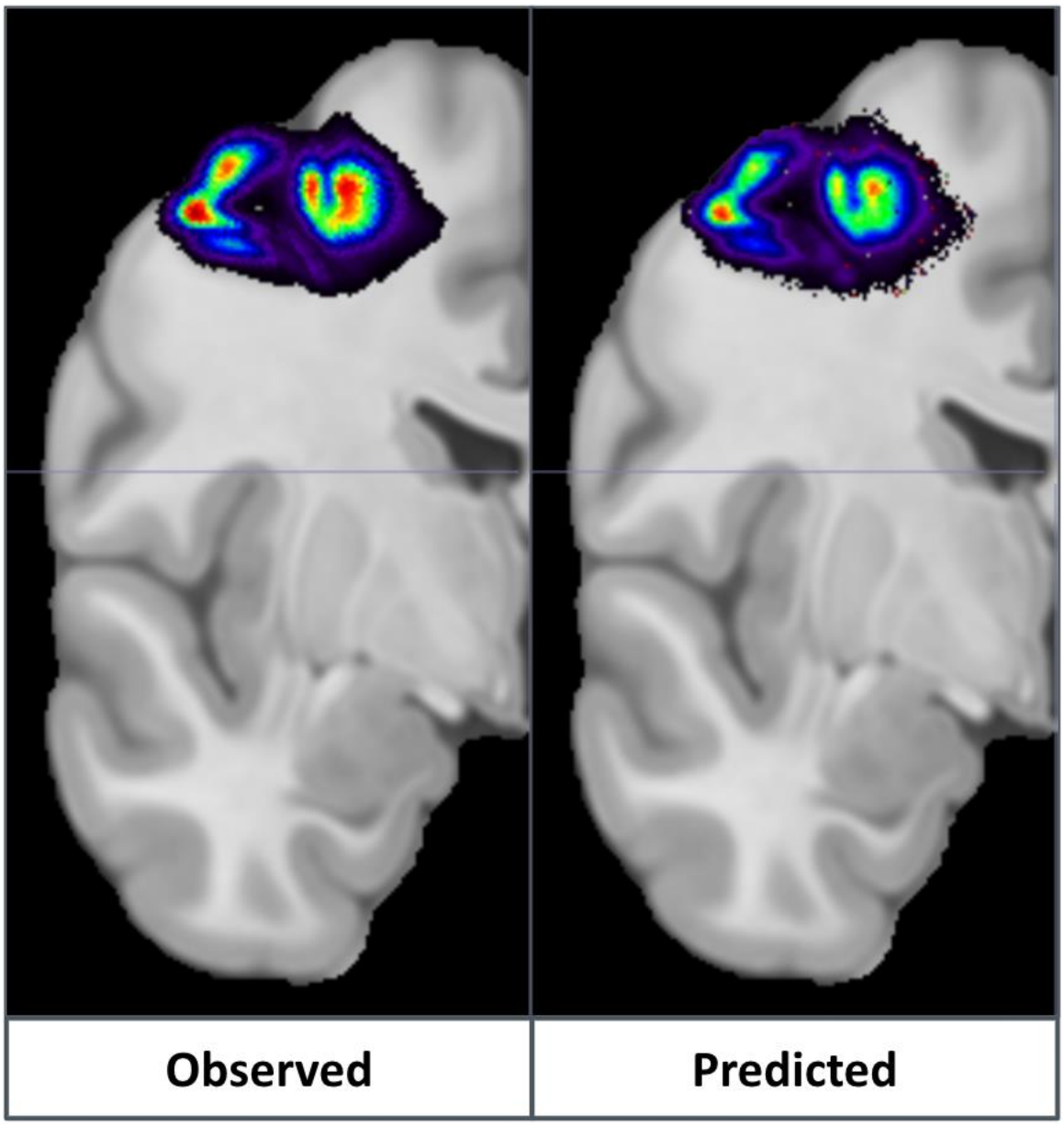


Figure 5.7.1: *NIfTI-based visualization of observed and model-predicted tau pathology (% AT8 positive for p-tau) in MFG for a participant.*

By saving predictions in both CSV and NIfTI formats, the process ensures that the results are stored in versatile formats that can be easily shared, analyzed, and visualized in a variety of tools.

5.7.2 Visualization Directory

Diagnostic plots help assess the model's accuracy and fit. These plots provide a visual representation of how well the predictions align with the observed values, as well as how the residuals (errors) are distributed.

The *plot_obs_pred_residuals* function generates key diagnostic plots for each spatial block during cross-validation and stores them in a dedicated visualization directory. These plots are saved as PNG files for easy viewing and future reference.

The types of diagnostic plots generated include:

- Observed vs Predicted Plot: This plot shows the relationship between the observed neuropathology values and the predicted values. It helps to visually assess whether the model is underestimating or overestimating neuropathology intensity.

- Predicted vs Residual Plot: This plot displays the predicted values versus the residuals (the difference between the observed and predicted values), providing insight into whether the model's errors are randomly distributed or if there are systematic biases.

These diagnostic plots are saved in the plots subdirectory, which is created within the output directory for each participant. For example, for participant 6966, the plots will be stored in a directory such as:

```
bash
Copy
/output/participant_6966/universal_kriging/plots/
```

The specific plot filenames might be:

- *observed_vs_predicted_6966.png*
- *predicted_vs_residuals_6966.png*
- *residuals_distribution_6966.png*

This directory allows researchers to easily access and review the diagnostic plots for each participant, ensuring that any issues with model performance can be quickly identified and addressed.

5.7.3 Logging and Outputs

Logging plays a crucial role in tracking the progress, storing detailed information about the model's execution, and debugging problems. The *logging* ***package*** is used throughout the *04_universal_kriging.R* script to record significant events and any issues that arise during model fitting, prediction, and validation.

Key aspects of logging include:

- Model Setup and Configuration: Logs are created when UK configuration files are loaded, ensuring that participant-specific settings (e.g., variogram model, n_splits, nmax, nmin) are applied correctly.
- Model Fitting and Variogram Fitting: Logs are generated when the variogram is calculated and fitted to the data, including the variogram model used and the fitting parameters (e.g., nugget, sill, range).
- Prediction and Cross-Validation: Logs are created during the prediction process for each spatial block, tracking the prediction process and the computation of cross-validation results. Information about performance metrics (e.g., RMSE, R^2) and outlier analysis is recorded.
- Error Handling: If any errors occur (e.g., in variogram fitting or prediction), the log files capture the details of the error message, allowing for easier troubleshooting and resolution.

The logs are saved in a central log file (e.g., *pipeline.log*), which can be reviewed to track the model's progress, identify any issues, and ensure that the steps are executed as intended. For example, the log file for participant 6966 might contain entries like:

```
[INFO] Loading Universal Kriging configuration for participant 6966
[INFO] Fitting variogram model: Spherical
[INFO] Fitting completed successfully, nugget=0.000, sill=0.860, range=11.036
[INFO] Starting prediction for spatial block 1 of 16
[INFO] Cross-validation results: RMSE=0.465, R²=0.643
```

By providing detailed logs throughout the process, researchers can review the progress of the UK model, monitor key stages of the analysis, and efficiently resolve any issues that arise.

Chapter 6. Results and analysis

This chapter presents the results of the NPP UK model pipeline, with a focused analysis of participant 6966, followed by a broader analysis of the results for all 10 participants in the study. The decision to begin with a deep dive into the results for participant 6966 is driven by the need to provide a clear illustration of the analysis pipeline and its underlying methods. By detailing the process for a single participant, I can guide the reader through a step-by-step procedure, from model setup to prediction and validation, ensuring that the methodology is fully understood before scaling it to the entire cohort. This approach provides a detailed explanation of how the Universal Kriging model operates, how cross-validation is conducted, and how model performance is evaluated using diagnostic plots and metrics. Once the individual analysis is explained, the reader can understand how these methods were applied across all participants, making the comparison of results across the whole group more insightful. The chapter will also delve into the analysis of primary predictors across all 10 participants, highlighting key findings related to variables that have the most influence in predicting neuropathology distribution in this group of participants for the middle frontal gyrus (MFG) region in the brain.

6.1 Results for participant 6966

The analysis and workflow for participant 6966 have been discussed in chapters 3, 4, and 5 to illustrate the setup of the NPP pipeline and its components. In this section, I will briefly summarize these components and provide a seamless transition from the pipeline's inception, starting with the quantitative AT8 positive-stained percentage values derived from Halo, their mapping to a standard brain volume, through preprocessing, and culminating in primary predictor analysis and the Universal Kriging model. For a more detailed breakdown of these components, please refer to chapters 3 and 5. The primary focus of this chapter is on the primary predictors, the Universal Kriging analysis, and the results.

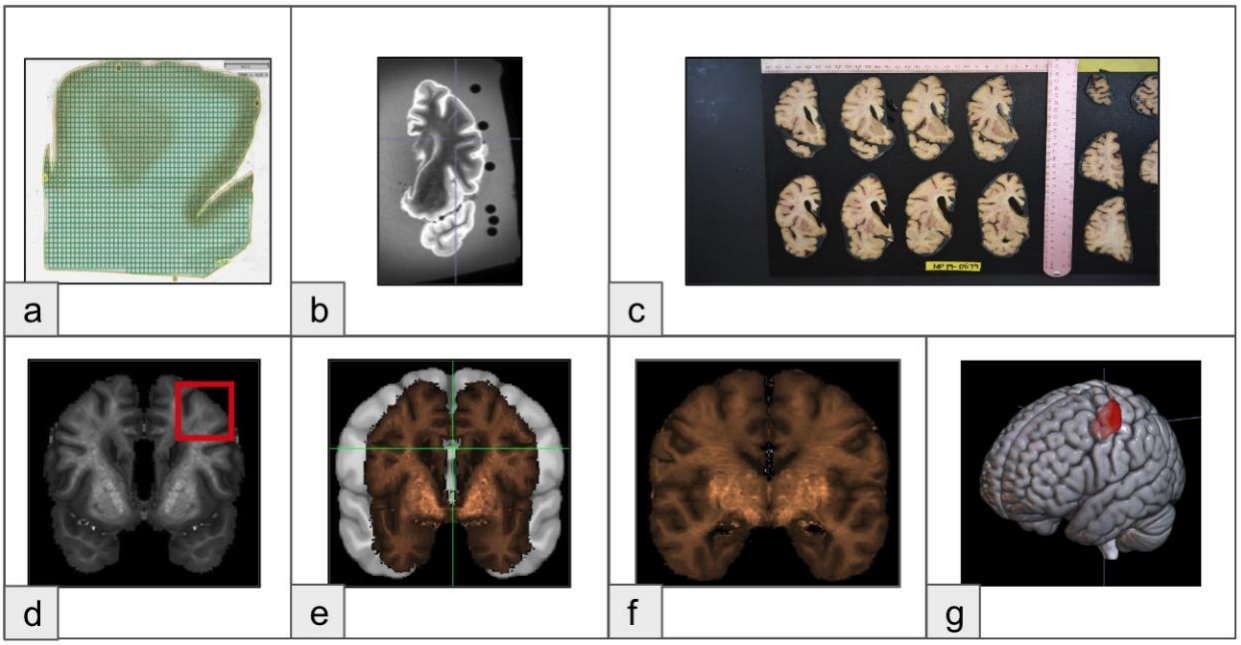


Figure 6.1.1: *In this figure, panels a, b, and c depict the input components for participant 6966, wherein panel a shows the quantitative halo extract, panel b depicts an ex vivo Flair image, and panel c shows the gross sliced tissue for the participant. Panels d, e, f, and g show the pre-processing steps for participant 6966. First, the MFG tissue sampling location is identified (panel d) using ex vivo flair and gross tissue images by a neuroanatomist. Next, ex vivo flair images for participant 6966 are co-registered to the MNI template in panels e and f. Finally, using the ANT transformation matrix from ex vivo to MNI registration, we map the co-located tissue sample to the MNI brain volume, extracting voxels that are further processed through I-BIS in panel g.*

After pre-processing participant 6966's quantitative neuropathology values and tissue location (Figure 6.1.1), voxels (n = 44,891) representing the tissue location and volume are used as inputs for the I-BIS pipeline. In this pipeline, covariates are calculated for each voxel for various variables of interest, as outlined in Chapter 3. These covariates include three categories. First, 'buffer zone' or BZ covariates, representing the density of variables around each voxel within a sphere of radius ranging from 1 mm to 15 mm. Second, distance-based or 'EDT' covariates represent the shortest absolute distance from each voxel to the nearest voxel corresponding to a variable. And the third, the variable values at each voxel location. Once these covariates are calculated, they are transformed and scaled using Z-scores. Finally, these covariates are used in a subsequent elastic net model for dimensionality reduction and identifying primary predictors after factor analysis, thereby removing multicollinearity.

The Elastic Net model is fitted to each participant's AT8 values (% positively stained by AT8 antibody for phosphor tau). As an example, I first applied a cross-validated elastic net to screen predictors for participant 6966. After removing features with >90% zeros

across 44,891 voxels, I modeled 398 covariates; cross-validation selected a ridge solution ($\alpha = 0.00$) with $\lambda = 0.0408$, consistent with high collinearity among anatomically related predictors. As summarized below, train–test metrics are closely aligned (e.g., $R^2 \approx 0.62$; RMSE ≈ 0.61), indicating limited overfitting and stable out-of-sample performance. Practically, this means the elastic-net stage captures a meaningful portion of variance while yielding a compact, interpretable set of 30 primary predictors (a mix of EDT and BZ variables) to serve as the universal-kriging drift, reducing noise from sparse features and improving downstream spatial modeling.

Elastic-net screening and performance for participant 6966:

Voxels analyzed	44,891
Covariates modeled (post-filter)	398
Selected regularization	$\alpha = 0.00$ (ridge), $\lambda = 0.0408$
Primary predictors	30 (EDT + BZ)

Predictive performance (participant 6966):

Metric	Train	Test
R^2	0.6340	0.6206
MSE	0.3664	0.3776
RMSE	0.6053	0.6145
MAE	0.4732	0.4801

Further factor analysis helped narrow down the list of primary predictors to 14. To narrow down the list, a combination of factors was considered. First, amongst any BZ covariates of the same variable, but different radii, the predictor with the larger radius was dropped. Next, a manual inspection of the correlation matrix (Figure 5.5.1) and the dendrogram plot (Figure 5.5.2) for participant 6966 helped remove multicollinearity in the set of predictors. Finally, the scree plot (Figure 5.5.3) and factor analysis results were taken into consideration. Given the subjective nature of manually inspecting these predictors, the final list of primary predictors is modifiable based on the use case. The list could have included 12, 13, 15, or 16 predictors; however, by testing various combinations, a list of 14 predictors was found to be the most suitable for participant 6966. It should also be noted that while the individual covariates have a slight influence on the model, they do not impact the spatial autocorrelation obtained by fitting a variogram in subsequent steps.

6.1.1 Primary predictors

The top 30 predictors, based on absolute coefficient values, are listed below. Amongst these 30 predictors, 21 belonged to the 'buffer zone' or BZ category, and the remaining nine belonged to the 'Euclidean Distance Transforms' or EDT (shortest absolute distance) category. Based on atlas categorization, 15 (50%) were derived from the white matter tracts atlas HCP1065[97], 11 (36.7%) were derived from the functional networks atlas Yeo17[98], two (6.7%) were derived from the structural MNI 2009b atlas[15], and one (3.4%) each from the vascular atlas Mouches & Forkert[69] and the regional parcellation based CerebrA[85] atlas.

1. v1_edt_dxFrom_CPT_F_R (Coefficient: 0.3242)
2. v1_edt_dxFrom_Yeo_5 (Coefficient: -0.3091)
3. bz_TR_S_R_7.5 (Coefficient: -0.3064)
4. bz_SLF2_R_7.5 (Coefficient: -0.2662)
5. bz_CPT_F_R_7.5 (Coefficient: 0.2577)
6. v1_edt_dxFrom_TR_A_L (Coefficient: 0.2440)
7. v1_edt_dxFrom_Yeo_7 (Coefficient: 0.2358)
8. bz_Yeo_12_15 (Coefficient: -0.2235)
9. v1_edt_dxFrom_Yeo_3 (Coefficient: -0.2234)
10. bz_CPT_F_R_10 (Coefficient: 0.2190)
11. bz_TR_S_R_10 (Coefficient: -0.2172)
12. bz_Yeo_13_7.5 (Coefficient: -0.2120)
13. v1_edt_dxFrom_Yeo_13 (Coefficient: -0.1995)
14. bz_FAT_R_15 (Coefficient: -0.1934)
15. bz_CC_10 (Coefficient: 0.1902)
16. bz_Yeo_13_5 (Coefficient: -0.1892)
17. bz_SLF2_R_10 (Coefficient: 0.1853)
18. bz_CPT_F_R_5 (Coefficient: 0.1808)
19. v1_edt_dxFrom_IFOF_R (Coefficient: 0.1755)
20. bz_CSF_bin_MFG_7.5 (Coefficient: 0.1713)
21. bz_CC_15 (Coefficient: 0.1699)
22. bz_Yeo_3_5 (Coefficient: 0.1675)
23. v1_edt_dxFrom_WM_MFG (Coefficient: -0.1672)
24. bz_Yeo_3_7.5 (Coefficient: 0.1668)
25. bz_AF_R_7.5 (Coefficient: 0.1667)
26. bz_vesRad_thr_0.5_MFG_15 (Coefficient: 0.1648)
27. bz_FAT_R_7.5 (Coefficient: 0.1636)
28. bz_mni_icbm152_CerebrA_09c_38_15 (Coefficient: -0.1592)
29. bz_Yeo_13_15 (Coefficient: 0.1589)
30. v1_edt_dxFrom_Yeo_6 (Coefficient: -0.1583)

These 30 predictors were further narrowed down to a list of 14 (Table 6.1.1) using a combination of correlation analysis, factor analysis, and manual selection. As an example of manual selection, three BZ predictors for CPT_F_R with radii of 5, 7.5, and 10 mm were selected in the top 30 predictor list. Only bz_CPT_F_R_5 was shortlisted for the universal kriging model. Finally, the shortlist of 14 primary predictors consists of five from white matter tracts (HCP1065[97]), four from functional networks (Yeo17[98]), two from tissue type (MNI 2009b[15]), one vasculature predictor (Mouches & Forkert[69]), and one region (CerebrA[85])predictor. It is possible to draw anatomical insights from this list of predictors for participant 6966; however, I decided to conduct a group analysis of these predictors, along with those from the other nine participants, as discussed further in Section 6.3 of this chapter.

Variable label	Atlas	Anatomical name	Variable description
bz_Yeo_13_5	Yeo17[98]	Control B network	Density of Yeo 13 network within a buffer zone with radius 5 mm
bz_vesRad_thr_0.5 MFG_15	Mouches & Forkert[69]	Vessel Radius 0.5mm or lower	Density of vessels less than 0.5 mm in radius within a buffer zone with radius 15 mm
bz_mni_icbm152 CerebrA_09c_38_15	CerebrA[85]	Superior Frontal region right	Density of CerebrA 38 region within a buffer zone with radius 15 mm
bz_Yeo_12_15	Yeo17[98]	Control A network	Density of Yeo 12 network within a buffer zone with radius 15 mm
bz_CSF_bin_MFG_7.5	MNI152 2009c[15]	Cerebrospinal Fluid	Density of CSF within a buffer zone with radius 7.5 mm
v1_edt_dxFrom_Yeo_6	Yeo17[98]	Dorsal Attention B network	Shortest absolute distance to Yeo 6 network
v1_edt_dxFrom_TR_A_L	HCP1065[97]	Thalamic Radiation Anterior Tract Left	Shortest absolute distance to TR_A_L white matter tract
v1_edt_dxFrom_IFOF_R	HCP1065[97]	Inferior Fronto Occipital Fasciculus Tract Right	Shortest absolute distance to IFOF_R white matter tract
bz_SLF2_R_7.5	HCP1065[97]	Superior Longitudinal Fasciculus2 tract Right	Density of SLF2_R tract within a buffer zone with radius 7.5 mm
v1_edt_dxFrom WM_MFG	MNI152 2009c[15]	Inferior Lateral Ventricle Right	Shortest absolute distance to CerebrA 5 region
v1_edt_dxFrom_Yeo_3	Yeo17[98]	Somatomotor A network	Shortest absolute distance to Yeo 3 network

v1_edt_dxFrom_Yeo_13	Yeo17[98]	Control B network	Shortest absolute distance to Yeo 13 network
bz_CPT_F_R_5	HCP1065[97]	Corticopontine Tract Frontal Right	Density of CPT_F_R tract within a buffer zone with radius 5 mm
bz_AF_R_7.5	HCP1065[97]	Arcuate Fasciculus tract Right	Density of AF_R tract within a buffer zone with radius 7.5 mm

Table 6.1.1: *This table lists 14 primary predictors selected for the universal kriging modelling step for participant 6966. Selected variables, corresponding atlases, variables' anatomical names, and their description in the context of NPP are listed in the table.*

6.1.2 Universal Kriging (UK)

UK was applied to participant 6966 as the final step in the NPP pipeline, to generate spatially continuous predictions of % AT8 positive values across the middle frontal gyrus (MFG). The trend model included fourteen covariates selected through Elastic Net regularization and subsequent factor analysis. Spatial coordinates (X, Y, Z) were included in the variogram detrending component to model residual spatial structure. The empirical variogram (Figure 5.6.1) was fit using a spherical model with optimized parameters: range = 10.88 mm, partial sill = 0.791, and nugget = 0. These values suggest substantial spatial autocorrelation beyond the variance explained by the covariates, justifying the need for spatial modeling.

Kriging Formula: % AT8 positive value ~ bz_SLF2_R_7.5 + v1_edt_dxFrom_TR_A_L + bz_Yeo_12_15 + v1_edt_dxFrom_Yeo_3 + v1_edt_dxFrom_Yeo_13 + bz_Yeo_13_5 + bz_CPT_F_R_5 + v1_edt_dxFrom_IFOF_R + bz_CSF_bin_MFG_7.5 + v1_edt_dxFrom_WM_MFG + bz_AF_R_7.5 + bz_vesRad_thr_0.5_MFG_15 + bz_mni_icbm152_CerebrA_09c_38_15 + v1_edt_dxFrom_Yeo_6

Variogram Formula: % AT8 positive value ~ X + Y + Z

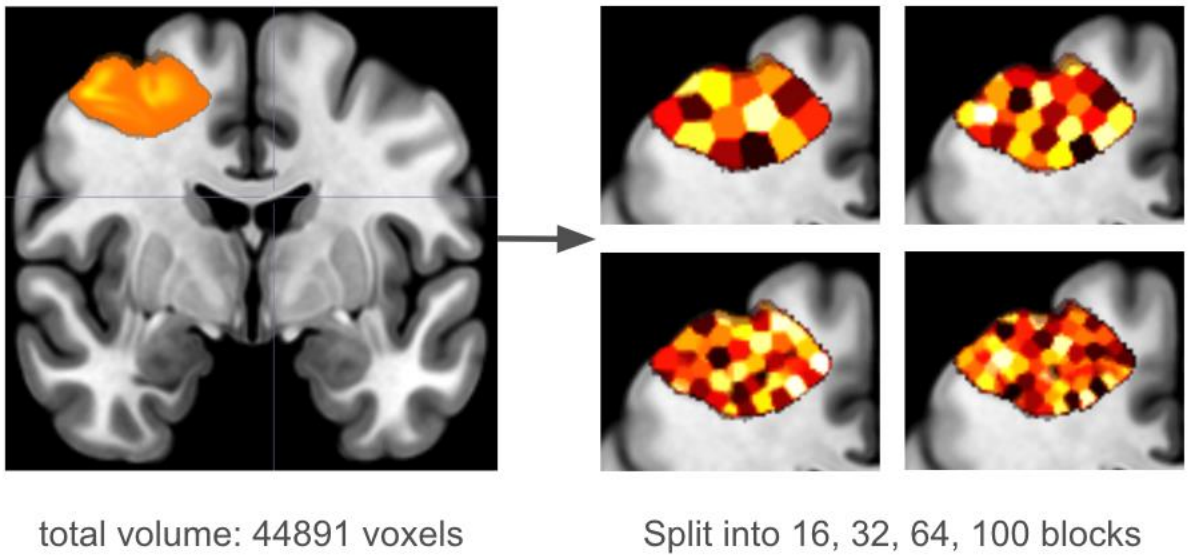


Figure 6.1.2: *This figure shows block splits for participant 6966's tissue sample. The splits were generated using the K-means clustering algorithm to produce continuous three-dimensional divisions with 16, 32, 64, and 100 blocks. The purpose of testing various split counts was to determine the best model fit for the UK.*

As described in Chapter 5, the model's generalizability was evaluated by assessing kriging predictions using block-wise (Figure 6.1.2) spatial cross-validation at increasing granularities (16, 32, 64, and 100 spatial blocks). These blocks were generated via K-Means clustering on voxel coordinates, ensuring spatially non-contiguous training and testing sets. The number of voxels per block varied across splits, and an average for each split level is shown in Figure 6.1.3. In this approach, at each split level, one block was selected as the test set, and the remaining blocks were designated as the training set.

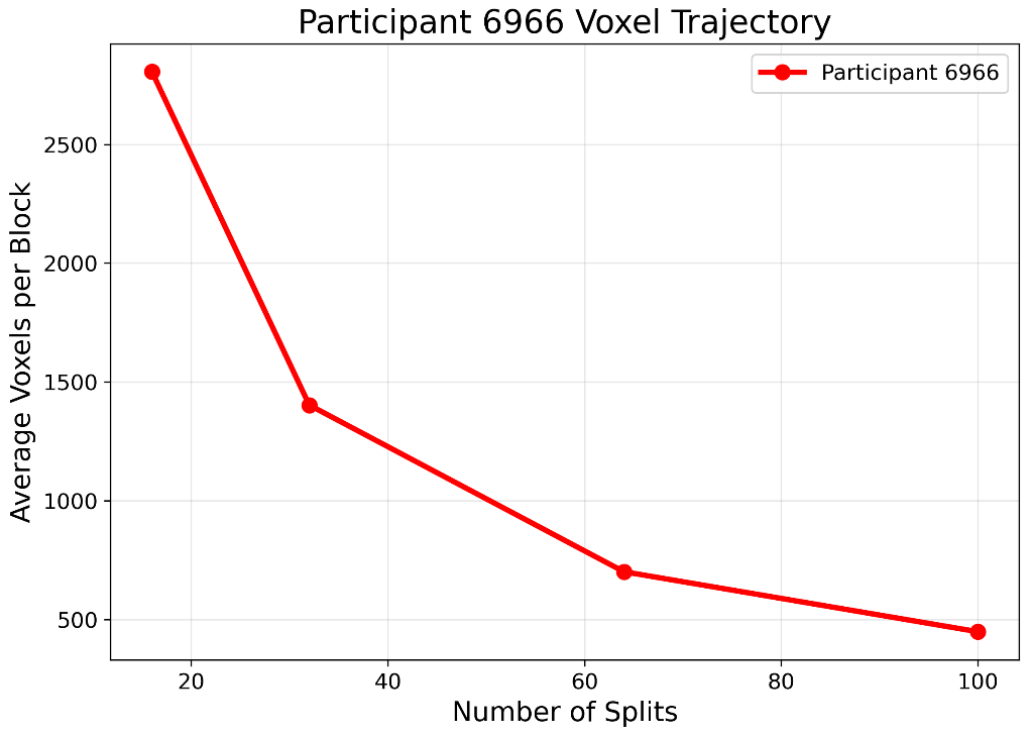


Figure 6.1.3: *This figure shows the average number of voxels per block across split levels for participant 6966. The block sizes slightly varied across each split level, given that K-means clustering yields slightly uneven block sizes.*

To ensure robust results, two exclusion criteria were used. In the first step, outlier exclusion (defined as predictions exceeding two standard deviations from the mean) was applied to all individual voxel-level predictions across all split levels and individual blocks. This affected ~3% of total predictions across block folds. Additionally, any predicted blocks across all split levels with R^2 values outside the range of 0 to 1 were excluded. As shown in Figure 6.1.4, at least 84% of the blocks were included in the analysis for participant 6966, for both 64- and 100-block splits. To further emphasize the significance of these exclusion criteria, Figure 6.1.5 compares RMSE and R^2 values across all four split levels, including all blocks and only valid blocks. Upon including blocks with valid R^2 values, there is a considerable change in the average split-level RMSE values; however, the average R^2 values stabilize. These exclusions ensured that extreme mismatches, potentially driven by noisy histological artifacts, edge effects, or complex matrix multiplication errors, did not distort the evaluation. These effects are discussed further in Chapter 7.

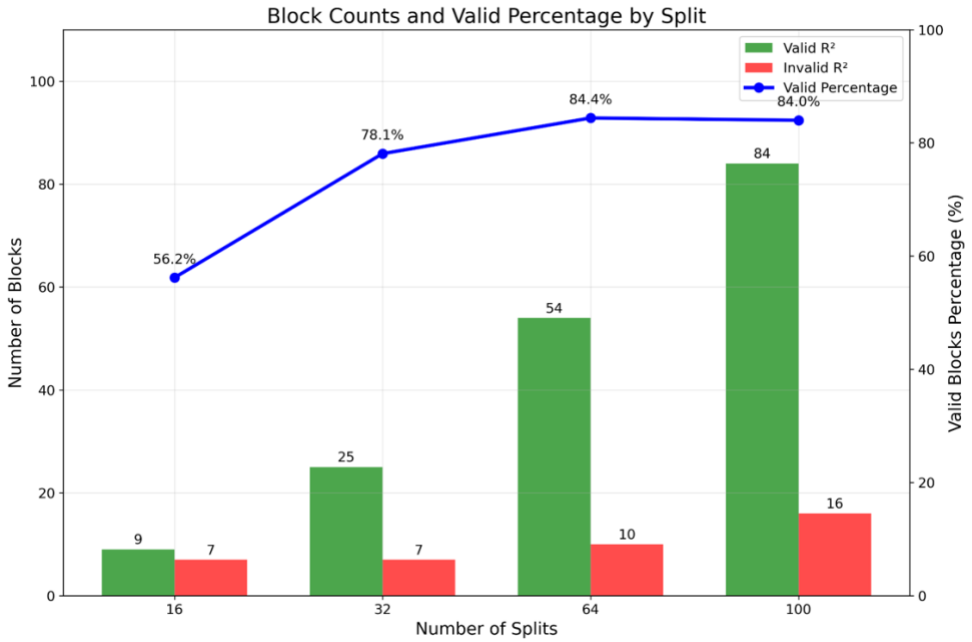


***Figure 6.1.4**: This figure illustrates, based on R² values, the total number of blocks included in the results for each split level, represented in green. The number of excluded blocks is shown using red bars. Additionally, the percentage of blocks included at each split level for participant 6966 is illustrated by the blue line.*

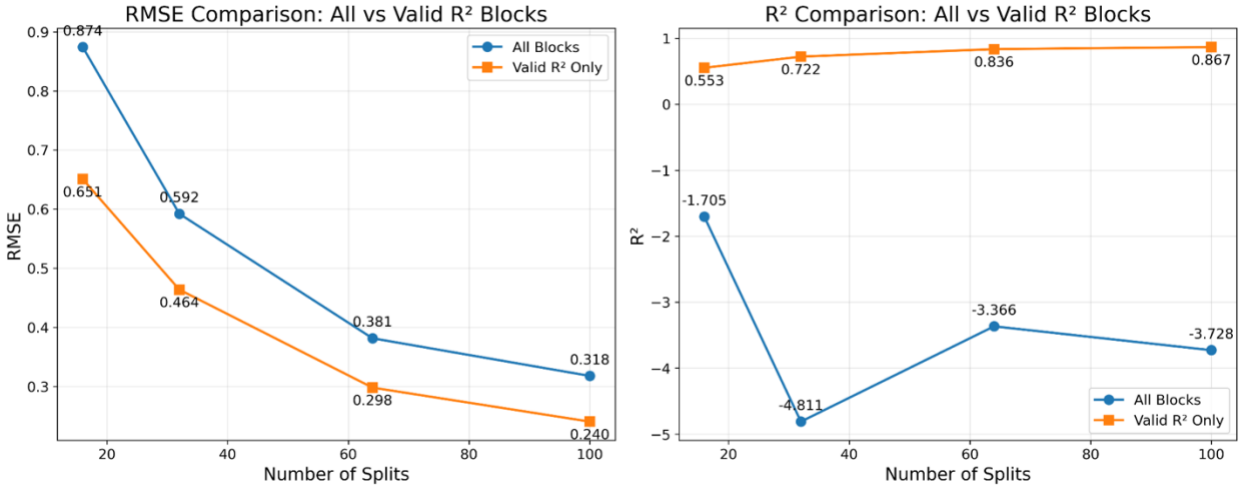


***Figure 6.1.5**: This figure compares average RMSE and R2 values for all four split levels for participant 6966 across two criteria, the first of which includes all blocks, and the second that excludes any blocks that have R2 values outside the range 0 to 1.*

To evaluate model performance, RMSE, MAE, and R² metrics are used to compare model fit across the four split levels. As shown in Figure 6.1.6, the test prediction accuracy improved significantly with finer block resolutions. At a 100-block resolution, the model

achieved an RMSE of 0.2403, an MAE of 0.0695, and an R^2 of 0.867, indicating a high level of agreement between the observed and predicted pathology intensities, with minimal residual bias. Even at 64-block resolution, the UK model maintained strong performance (RMSE = 0.298, R^2 = 0.838), though performance degraded at coarser resolutions (e.g., 32-block R^2 = 0.399), where blocks may have limited spatial autocorrelation ranges.

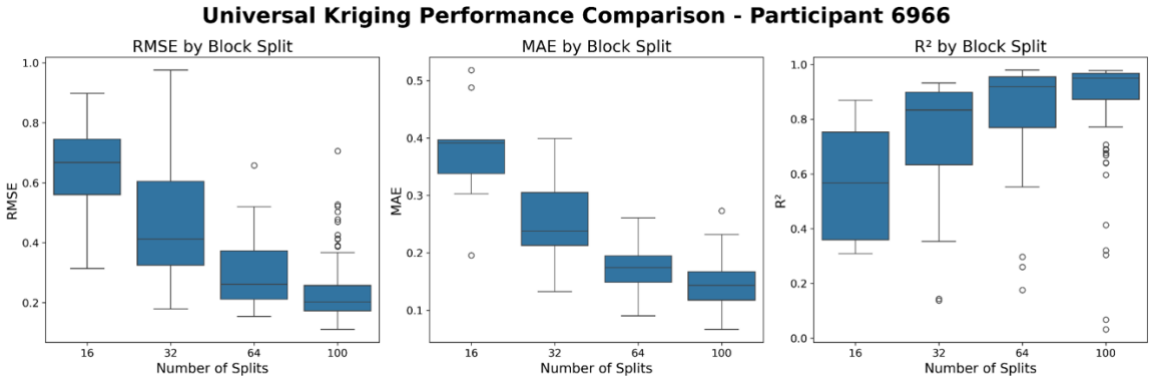


Figure 6.1.6: *This figure compares model metrics (RMSE, MAE, R2) across the four split levels for participant 6966. 100- and 64-block splits indicate a significantly better model performance, indicating a relationship between distance and spatial autocorrelation.*

To contextualize the UK performance, we compared results to two alternative regression baselines: (i) a linear regression model with the same 10 covariates used in the UK, and (ii) the Elastic Net model developed earlier in the pipeline. The linear regression model yielded an R^2 of 0.2889 and an RMSE of 0.843, underscoring its inability to capture spatially structured variance, even compared to the 16 block split UK model. In contrast, the Elastic Net model (which used a broader set of 30 predictors) achieved a test-set R^2 of 0.6206 (RMSE = 0.614) under 10-fold cross-validation.

These comparisons for participant 6966's data demonstrate that incorporating spatial autocorrelation via Universal Kriging yields substantial predictive gains over both regularized and non-regularized regression models. The high R^2 (>0.8) and low RMSE (<0.3) achieved by the UK for 64- and 100-block splits under spatial cross-validation indicate that spatial context, when correctly modeled, accounts for a significant proportion of pathology distribution not explained by anatomical or functional proximity alone. These findings are further confirmed by analyzing the data of the remaining nine participants, as described in the next section.

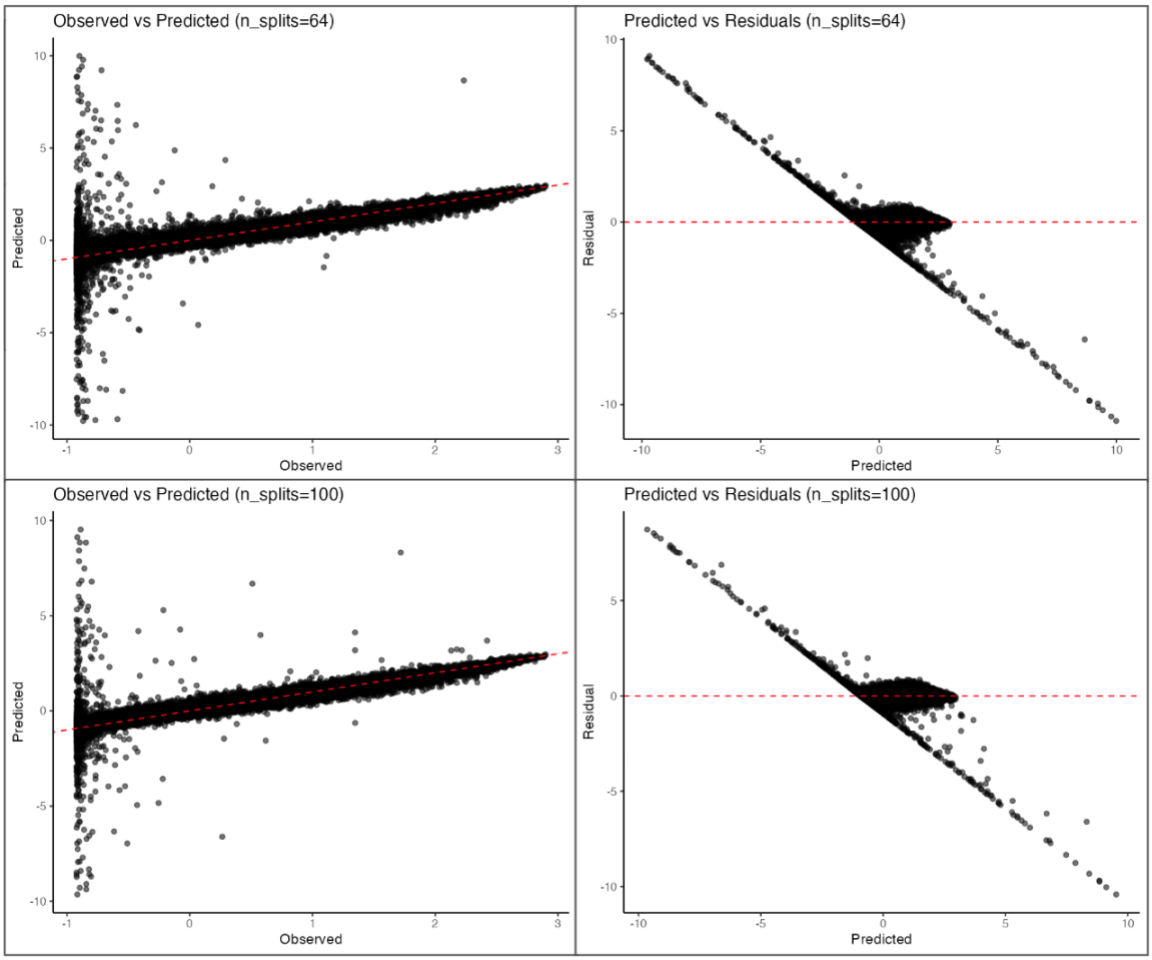


Figure 6.1.7: *In this figure, for participant 6966, observed vs predicted values are plotted on the left for n_splits = 64 and 100; Plots on the right compare predicted values and residuals for the same splits for participant 6966.*

6.2 All 10 participants modeled independently.

In the previous section, I described the analysis workflow and the results for one participant, participant 6966. Subsequently, I did the same analysis for the remaining nine participants. This section combines the results from all 10 participants to illustrate the efficacy and results of the NPP pipeline when applied to different participants. First, I discuss K-means clustering-based block split analysis across all participants, followed by the results of the Universal Kriging model.

6.2.1 Block split analysis

In Chapter 4, I mapped tau pathology (% AT8 antibody stained for phosphor-tau) for ten participants from the ACT study in the MFG region to the standard MNI brain volume.

Furthermore, I applied a 2mm Gaussian smoothing kernel, with a threshold of at least 1% positive AT8 value to account for positional variance and signal value when mapping from gross anatomy to ex vivo MRI to MNI space. This process yielded different tissue volumes and corresponding voxels for each participant. The number of voxels ranges from just 7,993 for participant 7144 to 97,007 for participant 7297. This variance in voxel numbers has implications for the models themselves. Hence, an analysis comparing voxel numbers and consequently split block volumes with respect to the number of primary predictors in the model was necessary.

Using Elastic net modeling and factor analysis, primary predictors for each participant were selected. The final number of primary predictors selected for the Universal Kriging model was based on a combination of factors, as described in section 6.1 for participant 6966. The number of selected primary predictors varies from 12 for participant 7157 to 18 for participants 7051, 7101, and 7144 (Figure 6.2.1). While eight of ten participants' voxel volumes and the number of primary predictors appear to follow a trend of approximately 4,000–6,000 voxels per significant predictor, there are two outliers. Participant 7101 with ~2257 voxels per significant predictor and participant 7144 with ~444 voxels per significant predictor.

Additionally, the number of voxels for each participant results in varying block sizes used in the leave-one-out cross-validation universal kriging model. Block sizes range from the smallest of 80 voxels for participant 7144 in 100 splits to 6063 voxels for participant 7297 in the 16-splits model. Block sizes here refer to the number of splits created out of the whole voxel volume for a participant. As shown in Figure 6.2.2, block sizes vary substantially for 16 splits; however, the absolute difference in voxels per split decreases as the split sizes increase to 64 and 100 blocks. This difference is important to note as I discuss results across participants and block-split sizes further in this section.

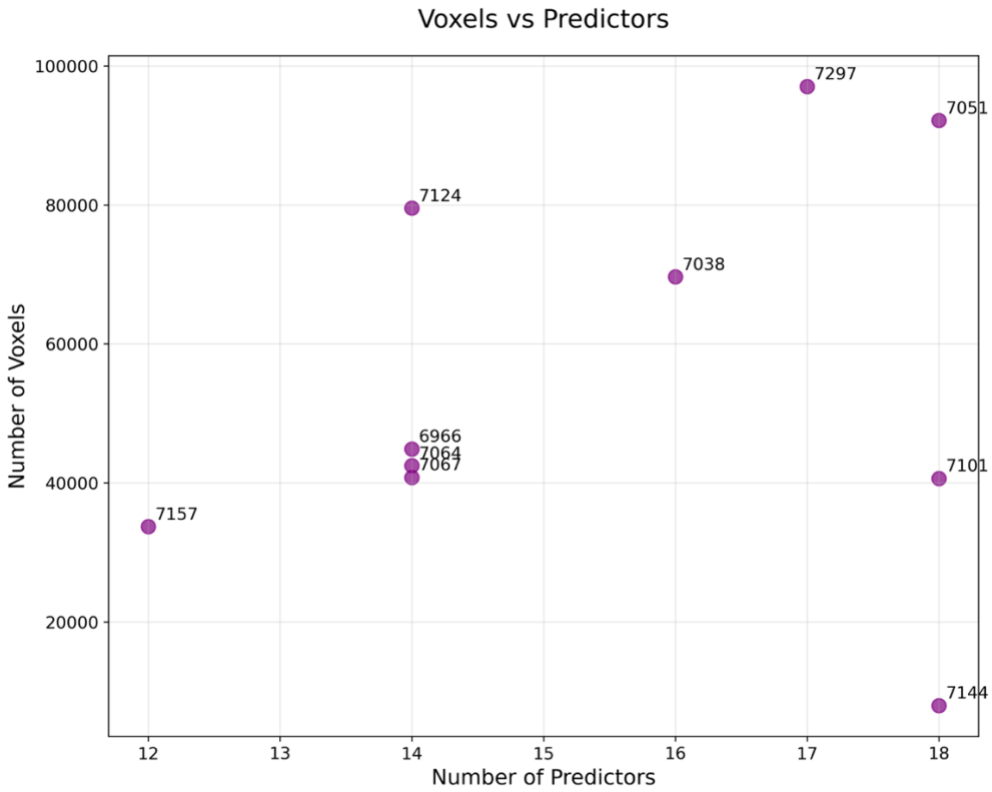


Figure 6.2.1: *This plot compares the number of voxels for each participant with the number of primary predictors included in the Universal Kriging model.*

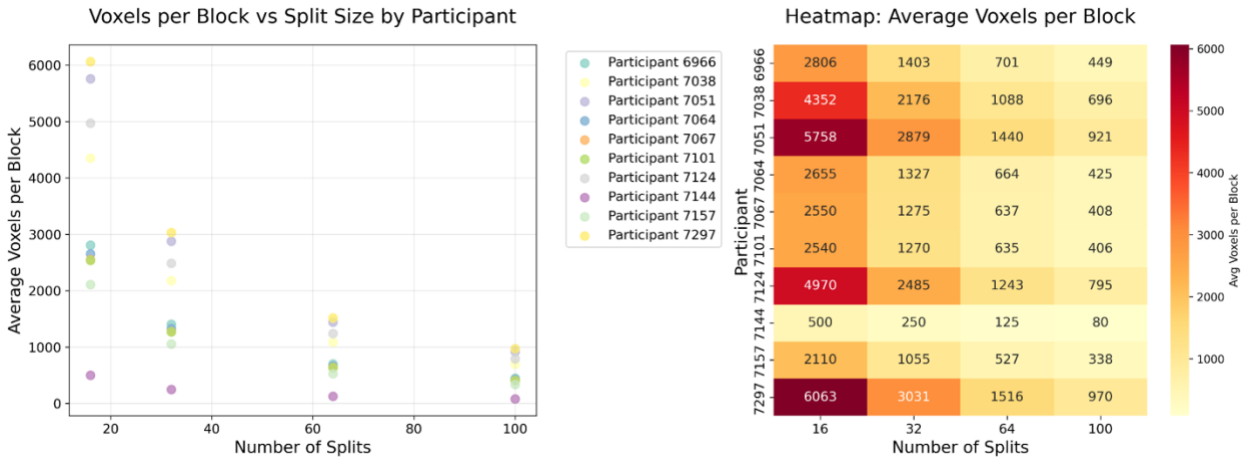


Figure 6.2.2*: These plots compare average numbers of voxels per block for each participant across the four split sizes of 16,32,64, and 100. The left panel directly compares average voxels per block per split size across the ten participants, and the panel on the right shows the same data as a heatmap.*

Further analysis revealed that the insufficiency of the number of voxels in each block and a lower number of voxels per significant predictor resulted in a higher number of excluded predictions and valid blocks ($R^2 < 0$ or $R^2 > 1$). These exclusion criteria are described in more detail in section 6.1. Nine out of ten participants had a block inclusion rate of 80% or higher for a split size of 100. Seven out of 10 participants had an inclusion rate of 90% or higher for split sizes of 64 and 100. The inclusion rate for participants 7144 and 7101 was lower than the others, primarily due to the voxel-to-significant predictor ratio. It is possible to test models for these participants with a lower number of primary predictors; however, for the scope of this dissertation, these test models were not deployed.

The block inclusion rate was much lower for split sizes of 32 and 16 blocks (Figure 6.2.3). The higher number of voxels per block for 16 and 32 splits compared to 64 and 100 splits suggests a model limitation in handling volume edges.

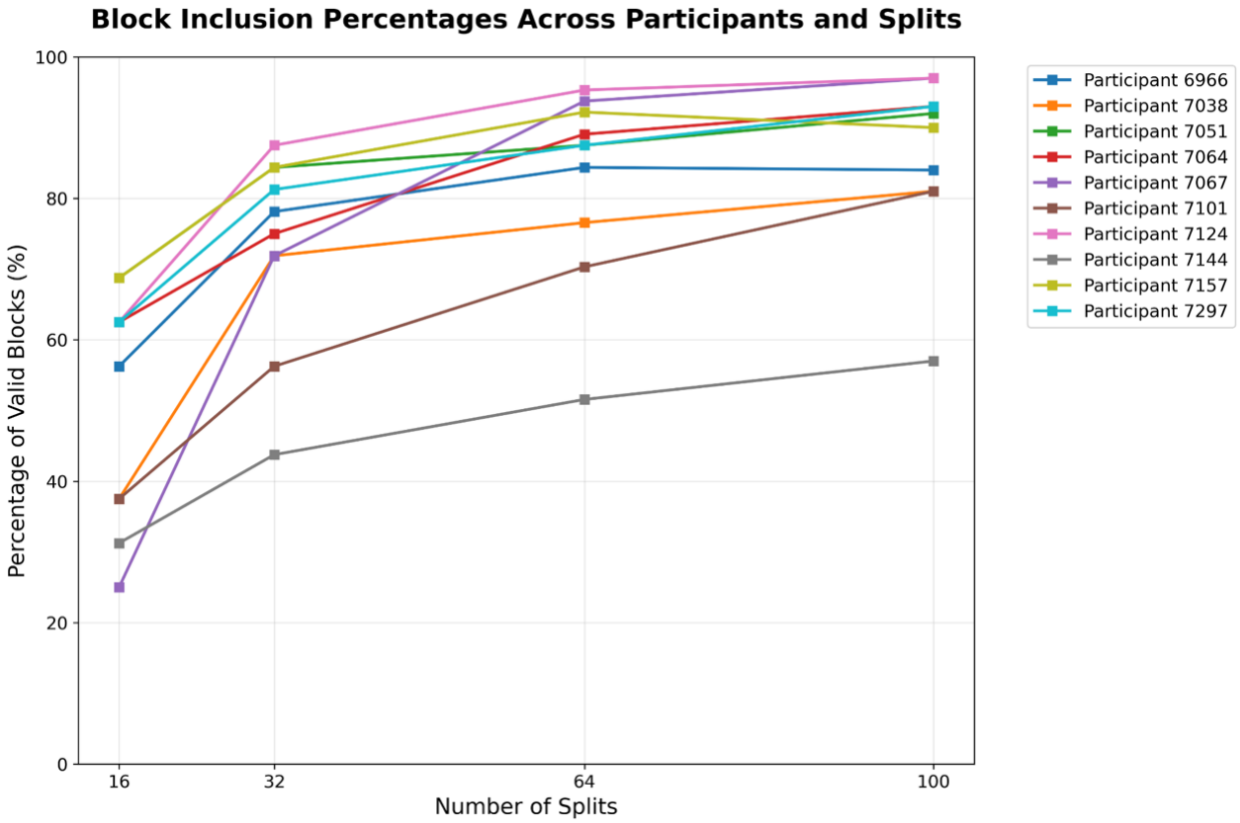


Figure 6.2.3: *In this figure, block inclusion percentages are mapped across all ten participants and the four split levels. The inclusion rate increases from 16 splits to 100 splits. Some participants have better block inclusion rates compared to others.*

6.2.2 Universal Kriging model

The same approach as demonstrated for participant 6966 was applied to the remaining nine participants. For each participant, a cross-validated elastic net model, followed by factor analysis, was used to pick primary predictors. For all ten participants, the alpha values for the final elastic net models were 0, indicating a ridge regression, with only an L2 penalty (Table 6.2.1). Further, a spatial variogram was fitted for each participant. As discussed in Chapter 5, these variogram parameters can affect the model's ability to capture the spatial variance in the data. A manual parameter tuning approach was employed to optimize the variogram parameters for cutoff, width, nmax, nmin, maxdist, and grid_size. Barring the cutoff parameter, the other parameters were found to be optimized at the same values: width = 2, nmax = 200, nmin = 30/40, maxdist = 100, and grid_size = 10. Cutoff values correspond to the maximum distance from the central voxel, up to which a variogram model should be applied. Based on individual variability in spatial autocorrelation, the values for the cutoff parameter ranged from 50 to 100.

An empirical variogram model was fitted for each participant to determine the appropriate parameter values for the nugget, psill, and range. These parameters have been discussed earlier in Chapter 5. Nugget values were zero for all 10 participants, psill values ranged from 0.52 to 0.95. The range parameter was found to be in the range of 7.1 to 16.19 mm across the 10 participants. These parameters were then used to fit the spatial component of the universal kriging model.

Participant id	n_voxels	Significant predictors	Elastic net		Variogram									
			Alpha	Lambda	Model	Nugget	psill	range	cutoff	width	nmax	nmin	maxdist	grid_size
6966	44891	14	0	0.0408	Spherical	0	0.7911	10.8802	50	2	200	30	100	10
7038	69637	16	0	0.046	Spherical	0	0.582	12.4067	50	2	200	40	100	10
7051	92132	18	0	0.0467	Spherical	0	0.5235	16.1925	80	2	200	40	100	10
7064	42478	14	0	0.0393	Spherical	0	0.6514	11.1883	50	2	200	40	100	10
7067	40793	14	0	0.0318	Spherical	0	0.7259	10.6657	100	2	200	40	100	10
7101	40634	18	0	0.0359	Spherical	0	0.5906	8.9995	50	2	200	40	100	10
7124	79525	14	0	0.0297	Spherical	0	0.6291	13.6438	50	2	200	40	100	10
7144	7993	18	0	0.0226	Spherical	0	0.9584	7.1027	100	2	200	40	100	10
7157	33752	12	0	0.0352	Spherical	0	0.7236	10.03	50	2	200	40	100	10
7297	97007	17	0	0.0352	Spherical	0	0.5455	14.9277	50	2	200	40	100	10

Table 6.2.1: *This table lists parameter values fitted for all 10 participants using the elastic net model and variogram model. Elastic net parameters, including alpha and lambda, help select primary predictors for the universal kriging model. Variogram parameters also help fit the spatial autocorrelation part of the universal kriging model.*

For the NPP UK model, RMSE and R^2 metrics have been used to assess the viability of the universal kriging modeling approach across the 10 participants. These metrics were

calculated for each participant across the four block split sizes: 16, 32, 64, and 100. The combined mean and range of RMSE values for all ten participants decreased from 16 split models to 100 split models (RMSE – 16 splits: 0.731, 32 splits: 0.487, 64 splits: 0.311, 100 splits: 0.24) (Figure 6.2.4). The combined mean R^2 values improved across split sizes (R^2-16 splits: 0.463, 32 splits: 0.686, 64 splits: 0.812, 100 splits: 0.85).

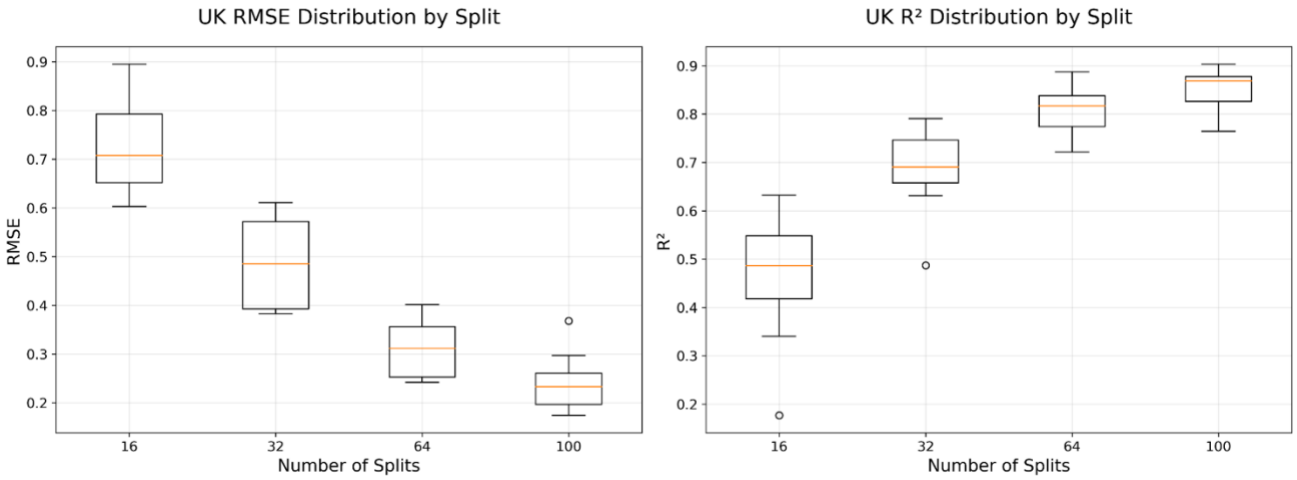


Figure 6.2.4: *This figure illustrates the distribution of boxplots for RMSE and R2 metrics across ten participants, encompassing four different block split sizes.*

The universal kriging model exhibits low RMSE values and high R^2 values, particularly for 64 and 100-block splits. This is further confirmed by plotting individual metric trajectories for each participant. Figures 2.2.5 and 6.2.6 demonstrate robust metrics across all participants. Participants 7144 and 7101 have slightly worse RMSE and R^2 values, which can be attributed to lower voxel volume and a lower voxel-to-significant predictor ratio, as discussed earlier.

These performance differences across participants and split sizes can be attributed to several potential factors. First, examining the differences between split sizes, the average voxel volume for 16 split blocks across all participants is 3430, decreasing to 549 for a 100-block split size. The universal kriging model accounts for spatial variance within the distance limits of the fitted variogram. The current model is set up to only fit spatial data within the range of the variogram. This model can be modified to predict successively and include prior predictions to increase the range of overall predictions; however, it might have compounded errors. Second, parameter tuning and the number of voxels to the number of primary predictors ratio influence the model results.

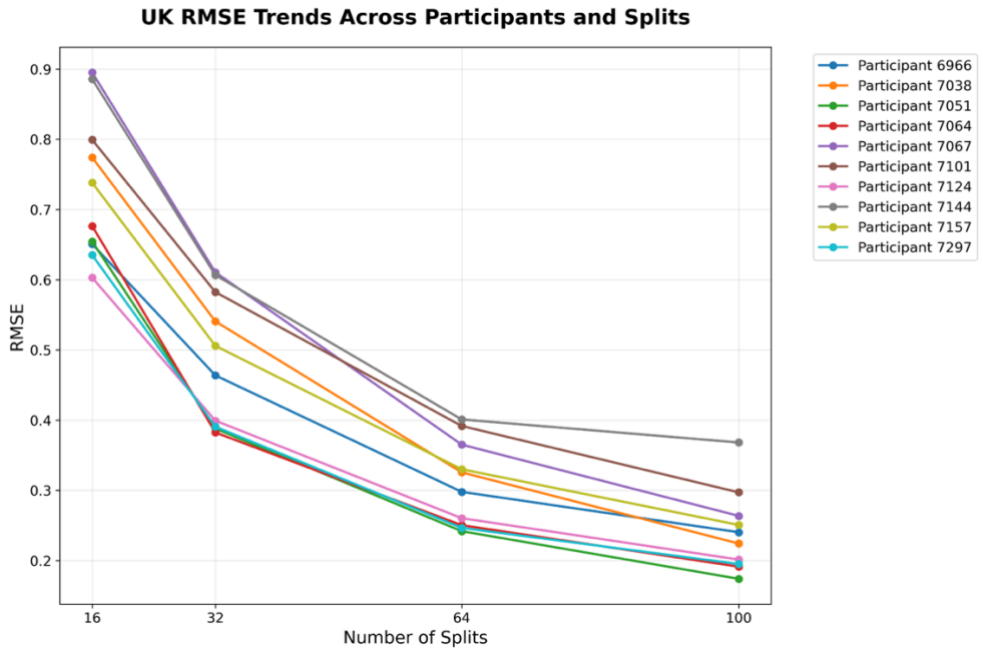


***Figure 6.2.5**: This figure depicts the Universal Kriging model's RMSE values for all ten participants across the four block split sizes.*

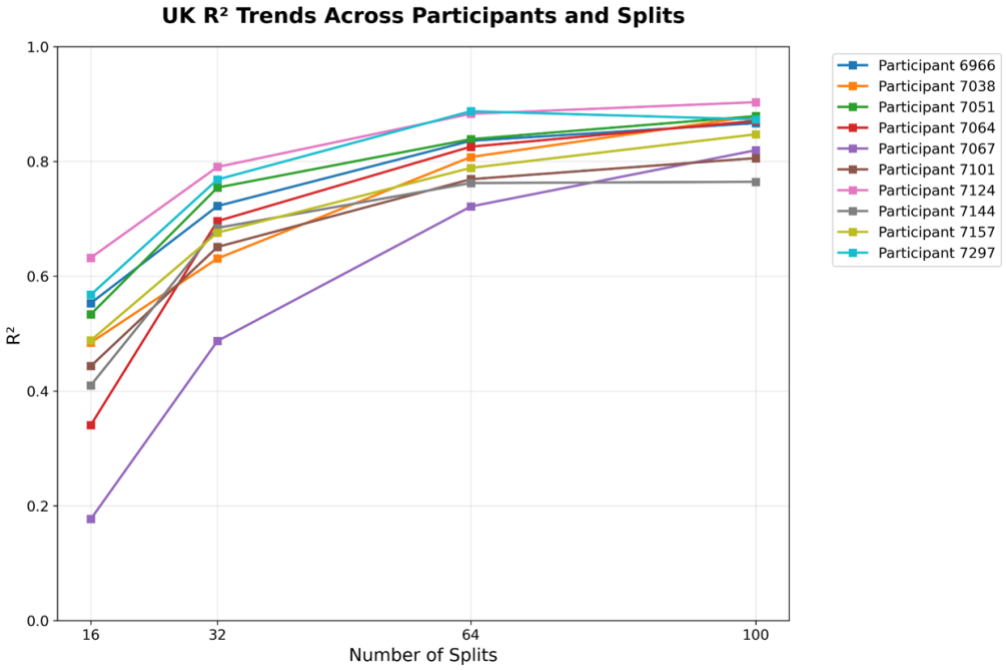


***Figure 6.2.6:** This figure depicts the Universal Kriging model's R2 values for all ten participants across the four block split sizes.*

I benchmarked universal kriging (UK) against two aspatial baselines: (i) linear regression using the same primary predictors and (ii) elastic net using the top 30 predictors per participant. Across all ten participants, UK consistently performed best (Fig. 6.2.7). Mean R^2 rose from 0.36 (linear) to 0.67 (elastic net) and to 0.85 with UK, indicating that explicitly modeling spatial autocorrelation captures meso-scale structure in tau deposition that the other models miss. Spatial autocorrelation should be treated as a first-order component in voxel-based tau modeling of MFG. Omitting spatial autocorrelation will likely underfit and blur the true pattern of disease spread.

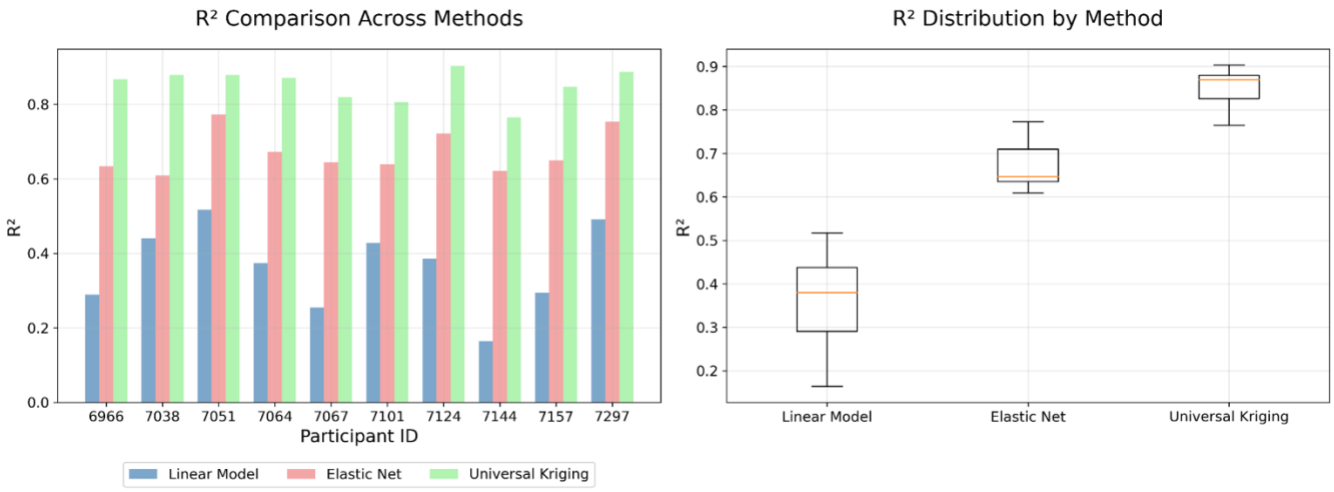


Figure 6.2.7*: In this figure, three models, linear regression, elastic net, and universal kriging (R2), are compared across all 10 participants. The R^2 values for Universal Kriging, for each participant, are derived from the best-performing split model. Participant 7297 had the best R2 values for 64 splits, and the remaining nine had the best R2 values for the 100-splits model.*

6.3 Significant predictor analysis

For all 10 participants, the top 30 predictors were selected based on absolute coefficient values using a 10-fold cross-validated elastic net model. Analysis of these predictors identified a total of 153 unique predictors across all participants. Amongst these predictors, 105 (68.6%) belonged to the 'buffer zone' or BZ category, and the remaining 48 (31.4%) belonged to the 'Euclidean Distance Transforms' or EDT (shortest absolute distance) category. Based on atlas categorization, 65 (42.5%) were derived from the white matter tracts atlas HCP1065[97], 39 (25.5%) were derived from the functional networks atlas Yeo17[98], 30 (19.6%) from regional parcellation based CerebrA[85] atlas, 12 (7.8%) were derived from the structural MNI 2009b atlas[15], and 7 (4.6%) from the vascular atlas Mouches & Forkert[69] atlas. Figure 6.3.1 further breaks down the contributing atlas's predictor frequency by each participant to the top 30 list. A recurring theme across participant 6966 and the groups' predictor analysis is that the white matter tract and

functional network atlases contribute the most predictors to the top 30 list. The regional parcellation atlas follows next. This may be due to the higher number of variables in the models from HCP1065[97] and CerebrA[85], 87 (37.7%) and 102 (44.2%), respectively. However, Yeo17[98] is an exception, contributing only 17 (7.4%) out of a total of 231 initial variables.

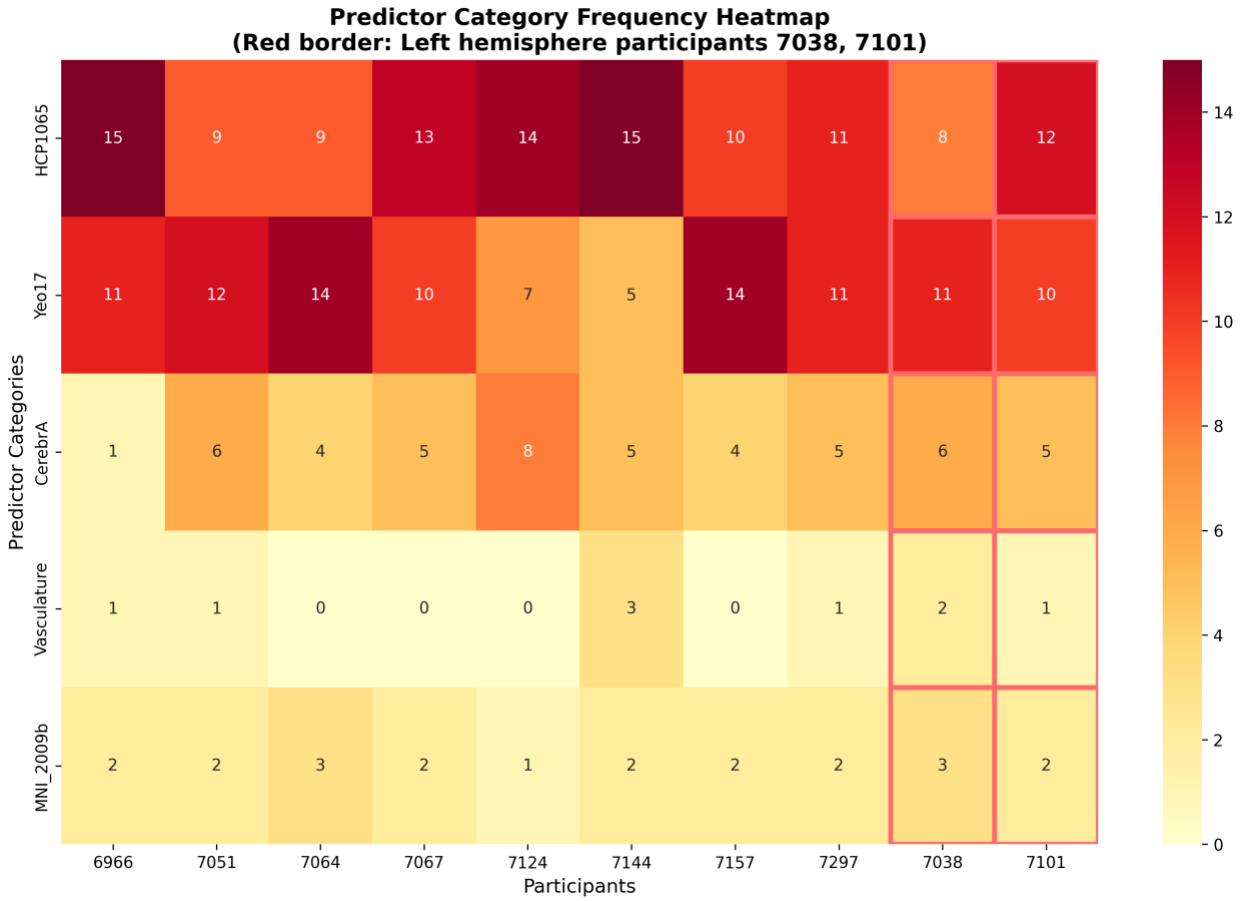


Figure 6.3.1: *In this figure, predictors are categorized based on base atlases and compared across the 10 participants using a frequency heatmap. Participants 7038 and 7101 are highlighted as they correspond to quantitative tau pathology sampled from the left hemisphere, while the remaining eight correspond to the right hemisphere.*

Next, the frequency of individual predictors was analyzed. BZ predictors of the same variables, but different radii, were combined into one with 'any' radius for this analysis. As an example, in the list derived from the top 30 predictors from each participant, there are five instances of 'bz_CC_10' and three instances of 'bz_CC_15', corresponding to the density of Corpus Callosum within a sphere with a radius of 10 and 15mm. These were combined into one, designated as 'bz_CC_any', with a total of nine instances across all

participants. After combining BZ predictors, the list of the top 30 most frequent predictors is shown in Figure 6.3.2. Out of these, 19 correspond to the BZ category and the remaining 9 to the EDT category. BZ predictors corresponding to the functional network Yeo 12 (Control A network[98]) and the white matter tract of the Corpus Callosum[97] are the most frequently selected, nine times. This is followed by the functional network Yeo 6 (Dorsal attention B[98]), which occurs eight times. These 30 most frequently selected predictors are further discussed in Table 6.3.1 below. Coefficient analysis was also conducted for these predictors; however, no meaningful insights were derived. Hence, the analysis is not included in this chapter.

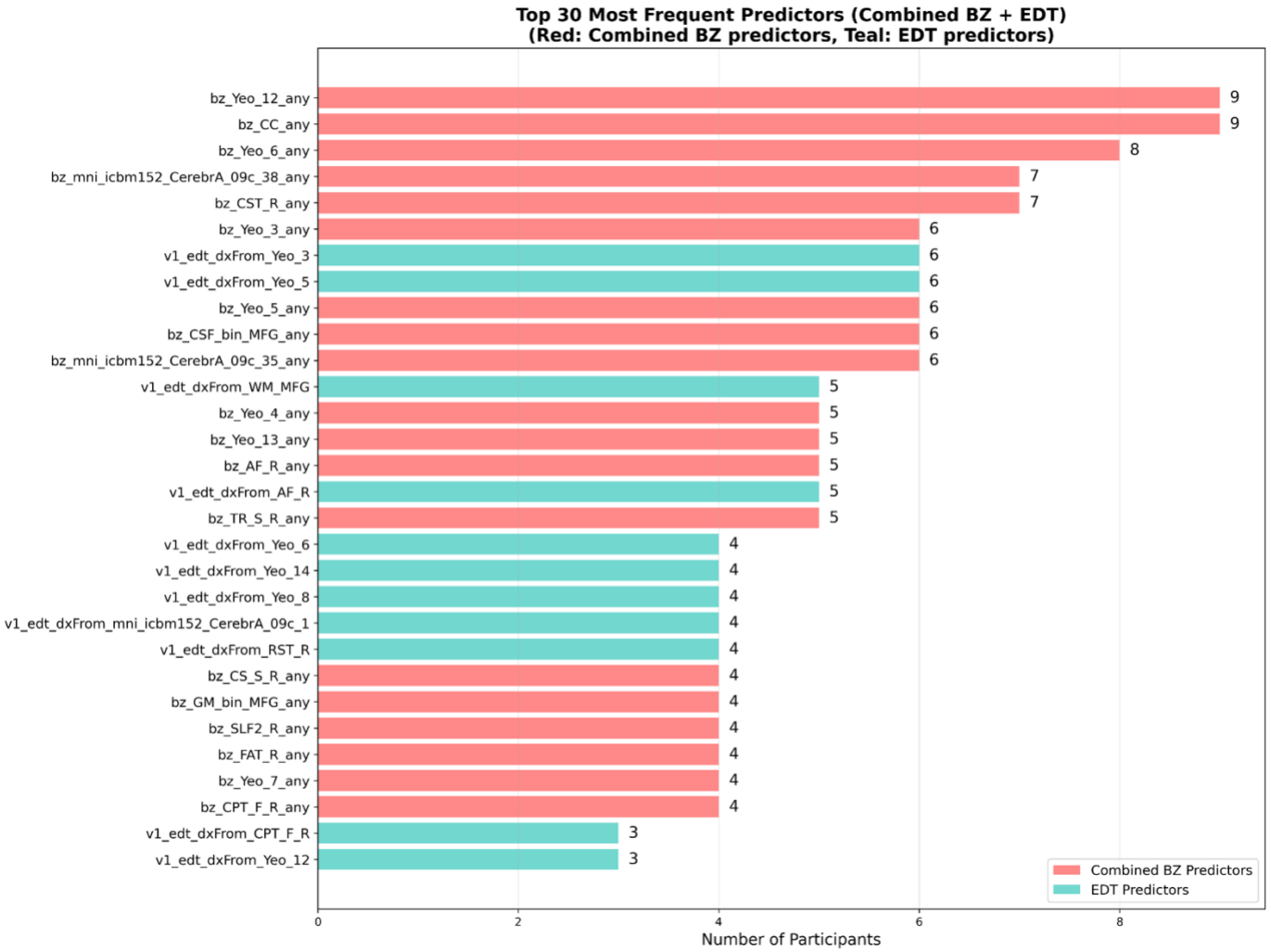


Figure 6.3.2: *Depicts the top 30 most frequent predictors selected from the list of all chosen predictors through a 10-fold cross-validated elastic net model. BZ predictors have been combined across any radii.*

***Table 6.3.1**: This table lists the top 30 most frequent predictors selected across all 10 participants: selected predictors, corresponding atlases, and variables' anatomical names listed in the table.*

Predictor label	Atlas	Anatomical name
bz_Yeo_12_any	Yeo17[98]	Control A network
bz_CC_any	HCP1065[97]	Corpus Callosum
bz_Yeo_6_any	Yeo17[98]	Dorsal Attention B network
bz_mni_icbm152 CerebrA_09c_38_any	CerebrA[85]	Superior Frontal right region
bz_CST_R_any	HCP1065[97]	Cortico-Spinal Tract Right
bz_Yeo_3_any	Yeo17[98]	Somatomotor A network
v1_edt_dxFrom_Yeo_3	Yeo17[98]	Superior Longitudinal Fasciculus2 tract Right
v1_edt_dxFrom_Yeo_5	Yeo17[98]	Dorsal Attention A network
bz_Yeo_5_any	Yeo17[98]	Dorsal Attention A network
bz_CSF_bin_MFG_any	MNI152 2009c[15]	Cerebrospinal Fluid
bz_mni_icbm152 CerebrA_09c_35_any	CerebrA[85]	Precentral region
v1_edt_dxFrom WM_MFG	MNI152 2009c[15]	White Matter
bz_Yeo_4_any	Yeo17[98]	Somatomotor A network
bz_Yeo_13_any	Yeo17[98]	Control B network
bz_AF_R_any	HCP1065[97]	Arcuate Fasciculus Right
v1_edt_dxFrom_AF_R	HCP1065[97]	Arcuate Fasciculus Right
bz_TR_S_R_any	HCP1065[97]	Thalamic Radiation Superior Right
v1_edt_dxFrom_Yeo_6	Yeo17[98]	Dorsal Attention B network
v1_edt_dxFrom_Yeo_14	Yeo17[98]	Temporal Parietal network
v1_edt_dxFrom_Yeo_8	Yeo17[98]	Salience / Ventral Attention B network
v1_edt_dxFrom_mni icbm152_CerebrA_09c_1	CerebrA[85]	Rostral Middle Frontal right region
v1_edt_dxFrom_RST_R	HCP1065[97]	Reticulospinal Tract Right
bz_CS_S_R_any	HCP1065[97]	Corticostriatal Tract Superior Right
bz_GM_bin_MFG_any	MNI152 2009c[15]	Grey matter
bz_SLF2_R_any	HCP1065[97]	Superior Longitudinal Fasciculus2 Right
bz_FAT_R_any	HCP1065[97]	Frontal Aslant Tract Right
bz_Yeo_7_any	Yeo17[98]	Salience / Ventral Attention A network
bz_CPT_F_R_any	HCP1065[97]	Corticopontine Tract Frontal Right
v1_edt_dxFrom_CPT_F_R	HCP1065[97]	Corticopontine Tract Frontal Right

v1_edt_dxFrom_Yeo_12	Yeo17[98]	Control A network

For the universal kriging model, these top 30 predictors were further reduced to yield primary predictors. As shown in Figure 6.2.1, these ranged from 12 to 18 primary predictors across participants. A total of 107 primary predictors were selected for all participants. Out of these 74 (69.2%) belonged to BZ category and the remaining 33 (30.8%) belonged to EDT category. 75 (70%) of these 107 were participant specific and did not repeat across participants. Figure 6.3.3 depicts the top 20 most frequent primary predictors selected for the UK modeling step. Amongst these 7 belong to EDT and the remaining 13 to BZ categories. Like earlier, BZ predictors corresponding to the functional network Yeo 12 (Control A network[98]) are the most frequent. This is followed by the Yeo 6 (Dorsal Attention B[68]) functional network, and Cortico-Spinal Right[70] white matter tract. Corpus Callosum is not as frequently selected due to multicollinearities.

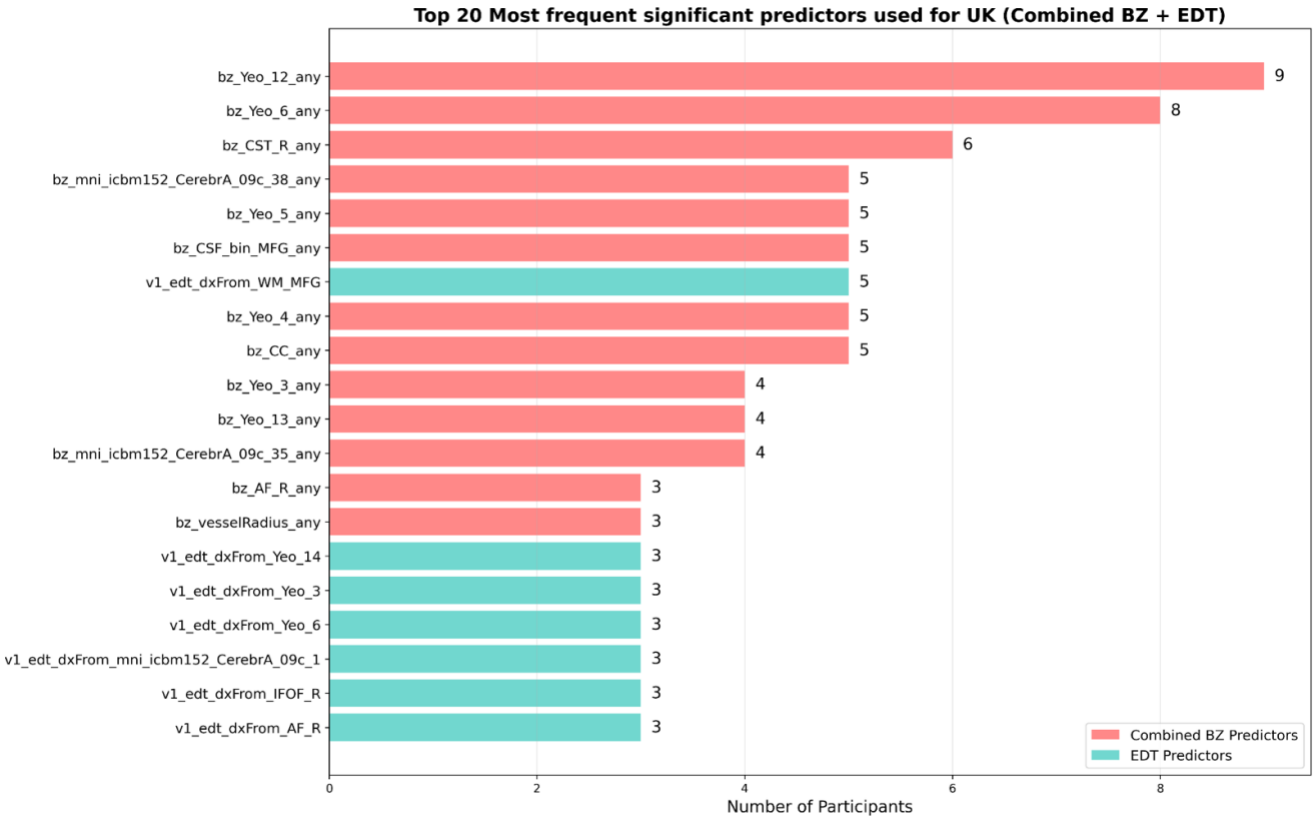


Figure 6.3.3: *Depicts the top 20 most frequent primary predictors selected for the UK modeling step, after factor analysis. BZ predictors have been combined across any radii.*

Chapter 7. Discussion & Conclusion

This final chapter synthesizes what the end-to-end NPP workflow contributes and why it matters. I open with a new Section 7.1 that interprets the Chapter 6 results, relating performance to variogram-defined spatial range, highlighting the centrality of white-matter and network predictors, and assessing robustness across donors. I then outline methodological extensions (7.2) and broader applications in AD (7.3), before briefly distilling the dissertation's contributions and constraints and closing with a forward-looking conclusion (7.6). Throughout, I tie these discussions back to the pipeline components developed earlier, QNPtoVox, I-BIS, and the UK-based modeling stack, so the chapter reads as a coherent bridge from proof-of-concept to a generalizable research program.

7.1 Interpreting results

In this section, I interpret the principal findings from Chapter 6, emphasizing what the results reveal about spatial scale, predictor biology, and model generalizability. Where helpful, I tie these interpretations to modelling choices described in Chapter 5 and to the QNPtoVox preprocessing decisions in Chapter 4.

7.1.1 Spatial block cross-validation and prediction distance.

Across all ten participants, universal kriging (UK) performance improved monotonically as spatial cross-validation became finer (16 → 32 → 64 → 100 blocks), with the strongest results at 64 and 100 splits (see Figure 6.2.4). This pattern accords with a finite spatial autocorrelation range: when the held-out test block lies within several millimeters of training observations, kriging borrows strength effectively; as the gap widens beyond the fitted range, predictive fidelity degrades. To make this distance intuition concrete, I report approximate linear edge-length estimates for the average test block under each split scheme. These linear edge-lengths are computed by treating each block as approximately cubic and converting average block volume (in voxels) into a linear scale using the 0.5-mm isotropic voxel size established in Chapter 3.

Table 7.1.1 lists, for each participant, the estimated edge length of an average block at each split along with the corresponding mean test R^2 and RMSE from Chapter 6. Edge lengths are derived from the participant's actual average block size under each split (voxel counts averaged across included blocks), converted to millimeters by multiplying by 0.5 mm in each dimension and taking the cubic root {[(average vox/block size)^(1/3)]*0.5}. Some blocks were excluded as I described in Chapter 6. The averages in this analysis are computed over the included blocks only.

Participant	Range	16: Edge (mm)	16: R^2	16: RMSE	32: Edge (mm)	32: R^2	32: RMSE	64: Edge (mm)	64: R^2	64: RMSE	100: Edge (mm)	100: R^2	100: RMSE
6966	10.8802	8.54	0.553	0.651	6.08	0.722	0.464	4.7	0.836	0.298	4.06	0.867	0.24
7038	12.4067	11.32	0.484	0.774	7.23	0.631	0.541	5.62	0.808	0.326	4.75	0.879	0.224
7051	16.1925	10.15	0.534	0.654	7.53	0.754	0.388	5.9	0.839	0.242	5	0.879	0.174
7064	11.1883	8.1	0.341	0.676	6.05	0.696	0.382	4.53	0.826	0.25	3.85	0.871	0.191
7067	10.6657	10.84	0.177	0.895	6.05	0.487	0.611	4.4	0.721	0.365	3.75	0.819	0.264
7101	8.9995	9.46	0.443	0.799	6.56	0.651	0.582	4.83	0.769	0.392	3.97	0.806	0.297
7124	13.6438	9.98	0.632	0.603	7.08	0.79	0.399	5.46	0.883	0.26	4.68	0.903	0.201
7144*	7.1027	5.85	0.41	0.886	4.15	0.684	0.607	3.12	0.762	0.401	2.6	0.764	0.368
7157	10.03	7.27	0.488	0.738	5.39	0.676	0.506	4.15	0.789	0.33	3.61	0.847	0.251
7297	14.9277	10.66	0.568	0.635	7.75	0.768	0.391	6	0.887	0.247	5.07	0.873	0.195
Average	**11.60371**	**9.217**	**0.463**	**0.7311**	**6.387**	**0.6859**	**0.4871**	**4.871**	**0.812**	**0.3111**	**4.134**	**0.8508**	**0.2405**

***Table 7.1.1**: Participant-level average block edge length (mm) and cross-validated performance at four split sizes. *Participant 7144 includes the least number of voxels, hence the smallest edge lengths and range (See Figure 6.2.2).*

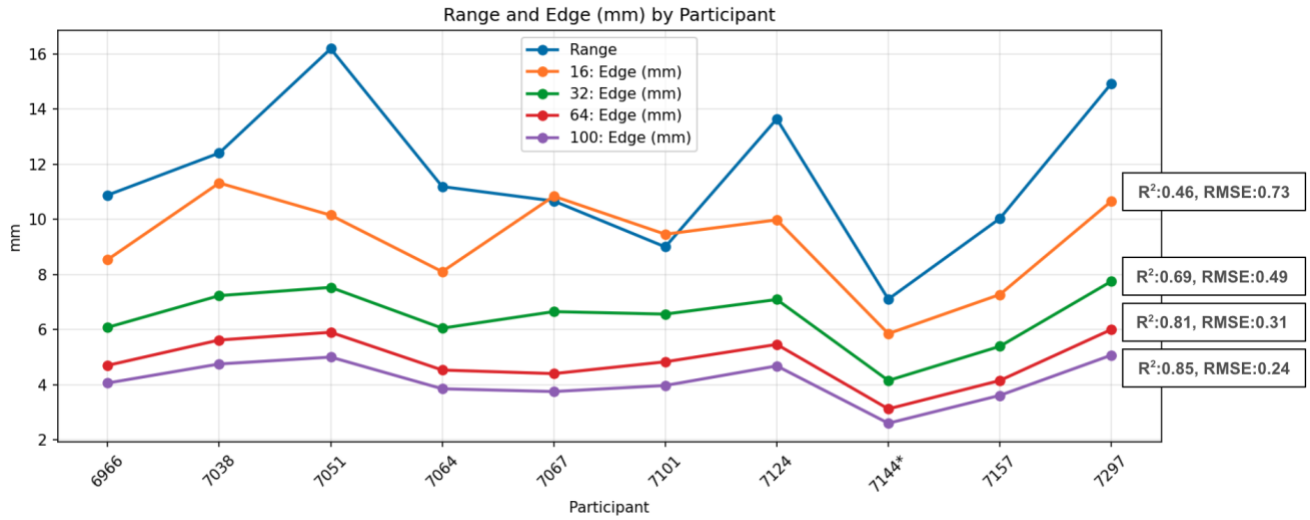


***Figure 7.1.1**: Variogram ranges (mm) and mean block edge lengths (mm) for each participant at four spatial cross-validation split sizes (16, 32, 64, 100), with corresponding average R^2 and RMSE values from UK models. *Participant 7144 includes the least number of voxels, hence the smallest edge lengths and range (See Figure 6.2.2).*

As shown in Table 7.1.1**,** the characteristic edge lengths at 16, 32, 64, and 100 splits are 9.21, 6.39, 4.87, and 4.13 mm, respectively, which align with the fitted variogram ranges (≈7.1–16.2 mm) reported in Chapter 6. With participant-specific edge-length analysis (see Figure 7.1.1), the cross-validation trend in Chapter 6 becomes mechanistically transparent: UK excels when the train–test separation resides within the empirical autocorrelation range inferred from the variogram. At coarser splits (16–32), the typical edge length approaches the lower bound of the fitted ranges, so predictions rely more heavily on the drift (external covariates), and errors rise. Conversely, at finer splits (64–100), the average block size is well within range, allowing the spatial variogram

component to dominate and performance to converge toward the upper bound observed in Chapter 6.

7.1.2 White-matter tracts and functional networks

The predictor analysis in Chapter 6 revealed a striking and consistent pattern: variables derived from white matter tracts (HCP1065[97]) and task activation-derived functional networks (Yeo 17[68]) emerged as the most frequently selected predictors across the ten participants, outpacing predictors tied to regional parcellations, vascular structures, or purely geometric distances. This trend may not be incidental and might reflect the spatial embedding of tau pathology within the brain's connectivity and activation infrastructure.

For participant 6966, this was particularly clear. The final model retained densities or distances linked to the superior longitudinal fasciculus (SLF), arcuate fasciculus (AF), inferior fronto-occipital fasciculus (IFOF), corticopontine tracts, and thalamic radiations. In parallel, control and attention networks from the Yeo-17 atlas featured prominently. These structures form high-capacity conduits for information and, by extension, potential vectors for protein propagation. The repeated selection of these covariates after both elastic net regularization and factor analysis suggests that they may not be statistical artefacts but instead may be robust correlates of pathology distribution.

This finding aligns with prior work demonstrating that tau pathology spreads preferentially along anatomically connected regions and that network connectivity can modulate vulnerability to deposition. The prominence of buffer-zone (local density) features over simple Euclidean distances further underscores that the micro-regional context, the composition and density of nearby fibers or network nodes, provides richer predictive information than proximity alone. This supports a model in which the tissue's immediate anatomical environment may affect both susceptibility and rate of spread, potentially through mechanisms such as axonal transport, local synaptic connectivity, and region-specific metabolic or inflammatory states.

7.1.3 Robustness across participants and the role of data "density"

When results were aggregated across all ten participants, the UK approach consistently outperformed both linear regression (with identical covariates) and the broader elastic-net baseline. At the finest spatial splits (64 and 100), mean R^2 values clustered in the mid-0.8s, with RMSE values markedly lower than at coarser splits. This stability in predictive performance across a heterogeneous cohort is noteworthy, as it suggests that the NPP framework captures generalizable relationships between spatial covariates and tau burden.

The progressive improvement in R^2 and reduction in RMSE from 16 to 100 splits held true for nearly all participants. Two exceptions, 7144 and 7101, showed weaker gains (see Figures 6.2.5 and 6.2.6) in R^2 and RMSE metrics when moving from coarser to finer splits. I suspect the lower metrics are due to low voxel counts per primary predictor, leading to insufficient within-block variability and coarser effective sampling. In such cases, the variogram fit becomes less reliable, and kriging solutions are more sensitive to noise.

From these observations, a pragmatic heuristic emerged: optimal stability seems to occur when there are approximately 4,000–6,000 voxels per selected predictor. This range appears to be a sweet spot balancing the richness of the covariate space with the statistical power available to fit spatial parameters. Falling below this threshold increases the likelihood of block exclusions (due to edge effects or insufficient coverage) and inflates uncertainty in predictions.

These findings have two important implications. First, they provide a quantitative basis for experimental design, allowing future studies to aim for tissue sampling densities that meet or exceed this voxels-per-predictor ratio. Second, they highlight the potential to adaptively adjust model complexity based on available data density, for example, by reducing the number of predictors or choosing finer split levels in cases where tissue volume is limited.

7.1.4 Variogram characteristics

Across participants, empirical variogram fitting yielded nugget terms close to zero, partial sills between approximately 0.5 and 0.95, and ranges spanning roughly 7–16 mm. The near-zero nugget values indicate that, after the 2-mm Gaussian smoothing applied in Chapter 4, almost all residual variation is spatially structured, with minimal uncorrelated noise at the microscale. This is important because a large nugget would suggest measurement error or unmodelled fine-scale variability that could undermine the kriging interpolation.

The fitted ranges align closely with neuroanatomical scales: they are comparable to the width of gyral crowns, and the separation between adjacent sulci in the middle frontal gyrus. This correspondence suggests that the spatial autocorrelation structure is capturing biologically meaningful propagation distances of tau pathology along connected cortical territories. Within these ranges, the kriging model interpolates effectively, combining the deterministic drift informed by covariates with the spatial component captured by the variogram. Beyond these ranges, the spatial term decays, and the model increasingly relies on drift alone, leading to higher predictive uncertainty. The

reproducibility of these variogram parameters across participants reinforces the interpretation that tau spread appears to follow a consistent spatial process, potentially shaped by underlying connectivity and tissue microstructure.

7.1.5 Added value of spatial modelling

Direct comparisons between universal kriging and non-spatial baselines in Chapter 6 illustrate the substantial contribution of spatial autocorrelation to model performance. When using identical covariate sets, a standard linear regression explained only a modest proportion of variance (mean $R^2 \approx 0.36$), and even a more flexible elastic-net regression plateaued around $R^2 \approx 0.67$. In contrast, universal kriging consistently exceeded $R^2 \approx 0.85$ at the finest spatial splits. This improvement is not simply statistical; in this study, it reflects the reality that tau pathology exhibits spatial dependence beyond what can be explained by covariates alone.

By explicitly modelling the spatial covariance structure through the variogram, the UK can capture residual variation arising from neighborhood effects, local tissue architecture, and mesoscale connectivity patterns. This not only improves predictive accuracy but also enhances the spatial coherence of predicted pathology maps, reducing spurious fluctuations and better matching the continuous nature of biological processes. These results underscore that spatial dependence is an informative signal, one that, if ignored, would lead to systematic underestimation of pathology in unmeasured regions.

7.1.6 Block inclusion patterns and edge effects

Analysis of block inclusion rates across spatial cross-validation configurations offers further insight into how spatial scale interacts with model evaluation. Inclusion rates were highest at finer splits (64 and 100) and lowest at coarser splits (16 and 32). At coarse splits, test blocks are larger and more irregularly shaped, increasing the likelihood that they intersect tissue boundaries, cross sulcal gaps, or extend beyond the spatial range over which the variogram supports reliable interpolation. Such conditions increase prediction difficulty and, in some cases, warrant excluding these blocks from performance calculations in Chapter 6.

This exclusion pattern is not just a statistical artifact; it mirrors the geometric and anatomical realities of the tissue samples. Coarser partitions can produce regions that are spatially isolated from training data by distances exceeding the fitted variogram range, effectively forcing extrapolation rather than interpolation. This contributes to the observed drop in R^2 and rise in RMSE at 16 and 32 splits. Conversely, at 64 and 100 splits, blocks are smaller and more likely to be fully surrounded by training data within the spatial range,

aligning with the kriging model's strengths and producing more stable, accurate predictions. This behavior corroborates the findings of Section 7.1.1, reinforcing the recommendation to match CV partition size to the model's effective prediction distance when designing spatial validation protocols.

7.2 Future Work (Methods)

In this section, I outline a forward-looking agenda for how the NPP framework could be extended by me and the broader community after this dissertation. The aim is to chart iterative improvements that make the approach more useful to the larger AD/neuroscience research space. Depending on opportunities and collaborations post-PhD, I may pursue a subset of these directions. Specifically, I describe priorities that include: (i) expanding from the current testbed to additional regions and ultimately brain-wide analysis; (ii) developing and benchmarking complementary models beyond universal kriging to address finite-range and non-stationarity constraints; (iii) integrating spatial omics (e.g., SEA-AD) to add cellular and proteomic context; and (iv) systematically enriching the I-BIS covariate library. I also outline engineering work for scalability and validation protocols to support reproducibility.

7.2.1 Regional expansion and brain-wide analysis

I envisage extending QNPtoVox, and subsequently the whole NPP framework to the full set of regions routinely sampled under NIA-AA guidelines[99], prioritizing hippocampus and its subfields, superior/middle temporal gyri, and the amygdala as tissue availability permits. A practical first tranche would likely begin with the hippocampus and superior/middle temporal gyrus to maximize comparability with existing research conducted using ACT tissue data repository.

Methodologically, this extension would:

1. Standardize regional tissue handling in QNPtoVox: harmonize slab-to-MRI co-location procedures, smoothing kernels, and ANTs settings across regions; implement region-specific quality-control (QC) metrics (e.g., slice-to-volume residuals, mutual-information thresholds) to detect misalignments early.

2. Scale to brain-wide analysis by registering regional volumes into MNI space and applying a multi-resolution prediction grid (2–5 mm) outside sampled tissue. Accuracy would be evaluated via spatial block CV within each region and leave-one-region-out validation across the mosaic to quantify out-of-sample generalization.

3. Adopt hierarchical modelling for multi-region data by fitting region-specific variograms and estimating shared hyper-priors for sill and range.

4. Harmonize data across donors by treating the participant as a random effect in a hierarchical UK model, enabling pooled inference while preserving individual differences.

7.2.2 Model development beyond UK

To complement the UK and mitigate finite-range limitations, a family of models would be implemented that leverages I-BIS covariates while capturing additional structure. According to a preliminary literature review, the following modeling approaches would be assessed:

- Tree-based ensembles (Random Forest, XGBoost): trained with spatial block CV, tuned via Bayesian optimization, and interpreted to identify stable predictor families across regions. These ensembles should be compared to UK using identical covariates.

- Graph-informed diffusion/GNNs**:** construct a voxel-wise graph using structural connectome priors (tract density, streamline probability) and functional network affiliation; train (i) Laplacian diffusion models and (ii) shallow graph neural networks to propagate signals along edges, explicitly modelling long-range spread that exceeds variogram ranges.

- Non-stationary/anisotropic Gaussian processes**:** replace stationary variograms with spatially varying length-scales or tensor kernels aligned with dominant tract orientations; evaluate via information criteria and coverage of predictive intervals.

- Kriging convolution networks**:** when multiple stains are available (e.g., tau, Aβ, TDP-43), use a kriging-derived convolution network to estimate interaction terms at the voxel level.

- Model stacking**:** combine UK with learned residuals from an ensemble (UK→RF/XGB residual modelling), thereby preserving geostatistical interpretability while capturing complex, non-linear effects.

Each model will be assessed under a matched spatial cross-validation framework, with uncertainty quantification and robustness checks (primary predictors, and region-wise sensitivity analyses).

7.2.3 Spatial omics integration (SEA-AD)

To add biological specificity, spatial omics could be generated within SEA-AD[100], e.g., spatial proteomics[101], high-plex RNA, and cell-type maps derived from single-cell references. Given that the ACT neuropathology core is already collaborating with the SEA-AD team to analyze the same set of participants' data, this would be a natural extension. The workflow will:

1. Co-register omics slides to QNPtoVox volumes via feature-based alignment anchored to block faces; propagate transforms to MNI.

2. Generate voxel-level omics covariates**:** cell-type abundance (astrocytes, oligodendrocytes, inhibitory/excitatory subclasses), microglial activation scores, synaptic and myelin protein panels, and local diversity indices. Both buffer-zone density and shortest distance features would be computed to match existing I-BIS design patterns.

3. Joint modelling: include omics covariates in UK drift; evaluate co-kriging where stains/omics share structure. Interaction models would be tested (e.g., does a specific cell type's density mediate the association between white matter tract proximity and tau burden?).

7.2.4 Enriching I-BIS covariates

To make I-BIS even more data-driven and robust, it would be expanded along four axes to capture micro- and mesoscale context more fully:

- Cytoarchitecture: incorporate layer-resolved covariates (e.g., using BigBrain/Julich laminar probabilities or Allen brain atlas) to compute layer-specific distances and buffer densities within cortex-constrained masks.

- Myelination and microstructure: add T1w/T2w myelin proxies, mapped to MNI 2009b.

- Geometry and topology: Moving beyond Euclidean absolute distances, include geodesic distance along the cortical surface and sulcal depth to better capture patterns observed in Chapter 6.

- Vascular context: include refined arterial and venous territory maps.

All additions will follow the existing I-BIS framework: voxel-wise point, distance, and buffer-zone density features at multiple radii (1–15 mm).

7.2.5 Pipeline engineering and scalability

I developed the NPP pipeline, including its subcomponents QNPtoVox, I-BIS, and UK model, with a focus on open-source sharing and reproducibility. To further advance these important scientific efforts, I would:

- Containerize the full pipeline with pinned dependencies, extending a CLI for batch runs and a YAML-based config system.
- Automate QC with metrics and visual reports (slice-to-volume error, Dice-like overlap on masks, transform Jacobian summaries) and a rules engine to flag reprocessing.
- Parallelize kriging and model training across nodes, cache intermediate NIfTI volumes, and adopt data versioning (DVC) and continuous integration for unit/regression tests.
- Provenance and audit trail**:** include hashes/logs of inputs/outputs, parameter files, and export companion JSON sidecars for every figure/table.

7.2.6 Validation and benchmarking

I would formalize validation along three tiers:

1. Internal validation: spatial block CV within region and cross-region (leave-one-region-out), with pre-registered primary endpoints (R^2, RMSE) and secondary structural-similarity metrics (NMI, cosine similarity)
2. External validation: apply trained models to independent donors/cohorts; harmonize intensities; report transport performance and calibration drift.
3. Sensitivity analysis**:** remove classes of covariates (e.g., tracts, networks) to quantify their marginal utility; perturb variogram parameters to assess stability; evaluate prediction under synthetic missingness (block drop-out experiments).

7.3 Generalizability and future applications

This section focuses on expanding the NPP framework to tackle key questions in AD research, which could have wider effects on our understanding of the disease and our capacity to model it. Building on the empirical takeaways established in 7.1 and the methodological roadmap described in 7.2, I now outline how NPP can be deployed to address central questions in Alzheimer's disease research. The emphasis is on scientifically meaningful use-cases that exploit QNPtoVox's marker-agnostic

preprocessing, the I-BIS covariate suite, and the spatially explicit inference afforded by universal kriging and its planned extensions.

7.3.1 Multi-pathology interaction mapping (Tau × Aβ × TDP-43)

Alzheimer's pathology rarely occurs in isolation. Mixed proteinopathies, most commonly tau with β-amyloid and the increasingly recognized TDP-43, are prevalent in community-based cohorts and have been associated with faster clinical decline, atypical symptom profiles, and PET signal confounding. Yet, the field lacks spatially resolved, quantitative maps of how these pathologies co-localize, compete, or synergize within the same tissue. NPP is uniquely positioned to help close this gap. Because QNPtoVox produces marker-agnostic, co-registered voxel volumes and I-BIS supplies biologically interpretable covariates (white matter, networks, vasculature), the framework can attribute regional burden not only to each protein individually but also to their non-additive interactions. The resulting "interaction maps" would clarify questions like whether Aβ accelerates tau along association tracts, or whether TDP-43 reshapes limbic vulnerability, and where co-pathology simply reflects shared anatomical conduits. Such maps would inform intervention development, refine clinicopathologic correlations, and guide multi-target trial strategies where combination therapies are contemplated.

Alzheimer's disease (AD) in the community is more often mixed than pure. Autopsy studies repeatedly show that a substantial proportion of older adults carry combinations of tau, β-amyloid, and TDP-43 (with or without α-synuclein), and these mixtures are associated with steeper decline, and atypical presentations. Despite this, most spatial analyses remain single-marker or rely on coarse regional scores that obscure where proteins truly co-localize or diverge within the cortex and medial temporal lobe. The lack of a common coordinate framework and quantitative intensity scales has made it difficult to establish whether interactions are mechanistic (one pathology facilitating spread of another along shared networks) or merely co-incidental (converging on the same vulnerable territories for independent reasons).

NPP could provide the missing infrastructure to adjudicate these possibilities. Because QNPtoVox is marker-agnostic, I can produce co-registered, voxel-level maps for tau, Aβ, and TDP-43 from the same donor and region, all aligned to MNI space. With I-BIS covariates, each voxel is also endowed with interpretable context, white-matter tract proximity and density, functional network connectivity, and vascular adjacency, allowing me to ask not only where co-pathologies meet but why they meet there. For example, if tau and Aβ co-burden concentrates along frontoparietal association fibers while TDP-43 remains limbic-centered, the data would support a model in which Aβ amplifies tau propagation in association cortex, whereas TDP-43 contributes a distinct limbic

vulnerability. Conversely, if all three co-localize, the vascular context may be the dominant driver. Finally, because the outputs are quantitative and standardized, they support meta-analysis across cohorts and enable rigorous tests of whether interaction patterns generalize across age, sex, and genetic backgrounds.

7.3.2 Reproducing and extending Braak staging at scale

The Braak and Braak[9] framework has shaped AD science for three decades, providing a clear, testable narrative of tau progression. Yet it rests on a small sample and ordinal regional ranks that compress rich spatial detail. Contemporary datasets, together with NPP's harmonized pipeline, create the conditions to revisit staging in a way that is continuous, probabilistic, and connectivity-aware. The central question is no longer simply whether pathology passes through the entorhinal cortex before the neocortex, but how the distribution unfolds within networks, along tracts, and across individuals.

NPP can contribute in decisive ways. First, it converts heterogeneous histology into comparable voxel-level maps, enabling direct pooling across donors. Second, by situating those maps within I-BIS, it embeds staging in anatomical constraints, allowing staging hypotheses to be evaluated against the brain's physical and functional infrastructure. Third, because the outputs are quantitative, I can express staging as distribution probability fields, the likelihood that a voxel or region is involved by a given "stage," and measure inter-individual variability rather than forcing all brains into a single sequence.

The result is a staging system that is both truer to biology and potentially useful for translational aims. Quantitative, connectivity-aware staging should better predict future involvement (what becomes positive next), track with in-vivo biomarkers (PET/MRI) and align with clinical course. It also provides a common language for comparing typical and atypical variants, e.g., hippocampal-sparing, or posterior-predominant trajectories, without discarding the coherence that made Braak staging so influential. In short, NPP enables a principled evolution from a milestone framework built on six brains to a robust atlas grounded in high-fidelity spatial data across many donors.

7.3.3 Subtype-specific trajectories and cross-modal subtypes

As described in Chapter 2, existing research suggests that AD is clinically and biologically heterogeneous. Limbic-predominant, hippocampal-sparing, posterior-predominant, dysexecutive, and language-led presentations[11,102,103] all appear under the umbrella of "AD," yet most subtype labels are derived from downstream manifestations, regional atrophy on MRI, PET uptake patterns, or domain-specific cognitive deficits. What is still missing is a mechanistic account anchored in the primary pathology distribution: where

abnormal tau accumulates first, how it travels through anatomical highways, and why particular networks succumb in some patients but not others. NPP can address this gap.

Because QNPtoVox delivers participant-level, voxel-wise pathology maps in a common spatial frame and I-BIS supplies interpretable covariates (white matter tracts, functional networks, vasculature), subtype heterogeneity can be examined at the level at which networks and tracts constrain spread. The results in section 7.1 already point to white-matter architecture and functional networks as dominant predictors of tau. By quantifying burden with millimeter-scale resolution and situating it within those anatomical constraints, NPP converts subtype hypotheses into spatially testable assertions.

The potential benefits extend across modalities. Voxel resolution pathology constitutes a ground truth against which in-vivo subtype algorithms, based on MRI, PET, or digital cognitive phenotyping, can be adjudicated. If an imaging-defined subtype purports to be "hippocampal-sparing," its correspondence with ex-vivo tau should be demonstrable in MNI space, not just by lobar averages. Conversely, where imaging and pathology diverge, NPP could provide the spatial context to interpret the discrepancy (e.g., off-target PET binding, partial-volume effects, or masking by atrophy). This cross-modal alignment allows for operationalizing definitions of subtypes, so they are portable across study designs and technologies.

NPP also adds a prognostic dimension. Subtype-resolved pathology maps can be translated into expected trajectories, which regions are likely to be involved next, given the current pattern, and into risk maps for cognitive decline tied to network functions (memory, executive, visuospatial, language). This supports precision stratification in trials: enrolling participants with comparable predicted trajectories reduces outcome variance and increases power, while revealing subtype-specific therapeutic windows.

Finally, a pathology-anchored view could help disentangle co-pathology and comorbidity, two common confounders of clinical heterogeneity. Because NPP is marker-agnostic, one can overlay Aβ and TDP-43 distributions on tau maps to determine whether a putative subtype reflects a true tau pathway or a mixed-protein signature that happens to present with similar symptoms.

7.3.4 Data-efficient corpora for diffusion and neural networks

Deep generative and predictive models can accelerate brain-wide neuropathology spatial distribution research. First, they learn non-linear, multi-scale dependencies implicit in tissue architecture (e.g., laminar layers, tract topology, network affiliation) and can therefore extrapolate beyond the local, finite-range regime that constrains kriging. Modern

architectures, 3D CNNs with dilated kernels, transformers with global attention, and graph neural networks aligned to tractography, can couple short-range anatomical variability with long-range connectivity, producing smooth, anatomically consistent fields over the whole brain. Second, diffusion models and other generative approaches provide a data engine: they can synthesize large amounts of biologically derived features, pathology volumes, and extrapolate to unsampled regions, and support high-throughput analysis at scales impractical with limited histology sampling alone. Third, once trained, these models run at inference speed, minutes instead of weeks, making it feasible to screen thousands of brain-wide scenarios (e.g., alternative seeding sites, covariate perturbations) and to iterate on study designs before tissue collection.

Applied to quantitative digital histopathology, these models can operate directly on voxel-based targets (continuous AT8/Aβ/TDP-43 intensity) rather than binary labels, preserving gradations of pathology burden. Grounding these models in NPP could substantially accelerate development and guard against failure modes:

- Standardized coordinates and covariates**:** QNPtoVox and I-BIS place every voxel in MNI space with a rich, biologically interpretable feature stack (tracts, networks, vasculature, geometry). This removes site-specific idiosyncrasies and turns disparate cases into a single interoperable corpus, enabling fair benchmarking rather than one-off demonstrations.

- Efficient data annotation: Kriged volumes can provide pseudo-labels with uncertainty, based on the distance of prediction measures. Networks can train with heteroscedastic losses (weighting errors by UK variance) so that uncertain regions exert less influence, improving robustness. The same uncertainty maps drive active learning, prioritizing new blocks where models and kriging disagree.

- Connectome-aware inductive bias**:** I-BIS features can be injected as channels or used to define a graph prior, steering networks toward mechanistically plausible spread rules (e.g., along SLF/IFOF or within control/default mode networks) rather than spuriously exploiting staining artefacts.

Diffusion and neural network models offer the capacity to generalize across whole-brain inference and generate data that amplifies discovery. By tying them to NPP's common coordinate system, interpretable covariates, and uncertainty-aware supervision, I can lower the data barrier, standardize evaluation, and align flexibility with biological realism, a combination likely to move digital quantified histopathology from promising demonstration to a reliable instrument for AD research.

7.4 Contributions

This dissertation makes several contributions to the study of Alzheimer's disease and, more broadly, to spatial modelling in human neuroscience. I organize them below as thematic subsections. Each contribution is grounded in the empirical findings of section 7.1 and in the forward-looking improvements and applications outlined in sections 7.2 and 7.3. Before I describe them in detail, here is a quick summary:

- Histology to neuroimaging bridge**:** QNPtoVox registers quantitative digital histopathology to standard coordinates, enabling pooled, cross-modal analyses.
- Biology-aware spatial modelling**:** I-BIS covariates encode tracts, networks, vasculature, and geometry; residuals capture local dispersion, together supporting mechanistic inference about spread.
- Kriging in brain space**:** Adapts universal kriging to MNI neuroanatomy, yielding calibrated, interpretable voxel-level maps of pathology from sparse histology.
- Open, modular stack**:** Three interoperable pipelines (QNPtoVox, I-BIS, NPP) support end-to-end studies or standalone use across research domains.

7.4.1 Mapping digital histopathology to neuroimaging coordinates

Quantitative digital histopathology offers cellular-scale signal but is typically limited to small, unevenly sampled slabs acquired under site-specific protocols. This dissertation contributes a robust, end-to-end method (QNPtoVox) to co-locate histology with ex-vivo MRI, reconstruct 3D pathology volumes from 2D sections, and register those volumes to MNI space. The mapping is not merely geometric: it preserves quantitative intensity, maintains orientation fidelity, and captures each step sufficient to retrace every transform. In practice, this turns slides and XML outputs into NIfTI volumes that can flow through the same analysis stack used for MRI/PET.

The approach addresses four persistent barriers. (i) Heterogeneity of acquisition, differences in staining, scanning resolution, and tissue handling are mitigated by standardizing downstream representation in a shared coordinate system. (ii) Limited field-of-view is countered by block-wise reconstruction and smoothing that yield coherent 3D signals suitable for spatial modelling. (iii) Tissue deformation during fixation and sectioning is handled via a staged registration (rigid -> affine -> SyN) with explicit QC, reducing local misalignment that can corrupt voxel-level analyses. (iv) Interoperability is achieved by emitting volumes and masks that can plug into atlas-based pipelines, enabling cross-modal statistics, parcellation summaries, and group comparisons.

Once in MNI space, histology gains leverage from decades of neuroimaging infrastructure, atlases, parcellations, tract/network maps, and morphometric tools become available. This unlocks analyses that were previously impractical. The method also facilitates multi-center science. Different laboratories can process their slides to a common MNI target and share derivative volumes rather than raw images, easing privacy and standardization concerns. Beyond AD, QNPtoVox can generalize to other neuropathologies, for example, α-synuclein in Parkinson's disease, TDP-43 in FTLD, demyelination in MS, and to the validation of imaging phenotypes by providing ground-truth voxel maps.

7.4.2 An anatomically derived way to study spatial patterns in the brain

NPP advances a data-driven, biologically grounded approach to modelling spatial pathology. The drift incorporates I-BIS covariates that encode white matter architecture, functional networks, vascular territories, and cortical geometry; the residual process captures local dispersion consistent with tissue microstructure. When white-matter and network features dominate, as they did across participants, results support hypotheses about connectomic spread; when residuals retain anisotropy after accounting for covariates, they point to micro-scale processes (local circuitry, glial responses) not yet fully captured by current atlases.

Importantly, NPP supports analyses at multiple spatial scales. Voxels resolve fine structure within gyri and sulci; regional summaries connect to clinical staging and trial endpoints; network-level aggregates capture mesoscale organization. This multiscale view clarifies how micro-scale pathology accumulates into macro-scale dysfunction, a linkage often asserted but rarely quantified. As the covariate library expands, the drift can absorb more biology, shrinking residual variance and sharpening mechanistic stories. In this sense, NPP is not a static model but a dynamic framework for spatial inference in human brain pathology.

A central contribution of this work is the development of I-BIS as a reusable and extensible covariate library for spatial modeling. I define a consistent methodology for generating point-based, distance-based, and buffer-density features across multiple radii, enabling reproducible feature engineering for diverse neuropathological analyses. Because I-BIS operates entirely in MNI space, it establishes a standardized reference frame for integrating more neuroanatomical features, including cytoarchitectonic laminar layers, myelination maps, geodesic distance matrices, and arterial maps. This also facilitates cross-laboratory model sharing without the need for re-registration.

The use of a common MNI coordinate framework enables one-time voxel-level covariate computation. This significantly reduces redundant processing when incorporating new datasets or conducting follow-up analyses. For this dissertation, I computed covariates for 231 variables of interest across more than 400,000 voxels in the middle frontal gyrus (MFG). These MNI-referenced voxels now constitute a ready-to-use covariate repository, allowing future analyses of the same region to proceed without repeating computationally expensive preprocessing steps. The framework is also engineered for computational efficiency. Distance and buffer operations are fully vectorized, memory-intensive processes are executed in spatial chunks to avoid resource bottlenecks, and intermediate results are cached to prevent redundant calculations across projects. This balance between methodological rigor and computational practicality ensures that I-BIS remains both scientifically robust and scalable to large neuroimaging datasets.

This standardization lowers the barrier to entry for spatial modelling and supports comparability and replication, two chronic pain points in pathology research. It also invites methodological innovation: connectome-aware kernels, graph features, or omics-derived scores can be added as new layers without rewriting downstream code. In time, I-BIS can serve as a community exchange format for voxel-based covariates, enabling multi-center harmonization and cumulative science.

7.4.3 Universal kriging in brain space

UK has a long history in geospatial and environmental exposure mapping, but its translation to the neuroanatomical frame of reference used in human brain research has been limited. This dissertation adapts UK to the MNI space and demonstrates that it is not merely an exotic alternative but a principled, practically useful estimator for turning sparse quantitative histology into continuous**,** brain maps. In contrast to ordinary kriging, which assumes a constant mean, UK explicitly models a drift, here, a biologically interpretable mean function built from I-BIS covariates and layers on a spatially correlated residual. This decomposition aligns naturally with current theories of disease spread: the drift captures variation explained by known anatomical constraints, while the residual represents local dispersion.

From a statistical standpoint, the residuals in the UK approach are parameterized by the variogram, whose fitted characteristics in Chapter 6 (near-zero nugget, partial sills ~0.5–0.95, ranges ~7–16 mm) are biologically interpretable and actionable. A near-zero nugget indicates that, after the 2-mm smoothing in QNPtoVox, most residual variance is spatially structured rather than measurement noise; the 7–16 mm range marks the distance over which information is shared, corresponding to gyral-scale anatomy. These parameters explain and predict the cross-validation behavior observed in 7.1.

The adaptation to brain space required several design choices that constitute methodological contributions. First, the I-BIS engine supplies point, distance, and buffer-zone density features at multiple radii, enabling the drift to encode hypotheses about connectomic spread, perivascular influence, and geometric constraints. Second, I adopted a dimensionality reduction approach to stabilize the drift, reduce multicollinearity across hundreds of covariates, and retain a compact set of predictors per participant. Third, I introduced a spatial cross-validation protocol tuned to kriging: fold geometry is created by clustering voxels into contiguous blocks.

The neurobiological interpretability of UK is another advantage. When white-matter and network features dominate the drift, as they did across participants (section 7.1.2), the resulting maps support mechanistic accounts of network-constrained spread. When substantial residual structure persists after accounting for these features, it motivates targeted enrichment of I-BIS (e.g., cytoarchitecture priors, myelination maps) and suggests micro-scale processes not yet captured by the drift. In this way, kriging becomes a lens for hypothesis generation as much as a predictor.

Finally, the work establishes guardrails for applying kriging in neuroanatomy. I show that (i) performance stabilizes when voxel-to-predictor ratios are on the order of 4–6K:1; (ii) mis-registration and edge effects must be monitored with automated QC; and (iii) uncertainty should always be reported and calibrated. These practices make UK both reliable and reproducible in the brain setting and provide a reference against which emerging alternatives, such as non-stationary Gaussian processes, graph diffusion, or deep networks, can be compared. In short, this dissertation elevates universal kriging from a niche import to a probable core method for brain-wide inference about microscopic pathology distribution, tightly integrated with neuroimaging coordinates and biological context.

7.4.4 Three open-source, semi-automated pipelines

I developed three interoperable pipelines, each semi-automated to allow researchers control, and each packaged with documentation, to support community use:

- **I-BIS covariate engine** (Chapter 3): produces biologically interpretable voxel-based features (point, distance, buffer zone density) from tracts, networks, vasculature, cytoarchitecture, and geometry. Beyond NPP, I-BIS can support stroke risk mapping, neurosurgical targeting, normative modelling of lesion distributions, or feature engineering for connectome-aware ML.

- **QNPtoVox** (Chapter 4): generates voxel-based pathology volumes aligned to MNI. Beyond NPP, labs can use it for quantitative lesion cartography (microinfarcts, CTE lesions), atlas-building from serial sections, teaching datasets, or validation of imaging phenotypes against ex-vivo ground truth. The tool exposes parameters for slice thickness, smoothing kernel, and registration techniques, so users can adapt to different stains and scanners.

- **NPP spatial modelling** (Chapters 5): couples UK with covariate drift and spatial CV to yield predictive maps. Beyond NPP, the UK can be used to model other pathological spatial associations in the human brain.

Modularity is deliberate: each pipeline solves a distinct class of problems while composing cleanly with the others. Together, they form a complete research stack, from raw slides to validated brain-wide predictions, but labs can adopt only the pieces they need. The public release includes setup, environment handling, and configuration files.

7.5 Limitations

In designing NPP, I made several deliberate choices to prioritize internal validity, to develop and stress-test the pipeline on a single brain region (MFG), to register quantitative histology to a common 0.5-mm coordinate framework, and to model spatial dependence with universal kriging. These decisions yielded a tractable proof-of-concept and enabled clear methodological inferences, but they also impose boundary conditions on what I can claim. In particular, the present findings reflect (i) the biology of ten donors drawn from a specific sampling protocol, (ii) the technical properties of our histology, MRI, and registration procedures, and (iii) the assumptions of the chosen model class (stationary residuals with a finite correlation range).

Throughout the thesis, I implemented safeguards, rigorous co-registration QC, harmonized preprocessing, and spatial block cross-validation at multiple granularities. These practices reduce risk but do not eliminate it: post-mortem factors and fixation history introduce variance that small cohorts cannot absorb; block selection and fold geometry can interact with spatial range; smoothing that stabilizes kriging can blur micro-scale contrasts; and semi-automated steps leave room for operator effects. As a result, the conclusions should be understood as general patterns supported by converging evidence, not as a final atlas of AD pathology.

I therefore organize the limitations below around four themes that most materially shape interpretation: (1) cohort size and sampling scope, which bear on external validity; (2) spatial resolution and scale mismatch between histology, MRI, and modelling; (3) model-class constraints inherent to kriging (finite range, stationarity) and the implications

for study design; and (4) automation and computational controls, which affect throughput and reproducibility. Each limitation directly motivates the extensions and safeguards laid out in sections 7.2 and 7.3.

7.5.1 Cohort size and sampling scope

The present analysis relies on ten post-mortem donors and focuses on a single cortical territory, the middle frontal gyrus (MFG). This was intentional: I treated MFG as a controlled testbed to evaluate QNPtoVox and the NPP modelling stack under tightly standardized conditions. Nevertheless, a cohort of this size cannot fully characterize inter-individual heterogeneity arising from age, sex, education, cortical atrophy, co-pathology burden, or genetic background. Nor can one region capture the diversity of tau trajectories that traverse medial temporal, parietal, occipital, and cingulo-insular systems. As such, the findings should be interpreted as proof-of-concept, not a definitive atlas of AD pathology.

Sampling scope also limits generalizability. Histological slabs were selected based on tissue availability and existing research priorities, which risks ascertainment bias toward neocortical association areas and away from subcortical or allocortical structures. Post-mortem interval, fixation history, and slide preparation introduce additional variance that a small cohort cannot average out. These constraints argue for replication across additional brain regions (hippocampus, entorhinal/transentorhinal, superior/middle temporal gyri, inferior parietal, cingulate, primary visual cortex) and across independent corpora (e.g., community-based resources such as ROS–MAP) that differ in demographics and collection protocols. Only with broader regional coverage and external datasets can we determine which relationships, particularly the dominance of white-matter and network covariates, are universal versus brain region- or cohort-specific.

7.5.2 Spatial resolution and scale mismatch

All volumes are expressed at 0.5-mm isotropic resolution. This choice balanced the need to register histology to MRI in a common space against the computational burden of micro-scale modelling, and it aligns with common ex-vivo MRI resolutions. Still, a 0.5-mm voxel averages over hundreds of microns of tissue containing many cellular processes; fine-grained gradients across laminar layers, columns, or perivascular sleeves are therefore partially averaged. The 2-mm smoothing applied during QNPtoVox further suppresses micro-scale fluctuations, which helps kriging (lower nugget) but may blunt subtle boundaries, especially near layer transitions or thin tracts.

As neuroimaging advances toward finer resolutions and as spatial omics becomes routine, the coordinate framework should evolve accordingly. Surface-based representations (geodesic distances via the cortical sheet) and finer-resolution modelling better capture micro-scale dynamics. Until then, results that hinge on sub-millimeter contrasts should be interpreted cautiously, and follow-up validation at native histology resolution is warranted.

7.5.3 Finite correlation distance and stationarity in UK

Universal kriging assumes a stationary residual process with a finite correlation range. Empirically, fitted ranges clustered around 7–16 mm, which explains why performance improved as spatial CV split size increased (64–100 blocks) and degraded at coarser splits. Beyond the fitted range, the UK transitions from interpolation to extrapolation, and the model is not good at extrapolation. This introduces two practical limitations. First, study designs that place large unsampled gaps between training and test regions will artificially depress UK performance, potentially conflating model weakness with sampling geometry. Second, strict stationarity ignores spatially varying length-scales and anisotropies introduced by tract architecture and cortical folding; even with I-BIS drift, residual structure may differ between deep white matter and gyral crowns.

Two broad remedies follow. One is design-oriented: adopt tissue sampling protocols that maintain inter-sample distances within the estimated range for the target region, thereby keeping UK in its interpolation regime. The other is model-oriented: embrace hybrid approaches that use UK where it excels (short-to-mid-range structure) and layer additional models for long-range or non-stationary effects, e.g., graph-informed diffusion along tractography, non-stationary Gaussian processes with spatially varying kernels, or neural networks that encode global context. These extensions are part of the future-work agenda as mentioned in section 7.2, but until implemented, they represent a ceiling on what the current model class can accomplish.

7.5.4 Automation, throughput, and computational controls

Although the pipelines are open-source, semi-automated, and customizable, several stages still require manual oversight: slab-to-MRI co-location, parameter tuning for ANTs registration, and human review of QC reports. Manual intervention is appropriate during method development but does not scale to high-throughput studies or multi-center deployment. Manual steps also introduce operator variance and raise barriers to exact replication.

A second limitation is the computational footprint. Voxel-based covariate generation, kriging solves with thousands of predictors, and spatial CV across dozens of folds are resource-intensive. Without containerization and strict dependency management, long-running jobs are vulnerable to environment drift; without task orchestration, reruns after minor configuration edits incur unnecessary cost.

The path forward, outlined in 7.2, includes (i) deeper automation wherever appropriate, (ii) containerized execution, (iii) job orchestration that parallelizes and caches intermediate outputs, and (iv) richer provenance capture so that every figure and table can be recreated verbatim. Until these improvements are complete, throughput and reproducibility remain limiting factors for large-scale deployment.

7.6 Conclusion

I began this dissertation with a simple premise and an ambitious goal. The premise was that the spatial distribution of neuropathological proteins, their location, and proximity to specific anatomical pathways, carries information essential for understanding Alzheimer's disease (AD). The goal was to transform sparse, heterogeneous quantitative histopathology into region-wide, interpretable maps that could be analyzed with the same spatial logic long available to neuroimaging. To do so, I designed and validated a complete workflow: QNPtoVox for mapping digital histology to MNI space; I-BIS for assembling biologically interpretable, voxel-based covariates; and NeuroPathPredict (NPP), centered on universal kriging (UK), for prediction with calibrated uncertainty. The work across Chapters 2–6 demonstrates that this stack functions end-to-end on real tissue and that it yields inferences about AD pathology that are both statistically rigorous and biologically meaningful.

From the outset, I framed three questions. First, can I co-locate quantitative histology with ex-vivo MRI in a way that preserves intensity, orientation, and provenance, so that volumes from different donors and stains are commensurate in a common coordinate system? Second, can I build a covariate library that embeds each voxel in the anatomical context most likely to shape disease spread, white-matter tracts, functional networks, vasculature, and cortical geometry, without sacrificing interpretability? Third, given those ingredients, can I recover spatially coherent pathology fields from limited samples, quantify uncertainty honestly, and explain prediction in terms that engage current theories of AD progression? The answers are affirmative. QNPtoVox produced high-quality MNI-registered volumes; I-BIS delivered a reusable feature stack with point, distance, and buffer zone density representations; and UK, paired with a tailored spatial cross-validation scheme, generated continuous maps whose performance and uncertainty behave in line with the fitted variograms.

Several results stand out. First, I found that white-matter architecture and functional networks dominate the predictor landscape. Features derived from major tracts (e.g., SLF, IFOF, thalamic radiations) and functional networks consistently survived selection, outranking purely regional labels. This pattern is consistent across donors and supports a propagation narrative in which tau burden respects mesoscale highways and network affiliations. Second, I showed that prediction quality depends on spatial scale in ways that are both empirically demonstrable and theoretically expected. Across ten participants, UK performance improved monotonically as cross-validation moved from coarse to fine partitions, with the strongest accuracy at 64–100 splits. This is precisely what one would predict if residual spatial correlation decays within a characteristic range: smaller held-out blocks lie closer to observed tissue, stay within range, and are therefore interpolated with higher fidelity. Variogram fits revealed near-zero nuggets and ranges on the order of 7–16 mm, an anatomical “scale bar” that aligns with gyral dimensions and short-range association fibers, and the cross-validation behavior mirrored those ranges. Third, the approach is robust across participants when data density is adequate. With finer partitions, median R^2 values reached the mid-0.8s, and the same families of predictors recurred, suggesting genuine regularities rather than idiosyncratic over-fitting. Together, these findings validate core assumptions behind NPP and provide quantitative guidance for future study design.

Equally important are the qualitative shifts this work enables. By anchoring histology in neuroimaging coordinates, the dissertation makes voxel-based pathology an important component in the analytic ecosystem. Histology can now be queried with atlases, parcellations, and connectomic context; compared across donors and studies without bespoke co-registration; and linked directly to MRI and PET. This bridge makes it possible to calibrate in vivo signals against ground truth, to adjudicate apparent discrepancies (e.g., off-target binding or partial volume artefacts), and to ask genuinely mechanistic questions: Does vasculature modulate burden independently of tracts? Do default-mode hubs act as amplifiers for tau spread? Does network membership predict where pathology will appear next? Because NPP produces both a mean prediction and a predictive variance, these questions can be pursued with an explicit sense of confidence and with active-learning strategies that target additional histology to the regions of greatest informational value.

The work also clarifies limits that matter for interpretation and for method development. UK delivers the best linear unbiased predictor under its assumptions, but those assumptions bring an explicit, finite correlation distance and (for the residual) stationarity. The fitted ranges explain why coarse spatial splits underperform and why predictions degrade as train–test separation increases. Rather than treating this as a weakness, I

interpret it as a scale constraint that can be engineered around by: (i) designing sampling protocols that keep inter-sample distances within range; (ii) using UK to model short- and mid-range structure while delegating long-range extrapolation to models with global context (e.g., graph-based diffusion or non-stationary Gaussian processes); and (iii) moving toward surface-aware variants that better respect cortical topology. In short, the dissertation both exploits and exposes the geometry of the problem, offering a transparent baseline against which more flexible models can be compared.

Within the broader field of AD research, the present work contributes to three ongoing conversations. The first concerns staging. Braak and Braak's[9] framework remains central but was derived from a small sample and expressed in coarse regional ranks. NPP makes it possible to revisit staging quantitatively at voxel resolution, in a way that is probabilistic and connectivity-aware. Even in the limited MFG testbed, the dominance of tract and network features hints that staging may be better understood as sequences constrained by anatomical infrastructure rather than as purely regional handovers. The second conversation concerns heterogeneity and subtypes. AD does not proceed uniformly; limbic--predominant, hippocampal-sparing, posterior-predominant, dysexecutive, and language-led variants are all described, often based on symptoms or imaging. NPP provides the substrate for a pathology-anchored view of subtypes: it can map subtype-specific trajectories in the coordinates of tracts and networks, link those trajectories to cognitive change, and, crucially, reveal when apparent subtypes reflect co-pathology rather than distinct tau pathways. The third conversation concerns methodological standards. The field often struggles with small samples, idiosyncratic preprocessing, and opaque evaluation. By releasing open pipelines, recommending spatially aware validation, and advocating uncertainty reporting alongside accuracy, the dissertation pushes toward practices that make results comparable and cumulative.

Although the experiments focus on tau in MFG, the framework is marker-agnostic and region-agnostic by design. The immediate next steps (outlined in sections 7.2–7.3) are to expand to the NIA-AA sampling set, hippocampus, entorhinal/transentorhinal cortex, temporal and parietal association areas, cingulate, and primary visual cortex, and to integrate additional stains, particularly β-amyloid and TDP-43, in the same donors and regions. Doing so will allow explicit quantification of interaction patterns: where do pathologies co-localize, where do they diverge, and along which anatomical conditions do interactions emerge? The same extension will enable a data-driven re-examination of Braak staging at scale and a rigorous evaluation of subtype trajectories. As the coordinate framework incorporates more anatomical maps, the drift will absorb more neurobiology, and the residual will shrink, bringing us closer to understanding the underlying mechanisms.

In parallel, the project opens a tractable path for data-efficient machine learning in post-mortem research. Deep generative and predictive models promise whole-brain inference at speed, but they have been hampered by limited labelled data and inconsistent preprocessing. NPP directly addresses both barriers: kriged volumes and their variances provide uncertainty-aware supervision for pre-training or distillation, and the common coordinate system turns disparate cases into an interoperable corpus. With this scaffolding, diffusion models, CNNs, and graph neural networks can learn multi-scale, connectivity-aware spatial rules, and their outputs can be evaluated fairly under shared splits and structure-sensitive metrics. The reward is not merely performance but generalization: models that carry predictive power across donors, regions, and cohorts because they are grounded in anatomy.

I am mindful that methodological advances must serve biological and clinical ends. The long-term vision is a research ecosystem in which ex vivo pathology, in vivo imaging, and computational models inform one another. In such an ecosystem, voxel-based pathology maps in MNI space will calibrate new phenotypes, anchor imaging-based subtypes, and provide mechanistic priors for prognostic models. Spatial atlases of multi-marker burden will identify candidate targets for combination therapies and reveal network territories where intervention might slow the spread most effectively. Uncertainty maps will guide tissue sampling and spatial omics, prioritizing experiments with the highest expected information gain. And, ultimately, harmonized pipelines and datasets will allow multi-center consortia to accumulate evidence about AD progression that is both granular and generalizable.

To close, I return to the motivating problem: we study a disease that unfolds across years and across the whole brain, yet we often measure it in thin slivers of tissue. This dissertation offers one route to reconcile the scales. By situating quantitative histopathology within a common spatial frame, enriching it with interpretable biological context, and modelling the remaining structure with methods that acknowledge distance and uncertainty, I have shown that it is possible to move from local slices to region-wide inferences without surrendering rigor. The immediate contributions are practical, open pipelines, design heuristics, and a set of empirically supported regularities about spatial scale and biological predictors. The larger contribution is conceptual: NPP demonstrates that spatial prediction in the brain can be both data-driven and mechanistically anchored, making it a credible instrument for answering some of the key questions that matter in AD. The path forward is clear: expand regions and markers, integrate spatial omics, develop hybrid models that link geostatistics with connectome-aware deep learning, and validate across independent cohorts. If we do so, we will not only refine our maps of where AD has been, we will also improve our ability to say where it is going and how, if at all, we might change its course.

Bibliography

1. 2024 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*. 2024;20(5):3708-3821. doi:10.1002/alz.13809

2. Vaz M, Silvestre S. Alzheimer's disease: Recent treatment strategies. *European Journal of Pharmacology*. 2020;887:173554-173554. doi:10.1016/J.EJPHAR.2020.173554

3. Busche MA, Hyman BT. Synergy between amyloid-β and tau in Alzheimer's disease. *Nature Neuroscience 2020 23:10*. 2020;23(10):1183-1193. doi:10.1038/s41593-020-0687-6

4. Meneses A, Koga S, O'Leary J, Dickson DW, Bu G, Zhao N. TDP-43 Pathology in Alzheimer's Disease. *Molecular Neurodegeneration 2021 16:1*. 2021;16(1):1-15. doi:10.1186/S13024-021-00503-X

5. van der Kant R, Goldstein LSB, Ossenkoppele R. Amyloid-β-independent regulators of tau pathology in Alzheimer disease. *Nature Reviews Neuroscience 2019 21:1*. 2019;21(1):21-35. doi:10.1038/s41583-019-0240-3

6. Josephs KA, Murray ME, Whitwell JL, et al. Staging TDP-43 pathology in Alzheimer's disease. *Acta Neuropathologica*. 2014;127(3):441-450. doi:10.1007/S00401-013-1211-9/METRICS

7. Gómez-Isla T, Hollister R, West H, et al. Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease. *Annals of Neurology*. 1997;41(1):17-24. doi:10.1002/ANA.410410106

8. Braak H, Braak E. Demonstration of Amyloid Deposits and Neurofibrillary Changes in Whole Brain Sections. *Brain Pathology*. 1991;1(3):213-216. doi:10.1111/j.1750-3639.1991.tb00661.x

9. Braak H, Braak E. Staging of alzheimer's disease-related neurofibrillary changes. *Neurobiology of Aging*. 1995;16(3):271-278. doi:10.1016/0197-4580(95)00021-6

10. La Joie R, Visani AV, Lesman-Segev OH, et al. Association of APOE4 and Clinical Variability in Alzheimer Disease With the Pattern of Tau- and Amyloid-PET. *Neurology*. 2021;96(5):e650-e661. doi:10.1212/WNL.0000000000011270

11. Vogel JW, Young AL, Oxtoby NP, et al. *Four Distinct Trajectories of Tau Deposition Identified in Alzheimer's Disease*. Vol 27.; 2021. doi:10.1038/s41591-021-01309-6

12. Uretsky M, Gibbons LE, Mukherjee S, et al. Longitudinal cognitive performance of Alzheimer's disease neuropathological subtypes. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*. 2021;7(1):e12201. doi:10.1002/TRC2.12201

13. Vogel JW, Hansson O. Subtypes of Alzheimer's disease: questions, controversy, and meaning. *Trends in Neurosciences*. 2022;45(5):342-345. doi:10.1016/j.tins.2022.02.001

14. Dubois B, Feldman HH, Jacova C, et al. Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *The Lancet Neurology*. 2014;13(6):614-629. doi:10.1016/S1474-4422(14)70090-0

15. Fonov V, Evans A, McKinstry R, Almli C, Collins D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*. 2009;47:S102. doi:10.1016/S1053-8119(09)70884-5

16. HALO | Quantitative Image Analysis for Pathology - Indica Labs. Accessed August 2, 2025. https://indicalab.com/halo/

17. Mercer LD, Szpiro AA, Sheppard L, et al. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric environment (Oxford, England : 1994)*. 2011;45(26):4412. doi:10.1016/J.ATMOSENV.2011.05.043

18. Prusiner SB. Biology and Genetics of Prions Causing Neurodegeneration. *Annu Rev Genet*. 2013;47:601-623. doi:10.1146/annurev-genet-110711-155524

19. Brundin P, Melki R, Kopito R. Prion-like transmission of protein aggregates in neurodegenerative diseases. *Nat Rev Mol Cell Biol*. 2010;11(4):301-307. doi:10.1038/nrm2873

20. Braak H, Alafuzoff I, Arzberger T, Kretzschmar H, Del Tredici K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol*. 2006;112(4):389-404. doi:10.1007/s00401-006-0127-z

21. Thal DR, Capetillo-Zarate E, Del Tredici K, Braak H. The Development of Amyloid β Protein Deposits in the Aged Brain. *Science of Aging Knowledge Environment*. 2006;2006(6):re1-re1. doi:10.1126/sageke.2006.6.re1

22. Thal DR, Rüb U, Orantes M, Braak H. Phases of A beta-deposition in the human brain and its relevance for the development of AD. *Neurology*. 2002;58(12):1791-1800. doi:10.1212/wnl.58.12.1791

23. Hampel H, Hardy J, Blennow K, et al. The Amyloid-β Pathway in Alzheimer's Disease. *Mol Psychiatry*. 2021;26(10):5481-5503. doi:10.1038/s41380-021-01249-0

24. Hyman BT, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement*. 2012;8(1):1-13. doi:10.1016/j.jalz.2011.10.007

25. Schöll M, Lockhart SN, Schonhaut DR, et al. PET Imaging of Tau Deposition in the Aging Human Brain. *Neuron*. 2016;89(5):971-982. doi:10.1016/j.neuron.2016.01.028

26. Maass A, Landau S, Baker SL, et al. Comparison of multiple tau-PET measures as biomarkers in aging and Alzheimer's disease. *NeuroImage*. 2017;157:448-463. doi:10.1016/j.neuroimage.2017.05.058

27. Therriault J, Pascoal TA, Lussier FZ, et al. Biomarker modeling of Alzheimer's disease using PET-based Braak staging. *Nat Aging*. 2022;2(6):526-535. doi:10.1038/s43587-022-00204-0

28. Schöll M, Lockhart SN, Schonhaut DR, et al. PET Imaging of Tau Deposition in the Aging Human Brain. *Neuron*. 2016;89(5):971-982. doi:10.1016/j.neuron.2016.01.028

29. Hyman BT, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement*. 2012;8(1):1-13. doi:10.1016/j.jalz.2011.10.007

30. Biel D, Brendel M, Rubinski A, et al. Tau-PET and in vivo Braak-staging as prognostic markers of future cognitive decline in cognitively normal to demented individuals. *Alzheimer's Research & Therapy*. 2021;13(1):137. doi:10.1186/s13195-021-00880-x

31. Marquié M, Verwer EE, Meltzer AC, et al. Lessons learned about [F-18]-AV-1451 off-target binding from an autopsy-confirmed Parkinson's case. *Acta Neuropathol Commun*. 2017;5:75. doi:10.1186/s40478-017-0482-0

32. Baker SL, Harrison TM, Maass A, Joie RL, Jagust WJ. Effect of Off-Target Binding on 18F-Flortaucipir Variability in Healthy Controls Across the Life Span. *Journal of Nuclear Medicine*. 2019;60(10):1444-1451. doi:10.2967/jnumed.118.224113

33. Schwarz CG, Therneau TM, Weigand SD, et al. Selecting software pipelines for change in flortaucipir SUVR: Balancing repeatability and group separation. *NeuroImage*. 2021;238:118259. doi:10.1016/j.neuroimage.2021.118259

34. Kotari V, Southekal S, Navitsky M, et al. Early tau detection in flortaucipir images: validation in autopsy-confirmed data and implications for disease progression. *Alzheimer's Research & Therapy*. 2023;15(1):41. doi:10.1186/s13195-023-01160-6

35. Montine TJ, Monsell SE, Beach TG, et al. Multisite assessment of NIA-AA guidelines for the neuropathologic evaluation of Alzheimer's disease. *Alzheimers Dement*. 2016;12(2):164-169. doi:10.1016/j.jalz.2015.07.492

36. Mirra SS, Gearing M, McKeel DW Jr, et al. Interlaboratory Comparison of Neuropathology Assessments in Alzheimer's Disease: A Study of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). *J Neuropathol Exp Neurol*. 1994;53(3):303-315. doi:10.1097/00005072-199405000-00012

37. Khodanovich MY, Anan'ina TV, Krutenkova EP, et al. Challenges and Practical Solutions to MRI and Histology Matching and Measurements Using Available ImageJ Software Tools. *Biomedicines*. 2022;10(7):1556. doi:10.3390/biomedicines10071556

38. Goubran M, Crukley C, de Ribaupierre S, Peters TM, Khan AR. Image registration of *ex-vivo* MRI to sparsely sectioned histology of hippocampal and neocortical temporal lobe specimens. *NeuroImage*. 2013;83:770-781. doi:10.1016/j.neuroimage.2013.07.053

39. Leuzy A, Raket LL, Villemagne VL, et al. Harmonizing tau positron emission tomography in Alzheimer's disease: The CenTauR scale and the joint propagation model. *Alzheimers Dement*. 2024;20(9):5833-5848. doi:10.1002/alz.13908

40. Masel J, Jansen VAA, Nowak MA. Quantifying the kinetic parameters of prion replication. *Biophysical Chemistry*. 1999;77(2-3):139-152. doi:10.1016/S0301-4622(99)00016-2

41. Payne RJH, Krakauer DC. The spatial dynamics of prion disease. *Proceedings of the Royal Society of London Series B: Biological Sciences*. 1998;265(1412):2341-2346. doi:10.1098/rspb.1998.0581

42. Raj A. Network models of dementia progression: the way forward? *http://dx.doi.org/102217/fnl1267*. 2012;7(6):659-662. doi:10.2217/FNL.12.67

43. Raj A, LoCastro E, Kuceyeski A, Tosun D, Relkin N, Weiner M. Network Diffusion Model of Progression Predicts Longitudinal Patterns of Atrophy and Metabolism in Alzheimer's Disease. *Cell Reports*. 2015;10(3):359-369. doi:10.1016/j.celrep.2014.12.034

44. Hu C, Hua X, Ying J, Thompson PM, Fakhri GE, Li Q. Localizing sources of brain disease progression with network diffusion model. *IEEE journal of selected topics in signal processing*. 2016;10(7):1214-1225.

45. Iturria-Medina Y, Sotero RC, Toussaint PJ, Evans AC, Initiative ADN. Epidemic spreading model to characterize misfolded proteins propagation in aging and associated neurodegenerative disorders. *PLoS computational biology*. 2014;10(11):e1003956.

46. Mezias C, Raj A. Analysis of Amyloid-β pathology spread in mouse models suggests spread is driven by spatial proximity, not connectivity. *Frontiers in Neurology*. 2017;8:653.

47. Oxtoby NP, Garbarino S, Firth NC, et al. Data-driven sequence of changes to anatomical brain connectivity in sporadic Alzheimer's disease. *Frontiers in neurology*. 2017;8:580.

48. Data-Driven Disease Progression Modeling | SpringerLink. Accessed August 16, 2025. https://link.springer.com/protocol/10.1007/978-1-0716-3195-9_17

49. Cohen MA, Adar SD, Allen RW, et al. Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environmental Science and Technology*. 2009;43(13):4687-4693. doi:10.1021/ES8030837/SUPPL_FILE/ES8030837_SI_001.PDF

50. Keller JP, Olives C, Kim SY, et al. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environmental Health Perspectives*. 2015;123(4):301-309. doi:10.1289/EHP.1408145

51. Zaleshina M, Zaleshin A. The brain as a multi-layered map. Scales and reference points for pattern recognition in neuroimaging. *European Journal of Geography*. 2017;8(1). Accessed August 16, 2025. https://www.eurogeojournal.eu/index.php/egj/article/download/278/222

52. Jerrett M, Arain A, Kanaroglou P, et al. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology*. 2005;15(2):185-204.

53. Hoek G, Beelen R, de Hoogh K, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*. 2008;42(33):7561-7578. doi:10.1016/j.atmosenv.2008.05.057

54. Keller JP, Olives C, Kim SY, et al. A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. *Environ Health Perspect*. 2015;123(4):301-309. doi:10.1289/ehp.1408145

55. Wackernagel H. *Multivariate Geostatistics*. Springer Berlin Heidelberg; 2003. doi:10.1007/978-3-662-05294-5

56. Matheron G. The theory of regionalised variables and its applications. *Les Cahiers du Centre de Morphologie Mathématique*. 1971;5:212.

57. Cressie N. *Statistics for Spatial Data*. John Wiley & Sons; 2015. Accessed August 16, 2025. https://books.google.com/books?hl=en&lr=&id=MzN_BwAAQBAJ&oi=fnd&pg=PP1&dq=Cressie,+N.+(1993).+Statistics+for+Spatial+Data+(rev.+ed.)&ots=NN_kx_250J&sig=24YWXn_HsBXEJJG1YyKAwPbF_e4

58. Chiles JP, Delfiner P. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons; 2012. Accessed August 16, 2025. https://books.google.com/books?hl=en&lr=&id=CUC55ZYqe84C&oi=fnd&pg=PR9&dq=Chil%C3%A8s,+J.-P.,+%26+Delfiner,+P.+(2012).+Geostatistics:+Modeling+Spatial+Uncertainty+(2nd+ed.).&ots=29nAT26Yvt&sig=hV8Rl_yLGUvARso2sWiDKKHgOz8

59. Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC; 2003. Accessed August 16, 2025. https://www.taylorfrancis.com/books/mono/10.1201/9780203487808/hierarchical-modeling-analysis-spatial-data-sudipto-banerjee-sudipto-banerjee-bradley-carlin-alan-gelfand

60. Crane PK, Groot C, Ossenkoppele R, et al. Cognitively defined Alzheimer's dementia subgroups have distinct atrophy patterns. doi:10.1002/alz.13567

61. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL. Unbiased Average Age-Appropriate Atlases for Pediatric Studies. *Neuroimage*. 2011;54(1):313-327. doi:10.1016/j.neuroimage.2010.07.033

62. Adult Changes in Thought (ACT) Study. ACT Study Home. https://actagingresearch.org/index.php?cID=308#data-sharing-policies

63. Mukhopadhyay S, Feldman MD, Abels E, et al. Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study). *The American Journal of Surgical Pathology*. 2018;42(1):39. doi:10.1097/PAS.0000000000000948

64. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1):16878. doi:10.1038/s41598-017-17204-5

65. Kapasi A, Poirier J, Hedayat A, et al. High-throughput digital quantification of Alzheimer disease pathology and associated infrastructure in large autopsy studies. *J Neuropathol Exp Neurol*. 2023;82(12):976-986. doi:10.1093/jnen/nlad086

66. Porzig R, Singer D, Hoffmann R. Epitope mapping of mAbs AT8 and Tau5 directed against hyperphosphorylated regions of the human tau protein. *Biochemical and Biophysical Research Communications*. 2007;358(2):644-649. doi:10.1016/j.bbrc.2007.04.187

67. Xia Y, Prokop S, Gorion KMM, et al. Tau Ser208 phosphorylation promotes aggregation and reveals neuropathologic diversity in Alzheimer's disease and other tauopathies. *Acta Neuropathol Commun*. 2020;8:88. doi:10.1186/s40478-020-00967-w

68. Yeo BTT, Krienen FM, Sepulcre J, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol*. 2011;106(3):1125-1165. doi:10.1152/jn.00338.2011

69. Mouches P, Forkert ND. A statistical atlas of cerebral arteries generated using multi-center MRA datasets from healthy subjects. *Scientific Data 2019 6:1*. 2019;6(1):1-8. doi:10.1038/s41597-019-0034-5

70. Yeh FC. Population-based tract-to-region connectome of the human brain and its hierarchical topology. *Nature communications*. 2022;13(1):4933.

71. Schwarz AJ, Yu P, Miller BB, et al. Regional profiles of the candidate tau PET ligand 18F-AV-1451 recapitulate key features of Braak histopathological stages. *Brain*. 2016;139(Pt 5):1539-1550. doi:10.1093/brain/aww023

72. Adult Changes in Thought Study: Dementia is an Individually Varying Convergent Syndrome with Prevalent Clinically Silent Diseases that may be Modified by Some Commonly Used Therapeutics - PMC. Accessed August 15, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC3409333/?utm_source=chatgpt.com

73. Marcum ZA, Keene CD, Larson EB. Leveraging neuropathological data in pharmacoepidemiology: A promising approach for dementia prevention? *Pharmacoepidemiology and Drug*. 2021;30(1):1-3. doi:10.1002/pds.5068

74. Latimer CS, Melief EJ, Ariza-Torres J, et al. Protocol for the Systematic Fixation, Circuit-Based Sampling, and Qualitative and Quantitative Neuropathological Analysis of Human Brain Tissue. In: Chun J, ed. *Alzheimer's Disease: Methods and Protocols*. Springer US; 2023:3-30. doi:10.1007/978-1-0716-2655-9_1

75. Montine TJ, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol*. 2012;123(1):1-11. doi:10.1007/s00401-011-0910-3

76. FreeSurfer. FreeSurfer. Accessed March 26, 2025. https://surfer.nmr.mgh.harvard.edu

77. McCarthy P. FSLeyes. Published online May 29, 2025. doi:10.5281/zenodo.15542963

78. SciPy 1.0: fundamental algorithms for scientific computing in Python | Nature Methods. Accessed August 5, 2025. https://www.nature.com/articles/s41592-019-0686-2

79. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*. 2008;12(1):26-41. doi:10.1016/j.media.2007.06.004

80. Horn A, Li N, Dembek TA, et al. Lead-DBS v2: Towards a comprehensive pipeline for deep brain stimulation imaging. *NeuroImage*. 2019;184:293-316. doi:10.1016/j.neuroimage.2018.08.068

81. Horn A. MNI T1 6thGen NLIN to MNI 2009b NLIN ANTs transform. Published online 2016:208235992 Bytes. doi:10.6084/M9.FIGSHARE.3502238.V1

82. Talairach coordinates. In: *Wikipedia*. ; 2025. Accessed March 20, 2025. https://en.wikipedia.org/w/index.php?title=Talairach_coordinates&oldid=1271977922

83. Model HB, Amunts K, Lepage C, et al. BigBrain: An Ultrahigh-Resolution 3D Human Brain Model. *Science*. 2013;340(June):1472-1475.

84. Revell AY, Silva AB, Arnold TC, et al. A framework For brain atlases: Lessons from seizure dynamics. *NeuroImage*. 2022;254:118986. doi:10.1016/j.neuroimage.2022.118986

85. Manera AL, Dadar M, Fonov V, Collins DL. CerebrA, registration and manual label correction of Mindboggle-101 atlas for MNI-ICBM152 template. *Sci Data*. 2020;7(1):237. doi:10.1038/s41597-020-0557-9

86. Buckner RL, Krienen FM, Castellanos A, Diaz JC, Yeo BTT. The organization of the human cerebellum estimated by intrinsic functional connectivity. *J Neurophysiol*. 2011;106(5):2322-2345. doi:10.1152/jn.00339.2011

87. Iglesias JE, Augustinack JC, Nguyen K, et al. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. *Neuroimage*. 2015;115:117-137. doi:10.1016/j.neuroimage.2015.04.042

88. Mori S, Oishi K, Jiang H, et al. Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. *NeuroImage*. 2008;40(2):570-582. doi:10.1016/j.neuroimage.2007.12.035

89. Glasser MF, Coalson TS, Robinson EC, et al. A multi-modal parcellation of human cerebral cortex. *Nature*. 2016;536(7615):171-178. doi:10.1038/nature18933

90. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature - PMC. Accessed August 14, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC2937159/

91. Dworetsky A, Seitzman BA, Adeyemo B, et al. Probabilistic mapping of human functional brain networks identifies regions of high group consensus. *NeuroImage*. 2021;237:118164-118164. doi:10.1016/J.NEUROIMAGE.2021.118164

92. Brett M, Markiewicz CJ, Hanke M, et al. nipy/nibabel: 5.1.0. Published online April 3, 2023. doi:10.5281/zenodo.7795644

93. contributors N, Chamma A, Frau-Pascual A, et al. nilearn. Published online January 20, 2025. doi:10.5281/zenodo.14697221

94. Neuroimaging in Python — NiBabel 5.4.0.dev1+g3b1c7b37 documentation. Accessed March 26, 2025. https://nipy.org/nibabel/coordinate_systems.html

95. Biswas A, Si BC, Biswas A, Si BC. Model Averaging for Semivariogram Model Parameters. In: *Advances in Agrophysical Research*. IntechOpen; 2013. doi:10.5772/52339

96. Spatial Interpolation Methods. Accessed June 11, 2025. https://iri.columbia.edu/~rijaf/CDTUserGuide/html/interpolation_methods.html

97. Yeh FC, Vettel JM, Singh A, et al. Quantifying Differences and Similarities in Whole-Brain White Matter Architecture Using Local Connectome Fingerprints. *PLOS Computational Biology*. 2016;12(11):e1005203. doi:10.1371/journal.pcbi.1005203

98. Thomas Yeo BT, Krienen FM, Sepulcre J, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*. 2011;106(3):1125. doi:10.1152/JN.00338.2011

99. Hyman BT, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement*. 2012;8(1):1-13. doi:10.1016/j.jalz.2011.10.007

100. Allen Institute for Brain Science, University of Washington Alzheimer's Disease Research Center, and Kaiser Permanente Washington Health Research Institute (2022). Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) -- MERFISH - middle temporal gyrus [Dataset]. https://registry.opendata.aws/allen-sea-ad-atlas/

101. Vijayaragavan K, Cannon BJ, Tebaykin D, et al. Single-cell spatial proteomic imaging for human neuropathology. *Acta Neuropathol Commun*. 2022;10(1):158. doi:10.1186/s40478-022-01465-x

102. Dickerson BC, McGinnis SM, Xia C, et al. Approach to atypical Alzheimer's disease and case studies of the major subtypes. *CNS Spectrums*. 2017;22(6):439-449. doi:10.1017/S109285291600047X

103. Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *The Lancet Neurology*. 2011;10(9):785-796. doi:10.1016/S1474-4422(11)70156-9

Index of Figures

Figure 1.2.1: *NeuroPathPredict (NPP) pipeline: QNPtoVox mapping, I-BIS covariate engineering, and universal-kriging based spatial prediction of tau pathology. 8*
Figure 2.2.1: *From city maps to brain maps. Panels A–C[17] illustrate the land-use/universal-kriging blueprint: sparse monitors (A), contextual covariates such as roads/topography (B), and the kriged continuous NOx surface (C). Panels D–F translate that logic to neuropathology: NIA–AA diagnostic sampling sites as sparse measurements (D)[35], a brain GIS in MNI space with tracts, networks, vasculature, and geometry as drift covariates (E), and a potential voxel-wise disease map (F[60]). 19*
Figure 2.3.1. *The ACT[62] "Living Laboratory": An integrated clinico-pathologic and neuroimaging repository enabling brain-wide spatial modeling. Image from ACT study's website. 22*
Figure 2.3.2. *Ex vivo MRI workflow in ACT[74]: (A) fixed brain, (B) agarose embedding, (C) specimen positioned for acquisition, and (D) post-mortem 3D FLAIR, reverse-contrast image, and representative cortical/subcortical segmentation. Image from Latimer et. al. 2023. ..... 23*
Figure 2.3.3. *Coronal slabbing and photographic documentation[74]: uniform 4 mm sections on a calibrated slicer, targeted block isolation, and slab mosaics photographed with a fixed camera rig for scale fidelity and alignment to MRI. Image from Latimer et. al. 2023. ..... 24*
Figure 3.1.1: *Overview of the QNPtoVox pipeline. This schematic outlines the key steps in transforming QNP data, extracted from Halo-stained histological slides, into voxel representations in MNI space, enabling spatial integration with ex vivo and standard neuroimaging volumes for use in the NPP pipeline. 26*
Figure 3.2.1: *Grid overlay of 0.5 × 0.5 mm tiles across Middle Frontal Gyrus (MFG) tissue sections from all 10 participants. The grid facilitates quantitative analysis of tau pathology and spatial alignment with voxel-based imaging data. Participant IDs are shown below each section. 29*
Figure 3.2.2: *Overlay of 0.5 × 0.5 mm analysis grid on a digitized histological section of the MFG for one participant. The magnified inset highlights the relative spatial arrangement and indexing of individual tiles used for coordinate extraction and quantitative analysis. 30*
Figure 3.2.3: *Example of grid tile-based quantification workflow. The left panel shows percent AT8-positive staining per tile as extracted from Halo's area quantification module. The right panel displays the corresponding XML annotation data, where each tile is defined by five polygon vertices in micron-scale X-Y coordinates. 30*
Figure 3.2.4: *Gross tissue image of fixed coronal brain slabs used for QNPtoVox processing. These 4 mm sections, photographed post-fixation and agarose embedding, provide anatomical reference for validating histological sampling and spatial alignment within the NPP pipeline. 31*
Figure 3.2.5: *The left panel shows the acquired postmortem FLAIR image of a single hemisphere embedded in agarose. The right panel displays the mirrored synthetic whole-brain volume generated originally for FreeSurfer-based surface reconstruction and spatial normalization. 32*
Figure 3.5.1: *Preprocessing workflow for ex vivo MRI volumes. Starting with a participant's original FLAIR scan (left), the image is mirrored and converted to FreeSurfer's .mgz format (center), followed by reorientation to LPI, upsampling to 0.5 mm isotropic resolution, and translation to the MNI grid (right). The final output serves as the anatomical reference for voxel-level integration with QNP data. 40*
Figure 3.5.2: *Preprocessing example of ex vivo MRI for participant 7038. The left panel shows the orientation and transformation matrix of the MNI2009b template used as a reference. The middle panel displays the manually reoriented ex vivo MRI scan in anatomical space. The right panel shows the final upsampled volume aligned to MNI space, ensuring consistent voxel orientation and spatial reference across the QNPtoVox pipeline. 41*
Figure 3.6.1: *Co-location of quantitative neuropathology with ex vivo MRI for participant 7038. The left panel shows the gross tissue slab used for histological sampling. The middle panel displays a coronal slice from the participant's upsampled ex vivo MRI volume. The right panel shows the corresponding Halo-annotated tissue section overlaid with a 0.5 × 0.5 mm tile grid. Red arrows illustrate the anatomical alignment process used to identify the Y-plane and apply coordinate offsets for accurate spatial mapping. 42*
Figure 3.7.1: *Transformation of QNP block to voxel space for participant 7038. (a) Upsampled and reoriented ex vivo MRI in native space. (b) Initial placement of the QNP block based on transformed coordinates prior to alignment. (c) Manually aligned QNP block adjusted to match anatomical boundaries using FSLeyes. (d) Final*

smoothed pathology map after application of a 2 mm Gaussian kernel, illustrating local diffusion and anatomical conformity. 44
Figure 3.8.1: *Registration of participant 7038's ex vivo MRI and QNP volume to MNI space using ANTs. (a) Pre-ANTs transformation view showing the native-space ex vivo MRI overlaid with the aligned QNP mask. (b) Post-ANTs transformation of the ex vivo MRI, nonlinearly warped and aligned to the MNI2009b template. (c) Final smoothed QNP mask registered to MNI space and visualized on the MNI2009b cortical surface.* 48
Figure 4.1.1: *Overview of the I-BIS pipeline. This schematic outlines the key steps of data input, pre-processing atlases, and generating custom covariates.* 50
Figure 4.2.1: *The MNI ICBM 2009c nonlinear symmetric template[15] displayed across multiple contrasts and tissue segmentations. Columns show: T1-weighted (T1w), T2-weighted (T2w), and proton density-weighted (PDw) volumes; segmented gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) probability maps; and a lobar parcellation map. Axial and sagittal views are shown for each modality. The template's 0.5 mm isotropic resolution and population-averaged anatomy provide high anatomical fidelity and broad compatibility with structural, functional, and microstructural atlases used in I-BIS.* 52
Figure 4.2.2: *Adapted from Revell et al. (2022). Framework for selecting brain atlases based on validity dimensions (descriptive, explanatory, predictive) and atlas features (structural, functional, multimodal, connectivity, cytoarchitectonic, boundary type, symmetry, coverage, source population). This framework guided the selection of atlases for I-BIS to ensure diverse coverage of anatomical and functional features relevant to spatial covariate generation.* 55
Figure 4.2.3: *Representative white matter tract, AF_L (Arcuate Fasciculus_L) derived from HCP1065 and placed in the MNI 2009b brain.* 59
Figure 4.2.4: *Forkert & Mouches (2019)[69] probabilistic cerebral artery atlas. Color intensity reflects the occurrence probability of major arteries, including ACA, MCA, and PCA branches. The atlas also encodes mean vessel radius, providing a high-resolution reference for defining arterial territories.* 61
Figure 4.2.5. *Comparison of cortical coverage across macrostructural parcellations. HCP-MMP1 (left) divides the cortex into 180 regions per hemisphere, offering fine-scale parcellation with multimodal definitions. CerebrA (right) defines 51 regions per hemisphere, manually corrected for anatomical accuracy and aligned to MNI space. The figure illustrates the difference in cortical coverage and regional granularity between the two atlases.* 63
Figure 4.2.6. *Yeo et al. (2011)[68] cortical functional networks. The 7-network and 17-network parcellations are shown with both "tight" and "liberal" masks in nonlinear MNI152 volumetric space. Tight masks (left in each pair) include only high-probability voxels for each network, while liberal masks (right) extend boundaries to include lower-probability voxels. The 17-network version was selected for I-BIS due to its finer functional subdivisions and enhanced utility for network-based neuropathology modeling.* 65
Figure 4.2.7. *Yeo17 cortical functional networks and Buckner17 cerebellar networks. The Yeo 17-network liberal mask (left) maps cortical functional organization into 17 distinct networks based on resting-state connectivity from 1,000 participants. The Buckner 17-network liberal mask (right) applies the same patient dataset and connectivity-based methodology to the cerebellum, enabling each cerebellar region to be assigned to its corresponding cortical network.* 66
Figure 4.2.8. *Eugenio HC hippocampal subfield atlas applied to MNI ICBM 2009b. Panel A: schematic representation of hippocampal anatomy, showing subfields including CA1–CA4, dentate gyrus, subiculum, and hippocampal proper. Panels B and C: volumetric representation of the 14 hippocampal subfields generated by applying the Eugenio HC pipeline to the MNI 2009b template, ensuring direct alignment with other I-BIS atlas layers. This enables subfield-specific covariates for modeling differential vulnerability patterns in neurodegenerative diseases.* 67
Figure 4.4.1*: Impact of operation order on ROI boundary quality. Left: Variable maps generated by first upsampling/co-registering multi-label atlases, then extracting variables—showing interpolation-induced graininess. Right: Maps generated by first extracting binary label masks, then upsampling/co-registering—yielding crisp boundaries and anatomically faithful representations.* 73
Figure 4.4.2*: Anatomy of an affine transform[94]. Example illustration showing how affine transformations adjust voxel coordinates during image registration. The transformation sequence includes scaling (green), rotation (yellow), and translation (red), relative to the original voxel grid (black). These operations correct for differences in orientation, position, and scale between source images and the MNI 2009b template, forming the first stage of atlas co-registration in the I-BIS pipeline.* 76

Figure 5.1.1: *This figure depicts the four primary components of the NPP UK model pipeline that enable the universal kriging modelling approach to predict the distribution of neuropathology. Preprocessing involves data standardization and transformation. CV elastic net and factor analysis are deployed to reduce covariate dimensionality, including cluster, scree, and correlation analyses. Finally, the UK model is applied with the help of a spatial variogram and leave-one-out block analysis. 84*
Figure 5.5.1*: Correlation plot for participant 6966, representing the top 30 predictors that are selected from the CV-Elastic net model. 95*
Figure 5.5.2: *Dendrogram plot for participant 6966, representing hierarchical clustering of the top 30 predictors that are selected from the CV-Elastic net model. 96*
Figure 5.5.3: *A scree plot for participant 6966, depicting the minimum number of factors or components needed to predict optimum variance from 30 predictors that are selected from the CV-Elastic net model. 96*
Figure 5.5.4*: Correlation plot for participant 6966, representing primary predictors that are selected after factor analysis. 98*
Figure 5.6.1: *An example of a fitted variogram that depicts the spatial autocorrelation of the percentage AT8 positive value distribution across voxels in MFG for participant 6966. ..... 100*
Figure 5.6.2: *A typical example of a semi-variogram showing different components[95]. 104*
Figure 5.6.3: *An example depicting how nmax, nmin, and maxdist are used to tune the spatial variogram and the UK model. [96] 104*
Figure 5.7.1: *NIfTI-based visualization of observed and model-predicted tau pathology (% AT8 positive for p-tau) in MFG for a participant. 107*
Figure 6.1.1: *In this figure, panels a, b, and c depict the input components for participant 6966, wherein panel a shows the quantitative halo extract, panel b depicts an ex vivo Flair image, and panel c shows the gross sliced tissue for the participant. Panels d, e, f, and g show the pre-processing steps for participant 6966. First, the MFG tissue sampling location is identified (panel d) using ex vivo flair and gross tissue images by a neuroanatomist. Next, ex vivo flair images for participant 6966 are co-registered to the MNI template in panels e and f. Finally, using the ANT transformation matrix from ex vivo to MNI registration, we map the co-located tissue sample to the MNI brain volume, extracting voxels that are further processed through I-BIS in panel g. 111*
Figure 6.1.2: *This figure shows block splits for participant 6966's tissue sample. The splits were generated using the K-means clustering algorithm to produce continuous three-dimensional divisions with 16, 32, 64, and 100 blocks. The purpose of testing various split counts was to determine the best model fit for the UK. 116*
Figure 6.1.3: *This figure shows the average number of voxels per block across split levels for participant 6966. The block sizes slightly varied across each split level, given that K-means clustering yields slightly uneven block sizes. 117*
Figure 6.1.4*: This figure illustrates, based on R^2 values, the total number of blocks included in the results for each split level, represented in green. The number of excluded blocks is shown using red bars. Additionally, the percentage of blocks included at each split level for participant 6966 is illustrated by the blue line. 118*
Figure 6.1.5*: This figure compares average RMSE and R2 values for all four split levels for participant 6966 across two criteria, the first of which includes all blocks, and the second that excludes any blocks that have R2 values outside the range 0 to 1. ..... 118*
Figure 6.1.6: *This figure compares model metrics (RMSE, MAE, R2) across the four split levels for participant 6966. 100- and 64-block splits indicate a significantly better model performance, indicating a relationship between distance and spatial autocorrelation. 119*
Figure 6.1.7: *In this figure, for participant 6966, observed vs predicted values are plotted on the left for n_splits = 64 and 100; Plots on the right compare predicted values and residuals for the same splits for participant 6966. .. 120*
Figure 6.2.1: *This plot compares the number of voxels for each participant with the number of primary predictors included in the Universal Kriging model. 122*
Figure 6.2.2*: These plots compare average numbers of voxels per block for each participant across the four split sizes of 16,32,64, and 100. The left panel directly compares average voxels per block per split size across the ten participants, and the panel on the right shows the same data as a heatmap. 122*
Figure 6.2.3: *In this figure, block inclusion percentages are mapped across all ten participants and the four split levels. The inclusion rate increases from 16 splits to 100 splits. Some participants have better block inclusion rates compared to others. 123*
Figure 6.2.4: *This figure illustrates the distribution of boxplots for RMSE and R2 metrics across ten participants, encompassing four different block split sizes 125*

***Figure 6.2.5**: This figure depicts the Universal Kriging model's RMSE values for all ten participants across the four block split sizes. 126*
***Figure 6.2.6:** This figure depicts the Universal Kriging model's R2 values for all ten participants across the four block split sizes..... 126*
***Figure 6.2.7**: In this figure, three models, linear regression, elastic net, and universal kriging (R2), are compared across all 10 participants. The R^2 values for Universal Kriging, for each participant, are derived from the best-performing split model. Participant 7297 had the best R2 values for 64 splits, and the remaining nine had the best R2 values for the 100-splits model. 127*
***Figure 6.3.1**: In this figure, predictors are categorized based on base atlases and compared across the 10 participants using a frequency heatmap. Participants 7038 and 7101 are highlighted as they correspond to quantitative tau pathology sampled from the left hemisphere, while the remaining eight correspond to the right hemisphere. 128*
***Figure 6.3.2:** Depicts the top 30 most frequent predictors selected from the list of all chosen predictors through a 10-fold cross-validated elastic net model. BZ predictors have been combined across any radii. 129*
***Figure 6.3.3**: Depicts the top 20 most frequent primary predictors selected for the UK modeling step, after factor analysis. BZ predictors have been combined across any radii. 131*
***Figure 7.1.1**: Variogram ranges (mm) and mean block edge lengths (mm) for each participant at four spatial cross-validation split sizes (16, 32, 64, 100), with corresponding average R^2 and RMSE values from UK models. *Participant 7144 includes the least number of voxels, hence the smallest edge lengths and range (See Figure 6.2.2). 133*

Index of Tables

Table 4.2.1: *List of atlases included in I-BIS, their corresponding anatomical variables of interest, base coordinate system, and base voxel resolution. 57*
Table 6.1.1: *This table lists 14 primary predictors selected for the universal kriging modelling step for participant 6966. Selected variables, corresponding atlases, variables' anatomical names, and their description in the context of NPP are listed in the table. 115*
Table 6.2.1: *This table lists parameter values fitted for all 10 participants using the elastic net model and variogram model. Elastic net parameters, including alpha and lambda, help select primary predictors for the universal kriging model. Variogram parameters also help fit the spatial autocorrelation part of the universal kriging model. 124*
Table 6.3.1*: This table lists the top 30 most frequent predictors selected across all 10 participants: selected predictors, corresponding atlases, and variables' anatomical names listed in the table. 130*
Table 7.1.1*: Participant-level average block edge length (mm) and cross-validated performance at four split sizes. *Participant 7144 includes the least number of voxels, hence the smallest edge lengths and range (See Figure 6.2.2). 133*